Electronic Journal of Statistics
Vol. 14 (2020) 1413–1448
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1693

Nonconcave penalized estimation in sparse vector autoregression model

Xuening Zhu[*]

School of Data Science
Fudan University, Shanghai, China.
e-mail: xueningzhu@fudan.edu.cn

Abstract: High dimensional time series receive considerable attention recently, whose temporal and cross-sectional dependency could be captured by the vector autoregression (VAR) model. To tackle with the high dimensionality, penalization methods are widely employed. However, theoretically, the existing studies of the penalization methods mainly focus on *i.i.d* data, therefore cannot quantify the effect of the dependence level on the convergence rate. In this work, we use the spectral properties of the time series to quantify the dependence and derive a nonasymptotic upper bound for the estimation errors. By focusing on the nonconcave penalization methods, we manage to establish the oracle properties of the penalized VAR model estimation by considering the effects of temporal and cross-sectional dependence. Extensive numerical studies are conducted to compare the finite sample performance using different penalization functions. Lastly, an air pollution data of mainland China is analyzed for illustration purpose.

MSC 2010 subject classifications: Primary 62M10, 62J07; secondary 62F12, 62H12.
Keywords and phrases: High dimensional time series, nonconcave penalization, vector autoregression, dependent data.

Received January 2019.

Contents

[*] Xuening Zhu is 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. It is also supported in part by China's National Key Research Special Program (no. 2016YFC0207704).

1. Introduction

Penalized estimation methods are fundamental to explore the dataset with high dimensions. Particularly, it is becoming an increasingly popular tool to analyze the data in the fields of genomics, economics, neuroscience with abundant information. To facilitate the analysis, a reasonable assumption is that only a small number of covariates are associated with the response. Under this framework, to select the important covariates, the penalized regression estimation is usually conducted as follows,

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2N}\big\|Y - X\beta\big\|^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{1.1}$$

where $Y \in \mathbb{R}^N$ is the response vector, $X \in \mathbb{R}^{N\times p}$ is the design matrix, and $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient. Here $p_\lambda(\cdot)$ is the penalization function with the tuning parameter λ. Popular choices of penalization functions include L_1 penalty [23], SCAD penalty [8], MCP penalty [26] and many others.

Despite the usefulness of the penalized regression method (1.1), its applications mainly focus on the *i.i.d* data [8, 10]. However, in practice, data with complex dependence structures are frequently encountered. One of the particular types, the high dimensional time series data receives great attention. For example, in economics, structural analysis and economic trend prediction typically require for a large number of macroeconomic variables [6, 1]; in finance, efficient risk management and quantification usually pull for numerous financial statements [11, 17]; in social network analysis, network influences are estimated among social media users by user dynamic behaviors [29, 27]. To model the dynamics of such type of the time series data, the vector autoregression (VAR) model is usually built and estimated. In a typical setting, a p-dimensional time series requires at least p^2 parameters of the transition matrix to be estimated, which are easily much larger than the number of time periods T (i.e., sample size). As a result, the estimation is infeasible unless we assume certain type of low dimensional structures. As we mentioned before, one of the most typical assumptions is the sparsity assumption. This enables us to incorporate the penalization regression framework (1.1) into the VAR model estimation.

Due to the special dependency structure of the high dimensional time series model, the theoretical framework of (1.1) should be reformulated. Particularly, the quantification of the temporal and cross-sectional dependence should be taken into consideration when studying the theoretical properties. In a recent paper of [2], they use the spectral properties to quantify the stability of a vector time series. Using the novel stability measure, they show that the convergence rate of the L_1-penalized (i.e., Lasso) VAR model closely relates to the dependence level of the time series.

Despite the computational attractiveness and competitive ability for prediction, the L_1-penalized (i.e., Lasso type) estimator requires more stringent conditions on the design matrix. As an alternative, the nonconcave estimators (e.g., SCAD and MCP) receive considerable attentions in recent years [10, 25]. Theoretically, such type of estimators enjoy the desirable oracle property. Namely, it could identify the zero coefficients with probability tending to 1, and in the meanwhile estmate the nonzero coefficients as efficiently as if the sparsity pattern is known in advance. For the *i.i.d* data, the theoretical properties of the nonconcave penalized estimator are sufficiently studied [8, 12, 13]. Particularly, the feature dimensionality is allowed to grow exponentially fast with the sample size [10]. However, for the data with dependency structures, the theoretical properties are unknown and need to be investigated.

To fill this gap, in this work, we study the nonconcave penalized VAR model estimation methods. Particularly, we follow [2] to use the tool of spectral density to quantify the temporal and cross-sectional dependences of the high-dimensional time series. By using the novel measure of stability, we manage to explain how the convergence rate of the estimator is related to the dependence level of the time series. We contribute to the existing theory in the following three folds: (a) we establish the selection consistency for the high dimensional sparse stable VAR model by using nonconcave penalty functions, which assumes weaker condition than the irrepresentable condition; (b) we consider and involve the dependence measures in the convergence rate of the estimated parameters beyond the *i.i.d* setting in nonconcave penalized estimation; (c) we establish the oracle properties for the estimated parameters and the asymptotic normality results are proved, which is not achievable in the L_1-penalty setting. Lastly, a real data example about an air pollution index in mainland China, i.e., PM$_{2.5}$, is analyzed using the proposed method.

The rest of the article is organized as follows. Section 2 introduces the nonconcave penalization methods for VAR model estimation. Section 3 investigates the theoretical properties of the variable selection as well as the parameter estimation. Numerical studies are given in Section 4. The article is concluded with a brief discussion in Section 5. All technical details are left to the Appendix.

2. Nonconcave penalization methods in vector autoregression

2.1. Model and notations

Consider a p-dimensional stationary time series vector $\{X_t\} = \{(X_{1t}, \cdots, X_{pt})^\top \in \mathbb{R}^p\}$, which are collected at time points $t = 1, \cdots, T$. To model the dynamics of the X_t, we consider a vector autoregression (VAR) model of lag d [VAR(d)] with serially uncorrelated random errors as

$$X_t = A_1 X_{t-1} + \cdots + A_d X_{t-d} + \mathcal{E}_t \tag{2.1}$$

where $\mathcal{E}_t = (\varepsilon_{1t}, \cdots, \varepsilon_{pt})^\top \in \mathbb{R}^p$ independently follows multivariate normal distribution with mean $\mathbf{0}$ and covariance $\mathrm{cov}(\mathcal{E}_t) = \Sigma_e \in \mathbb{R}^{p\times p}$. Here $A_1, \cdots, A_d$ are $p \times p$-dimensional unknown transition matrices. They provide deep insights about the temporal and cross-sectional relationships among the p time series. In practice, the time series is usually of high dimension. As a result, the number of estimation parameters, i.e., dp^2 could grow polynomially fast with p. To estimate the model (2.1), we consider here a penalized least squares estimation method.

To facilitate the discussion, we first rewrite the model (2.1) in a vector form as follows. Define $\widetilde{X}_t = (X_t^\top, X_{t-1}^\top, \cdots, X_{t-d+1}^\top)^\top \in \mathbb{R}^{dp}$. Let $\mathcal{Y} = (X_T, \cdots, X_d)^\top \in \mathbb{R}^{N\times p}$, $\mathcal{X} = (\widetilde{X}_{T-1}, \cdots, \widetilde{X}_{d-1})^\top \in \mathbb{R}^{N\times dp}$, and $\mathcal{E} = (\mathcal{E}_T, \cdots, \mathcal{E}_d)^\top \in \mathbb{R}^{N\times p}$, where $N = T - d + 1$. In addition, define $B = (A_1, \cdots, A_d)^\top \in \mathbb{R}^{dp\times p}$ to be the parameter matrix. Then we have

$$\mathrm{vec}(\mathcal{Y}) = \mathrm{vec}(\mathcal{X}B) + \mathrm{vec}(\mathcal{E}) = (I_p \otimes \mathcal{X})\beta + \mathrm{vec}(\mathcal{E}), \tag{2.2}$$

where $\beta = \mathrm{vec}(B)$. By the vector form in (2.2), the VAR model could be represented in a general regression form. Define $Y = \mathrm{vec}(\mathcal{Y})$ and $Z = I_p \times \mathcal{X} \in \mathbb{R}^{p(T-d+1)\times q}$, where $q = dp^2$. Then one could rewrite the VAR model (2.1) as a regression model with q-dimensional predictors as $Y = Z\beta + \mathrm{vec}(\mathcal{E})$. To estimate the parameter, we minimize the following regularized least squares type objective function, which yields the penalized least squares estimator as

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^q} \frac{1}{N}\|Y - Z\beta\|^2 + \sum_{j=1}^{q} p_\lambda(|\beta_j|), \tag{2.3}$$

where $p_\lambda(\cdot)$ is the penalization function and $\lambda \geq 0$ is the corresponding regularization parameter. A popular choice of $p_\lambda(\cdot)$ is the L_1-regularization, i.e., $p_\lambda(\delta) = \lambda|\delta|$, which could result in a Lasso type estimator. This approach and corresponding statistical properties are studied by [2]. In this work, we restrict $p_\lambda(\cdot)$ in the family of nonconcave penalization functions. Popular nonconcave penalization functions include the SCAD penalty [8], MCP penalty [26] and many others. Their function forms are visualized in Figure 1 and given in the numerical studies; see equation (4.1) and (4.2). See [21] for a comprehensive discussion. To facilitate the discussion, we define the following notations.

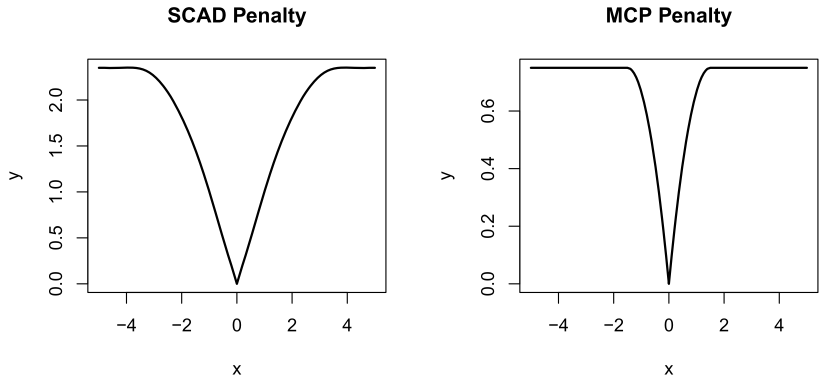

Fig 1. *The SCAD ($a = 3.7$) and MCP ($a = 1.5$) penalty functions with the regularization parameter $\lambda = 1$. The penalty function forms are given by (4.1) and (4.2) respectively.*

Notation Throughout this paper, we denote the cardinality of a set $\mathcal{S}$ by $|\mathcal{S}|$. In addition, let $\mathcal{S}^c$ be the complement of the set $\mathcal{S}$. For a vector $v = (v_1, \cdots, v_p)^\top \in \mathbb{R}^p$, let $\|v\|_q = (\sum_{j=1}^p v_j^q)^{1/q}$ for $q > 0$. For convenience we omit the subindex when $q = 2$. Denote $\text{supp}(v)$ as the support of the vector. Particularly, we use $\|v\|_0$ to denote $|\text{supp}(v)| = \sum_{j=1}^p \mathbf{1}(v_j \neq 0)$ and $\|v\|_{\max}$ to denote $\max_j v_j$. In addition, denote $v_{\mathcal{S}}$ as a sub-vector of v as $v_{\mathcal{S}} = (v_j : j \in \mathcal{S})^\top \in \mathbb{R}^{|\mathcal{S}|}$. For arbitrary matrix $M = (m_{ij}) \in \mathbb{R}^{p_1 \times p_2}$, denote $M_{\mathcal{S}} = (m_{i,j} : 1 \leq i \leq p_1, j \in \mathcal{S})$ as the sub-matrix with columns in $\mathcal{S}$. In addition, let $M^{(\mathcal{S}_1, \mathcal{S}_2)}$ be the sub-matrix of M as $M^{(\mathcal{S}_1, \mathcal{S}_2)} = (m_{ij} : i \in \mathcal{S}_1, j \in \mathcal{S}_2)$ for two sets $\mathcal{S}_1$ and $\mathcal{S}_2$. We further write $M^{(\mathcal{S}, \mathcal{S})}$ as $M^{(\mathcal{S})}$ for simplicity. Furthermore, denote $\|M\|_\infty = \max_{1 \leq i \leq p_1} (\sum_{j=1}^{p_2} |m_{ij}|)$ and $\|M\|_{\max} = \max_{1 \leq i \leq p_1, 1 \leq j \leq p_2} |m_{ij}|$. For a symmetric or Hermitian matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ respectively as its maximum and minimum eigenvalues. For arbitrary two sequences $\{a_N\}$ and $\{b_N\}$, denote $a_N \gg b_N$ to mean that $a_N / b_N \to \infty$ as $N \to \infty$ and $\ll$ the otherwise. Lastly, we use e_i to denote the ith unit vector.

2.2. Measure of stability

Consider a p-dimensional VAR(d) time series model (2.1). Assume it is centered and covariance stationary. As a result, we could define the autocovariance function as $\Gamma_X(h) = \text{cov}(X_t, X_{t-h})$. Generally, the autocovariance function characterizes the temporal and cross-sectional dependence for the VAR(d) model.

Since X_t is a vector time series, it is of particular interest to investigate the stability properties of $\{X_t\}$. In the classical time series analysis, the temporal dependence is usually controlled by imposing some mixing conditions [14]. In the context of VAR model, this sums to assuming that $\lambda_{\max}^{1/2}(A^\top A) < 1$ [20, 19, 16]. However, this condition might be restrictive and can be violated even by many stable VAR models. In addition, it fails to capture all the cross-sectional dependence in the high dimensional setting. In this work, we adopt the idea of [2] to use the spectral density of $\{X_t\}$ to establish the *measure of stability*. Specifically, we assume the VAR(d) model satisfies the following condition.

(C1) The spectral density function

$$f_X(\theta) = \frac{1}{2\pi}\sum_{l=-\infty}^{\infty}\Gamma_X(l)\exp(-il\theta), \quad \theta \in [-\pi, \pi]$$

exists, and its maximum eigenvalue is bounded on $[-\pi, \pi]$, i.e.,

$$\mathcal{M}(f_X) = \operatorname{ess\,sup}_{\theta\in[-\pi,\pi]}\lambda_{\max}(f_X(\theta)) < \infty. \tag{2.4}$$

The spectral density function $f_X(\theta)$ has close relationship with the autocovariance function $\Gamma_X(l)$. If we have $\sum_{l=0}^{\infty}\|\Gamma_X(l)\|^2 < \infty$, then the spectral density exists. Furthermore, the spectral density is bounded, continuous, and the essential supremum, i.e., $\mathcal{M}(f_X)$, is actually its maximum. Generally, the existence of the spectral density will facilitate the following representation,

$$\Gamma_X(l) = \int_{-\pi}^{\pi} f_X(\theta)e^{il\theta}d\theta.$$

The maximum eigenvalue of the spectral density function (2.4) could be a *measure of stability* of the process. Typically higher $\mathcal{M}(f_X)$ implies a less stable process.

For the VAR(d) process, the spectral density function has a closed form. Define $\mathcal{A}(z) = I_p - \sum_{j=1}^{d} A_j z^j$. For the VAR($d$) process [22, 3] it holds

$$f_X(\theta) = \frac{1}{2\pi}\big(\mathcal{A}^{-1}(e^{-i\theta})\big)\Sigma_e\big(\mathcal{A}^{-1}(e^{-i\theta})\big)^*.$$

Other than $\mathcal{M}(f_X)$, the lower extremum of the spectral density is also crucial when dealing with the dependence of the design matrix in the high dimensional setting, i.e.,

$$\mathfrak{m}(f_X) = \operatorname{ess\,inf}_{\theta\in[-\pi,\pi]}\lambda_{\min}(f_X(\theta)).$$

The maximum and minimum of eigenvalue of the spectral density provide important bounds on the dependence of the process $\mathbb{X} = (X_1, \cdots, X_T)^\top \in \mathbb{R}^{T\times p}$. Define $\Upsilon_X = \operatorname{cov}(\operatorname{vec}(\mathbb{X}^\top)) \in \mathbb{R}^{(Tp)\times(Tp)}$. Then it holds that [2]

$$2\pi\mathfrak{m}(f_X) \le \lambda_{\min}(\Upsilon_X) \le \lambda_{\max}(\Upsilon_X) \le 2\pi\mathcal{M}(f_X), \tag{2.5}$$

which is free from the sample size N. Furthermore, for a stationary VAR(d) process $\mathcal{M}(f_X)$ and $\mathfrak{m}(f_X)$ could be further bounded by a closed form as

$$\mathcal{M}(f_X) \le \frac{1}{2\pi}\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})}, \quad \mathfrak{m}(f_X) \ge \frac{1}{2\pi}\frac{\lambda_{\min}(\Sigma_e)}{\mu_{\max}(\mathcal{A})}, \tag{2.6}$$

where

$$\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)), \quad \mu_{\min}(\mathcal{A}) = \min_{|z|=1} \lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z)).$$

Note that the closed form of $\mathcal{M}(f_X)$ and $\mathfrak{m}(f_X)$ of VAR(d) model clearly separates the two types of dependencies of $\{X_t\}$: the temporal dependence captured by the transition matrices A_j, and the additional cross-sectional dependence characterized by Σ_e.

Remark 1. Note that the VAR(d) model could be re-expressed as a VAR(1) model as follows. Specifically, we have

$$\widetilde{X}_t = \widetilde{A}_1 \widetilde{X}_{t-1} + \widetilde{\mathcal{E}}_t$$

where

$$\widetilde{A}_1 = \begin{pmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_p & \mathbf{0} \end{pmatrix}, \quad \widetilde{\mathcal{E}}_t = \begin{pmatrix} \mathcal{E}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

Define $\widetilde{\mathcal{A}} = I_{dp} - \widetilde{A}_1 z$ and $\mu_{\max}(\widetilde{\mathcal{A}})$ and $\mu_{\min}(\widetilde{\mathcal{A}})$ could be defined accordingly. The process $\{\widetilde{X}_t\}$ is stable if and only if the process $\{X_t\}$ is stable [20].

3. Theoretical properties

3.1. Local optimality

In a general framework, note that the VAR estimation is equivalent to the following optimization

$$\arg\min_{\beta \in \mathbb{R}^q} -\beta^\top \widehat{\gamma}_Z + 2^{-1} \beta^\top \widehat{\Gamma}_Z \beta, \tag{3.1}$$

where $\widehat{\gamma}_Z = (I_p \otimes \mathcal{X}^\top) Y/N$, and $\widehat{\Gamma}_Z = Z^\top Z/N = (I \otimes \mathcal{X}^\top \mathcal{X}/N)$. By using the regularization, it leads to

$$Q_p(\beta) = -\beta^\top \widehat{\gamma}_Z + 2^{-1} \beta^\top \widehat{\Gamma}_Z \beta + \sum_{j=1}^{q} p_\lambda(|\beta_j|). \tag{3.2}$$

The solution is then given by $\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^q} Q_p(\beta)$.

In this section, we discuss the theoretical properties of the penalized least squares estimator. The first concern is whether the resulting estimator attains local optimality. We first discuss the existence of the local minimizer of (3.2). It is closely related to the form of the penalty function $p_\lambda(\cdot)$. Define $\rho_\lambda(t) = \lambda^{-1} p_\lambda(t)$. We then give the characterization of the function $\rho_\lambda(t)$ as in the condition (C2).

(C2) Assume $\rho_\lambda(t)$ is increasing and concave in $[t,\infty)$ with a continuous derivative $\rho_\lambda'(t)$ and $\rho_\lambda(0+) > 0$. In addition, $\rho_\lambda'(t)$ is increasing in $\lambda \in (0,\infty)$ and $\rho_\lambda'(0+)$ is not related to λ.

Following [21] and [26], we define the local concavity of the penalty function $\rho_\lambda(t)$ at $v = (v_1,\cdots,v_q)^\top \in \mathbb{R}^q$ as

$$\kappa(\rho;v) = \lim_{\epsilon\to 0+} \max_{1\le j\le q} \sup_{t_1<t_2\in(|v_j|-\epsilon,|v_j|+\epsilon)} -\frac{\rho_\lambda'(t_2)-\rho_\lambda'(t_1)}{t_2-t_1}. \tag{3.3}$$

Since it is assumed that $\rho_\lambda(\cdot)$ takes a concave form by Condition (C2), we have $\kappa(\rho;v)\ge 0$. Moreover, if the second order derivative of $\rho_\lambda(t)$ exists, one could derive that $\kappa(\rho;v) = \max_{1\le j\le q} -\rho''(|v_j|)$. Define $\overline{\rho}(v) = (\overline{\rho}(v_1),\cdots,\overline{\rho}(v_q))^\top \in \mathbb{R}^q$, where $\overline{\rho}(v_j) = \mathrm{sgn}(v_j)\rho'(|v_j|)$. The following proposition establishes conditions for the strict local minimizer of the objective function.

Proposition 1 (Local Minimizer). *Assume $p_\lambda(\cdot)$ satisfies Condition (C2). Then $\widehat{\beta}\in\mathbb{R}^q$ is a strict local minimizer with probability tending to 1 if*

$$Z_{\mathcal{S}}^\top Y - Z_{\mathcal{S}}^\top Z\widehat{\beta} - N\lambda_N\overline{\rho}(\widehat{\beta}_{\mathcal{S}}) = 0 \tag{3.4}$$

$$(N\lambda_N)^{-1}\|Z_{\mathcal{S}^c}^\top(Y - Z\widehat{\beta})\|_\infty < \rho'(0+) \tag{3.5}$$

$$\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) > N\lambda_N\kappa(\rho,\widehat{\beta}_{\mathcal{S}}). \tag{3.6}$$

as $N\to\infty$, where $\mathcal{S} = supp(\beta)$ is the index set for the nonzero coefficients.

The proof is given in Appendix A.2. Specifically, Condition (3.4) and (3.6) ensure that $\widehat{\beta}$ is a strict local minimizer when constraint on the $\|\beta\|_0$-dimensional subspace $\{\beta\in\mathbb{R}^p : \beta_{\mathcal{S}^c} = \mathbf{0}\}$. Condition (3.5) further makes sure that the sparse solution $\widehat{\beta}$ is strict local minimizer on the whole space $\mathbb{R}^q$.

Under the framework of strict local minimizer, we discuss the theoretical properties of the nonconcave penalized VAR estimator. The main difficulty here from the *i.i.d* case is that one need to take the temporal as well as the cross-sectional dependency into consideration. To this end, the deviation bounds for the design matrix and error terms are firstly established in Section 3.2, which are essential tools to develop the oracle properties. Next, the oracle properties are given in Section 3.3.

3.2. Deviation bounds

In this section, we establish some deviation bounds for both the design matrix and the error terms. The deviation bounds are important for deriving the asymptotic properties of the nonconcave penalized estimator. Particularly, it gives conditions for the design matrix and error terms to behave properly. We first give the results of the design matrix.

By (3.6), we know that the strict local minimizer requires the minimum eigenvalue of $Z_{\mathcal{S}}^\top Z_{\mathcal{S}}$ is well bounded. Define $\omega = c_{1\lambda}/c_{2\lambda}$, where $c_{1\lambda} = \lambda_{\max}^{-1}(\Sigma_e)\mu_{\min}(\mathcal{A})$ and $c_{2\lambda} = \lambda_{\min}^{-1}(\Sigma_e)\mu_{\max}(\mathcal{A})$. We have the following proposition.

Proposition 2. *Consider a stable VAR(d) model (2.1) and its vectorization form (2.2). Assume Condition (C1) holds. We have*

$$P\Big\{\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) > 2^{-1}Nc_{2\lambda}^{-1}\Big\} \geq 1 - 2\exp\{-cN\min(1,\omega^2) + s\log 21\} \tag{3.7}$$

$$P\Big\{\lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) < 2Nc_{1\lambda}^{-1}\Big\} \geq 1 - 2\exp\{-cN + s\log 21\} \tag{3.8}$$

as $N \to \infty$, where c is a positive constant.

The proof of Proposition 2 is given in Appendix A.3. By $s = o(N\min\{1,\omega^2\})$, one could conclude $2^{-1}Nc_{2\lambda}^{-1} < \lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) \leq \lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) < 2Nc_{1\lambda}^{-1}$ by (3.7) and (3.8) with probability tending to 1.

Besides the eigenvalue bound of the design matrix, the maximum absolute value bound is another important bound we need to quantify. This could be critical for deriving the variable selection consistency. Let $\Gamma_{\widetilde{X}}(0) \in \mathbb{R}^{dp\times dp}$ be the covariance of $\widetilde{X}_t$. Recall that $Z = I_p \otimes \mathcal{X}$ and define $\Gamma_Z = N^{-1}E(Z^\top Z)$. Given the results in Proposition 2, we are able to derive the bounds for the maximum absolute value of $(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}$ and $Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}$ in the following proposition.

Proposition 3. *Consider a stable VAR(d) model (2.1) and its vectorization form (2.2). Assume Condition (C1) holds. We have*

$$P\Big\{\|(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty \leq c_\mu\Big\} \geq 1 - 2\exp\{-cN\min(1,\omega^2) + s\log 21\} \tag{3.9}$$

$$P\Big\{\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max} \leq c_\Gamma\Big\} \geq 1 - 6\exp\{-cN\min(\nu^2,\nu) + \log q + \log s\} \tag{3.10}$$

where

$$c_\mu = 2N^{-1}s^{1/2}c_{2\lambda}, \quad c_\Gamma = N\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} + N\max\{\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}, c_{1\lambda}^{-1}N^{-\delta}\} \tag{3.11}$$

$$\nu = 1/3\max\{c_{1\lambda}\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}, N^{-\delta}\},$$

with $0 \leq \delta < 1/2$ being a positive constant. The autocovariance matrix Γ_Z can be expressed as $\Gamma_Z = I_p \otimes \Gamma_{\widetilde{X}}(0)$, with $vec(\Gamma_{\widetilde{X}}(0)) = (I - \widetilde{A}_1 \otimes \widetilde{A}_1)^{-1}vec(\widetilde{\Sigma}_e)$ and $\widetilde{\Sigma}_e = diag(\Sigma_e, \mathbf{0}_{p\times p}, \cdots, \mathbf{0}_{p\times p}) \in \mathbb{R}^{(pd)\times(pd)}$.

The proof of Proposition 3 is given in Appendix A.4. Proposition 3 establishes the upper bound of $\|(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty$ and $\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max}$ respectively, which essentially restricts the covariance level among the covariates. Furthermore, one could obtain that $\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty \leq \|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max}\|(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty$ is bounded by $c_\mu c_\Gamma$ with high probability under certain conditions, where c_μ and c_Γ involve the dependence measures of the VAR model. For simple stable VAR(1) models with no dynamic and cross-sectional dependences (e.g., $A_1 = 0.5I_p$ and $\Sigma_e = \sigma_e^2 I_p$), it can be verified that $c_\mu = O(s^{1/2}/N)$ and $c_\Gamma = O(N)$, thus $c_\mu c_\Gamma = O(s^{1/2})$.

Note that the positive constant δ is involved in the left side as well as the right side of (3.10). A higher δ will possibly result in a tighter upper bound of $\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max}$, and in the meanwhile smaller probability in the right side. If

$c_{1\lambda}\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} \le N^{-\delta}$ and $\log q = o(N^{1-2\delta})$, then one could conclude that $c_\Gamma \le 2c_{1\lambda}^{-1}N^{1-\delta}$ and the probability for the event $\{\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max} \le c_\Gamma\}$ will tend to 1 as $N \to \infty$. Next, one may note that $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$ could be very related to the irrepresentable assumption of the *i.i.d* case [28], while for the VAR model it involves the dependence structures explicitly. We further give a comment about the $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$ in the following remark.

Remark 2. Note that Γ_Z takes a block diagonal structure. Generally, we have $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} \le \lambda_{\max}(\Gamma_Z) \le c_{1\lambda}^{-1}$. Therefore we could obtain that $\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max}$ is bounded by $2Nc_{1\lambda}^{-1}$ with high probability, where higher $c_{1\lambda}^{-1}$ implies higher dependence levels. The upper bound given by (3.10) is tighter. To see this, one could express $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$ by

$$\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} = \max_{1\le j\le p}\Big\{\max_{i_1\ne i_2, B_{i_1 j}\ne 0, B_{i_2 j}\ne 0} |\Gamma_{\widetilde{X},i_1 i_2}(0)|\Big\}.$$

It can be then concluded that, if for any column j of B, there exists at most one element $B_{ij} \ne 0$ for $1 \le i \le pd$, then we have $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} = 0$ and we then have the upper bound as $c_\Gamma = c_{1\lambda}^{-1}N^{1-\delta}$ in (3.11). Hence, as one can see, the dependence level of the VAR model is explicitly involved here compared to the irrepresentable assumptions.

Lastly, we establish the upper bound for the error terms. For the vector formed stable VAR(d) model (2.2), define $\xi = Z^\top(Y - Z\beta) = \mathrm{vec}(\mathcal{X}^\top\mathcal{E})$. Then the convergence rate of the penalized VAR model estimator is largely determined by how concentrate ξ is around $\mathbf{0}$. In addition note $\xi \in \mathbb{R}^q$ naturally constitutes a high dimensional vector. As a result, it is important to control the diverging rate of $\|\xi\|_\infty$. Specifically, we have the following proposition.

Proposition 4. *Assumes Condition (C1) holds. We then have*

$$P\Big\{\|\xi_{\mathcal{S}}\|_\infty > c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}\Big\} \le 6s/N \tag{3.12}$$

$$P\Big\{\|\xi_{\mathcal{S}^c}\|_\infty > \mathbb{Q}(\beta,\Sigma_e)u_N\sqrt{N}\Big\} \le 6(q-s)\exp(-c_1u_N^2), \tag{3.13}$$

as $N \to \infty$, where $u_N = c_1^{-1/2}N^\alpha(\log N)^{1/2}$ with $\alpha \in [0, 1/2]$ and

$$\mathbb{Q}(\beta,\Sigma_e) = c_0\Big\{\lambda_{\max}(\Sigma_e) + \frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})} + \frac{\lambda_{\max}(\Sigma_e)\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\Big\}$$

with c_0 as a finite positive constant.

The proof of Proposition 4 is given in Appendix A.5. By Proposition 4, the concentration of ξ around $\mathbf{0}$ is affected not only by the sample size and the parameter dimension, but also by the temporal and contemporaneous dependence of the time series. As we mentioned before, the temporal and contemporaneous dependencies are characterized by $\mathcal{A}$ and Σ_e respectively. Particularly, the error bound is lower when the eigenvalues of $\mathcal{A}$ and Σ_e behave more uniformly, i.e.,

$\lambda_{\max}(\Sigma_e)$ and $\mu_{\max}(\mathcal{A})$ are smaller, and in the meanwhile $\lambda_{\min}(\Sigma_e)$ and $\mu_{\min}(\mathcal{A})$ are larger. This results in a less spiky spectrum and leads to lower temporal and cross-sectional dependencies, i.e., lower $\mathbb{Q}(\beta, \Sigma_e)$ [2].

3.3. Oracle property

Given the bounds on the design matrix as well as the error bounds, we establish the oracle properties of the nonconcave penalized VAR estimator in this section. The weak oracle property is firstly given, which is introduced by [21]. It has two folds of meaning: (a) first, with probability tending to 1 the nonzero coefficients could be estimated to be exact $\mathbf{0}$ (i.e., $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$), and (b) second, the estimator is consistent in the sense of L_∞ loss. Although the weak property does not give the asymptotic distribution of the estimator, it provides insights about the asymptotic behaviors of the penalized estimator. The following conditions are assumed.

(C3) (Dependence) Assume $c_\mu c_\Gamma =$

$$2c_{2\lambda}s^{1/2}\Big[\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} + \max\{\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}, c_{1\lambda}^{-1}N^{-\delta}\}\Big] \le \min\Big\{C\frac{\rho'(0+)}{\rho'(d_N)}, O(N^\alpha)\Big\}, \tag{3.14}$$

where $C \in (0,1)$, $0 < \delta < 1/2 - \alpha$, and $0 \le \alpha < 1/2$.

(C4) (Minimum Signal) Assume

$$d_N \ge N^{-\gamma}\log N\mathbb{Q}(\beta, \Sigma_e), \quad \lambda_N \gg N^{-(1/2-\alpha)}\log N\mathbb{Q}(\beta, \Sigma_e)$$
$$p'_{\lambda_N}(d_N) = o\{c_{2\lambda}^{-1}\mathbb{Q}(\beta_{\mathcal{S}}, \Sigma_e)s^{-1/2}N^{-\gamma}\log N\}.$$

(C5) (Penalty Concavity) Define $\kappa_0 = \max_{\delta\in\mathcal{N}_0}\kappa(\rho;\delta)$, where $\mathcal{N}_0 = \{\theta \in \mathbb{R}^s : \|\theta - \beta_{\mathcal{S}}\|_\infty \le d_N\}$. Assume $\lambda_N\kappa_0 = o(c_{2\lambda}^{-1})$.

First, Condition (C3) imposes restrictions on the correlation level between $Z_{\mathcal{S}}$ and $Z_{\mathcal{S}^c}$ as well as the temporal and cross-sectional dependence. In the $i.i.d$ case, it is typical to directly assume $\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty$ is bounded by the right side [10]. For the VAR model, we could obtain the upper bound from Proposition 3, which involves $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$ explicitly. Generally, $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$ is bounded by $c_{1\lambda}^{-1}$. Therefore the left hand (i.e., $c_\mu c_\Gamma$) is bounded by $4s^{1/2}\omega^{-1}$, where larger ω^{-1} implies higher dependence. For a simple VAR(1) model with $A_1 = 0.5I_p$ and $\Sigma_e = \sigma_e^2 I_p$, we will have $\omega = O(1)$. In this situation, Condition (C3) requires $4s^{1/2} \le \min\{c\rho'(0+)/\rho'(d_N), O(N^\alpha)\}$. However, for general stable VAR models, the situations can be more complicated. According to [2], even with the same spectral radius of the transition matrix, the dependence measures (e.g., $c_{2\lambda}$ and $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}$) could still be quite different and might lead to different convergence behaviours of the estimated parameters. Next, the upper bound $C\rho'(0+)/\rho'(d_N)$ in (3.14) is closely related to the penalty form. If the L_1 penalty is used, then it requires the left side of (3.14) is strictly less than 1.

In such a case, this condition could lead to the strong irrepresentable condition proposed by [28], which is restrictive in practice. While for nonconcave penalty function, e.g., SCAD penalty, the upper bound could grow to ∞, which makes this condition easier to satisfy.

Next, Condition (C4) sets the assumption about the minimum signal and the tuning parameter λ_N, which is critical for deriving the converging rate of the estimator. The requirements are mostly the same with the *i.i.d* case, except that it further includes terms related to the dependency measures, e.g., $\mathbb{Q}(\beta_{\mathcal{S}}, \Sigma_e)$ [10]. Specifically, in the VAR model setting, the condition for the minimum signal d_N is more restrictive. For instance, if the dynamic as well as the cross sectional dependence is higher, then it will lead to higher $\mathbb{Q}(\beta_{\mathcal{S}}, \Sigma_e)$, thus requires the signal strength stronger enough to be detected.

Lastly, Condition (C5) guarantees the local optimality of the estimator. The condition corroborates with the $\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})$ condition in the *i.i.d* case [10], which regularizes the multilinearity of $Z_{\mathcal{S}}$. In the VAR model, the $\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})$ is connected to the dependence measure $c_{2\lambda}$ by (3.7), which results in Condition (C5). This condition can be easily satisfied by the SCAD type condition when $\lambda_N \ll d_N$, which leads to $\kappa_0 = 0$ with sufficiently large N. We then establish the weak oracle property as follows.

Theorem 1 (Weak Oracle Property)**.** *Let* $\log q = O(N^{2\alpha})$ *and* $s = o\{\min(N\omega^2, c_{2\lambda}^{-2} N^{1-2\gamma} \log N)\}$ *and* $0 < \gamma \leq 1/2$ *and* α *defined in Condition (C3). Assume Conditions (C1)–(C5) hold. Then there exists a nonconcave penalized least squares estimator* $\widehat{\beta}$ *such that for sufficiently large* N*, with probability tending to 1, we have*
(a) (Sparsity). $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$*;*
(b) (L_∞ *loss).* $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|_\infty \leq \mathbb{Q}(\beta, \Sigma_e) N^{-\gamma} \log N$*.*

The proof of Theorem 1 is given in Appendix B.1. Note that the parameter dimension q is allowed to grow exponentially fast with the sample size N, and the growth rate is controlled by α. In addition, the dependence term $\mathbb{Q}(\beta, \Sigma_e)$ is involved in the upper bound of the L_∞ loss of $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|_\infty$. This implies that the time process with lower temporal and cross-sectional dependencies could lead to a tighter upper bound. When $\gamma = 1/2$, this result corroborates with the finding of [2] for the L_1 penalty. Lastly, one might note the convergence rate is slightly slower than $O_p(\sqrt{s/N})$ in the *i.i.d* case under L_2 loss. This is because here we use the L_∞ loss in the derivation of the consistency rate.

Next, we establish the oracle properties of the nonconcave penalized VAR estimator. Let $\alpha_\lambda = \mathbb{Q}(\beta, \Sigma_e)\lambda_{\max}(\Sigma_e)c_{2\lambda}$. We first establish the existence of the nonconcave penalized VAR estimator and its convergence rate. To this end, we need the following condition.

(C6) Assume

$$d_N \gg \lambda_N \gg \alpha_\lambda \max\{c_{1\lambda}^{-1}\sqrt{s/N}, N^{(\alpha-1)/2}(\log N)^{1/2}\}.$$

and $p'_{\lambda_N}(d_N) = O(N^{-1/2}\alpha_\lambda c_{2\lambda}^{-1/2})$. In addition, assume $\lambda_N \kappa_0 = o(c_{2\lambda}^{-1})$.

The Condition (C6) gives the minimal signal strength for the nonconcave penalized VAR estimator to achieve $\sqrt{s/N}$-consistency. Due to the assumption $d_N \gg \lambda_N$, the SCAD type penalties satisfy the condition on $p'_{\lambda_N}(d_N)$ since $p'_{\lambda_N}(d_N) = 0$. However, for L_1 penalty, we have

$$p'_{\lambda_N}(d_N) = \lambda_N \gg \alpha_\lambda \max\{c_{1\lambda}^{-1}\sqrt{s/N}, N^{(\alpha-1)/2}(\log N)^{1/2}\}.$$

This contradicts with the assumption $p'_{\lambda_N}(d_N) = O(N^{-1/2}\alpha_\lambda c_{2\lambda}^{-1/2})$. This explains that the L_1 penalized estimator could not achieve the consistency rate of $O_p(\sqrt{s/N})$ even with low dependence levels. Compared to the *i.i.d* case [10], the condition involves further requirements on the minimum signal with respect to the dependence levels, i.e., α_λ. Specifically, higher dependence levels will result in a higher α_λ, thus will require d_N to be larger to be detected. We give the convergence rate for the nonconcave penalized VAR estimator as follows.

Theorem 2 (Existence of Nonconcave Penalized VAR Estimator)**.** *Assume $\log q = O(N^\alpha)$ and $s = o(N\min(1,\omega^2))$, where $\alpha \in [0, 1/2)$. In addition, assume the VAR(d) process (2.1) is stable and Conditions (C1)–(C6) hold. There exists a strict local minimizer $\widehat{\beta}$, then with probability tending to 1 as $N \to \infty$, the nonconcave penalized likelihood estimator satisfies $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$; and $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\| \le \alpha_\lambda\sqrt{s}N^{-1/2}$ with probability tending to 1.*

The proof of Theorem 2 is given in Appendix B.2. Under the signal strength d_N, the Theorem 2 states the parameter dimension that the nonconcave penalized VAR estimation method could handle. Note that the dependence term α_λ is also involved in the upper bound of $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|$. This partially explains how the temporal and cross-sectional dependence affects the convergence rate of $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|$. We further investigate the asymptotic normality of the nonconcave penalized VAR estimator, which is stated as follows.

Theorem 3 (Oracle Property)**.** *Assume Conditions (C1)–(C6) hold. In addition, assume $p'_{\lambda_N}(d_N) = o(s^{-1/2}N^{-1/2}c_{2\lambda}^{-1}\lambda_{\min}(\Sigma_e))$ and $s = o(N\min(1,\omega^2))$. Under the conditions of Theorem 2, with probability tending to 1 and $N \to \infty$, for a stable VAR(d) process, the nonconcave penalized likelihood estimator satisfies:*
(a) (Sparsity). *$\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$;*
(b) (Asymptotic Normality). *Let $A_N \in \mathbb{R}^{m\times s}$ satisfying $A_N A_N^\top \to G$ as $N \to \infty$, where m is any fixed integer, and G is a $m \times m$ nonnegative symmetric matrix. It holds that as $N \to \infty$*

$$\sqrt{N}A_N\Sigma_{\mathcal{S}}^{-1/2}(\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}) \to_d N(\mathbf{0}, G), \tag{3.15}$$

where $\Sigma_{\mathcal{S}} = (\Gamma_Z^{(\mathcal{S})})^{-1}\Gamma_{Ze}^{(\mathcal{S})}(\Gamma_Z^{(\mathcal{S})})^{-1}$ with $\Gamma_{Ze} = \Sigma_e \otimes \Gamma_{\widetilde{X}}(0)$.

The proof of Theorem 3 is given in Appendix B.3. Note that from Theorem 3, if s is finite, then by setting $A_N = I_s$, we could obtain the asymptotic normality of $\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}$ directly.

In practice, one may consider to further conduct inference based on (3.15). Define $Z_t = I_p \otimes \widetilde{X}_t \in \mathbb{R}^{dp^2 \times p}$, $\widehat{\mathcal{E}}_t = X_t - Z_{t-1}^{(\mathcal{S})\top} \widehat{\beta}_{\mathcal{S}}$ and $\widehat{\mathcal{E}} = (\widehat{\mathcal{E}}_T, \cdots, \widehat{\mathcal{E}}_{T-d+1})^\top \in \mathbb{R}^{N \times p}$. Let $\widehat{\Gamma}_Z = I_p \otimes \widehat{\Gamma}_{\widetilde{X}}(0)$ and $\widehat{\Gamma}_{Ze} = \widehat{\Sigma}_e \otimes \widehat{\Gamma}_{\widetilde{X}}(0)$, where $\widehat{\Gamma}_{\widetilde{X}}(0) = \mathcal{X}^\top \mathcal{X}/N$ and $\widehat{\Sigma}_e = \widehat{\mathcal{E}}^\top \widehat{\mathcal{E}}/N$. Then a natural estimator of $\Sigma_{\mathcal{S}}$ is $\widehat{\Sigma}_{\mathcal{S}} = (\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1} \widehat{\Gamma}_{Ze}^{(\mathcal{S})} (\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1}$. We then establish the following theorem for the consistency of the $\widehat{\Sigma}_{\mathcal{S}}$.

Theorem 4. *Assume all the conditions in Theorem 3 hold. Further assume that $p = o(N)$ and $s = o\{N \min(1, \omega^2, c_{1\lambda} \alpha_\lambda^{-1} \lambda_{\max}^{-1}(\Sigma_e), c_\lambda^4 c_{2\lambda}^{-6} \lambda_{\max}^{-1}(\Sigma_e))\}$. Then we have $A_N \widehat{\Sigma}_{\mathcal{S}} A_N^\top - A_N \Sigma_{\mathcal{S}} A_N^\top \to_p 0$ as $N \to \infty$, where $A_N \in \mathbb{R}^{m \times s}$ satisfying $A_N A_N^\top \to G$ as $N \to \infty$, m is any fixed integer, and G is a $m \times m$ nonnegative symmetric matrix.*

Then proof of Theorem 4 is given in Appendix B.4. Note that it further requires the condition of $p = o(N)$ and $s = o\{N \min(1, \omega^2, c_{1\lambda} \alpha_\lambda^{-1} \lambda_{\max}^{-1}(\Sigma_e), c_\lambda^4 c_{2\lambda}^{-6} \lambda_{\max}^{-1}(\Sigma_e))\}$. That is mainly because the estimation of Σ_e requires to estimate $O(p^2)$ parameters. It then needs larger sample size to obtain a consistent estimate. In addition, higher dependence levels also have influences. Specifically, higher dependences will imply larger α_λ, $\lambda_{\max}(\Sigma_e)$, $c_{2\lambda}$, $c_{1\lambda}^{-1}$ values, thus restrict s to be smaller.

To examine the performance of the nonconcave penalized VAR estimation, we conduct a number of numerical studies, which are discussed in details in the following section.

4. Numerical study

4.1. Simulation models

In this section, we evaluate the finite sample performance of the nonconcave penalized VAR model. To generate the data, we first generate the transition matrix $A_1 = (a_{1,ij}) \in \mathbb{R}^{p \times p}$, which has about 5% nonzero off-diagonal entries. Specifically, we set the off-diagonal nonzero elements to be 0.3, and all the diagonal elements to be 0.5. Following [2], we generate the innovation process using Gaussian process with the following three covariance structures Σ_e:
(1) Block I: $\Sigma_e = (\sigma_{e,ij}) \in \mathbb{R}^{p \times p}$ with $\sigma_{e,ii} = 1$, $\sigma_{e,ij} = \rho$ if $1 \le i \ne j \le p/2$, and $\sigma_{e,ij} = 0$ otherwise.
(2) Block II: $\Sigma_e = (\sigma_{e,ij}) \in \mathbb{R}^{p \times p}$ with $\sigma_{e,ii} = 1$, $\sigma_{e,ij} = \rho$ if $1 \le i \ne j \le p/2$ or $p/2 < i \ne j \le p$, otherwise $\sigma_{e,ij} = 0$.
(3) Toeplotz: $\Sigma_e = (\sigma_{e,ij}) \in \mathbb{R}^{p \times p}$ with $\sigma_{e,ij} = \rho^{|i-j|}$. Here larger ρ values indicate that the innovation processes have higher correlation with each other.

4.2. Implementation of the algorithm

For comparison, we implement the following algorithms. They are, the Lasso estimator [23, 2], the adaptive Lasso (ALasso) estimator [30, 4], the SCAD estimator ((4.1) with $a = 3.7$) proposed by [8], and MCP estimator ((4.2) with

$a = 1.5$) proposed by [26]. Specifically, the penalty functions of SCAD and MCP are given as below, namely

$$p^{\text{SCAD}}_{\lambda,a}(\theta) = \begin{cases} \lambda\theta & \theta \leq \lambda, \\ (a-1)^{-1}\{a\lambda\theta - 0.5(\theta^2 + \lambda^2)\} & \lambda < \theta \leq a\lambda, \\ 2^{-1}(a-1)^{-1}\lambda^2(a^2-1) & \theta > a\lambda, \end{cases} \tag{4.1}$$

and

$$p^{\text{MCP}}_{\lambda,a}(\theta) = \begin{cases} \lambda\theta - (2a)^{-1}\theta^2 & \theta \leq a\lambda, \\ 2^{-1}a\lambda^2 & \theta > a\lambda. \end{cases} \tag{4.2}$$

Furthermore, in the numerical study, to make the computation more feasible and increase the efficiency, we use screening method to reduce the parameter dimension before we conduct the model selection and estimation. Specifically, we use the SIS method proposed by [9] on the regression form (2.2) of the VAR model and keep the $q^* = 200$ covariates with highest absolute correlations with the response variable.

Lastly, to optimize the objective function (3.2), we use the local linear approximation (LLA) algorithm proposed by [31]. To choose the tuning parameter, the HBIC criterion [25] is employed, which expresses as

$$\text{HBIC}(\lambda) = \ell(\widehat{\theta}_{\mathcal{M}}) + |\mathcal{M}_\lambda| \frac{C_N \log(q)}{N}, \tag{4.3}$$

where $\mathcal{M}_\lambda = \{(i,j) : \widehat{a}_{1,ij} \neq 0\}$ denotes the selected set, and $C_N = \log\{\log(N)\}$ is slowly diverging with N.

4.3. Performance measurements and simulation results

We then evaluate the sparse recovery and estimation accuracy of the nonconcave penalized VAR model. For each simulation setting, we consider (1) Medium VAR ($p = 30, d = 1, T = 80, 120$), and (2) Large VAR ($p = 50, d = 1, T = 100, 150$) respectively. In addition, we set $\rho = 0.5, 0.6$ respectively to reflect different levels of dependences. To obtain a reliable evaluation, the experiment is replicated for $R = 500$ times.

Denote the estimation of the transition matrix A_1 of the rth replication as $\widehat{A}_1^{(r)} = (\widehat{a}^{(r)}_{1,ij})$. We first investigate the sparse recovery property of the nonconcave penalized VAR model. With to this regard, we consider two measurements. First, the true positive value (TP) is defined by the number of nonzero edges estimated to be nonzero, i.e., $\text{TP} = \sum_{r=1}^R \sum_{i,j} I(\widehat{a}^{(r)}_{1,ij} a_{1,ij} \neq 0)$. Second, the false positive value (FP) is defined as the number of zero edges incorrectly estimated to nonzero, i.e., $\text{FP} = \sum_{r=1}^R \sum_{i,j} I(a_{1,ij} = 0, \widehat{a}_{1,ij} \neq 0)$. Next, we evaluate the estimation accuracy. To this end, we calculate the root mean square error (RMSE) for the transition matrix A_1 as $\text{RMSE}_A = (\|\widehat{A}_1 - A_1\|_F^2/p^2)^{1/2}$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Table 1
Simulation Results with 500 Replications for Example 1. The true positive number (TP), false positive number (FP), and the RMSE$_A$ are reported for different dependence levels (i.e., $\rho = 0.5$ and $\rho = 0.6$).

			$\rho = 0.5$			$\rho = 0.6$		
p	T		TP	FP	RMSE$_A$	TP	FP	RMSE$_A$
30	80	Lasso	34.3	60.3	0.054	33.1	53.9	0.056
		SCAD	29.4	5.3	0.050	28.4	5.7	0.053
		MCP	27.9	3.6	0.051	27.0	4.4	0.054
		ALasso	29.5	4.9	0.048	28.5	5.5	0.051
30	120	Lasso	35.9	41.9	0.045	35.2	38.8	0.046
		SCAD	33.4	2.0	0.031	32.7	2.3	0.034
		MCP	33.0	2.0	0.032	32.3	2.5	0.035
		ALasso	33.7	2.2	0.031	32.9	2.4	0.034
50	100	Lasso	53.3	57.5	0.040	49.9	55.1	0.043
		SCAD	49.0	5.4	0.034	45.1	6.4	0.040
		MCP	48.1	6.5	0.036	44.3	7.7	0.041
		ALasso	50.2	6.7	0.033	46.1	7.4	0.039
50	150	Lasso	57.2	52.2	0.031	54.5	49.9	0.033
		SCAD	55.2	2.6	0.022	52.3	3.5	0.027
		MCP	54.9	3.0	0.023	51.8	4.2	0.028
		ALasso	55.6	2.8	0.022	52.5	3.0	0.026

The simulation results are given in Tables 1–3. First, under higher dependency levels, the performances of both sparsity recovery as well as model estimation are affected and worse than the lower dependence levels. That corroborates with the theoretical properties established in this work. Second, comparably speaking, while the true positive numbers are similar for all the methods, the nonconcave penalization methods are capable to achieve lower false positive number, which leads to a more parsimonious model. For instance, in Example 2 with $p = 50$, $T = 100$ and $\rho = 0.5$ (i.e., Table 2), the SCAD penalty is able to control the FP at about 5.8, while the FP for the Lasso method is 58.3, which is almost 10 times larger than the SCAD method. In the meanwhile, in terms of the estimation accuracy, the nonconcave penalization methods could obtain relatively lower estimation errors. For example, the RMSE for the MCP penalized VAR model is about 0.020 with $p = 50$, $T = 150$, and $\rho = 0.6$ in Example 3 (i.e., Table 3), which is much smaller than the Lasso method with RMSE $=$ 0.032.

Lastly, one could observe that the adaptive Lasso is also a competitive method, which also has better performance than the Lasso method in terms of the finite sample performance. Compared to the nonconcave penalization method, we would like to comment that the nonconcave penalty is typically more flexible than the adaptive Lasso approach. That is because, an element being estimated as zero (e.g., by SCAD penalty) can escape from zero in the next iteration; while the adaptive Lasso absorbs zeros in each iteration, and always results in sparser solutions than the initial values [7]. Therefore in practice the adaptive Lasso method requires to set the initial values more carefully.

Table 2
Simulation Results with 500 Replications for Example 2. The true positive number (TP), false positive number (FP), and the RMSE$_A$ are reported for different dependence levels (i.e., $\rho = 0.5$ and $\rho = 0.6$).

			$\rho = 0.5$			$\rho = 0.6$		
p	T		TP	FP	RMSE$_A$	TP	FP	RMSE$_A$
30	80	Lasso	33.4	54.7	0.057	31.6	50.1	0.061
		SCAD	28.1	5.1	0.053	26.4	6.3	0.060
		MCP	26.6	4.6	0.056	25.1	6.3	0.062
		ALasso	28.0	5.4	0.052	26.1	6.3	0.059
30	120	Lasso	35.5	42.2	0.047	34.5	41.6	0.050
		SCAD	32.6	2.0	0.034	31.2	2.7	0.040
		MCP	32.3	2.4	0.035	30.8	3.4	0.042
		ALasso	32.7	2.1	0.034	31.1	2.6	0.040
50	100	Lasso	49.9	58.3	0.044	44.1	54.7	0.049
		SCAD	45.0	5.8	0.040	38.8	6.7	0.048
		MCP	44.3	7.4	0.041	38.2	8.6	0.049
		ALasso	45.9	6.6	0.038	39.5	7.1	0.046
50	150	Lasso	55.5	56.9	0.034	50.7	52.7	0.038
		SCAD	52.5	2.4	0.026	47.0	2.9	0.034
		MCP	52.2	3.4	0.027	46.6	4.0	0.035
		ALasso	52.9	2.4	0.026	47.3	2.6	0.034

Table 3
Simulation Results with 500 Replications for Example 3. The true positive number (TP), false positive number (FP), and the RMSE$_A$ are reported for different dependence levels (i.e., $\rho = 0.5$ and $\rho = 0.6$).

			$\rho = 0.5$			$\rho = 0.6$		
p	T		TP	FP	RMSE$_A$	TP	FP	RMSE$_A$
30	80	Lasso	35.3	72.1	0.053	34.9	67.2	0.054
		SCAD	30.1	5.1	0.047	29.5	5.1	0.050
		MCP	28.6	2.8	0.048	28.0	3.3	0.050
		ALasso	30.2	4.4	0.045	29.5	4.9	0.048
30	120	Lasso	36.4	48.5	0.043	36.3	46.3	0.045
		SCAD	33.7	1.3	0.028	33.3	1.6	0.031
		MCP	33.6	1.3	0.029	33.0	1.6	0.032
		ALasso	34.0	1.7	0.029	33.5	1.9	0.031
50	100	Lasso	56.1	50.4	0.043	55.7	55.7	0.043
		SCAD	52.4	5.0	0.029	51.5	5.6	0.031
		MCP	51.3	5.2	0.031	50.2	6.0	0.033
		ALasso	54.0	6.9	0.028	53.2	7.6	0.030
50	150	Lasso	58.9	28.5	0.038	59.2	53.1	0.032
		SCAD	57.2	2.2	0.018	56.9	2.7	0.019
		MCP	57.2	2.8	0.019	56.5	3.0	0.020
		ALasso	58.2	2.8	0.018	57.8	3.1	0.019

4.4. Air pollution data analysis

In recent years, the air pollution has become a more and more serious problem in mainland China. Along with this issue, the PM$_{2.5}$ index becomes a popular tool to quantify the air pollution level. It refers to the particle with aerodynamic diameters less than 2.5 micrometers. Particularly, high concentration of PM$_{2.5}$ could lead to severe clinical symptoms, such as lung morbidity, respiratory and so on [18, 5]. It yields an important question to understand the distribution and diffusion pattern of the $\mathrm{PM}_{2.5}$ spatially and temporally.

4.4.1. Data description

The $\mathrm{PM}_{2.5}$ data are collected from $p = 29$ provincial capital cities in mainland China. The daily $\mathrm{PM}_{2.5}$ index (unit: $\mu\mathrm{g/m}^3$) is reported for one year from 2015-01-01 to 2015-12-31 with $T = 365$. Figure 2 illustrates the time series of average $\mathrm{PM}_{2.5}$ over all cities in the whole year. A higher level of $\mathrm{PM}_{2.5}$ index could be captured during January to March and October to December. To visualize the spatial distribution pattern of the $\mathrm{PM}_{2.5}$, we plot the average $\mathrm{PM}_{2.5}$ of each city in the year of 2015 in Figure 3. One could observe heavier concentration levels of $\mathrm{PM}_{2.5}$ in northeastern area of China.

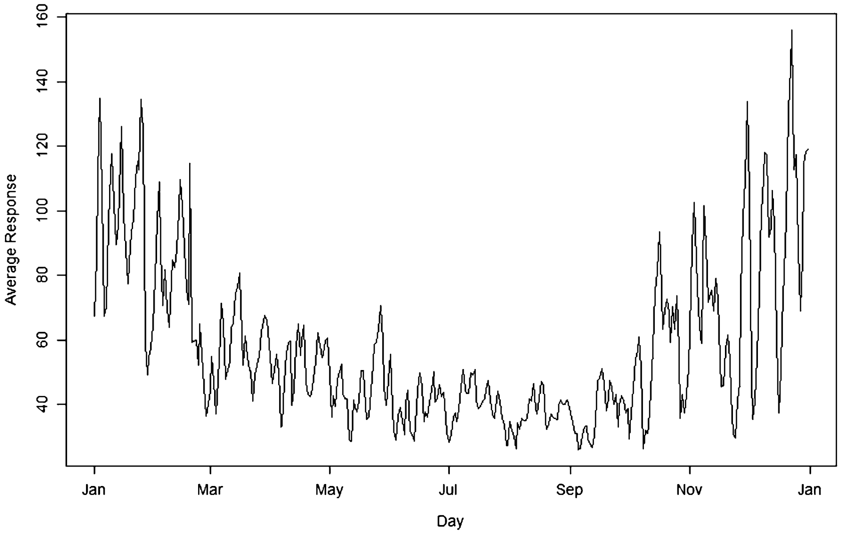

FIG 2. *Daily average $PM_{2.5}$ of $p = 29$ provincial capital cities in the year of 2015. A higher level of $PM_{2.5}$ index is observed during January to March and October to December.*

4.4.2. Model estimation and exploration

According to the concentration levels of the $\mathrm{PM}_{2.5}$ index, we split the data into 3 time periods: from January to March (PERIOD I), from April to September (PERIOD II), from October to December (PERIOD III). The PERIOD II, which mostly ranges from summer to early autumn, has lower $\mathrm{PM}_{2.5}$ levels than the other two periods. Specifically, we use the log-transformed $\mathrm{PM}_{2.5}$ levels as responses, which are centered with mean 0 for each city. Then, for each time

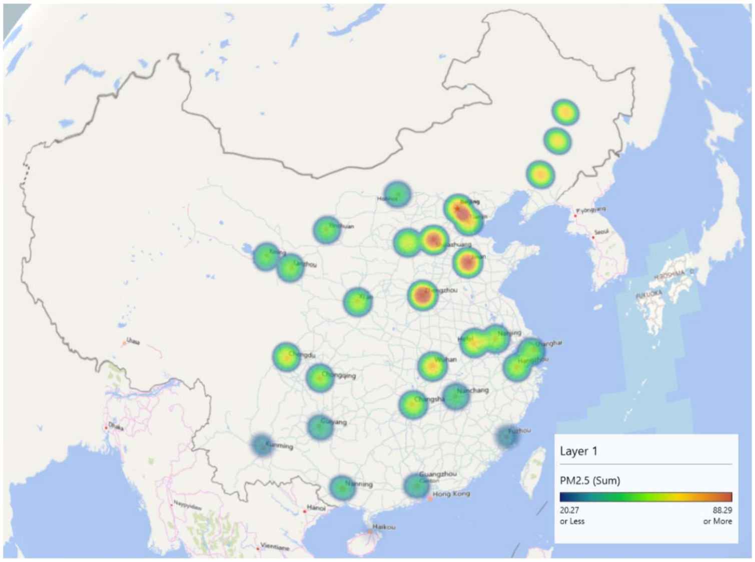

Fig 3. *Average* $PM_{2.5}$ *in the year of 2015 of* $p = 29$ *provincial capital cities. Heavier concentration levels of* $PM_{2.5}$ *are exposed in northeastern area of China.*

period, we implement the nonconcave penalized VAR model with the SCAD penalty[1] to obtain the results. Here we only consider the lag-1 response to maintain the model simplicity. To further save the computational complexity, a SIS screening procedure [9] is firstly conducted to keep 300 edges with highest absolute correlations with the response. The HBIC criterion (4.3) is used for model selection.

We visualize the estimated transition matrix $\widehat{A}_1$ using heatmap in Figure 4–6 for the three time periods. The estimated coefficients are all within $[0, 1]$. The cities in the figure are ordered roughly from north to south and east to west. Therefore the neighbouring cities are close in spatial distance to each other. First, we observe that the patterns of Period I and Period III are similar and slightly different from Period II. In Period II, we observe relatively stronger momentum effects (i.e., higher diagonal elements in $\widehat{A}_1$), and lower between-city influences (i.e., sparser off-diagonal elements in $\widehat{A}_1$). While in Period I and Period III the transition matrices exhibit denser between-city edges. This provides evidence of the air pollution diffusion effects in neighbouring cities, especially in Spring and Winter. Next, one could observe that the estimated $\widehat{A}_1$ is not symmetric. In Period I and Period III, we have more estimated nonzero elements in lower triangle of $\widehat{A}_1$ than the upper triangle. This implies that the air diffusion direction is from North to South (recall that the cities are

[1] The performances of the MCP penalization and adaptive Lasso are similar. The result of SCAD penalty is presented here for illustration.

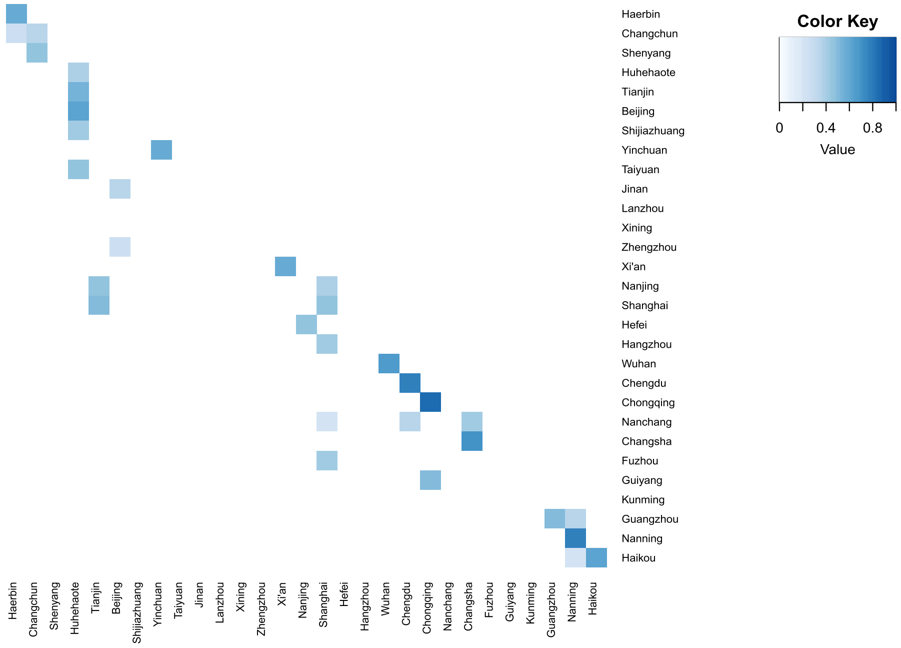

FIG 4. *The estimated transition matrix $\widehat{A}_1$ in* PERIOD I.

ordered roughly from North to South). Lastly, we further visualize the between-city coefficients of $\widehat{A}_1$ in the map; see Figure 7. By the connection pattern, one could further confirm that the influences in PERIOD I and PERIOD III are more intense than PERIOD II. In addition, it is found the between-city connections in PERIOD II are mostly local and among neighbouring cities than the other two periods.

Lastly, we compare the VAR model estimation using the nonconcave penalization methods and the Lasso penalty. The HBIC criterion (4.3) is used for model selection. We compare the performances in terms of the model sparsity, i.e., measured by number of nonzero estimates, and fitted level, i.e., measured by the fitted RMSE. The measurements are reported in Table 4. First, compared to the Lasso method, the nonconcave methods (e.g., SCAD and MCP) and the adaptive Lasso method are able to obtain VAR models with less nonzero parameters. For example, for PERIOD III, the number of nonzero estimates of Lasso is 91, while the nonzero estimates of SCAD is 43, which is much smaller than the Lasso method. In the meanwhile, although the model achieved by the nonconcave methods is more parsimonious, better fitting levels could be obtained by the nonconcave methods. For instance, the fitted RMSE for the SCAD method of PERIOD II is 0.372, while the RMSE for the Lasso method in the same period is 0.388.

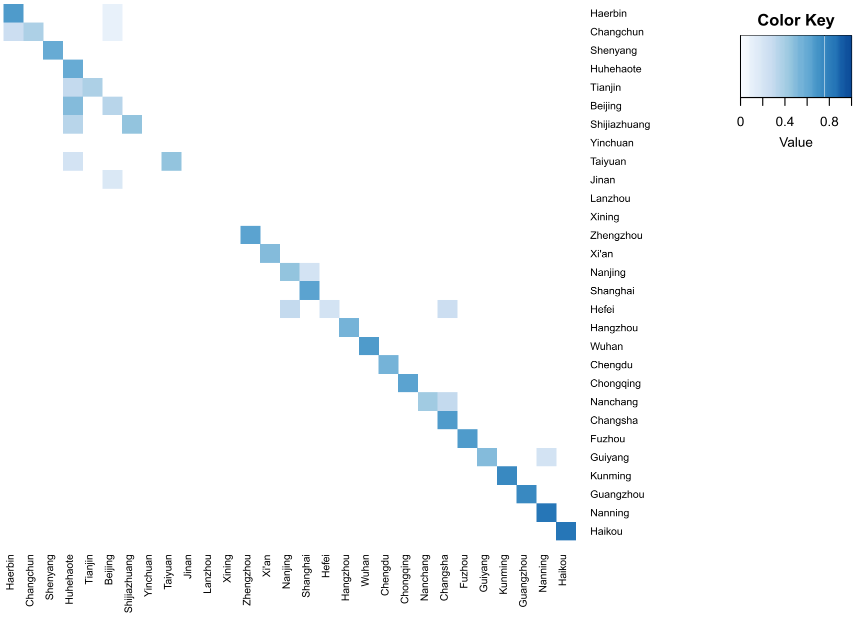

Fig 5. *The estimated transition matrix* $\widehat{A}_1$ *in* Period II.

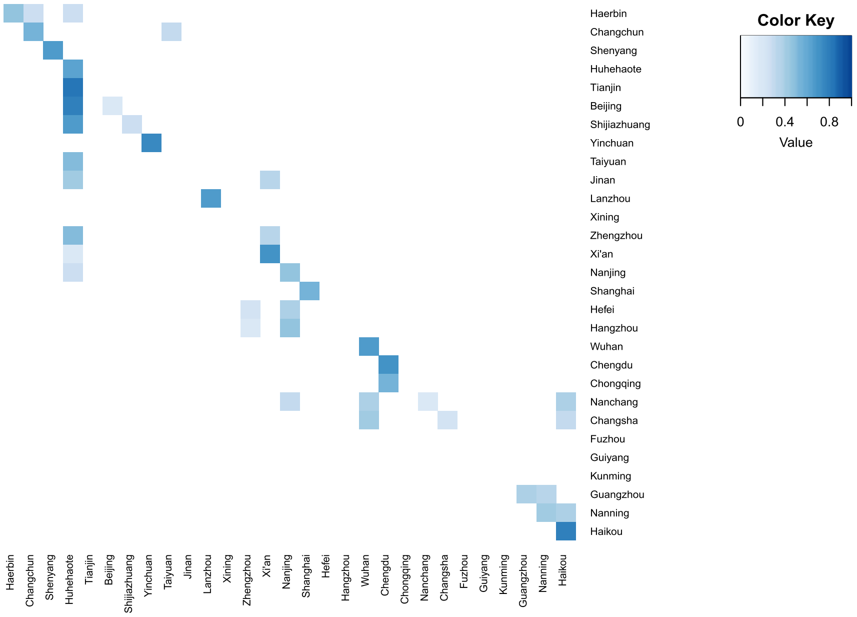

Fig 6. *The estimated transition matrix* $\widehat{A}_1$ *in* Period III.

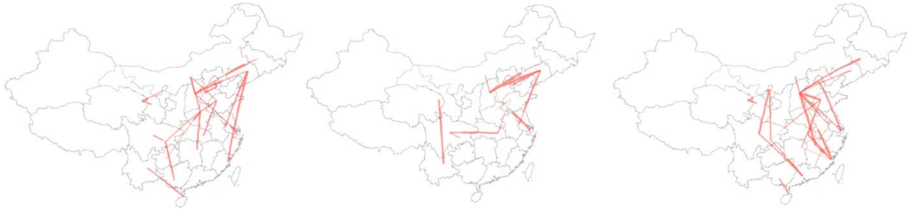

Fig 7. *The estimated edges among the $p = 29$ cities. The bolder lines indicate larger coefficients.*

Table 4

The performance of the VAR(1) model estimation with Lasso, SCAD, MCP, and adaptive Lasso penalization. The number of nonzero estimates and fitted RMSE are summarized and compared. The nonconcave penalization methods are able to achieve a parsimonious model with better fitting level.

	Number of Nonzero Estimates				Fitted RMSE			
	Lasso	SCAD	MCP	ALasso	Lasso	SCAD	MCP	ALasso
Period I	54	33	35	41	0.478	0.442	0.441	0.430
Period II	61	38	26	37	0.388	0.372	0.386	0.369
Period III	91	43	50	46	0.471	0.458	0.456	0.448

5. Conclusion

In this work, we investigate the nonconcave penalized VAR model estimation methods. Specifically, the estimation properties are established under the influence of both temporal and cross-sectional dependences. The oracle properties are given for the nonconcave penalties, where the feature dimensionality is allowed to grow exponentially fast with the sample size. Lastly, an air pollution dataset is analyzed for illustration propose, it is found that the influence patterns among different cities are highly related to the specific areas as well as the seasons in mainland China.

To conclude this work, we consider the following directions as future research topics. First, although the regularized VAR model estimation could help to recover the connection patterns among the nodes, however, the regularization level is not clear. Therefore, certain type of criterions (e.g., BIC) should be designed to select the true model efficiently. Second, as mentioned in the numerical study, before we conduct penalization estimation, screening methods could be firstly used to reduce the computational complexity. As a result, efficient screening algorithms should be proposed to better suit the complex dependence structure of the data. In addition, we could observe that from the numerical performance the adaptive Lasso is also a competitive method. It is then of great interest to study the theoretical properties of other shrinkage methods. Lastly, the penalization methods could be revised when new information of the nodes is obtained. For example, when the network relationships are observed among the nodes, new mechanisms to combine such known structure information should be further investigated.

Appendix A: Proof of derivation bounds

A.1. Useful lemmas

Lemma 1. *Assume all the conditions in Proposition 2. Define $\mathcal{K} = \{v \in \mathbb{R}^q : \|v\| \le 1, supp(v) \in \mathcal{S}\}$. In addition, let $c_{1\lambda} = \mu_{\min}(\mathcal{A})/\lambda_{\max}(\Sigma_e)$ and $c_{2\lambda} = \lambda_{\min}^{-1}(\Sigma_e)\mu_{\max}(\mathcal{A})$. Then we have*

$$P\Big\{ \sup_{v\in\mathcal{K}} |v^\top(\widehat{\Gamma}_Z - \Gamma_Z)v| > 2\eta c_{1\lambda}^{-1} \Big\} \le 2\exp\{-cN\min(\eta,\eta^2) + s\log(21)\}. \quad \text{(A.1)}$$

Further assume $s = o\{N\min(1,\omega^2)\}$. Then we have

$$P\Big\{ \sup_{v\in\mathcal{K}} |v^\top(\widehat{\Gamma}_Z - \Gamma_Z)v| > 2\eta c_{1\lambda}^{-1} \Big\} \le 2\exp\{-cN\min(\eta,\eta^2) + s\log(21)\}, \quad \text{(A.2)}$$

$$\max_i |\lambda_i(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})})| = o_p\Big(c_{2\lambda}^{-1}\Big). \quad \text{(A.3)}$$

Proof of Lemma 1. Note that $\widehat{\Gamma}_Z - \Gamma_Z = I_p \otimes (\widehat{\Gamma}_{\widetilde{X}} - \Gamma_{\widetilde{X}})$. We then have the results of Proposition 2.4 of [2] also hold for $\widehat{\Gamma}_Z - \Gamma_Z$. By Proposition 2.4 of [2], we have for any $v \in \mathbb{R}^q$, $\|v\| \le 1$, and any $\eta > 0$,

$$P\Big\{ |v^\top(\widehat{\Gamma}_Z - \Gamma_Z)v| > \eta\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})} \Big\} \le 2\exp\{-cN\min(\eta,\eta^2)\}. \quad \text{(A.4)}$$

This result is for a single $v \in \mathbb{R}^q$. Let $D = \widehat{\Gamma}_Z - \Gamma_Z$. Then we consider the set $\mathcal{K} = \{v \in \mathbb{R}^q : \|v\| \le 1, \text{supp}(v) \in \mathcal{S}\}$.

Choose $\mathcal{K}^* = \{u_1, \cdots, u_m\}$ as a 1/10-net of $\mathcal{K}$. By Lemma 3.5 of [24], $|\mathcal{K}^*| \le 21^s$. For every $v \in \mathcal{K}^*$, there exists some $u_i \in \mathcal{K}^*$ such that $\|\Delta v\| \le 1/10$, where $\Delta v = v - u_i$. Then we have

$$\gamma \overset{\text{def}}{=} \sup_{v\in\mathcal{K}} |v^\top Dv| \le \max_i |u_i^\top Du_i| + 2\sup_{v\in\mathcal{K}} |\max_j u_i^\top D\Delta v| + \sup_{v\in\mathcal{K}} |(\Delta v)^\top D(\Delta v)|.$$

Since $10(\Delta v) \in \mathcal{K}$, the third term is bounded by $\gamma/100$. Next, by Cauchy's inequality, we have

$$2\sup_{v\in\mathcal{K}} |\max_j u_i^\top D\Delta v| \le 2\sqrt{\max_i(u_i^\top Du_i)\sup_{v\in\mathcal{K}}\{(10\Delta v)^\top D(10\Delta v)\}} \le 2/10\gamma.$$

It could be concluded that γ is bounded by $\gamma \le 2\max_i u_i^\top Du_i$. Together with (A.4) we have

$$P\Big\{ \sup_{v\in\mathcal{K}} |v^\top Dv| > 2\eta\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})} \Big\} \le 2\exp\{-cN\min(\eta,\eta^2) + s\log(21)\}.$$

Further set $\eta = c_{1\lambda}/c_{2\lambda} = \omega$ and note that $\min(\omega,\omega^2) \ge \min(1,\omega^2)$, we then have (A.3). □

Lemma 2. *Assume the same conditions in Proposition 3. Then we have*

$$P\Big\{\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c}\|_{\max} \le 2c_{1\lambda}^{-1}\Big\} \ge 1 - 6\exp\{-cN + 2\log q\}. \tag{A.5}$$

Proof of Lemma 2. The proof is similar to the proof of (3.10). By Proposition 2.4 (2.9), and (2.6) of [2], we have

$$P\Big\{|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_j/N - e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S}^c)}e_j| > 3\eta\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})}\Big\} \le 6\exp\{-c^*N\min(\eta^2,\eta)\}.$$

where c^* is a finite positive constant. Therefore, it can be concluded with probability at least $1-6\exp\{-c^*N\min(\eta^2,\eta)\}$, we have $|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_j/N - e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S}^c)}e_j| \le 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A})$. It leads to

$$\begin{aligned}|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_j/N| &\le |e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S}^c)}e_j| + 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A})\\ &\le \|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} + 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A}).\end{aligned}$$

Further note that $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} \le \lambda_{\max}(\Gamma_Z) = \lambda_{\max}(\Gamma_{\widetilde{X}}(0)) \le c_{1\lambda}^{-1}$ by (2.5). By summing over all $1\le i,j\le q-s$, and letting $\eta = 1/3$, we have

$$\begin{aligned}&P\Big\{\max_{i,j}|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_j/N| \ge 2c_{2\lambda}^{-1}\Big\}\\ &\le \sum_{1\le i,j\le q-s} 6\exp(-c^*N/9) \le 6\exp\{-c^*N/9 + 2\log q\},\end{aligned}$$

which yields the result by letting $c = c^*/9$. □

Lemma 3. *Assume the conditions in Theorem 4 hold. Then we have*

$$\max_i|\lambda_i(\widehat{\Gamma}_{Ze}^{(\mathcal{S})} - \Gamma_{Ze}^{(\mathcal{S})})| = o_p\{c_{2\lambda}^{-2}\}.$$

Proof of Lemma 3. Note that $\widehat{\Sigma}_{Ze} = \widehat{\Sigma}_e \otimes \widehat{\Gamma}_{\widetilde{X}}(0) = (\widehat{\Sigma}_e \otimes I_{dp})\widehat{\Gamma}_Z$. Then we have $\max_i|\lambda_i(\widehat{\Gamma}_{Ze} - \Gamma_{Ze})| \le \max_i|\lambda_i(\widehat{\Sigma}_e - \Sigma_e)|\max_i|\lambda_i(\widehat{\Gamma}_Z - \Gamma_Z)|$. Consider the events $\mathcal{H}_1 = \{\max_i|\lambda_i(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})})| \le 2\lambda_{\max}^{-1}(\Sigma_e)c_{2\lambda}^{-2}\}$, $\mathcal{H}_2 = \{\max_i|\lambda_i(\widehat{\Sigma}_e - \Sigma_e)| \le \lambda_{\max}(\Sigma_e)\}$. On the events $\mathcal{H}_1$ and $\mathcal{H}_2$, we have $\max_i|\lambda_i(\widehat{\Gamma}_{Ze} - \Gamma_{Ze})| = o(c_{2\lambda}^{-2})$. We then prove $P(\mathcal{H}_1^c)\to 0$ and $P(\mathcal{H}_2^c)\to 0$ respectively as follows.

1. Proof of $P(\mathcal{H}_1^c)\to 0$.

By (A.2), we already have

$$P\Big\{\sup_{v\in\mathcal{K}}|v^\top(\widehat{\Gamma}_Z - \Gamma_Z)v| > 2\eta c_{1\lambda}^{-1}\Big\} \le 2\exp\{-cN\min(\eta,\eta^2) + s\log(21)\}.$$

Then set $\eta = c_{1\lambda}c_{2\lambda}^{-2}\lambda_{\max}^{-1}(\Sigma_e)$, and note $s \ll Nc_{1\lambda}c_{2\lambda}^{-2}\lambda_{\max}^{-1}(\Sigma_e)$, then we have $P(\mathcal{H}_1^c)\to 0$.

2. Proof of $P(\mathcal{H}_2^c)\to 0$.

Define $\widehat{B}$ is the matrix form estimator obtained from $\widehat{\beta}$. Then we have $\widehat{\Sigma}_e = N^{-1}(\mathcal{Y}-\mathcal{X}\widehat{B})^\top(\mathcal{Y}-\mathcal{X}\widehat{B}) = N^{-1}\mathcal{E}^\top\mathcal{E}+N^{-1}\{\mathcal{X}(B-\widehat{B})\}^\top\mathcal{E}+\mathcal{E}^\top\{\mathcal{X}(B-\widehat{B})\}+(B-\widehat{B})^\top\mathcal{X}^\top\mathcal{X}(B-\widehat{B})$. Then it suffices to show that

$$\sigma_{\max}(N^{-1}\mathcal{E}^\top\mathcal{E}-\Sigma_e) = o_p(\lambda_{\max}(\Sigma_e)), \tag{A.6}$$

$$\sigma_{\max}(N^{-1}\{\mathcal{X}(B-\widehat{B})\}^\top\mathcal{E}) = o_p(\lambda_{\max}(\Sigma_e)), \tag{A.7}$$

$$\lambda_{\max}\{N^{-1}(B-\widehat{B})^\top\mathcal{X}^\top\mathcal{X}(B-\widehat{B})\} = o_p(\lambda_{\max}(\Sigma_e)). \tag{A.8}$$

Note that the second one (A.7) could be implied by (A.6) and (A.8). We only show the proofs of (A.6) and (A.8) respectively as follows.

First by (A.1) we have

$$P\Big\{\sup_{v\in\mathbb{R}^p}|v^\top(N^{-1}\mathcal{E}^\top\mathcal{E}-\Sigma_e)v| > 2\eta\lambda_{\max}(\Sigma_e)\Big\}$$
$$\le 2\exp\{-cN\min(\eta,\eta^2)+p\log(21)\}.$$

By letting $\eta=1/2$ and $p\ll N$, the results can be obtained.

Next, we have $\lambda_{\max}\{N^{-1}(B-\widehat{B})^\top\mathcal{X}^\top\mathcal{X}(B-\widehat{B})\} \le (\widehat{\beta}-\beta)^\top(N^{-1}Z^\top Z)(\widehat{\beta}-\beta) = (\widehat{\beta}_{\mathcal{S}}-\beta_{\mathcal{S}})^\top\widehat{\Gamma}_Z^{(\mathcal{S})}(\widehat{\beta}_{\mathcal{S}}-\beta_{\mathcal{S}}) \le \lambda_{\max}(\widehat{\Gamma}_Z^{(\mathcal{S})})\|\widehat{\beta}_{\mathcal{S}}-\widehat{\beta}_{\mathcal{S}}\|^2 \le 2\alpha_\lambda c_{1\lambda}^{-1}s/N$ by setting $\eta=1$ in (A.2). Further note $s\ll N\alpha_\lambda^{-1}c_{1\lambda}\lambda_{\max}(\Sigma_e)$, then we have (A.8). □

A.2. Proof of Proposition 1

The proof is basically the same as Theorem 1 of [10]. For the completeness, we state the basic idea as follows. First we give the necessary conditions. Define the unpenalized objective function as $Q(\beta) = -\beta^\top\widehat{\gamma}_Z+2^{-1}\beta^\top\widehat{\Gamma}_Z\beta$, where $\widehat{\gamma}_Z = (I_p\otimes\mathcal{X}^\top)Y/N$, and $\widehat{\Gamma}_Z = Z^\top Z/N = (I\otimes\mathcal{X}^\top\mathcal{X}/N)$. Then we have

$$\dot{Q}(\beta) = \widehat{\Gamma}_Z\beta-\widehat{\gamma}_Z, \quad \ddot{Q}(\beta) = \widehat{\Gamma}_Z$$

By the classical optimization theory, if $\widehat{\beta} = (\widehat{\beta}_1,\cdots,\widehat{\beta}_q)^\top$ is the local minimizer of the penalized objective function (3.2), then the Karush-Kuhn-Tucker (KKT) conditions hold. Namely,

$$\widehat{\Gamma}_Z\widehat{\beta}-\widehat{\gamma}_Z+\lambda_N v = 0, \tag{A.9}$$

where $v = (v_1,\cdots,v_q)^\top \in \mathbb{R}^q$ with $v_j = \overline{\rho}(\widehat{\beta}_j)$ for $\widehat{\beta}_j \ne 0$ and $v_j \in [-\rho'(0+),\rho'(0+)]$ for $\widehat{\beta}_j = 0$. Note that $\widehat{\beta}$ is also a local minimizer of of (3.2) on the subspace $\{\beta\in\mathbb{R}^q : \beta_{\mathcal{S}^c} = \mathbf{0}\}$. From the second order condition, we have

$$\lambda_{\min}(\widehat{\Gamma}_Z^{(\mathcal{S})}) \ge \lambda_N\kappa(\rho;\widehat{\beta}_{\mathcal{S}}),$$

where $\kappa(\rho;\widehat{\beta}_{\mathcal{S}})$ is defined in (3.3). Equivalently, (A.9) can be expressed as

$$Z_{\mathcal{S}}^\top(Y-Z\widehat{\beta})-N\lambda_N\overline{\rho}(\widehat{\beta}_{\mathcal{S}}) = 0,$$

$$(N\lambda_N)^{-1}\|Z_{\mathcal{S}^c}^\top(Y - X\widehat{\beta})\|_\infty \le \rho'(0+),$$

which correspond to (3.4) and (3.5).

Next, we show the sufficient conditions. First, constrain the penalized objective function (3.2) on subspace defined by $\mathcal{S}$ as $\mathcal{B} = \{\beta \in \mathbb{R}^q : \beta_{\mathcal{S}^c} = \mathbf{0}\}$. From (3.6), we could conclude that the penalized objective function is strictly convex in a ball $\mathcal{N}_0$ in the subspace $\mathcal{B}$ centered at $\widehat{\beta}$. Along with (3.4), this implies that $\widehat{\beta}$ is a critical point of $Q_p(\beta)$ and also the unique minimizer of $Q_p(\beta)$.

Then we only need to prove that the sparse vector $\widehat{\beta}$ is ia strict local minimizer of $Q_p(\beta)$ on the space $\mathbb{R}^q$. To this end, define a sufficiently small ball $\mathcal{N}_1$ in $\mathbb{R}^q$ centered at $\widehat{\beta}$ such that $\mathcal{N}_1 \cap \mathcal{B} \subset \mathcal{N}_0$. It then suffices to show that $Q_p(\widehat{\beta}) < Q_p(\gamma_1)$ for any $\gamma_1 \in \mathcal{N}_1 \backslash \mathcal{N}_0$. Let γ_2 be the projection of γ_1 on the subspace $\mathcal{B}$. Then we have $\gamma_2 \in \mathcal{N}_0$, which implies $Q_p(\widehat{\beta}) < Q_p(\gamma_2)$ if $\gamma_2 \neq \widehat{\beta}$ due to that $\widehat{\beta}$ is the strict minimizer of $Q_p(\beta)$ in $\mathcal{N}_0$. Therefore it suffices to show that $Q_p(\gamma_2) < Q_p(\gamma_1)$.

Using the mean-value theorem, we have

$$Q_p(\gamma_2) - Q_p(\gamma_1) = \dot{Q}_p^\top(\gamma_0)(\gamma_2 - \gamma_1), \tag{A.10}$$

where γ_0 lies on the line joining γ_1 and γ_2. For $j \in \mathcal{S}$, we have $\gamma_{2j} - \gamma_{1j} = 0$. For $j \notin \mathcal{S}$, we have the sign of γ_{0j} is the same as γ_{1j}. Consequently, the right hand of (A.10) can be written as

$$N^{-1}\Big\{Z_{\mathcal{S}^c}^\top\big(Y - Z\gamma_0\big)\Big\}^\top \gamma_{1,\mathcal{S}^c} - \lambda_N \sum_{j \neq \mathcal{S}} \rho'(|\gamma_{0j}|)|\gamma_{1,j}|. \tag{A.11}$$

Since $\gamma_1 \in \mathcal{N}_1 \backslash \mathcal{N}_0$, we have $\gamma_{1,\mathcal{S}^c} \neq \mathbf{0}$.

By condition (C2), $\rho'(t)$ is decreasing in $t \in [0, \infty)$. By (3.5) and the continuity of $\rho'(t)$, there exists $\delta > 0$ and for any $\beta \in \mathcal{N}_\delta$ (where $\mathcal{N}_\delta$ is a ball in $\mathbb{R}^q$ centered at $\widehat{\beta}$ with radius δ), such that

$$\Big\|(N\lambda_N)^{-1}Z_{\mathcal{S}^c}^\top(Y - Z\beta)\Big\|_\infty < \rho'(\delta).$$

Next we shrink the radius of the ball $\mathcal{N}_1$ to be less than δ such that $|\gamma_{0j}| \le |\gamma_{1j}| < \delta$ for $j \notin \mathcal{S}$. Since $\gamma_0 \in \mathcal{N}_1$, we have (A.11) is strictly less than $\lambda_N \rho'(\delta)\|\gamma_{1,\mathcal{S}^c}\|_1 - \lambda_N \rho'(\delta)\|\gamma_{1,\mathcal{S}^c}\|_1 = 0$, where the second term is due to $\rho'(|\gamma_{0j}|) > \rho'(\delta)$ by using the monotonicity of $\rho'(\cdot)$. This proves the result.

A.3. Proof of Proposition 2

Define $\widehat{\Gamma}_{\widetilde{X}}(0) = \mathcal{X}^\top\mathcal{X}/N$ and $\Gamma_{\widetilde{X}}(0) = E(\widehat{\Gamma}_{\widetilde{X}}(0))$. Recall $\widehat{\Gamma}_Z \overset{\text{def}}{=} Z^\top Z/N = I_p \otimes \widehat{\Gamma}_{\widetilde{X}}(0)$. One could easily verify that $\Gamma_Z = I_p \otimes \Gamma_{\widetilde{X}}(0)$.

Proof of (3.7). Let $D = \widehat{\Gamma}_Z - \Gamma_Z$. Set $\eta = \omega/4$ (recall that $\omega = c_{1\lambda}/c_{2\lambda}$, where $c_{1\lambda} = \lambda_{\max}^{-1}(\Sigma_e)\mu_{\min}(\mathcal{A})$ and $c_{2\lambda} = \lambda_{\min}^{-1}(\Sigma_e)\mu_{\max}(\mathcal{A})$). By (A.1) in Lemma 1, we then have at least with probability $1 - 2\exp\{-cN\min(1,\omega^2) + s\log(21)\}$, $|v^\top Dv| \leq 1/2\lambda_{\min}(\Sigma_e)/\mu_{\max}(\mathcal{A})$. Then we have

$$(v^\top \widehat{\Gamma}_Z v)/N \geq v^\top \Gamma_Z v - 1/2\lambda_{\min}(\Sigma_e)/\mu_{\max}(\mathcal{A}).$$

Since $\Gamma_Z = I_p \otimes \Gamma_{\widetilde{X}}(0)$, we have $v^\top \Gamma_Z(0)v \geq \lambda_{\min}(\Gamma_{\widetilde{X}}(0))$. By Proposition 2.3 of [2] and (2.6) we have

$$\lambda_{\min}(\Gamma_{\widetilde{X}}(0)) \geq \lambda_{\min}(\Upsilon_X) \geq \frac{\lambda_{\min}(\Sigma_e)}{\mu_{\max}(\mathcal{A})}.$$

Combining the results, we have

$$P\Big\{\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) > \frac{N}{2}\frac{\lambda_{\min}(\Sigma_e)}{\mu_{\max}(\mathcal{A})}\Big\} \geq 1 - 2\exp\{-cN\min(1,\omega^2) + s\log 21\}.$$

□

Proof of (3.8). Similarly, for any $v \in \mathbb{R}^q$, by (A.4) we have

$$P\Big\{v^\top \widehat{\Gamma}_Z v > (1+\eta)\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})}\Big\} \leq 2\exp\{-cN\min(\eta,\eta^2)\}.$$

By the same technique in proof of (3.7) and letting $\eta = 1$, the results can be obtained. □

A.4. Proof of Proposition 3

Proof of (3.9). Note that for an arbitrary matrix $M \in \mathbb{R}^{s\times s}$, we have $\|M\|_\infty \leq \sqrt{s}\lambda_{\max}(M)$. Consequently, (3.9) can be directly obtained from (3.7). □

Proof of (3.10). Recall $\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}$ be the submatrix of Γ_Z with row and column index in $\mathcal{S}^c$ and $\mathcal{S}$ respectively. Without loss of generality, we assume $\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} > 0$. By Proposition 2.4 (2.9), and (2.6) of [2], we have

$$P\Big\{|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})e_j/N - e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}e_j| > 3\eta\frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})}\Big\} \leq 6\exp\{-cN\min(\eta^2,\eta)\}.$$

Therefore with probability at least $1 - 6\exp\{-cN\min(\eta^2,\eta)\}$, we have

$$|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})e_j/N - e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}e_j| \leq 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A}).$$

It leads to $|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})e_j/N| \leq |e_i^\top\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}e_j| + 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A}) \leq \|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} + 3\eta\lambda_{\max}(\Sigma_e)\mu_{\min}^{-1}(\mathcal{A})$. By summing over all $1 \leq i \leq q-s$ and $1 \leq j \leq s$, and letting $\eta = \nu$, we have

$$P\Big\{\frac{1}{N}\max_{i,j}|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})e_j| \geq \|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max} + \max\Big(\|\Gamma_Z^{(\mathcal{S}^c,\mathcal{S})}\|_{\max}, \frac{\lambda_{\max}(\Sigma_e)}{\mu_{\min}(\mathcal{A})N^\delta}\Big)\Big\}$$
$$\leq \sum_{1\leq i\leq q-s, 1\leq j\leq s} 6\exp\{-cN\min(\nu^2,\nu)\} \leq 6\exp\{-cN\min(\nu^2,\nu) + \log q + \log s\},$$

which yields the result. □

A.5. Proof of Proposition 4

Note that for the any vector ξ, $\|\xi\|_\infty = \|\xi\|_{\max}$. Therefore we only show the results hold for the maximum norm of $\xi_{\mathcal{S}}$ and $\xi_{\mathcal{S}^c}$.

By Proposition 2.4 (2.11) of [2], we have

$$P\Big\{|e_i^\top(\mathcal{X}^\top\mathcal{E})e_j| > N\mathbb{Q}(\beta,\Sigma_e)\eta\Big\} \le 6\exp\Big(-c_1N\min(\eta,\eta^2)\Big)$$

First, by letting $\eta = c_1^{-1/2}\sqrt{\log N/N}$, we have $P\{|\xi_j| > c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}\} \le 6\exp(-\log N) = 6N^{-1}$ for $j \in \mathcal{S}$. Therefore, we have

$$\begin{aligned}&P(\|\xi_{\mathcal{S}}\|_\infty > c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N})\\ &\le \sum_{j\in\mathcal{S}} P(|\xi_j| > c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}) \le 6s/N.\end{aligned}$$

This proves (3.12).

Next, by letting $\eta = c_1^{-1/2}u_N/\sqrt{N}$, we then have

$$P\{|\xi_j| > \mathbb{Q}(\beta,\Sigma_e)u_N\sqrt{N}\} \le 6\exp\{-N^{2\alpha}\log N\}$$

for $j \in \mathcal{S}^c$. Similarly, we have $P(\|\xi_{\mathcal{S}^c}\|_\infty > \mathbb{Q}(\beta,\Sigma_e)u_N\sqrt{N}) \le$

$$\sum_{j\in\mathcal{S}^c} P(|\xi_j| > \mathbb{Q}(\beta,\Sigma_e)u_N\sqrt{N}) \le 6(q-s)\exp\{-N^{2\alpha}\log N\}.$$

This proves (3.13).

Appendix B: Proof of the main results

B.1. Proof of Theorem 1

Recall $\xi = Z^\top(Y - Z\beta) = \text{vec}(\mathcal{X}^\top\mathcal{E})$ and $u_N = c_1^{-1/2}N^\alpha(\log N)^{1/2}$. Consider the events

$$\begin{aligned}\mathcal{H}_1 &= \Big\{\|\xi_{\mathcal{S}}\|_\infty \le c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}\Big\},\\ \mathcal{H}_2 &= \Big\{\|\xi_{\mathcal{S}^c}\|_\infty \le \mathbb{Q}(\beta,\Sigma_e)u_N\sqrt{N}\Big\},\\ \mathcal{H}_3 &= \Big\{\|(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty \le c_\mu\Big\},\ \mathcal{H}_4 = \Big\{\|Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}\|_{\max} \le c_\Gamma\Big\},\\ \mathcal{H}_5 &= \Big\{\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) > \frac{N}{2c_{2\lambda}}\Big\}.\end{aligned}$$

By Bonferroni's inequality, (3.6), and Proposition 3, 4, $P(\mathcal{H}_1 \cap \mathcal{H}_2 \cap \mathcal{H}_3 \cap \mathcal{H}_4 \cap \mathcal{H}_5) \ge 1 - \sum_{j=1}^5 P(\mathcal{H}_j^c) = 1 - 6[s/N + \exp(-c_1N^{2\alpha}\log N + \log q) + 2\exp\{-cN\min(1,\omega^2) + s\log 21\} + \exp\{-cN\min(\nu,\nu^2) + \log q + \log s\}]$. By

$\log q = O(N^{2\alpha}) = o(N^{1-2\delta})$ in Condition (C3) and $s = o(N \min(1, \omega^2))$ in Theorem 1, we have $P(\mathcal{H}_j^c) \to 0$ for $1 \le j \le 5$ as $N \to \infty$. Consequently, in the following, we will show that under the event $\mathcal{H}_1 \cap \mathcal{H}_2 \cap \mathcal{H}_3 \cap \mathcal{H}_4 \cap \mathcal{H}_5$, there exists a solution $\widehat{\beta}_{\mathcal{S}}$ satisfying $\mathrm{sgn}(\widehat{\beta}_{\mathcal{S}}) = \mathrm{sgn}(\beta_{\mathcal{S}})$ and $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|_\infty = O(\mathbb{Q}(\beta, \Sigma_e) N^{-\gamma} \log N)$.

Step 1: Existence of Solution to (3.4)

We first prove that for sufficiently large N, there exists a solution $\widehat{\beta}_{\mathcal{S}}$ insider the cube

$$\mathcal{N} = \big\{\theta \in \mathbb{R}^s : \|\theta - \beta_{\mathcal{S}}\| = \nu_N\big\}.$$

where $\nu_N = \mathbb{Q}(\beta, \Sigma_e) N^{-\gamma} \log N$. For any $\theta = (\theta_1, \cdots, \theta_s)^\top \in \mathcal{N}$, we have

$$\min_{j \in \{1, \cdots, s\}} |\theta_j| \ge \min_{j \in \mathcal{S}} |\beta_{0,j}| - d_N = d_N \tag{B.1}$$

and $\mathrm{sgn}(\theta) = \mathrm{sgn}(\beta_{\mathcal{S}})$ due to that $d_N \ge \mathbb{Q}(\beta, \Sigma_e) N^{-\gamma} \log N$. Let $\eta = N\lambda_N \overline{\rho}(\theta)$. By the monotone condition of $\rho'(t)$ and (B.1) we have $\|\eta\|_\infty \le N\lambda_N \rho'(d_N)$. Along with the definition of $\mathcal{H}_1$, it yields

$$\|\xi_{\mathcal{S}} - \eta\|_\infty \le c_1^{-1/2} \mathbb{Q}(\beta, \Sigma_e) \sqrt{N \log N} + N\lambda_N \rho'(d_N) \tag{B.2}$$

Define the vector valued function

$$\begin{aligned}
&\gamma(\theta) = Z^\top Z_{\mathcal{S}} \theta, \quad \theta \in \mathbb{R}^s, \\
&\Psi(\theta) = \gamma_{\mathcal{S}}(\theta) - \gamma_{\mathcal{S}}(\beta_{\mathcal{S}}) - (\xi_{\mathcal{S}} - \eta), \quad \theta \in \mathbb{R}^s
\end{aligned}$$

In addition, let

$$\overline{\Psi}(\theta) = (Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1} \Psi(\theta) = (\theta - \beta_{\mathcal{S}}) + u, \tag{B.3}$$

where $u = -(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}(\xi_{\mathcal{S}} - \eta)$. By (3.9) and (B.2), we then have

$$\begin{aligned}
\|u\|_\infty &\le \|(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty \|\xi_{\mathcal{S}} - \eta\|_\infty \\
&= O\big\{c_{2\lambda} \mathbb{Q}(\beta, \Sigma_e) s^{1/2} N^{-1/2} \sqrt{\log N} + s^{1/2} c_{2\lambda} \lambda_N \rho'(d_N)\big\}
\end{aligned}$$

By conditions of Theorem 1, we have $s^{1/2} c_{2\lambda} = o(N^{1/2-\gamma} (\log N)^{1/2})$. Thus for the first term we should have Furthermore, by Condition (C4), we have $|s^{1/2} c_{2\lambda} \lambda_N \rho'(d_N)| = o\{\mathbb{Q}(\beta_{\mathcal{S}}, \Sigma_e) N^{-\gamma} \log N\} = o(\nu_N)$. Therefore we have $\|u\|_\infty = o(\nu_N)$. If $(\theta - \beta_{\mathcal{S}})_j = \nu_N$, we have $\overline{\Psi}_j(\theta) \ge \nu_N - \|u\|_\infty \ge 0$. If $(\theta - \beta_{\mathcal{S}})_j = -\nu_N$, we have $\overline{\Psi}_j(\theta) \le -\nu_N + \|u\|_\infty \le 0$. By the Miranda's existence theorem, there exists a solution $\widehat{\beta}_{\mathcal{S}}$ in $\mathcal{N}$ such that $\overline{\Psi}(\theta) = \mathbf{0}$, which is also the solution of $\Psi(\theta) = \mathbf{0}$.

Step 2: Verification of Condition (3.5)

Let $\widehat{\beta} \in \mathbb{R}^q$ with $\widehat{\beta} \in \mathcal{N}$ being a solution to (3.4) and $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$. We next show that $\widehat{\beta}$ satisfies (3.5). Note that

$$\begin{aligned}
z &= (N\lambda_N)^{-1} Z_{\mathcal{S}^c}^\top (Y - Z_{\mathcal{S}} \widehat{\beta}) \\
&= (N\lambda_N)^{-1} \Big\{\xi_{\mathcal{S}^c} - Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}} (\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}})\Big\}.
\end{aligned}$$

On the event $\mathcal{H}_2$, we have $\|(N\lambda_N)^{-1}\xi_{\mathcal{S}^c}\|_\infty = O(N^{-1/2}\lambda_N^{-1}u_N\mathbb{Q}(\beta,\Sigma_e))$. Recall that $u_N = c_1^{-1/2}N^\alpha(\log N)^{1/2}$. Together with Condition (C4) we have $\|(N\lambda_N)^{-1}\xi_{\mathcal{S}^c}\|_\infty = O((\log N)^{-1}) = o(1)$. Recall that $\widehat{\beta}_{\mathcal{S}}$ solves the equation that $\overline{\Psi}(\theta) = \mathbf{0}$. As a result, we have

$$\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}} = (Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}(\xi_{\mathcal{S}} - \eta).$$

Consequently, by (B.2) we have

$$\begin{aligned}\|z\|_\infty &\le o(1) + (N\lambda_N)^{-1}\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty\|\xi_{\mathcal{S}} - \eta\|_\infty\\ &\le o(1) + (N\lambda_N)^{-1}\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty\\ &\quad\times\Big\{c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N} + N\lambda_N\rho'(d_N)\Big\}\end{aligned}$$

Furthermore, we have $\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty \le c_\mu c_\Gamma$ on events $\mathcal{H}_3$ and $\mathcal{H}_4$. Further by Condition (C3), we have $c_\mu c_\Gamma \le \min\{C\rho'(0+)/\rho'(d_N), O(N^\alpha)\}$. Therefore for the second and third terms in above $\|z\|_\infty$ we have by Condition (C3)

$$\begin{aligned}&(N\lambda_N)^{-1}\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty\Big\{c_1^{-1/2}\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}\Big\}\\ &\qquad\le (N\lambda_N)^{-1}O(N^\alpha\mathbb{Q}(\beta,\Sigma_e)\sqrt{N\log N}) = o((\log N)^{-1/2}) = o(1)\\ &\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})^{-1}\|_\infty\rho'(d_N) \le C\rho'(0) < \rho'(0+)\end{aligned}$$

where the first inequality is due to $N\lambda_N = N^{\alpha+1/2}\log N\mathbb{Q}(\beta,\Sigma_e)$ of Condition (C4), and the second is due to Condition (C3) for sufficiently large N.

Step 3: Verification of (3.6)

Lastly, (3.6) is guaranteed by event $\mathcal{H}_5$ and Condition (C5). This completes the proof of Theorem 1.

B.2. Proof of Theorem 2

Denote the objective function as $Q(\beta) = -2\beta^\top\widehat{\gamma} + \beta^\top\widehat{\Gamma}\beta + \sum_{j=1}^q p_\lambda(|\beta_j|)$. To prove the result, it suffices to show that under the given regularity conditions, there exists a strict local maximizer $\widehat{\beta}$ of $Q(\beta)$ such that (1) $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$ with probability tending to 1 as $N \to \infty$ (i.e., sparsity), and (2) $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\| = O_p(\alpha_\lambda\sqrt{s/N})$ (i.e., $\alpha_\lambda\sqrt{s/N}$-consistency).

Step 1: (Consistency) We first constrain $Q(\beta)$ on the s-dimensional subspace $\{\beta \in \mathbb{R}^q : \beta_{\mathcal{S}^c} = \mathbf{0}\}$. The constrained penalized likelihood is given by

$$\overline{Q}(\theta) = -2\theta^\top\widehat{\gamma}_{\mathcal{S}} + \theta^\top\widehat{\Gamma}_{\mathcal{S}}\theta + \sum_{j=1}^q p_\lambda(|\theta_j|), \tag{B.4}$$

where $\theta = (\theta_1, \cdots, \theta_s)^\top \in \mathbb{R}^s$. We then show there exists a strict local minimizer $\widehat{\beta}_{\mathcal{S}}$ of $Q(\theta)$ such that $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\| = O_p(\alpha_\lambda\sqrt{s/N})$. To this end, define

$$\mathcal{O}_1 = \Big\{\overline{Q}(\beta_{\mathcal{S}}) < \min_{\theta\in\partial N_\tau}\overline{Q}(\theta)\Big\}, \tag{B.5}$$

where ∂N_τ denotes the boundary of a closed set $N_\tau = \{\theta \in \mathbb{R}^s : \|\theta - \beta_{\mathcal{S}}\| \leq \alpha_\lambda\sqrt{s/N}\tau\}$ and $\tau \in (0,\infty)$. It suffices to show $P(\mathcal{O}_1)$ converges to 1 as $N \to \infty$. To this end, we need to analyze the function $\overline{Q}(\theta)$ on the boundary ∂N_τ.

Let N be sufficiently large that $\alpha_\lambda\sqrt{s/N}\tau \leq d_N$ since $d_N \gg \alpha_\lambda\sqrt{s/N}$ by Condition (C6). By Taylor's expansion we have

$$\overline{Q}(\theta) - \overline{Q}(\beta_{\mathcal{S}}) = -2(\theta - \beta_{\mathcal{S}})^\top v + (\theta - \beta_{\mathcal{S}})^\top D(\theta - \beta_{\mathcal{S}}), \tag{B.6}$$

where

$$v = \widehat{\gamma}_{\mathcal{S}} - N^{-1}Z_{\mathcal{S}}^\top Z_{\mathcal{S}}\beta_{\mathcal{S}} - \overline{p}_{\lambda_N}(\beta_{\mathcal{S}}), \quad D = N^{-1}Z_{\mathcal{S}}^\top Z_{\mathcal{S}} + \operatorname{diag}\{p''_\lambda(|\theta^*|)\},$$

where $\theta^* = Z_{\mathcal{S}}\theta^*$, and θ^* lies on the line segment joining θ and $\beta_{\mathcal{S}}$. Note that the second order derivative of the penalty p_λ does not necessarily exist. One could verify that the second part of D can be replaced by a diagonal matrix with maximum absolute element bounded by $\lambda_N\kappa_0$. Recall that for any $\theta \in \partial N_\tau$, we have $\|\theta - \beta_{\mathcal{S}}\| = \alpha_\lambda\sqrt{s/N}\tau$. Since $\theta^* \in \partial N_\tau$ by Condition (C6), we then have $\theta^* \in \mathcal{N}_0$, where $\mathcal{N}_0$ is defined in Condition (C4) as $\mathcal{N}_0 = \{\theta \in \mathbb{R}^s : \|\theta - \beta_{\mathcal{S}}\|_\infty \leq d_N\}$. Consider the event $\mathcal{O}_2 = \{\lambda_{\min}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}/N) > c_{2\lambda}^{-1}/2\}$. By (3.7), we have $P(\mathcal{O}_2) \geq 1 - 2\exp\{-cN\min(1,\omega^2) + s\log 21\}$. Consequently on $\mathcal{O}_2$ we have $\lambda_{\min}(D) \geq 2^{-1}c_{2\lambda}^{-1} - \lambda_N\kappa_0 \geq c_{2\lambda}^{-1}/4$. Thus by (B.6), we have

$$\min_{\theta\in\partial N_\tau} \overline{Q}(\theta) - \overline{Q}(\beta_{\mathcal{S}}) \geq -2\alpha_\lambda\sqrt{s/N}\tau\|v\| + \alpha_\lambda^2\tau^2(s/N)c_{2\lambda}^{-1}/4 \tag{B.7}$$

Consequently, we have

$$P(\mathcal{O}_1) \geq P\Big(\|v\|^2 < \frac{c_{2\lambda}^{-2}\alpha_\lambda^2 s\tau^2}{64N}\Big) \geq 1 - \frac{64NE\|v\|^2}{c_{2\lambda}^{-2}\alpha_\lambda^2 s\tau^2} \tag{B.8}$$

by the Markov inequality. Further consider the event $\mathcal{O}_3 = \{\lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) < 2Nc_{1\lambda}^{-1}\}$. We have $P(\mathcal{O}_3) \geq 1 - 2\exp\{-cN + s\log 21\} \to 1$. It can be derived that on the event $\mathcal{O}_3$ that

$$\begin{aligned}
E\|v\|^2 &= N^{-2}E\|Z_{\mathcal{S}}^\top(Y - Z_{\mathcal{S}}\beta_{\mathcal{S}})\|^2 + \|\overline{p}_\lambda(\beta_{\mathcal{S}})\|^2\\
&\leq N^{-2}E\{\operatorname{tr}(Z_{\mathcal{S}}^\top(\Sigma_e \otimes I_N)Z_{\mathcal{S}})\} + sp'_\lambda(d_N)^2\\
&\leq N^{-2}s\lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})\lambda_{\max}(\Sigma_e) + sp'_\lambda(d_N)^2 = O(N^{-1}s\alpha_\lambda c_{2\lambda}^{-1})
\end{aligned}$$

due to that $p'_\lambda(t)$ is decreasing in $t \in [0,\infty)$ and $p'_{\lambda_N}(d_N) = O(N^{-1/2}\alpha_\lambda c_{2\lambda}^{-1/2})$ in Condition (C6). Therefore, one could derive $P(\mathcal{O}_1) \geq 1 - O(\tau^{-2}\alpha_\lambda^{-1}c_{2\lambda})$, which leads to the result that $\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\| = O_p(\alpha_\lambda\sqrt{s/N})$.

Step 2. (Sparsity) Let $\widehat{\beta} \in \mathbb{R}^q$ with $\widehat{\beta}_{\mathcal{S}} \in N_\tau \subset \mathcal{N}_0$ and $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$. It suffices to show that $\widehat{\beta}$ is a strict local minimizer of $Q(\beta)$ on the space of $\mathbb{R}^q$. From (3.5), it suffices to check that $\|z\|_\infty < \rho'(0+)$, where

$$z = (N\lambda)^{-1}Z_{\mathcal{S}^c}^\top(Y - Z_{\mathcal{S}}\widehat{\beta}_{\mathcal{S}}) = (N\lambda)^{-1}\Big\{\xi_{\mathcal{S}^c} - Z_{\mathcal{S}^c}^\top(Z_{\mathcal{S}}\widehat{\beta}_{\mathcal{S}} - Z_{\mathcal{S}}\beta_{\mathcal{S}})\Big\}$$

where $\xi = Z^\top(Y - Z\beta)$. We deal with the two parts in z respectively. First we consider the event

$$\mathcal{T}_1 = \Big\{\|\xi_{\mathcal{S}^c}\|_\infty \le u_N\sqrt{N}\mathbb{Q}(\beta, \Sigma_e)\Big\}, \tag{B.9}$$

where $u_N = c_1^{-1/2}N^{\alpha/2}\sqrt{\log N}$. On the event $\mathcal{T}_1$,

$$(N\lambda_N)^{-1}\|\xi_{\mathcal{S}^c}\|_\infty = O(N^{-1/2}\lambda_N^{-1}u_N\alpha_\lambda) = o(1)$$

by $\lambda_N \gg N^{\alpha/2-1/2}(\log N)^{1/2}\alpha_\lambda$ in Condition (C6) and

$$P(\mathcal{T}_1) \ge 1 - 6\exp(-N^\alpha \log N + \log q) \to 1$$

by $\log q = O(N^\alpha)$ in Condition (C4). For the second part we have

$$\begin{aligned}&(N\lambda_N)^{-1}\|(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}})(\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}})\|_\infty\\ &\le (N\lambda_N)^{-1}\Big\{\max_i |e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}Z_{\mathcal{S}}^\top Z_{\mathcal{S}^c})e_i|^{1/2}\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|_2\Big\}.\end{aligned}$$

Here we consider the event

$$\mathcal{T}_2 = \Big\{\max_i |e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}Z_{\mathcal{S}}^\top Z_{\mathcal{S}^c})e_i|/N^2 \le 4c_{1\lambda}^{-2}\Big\}$$

For each i we have $|e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}Z_{\mathcal{S}}^\top Z_{\mathcal{S}^c})e_i| \le \lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}})\max_i |e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_i|$. By (3.8) we have $P\{N^{-1}\lambda_{\max}(Z_{\mathcal{S}}^\top Z_{\mathcal{S}}) \ge 2c_{1\lambda}^{-1}\} \le 2\exp\{-cN + s\log 21\}$. In addition, by (A.5) of (2), $P\{\max_i |e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}^c})e_i| \ge 2c_{1\lambda}^{-1}\} \le 2\exp\{-cN + 2\log q\}$. By summing over $i = 1, \cdots, q - s$, we have $P(\mathcal{T}_2) \ge 1 - 2\exp(-cN + s\log 21 + 3\log q)$, which converges to 1 by the assumption that $\log q = o(N^\alpha)$ in Condition (C6) and $s = o(N\min(1, \omega^2))$. Under the event $\mathcal{T}_2$, we have

$$\begin{aligned}&(N\lambda_N)^{-1}\Big\{\max_i |e_i^\top(Z_{\mathcal{S}^c}^\top Z_{\mathcal{S}}Z_{\mathcal{S}}^\top Z_{\mathcal{S}^c})e_i|^{1/2}\|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\|_2\Big\}\\ &= O(\lambda_N^{-1}\sqrt{s/N}\alpha_\lambda c_{1\lambda}^{-1}\tau) = o(1),\end{aligned}$$

given the condition $\lambda_N \gg \sqrt{s/N}\alpha_\lambda c_{1\lambda}^{-1}$ in Condition (C6).

B.3. Proof of Theorem 3

By Theorem 2, we only need to prove the asymptotic normality of $\widehat{\beta}_{\mathcal{S}}$. It has been shown that $\widehat{\beta}_{\mathcal{S}}$ is a strict local minimizer and $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$. As in the proof of Theorem 2, $\widehat{\beta}_{\mathcal{S}}$ is a strict local minimizer of $\overline{Q}(\theta)$ and $\widehat{\beta}_{\mathcal{S}^c} = \mathbf{0}$. Therefore we have $\partial\overline{Q}(\widehat{\beta}_{\mathcal{S}}) = \mathbf{0}$. It can be derived $\partial\overline{Q}(\theta) = -2\widehat{\gamma} + 2\widehat{\Gamma}_Z\theta + \overline{p}_\lambda(\theta)$, where $\overline{p}_\lambda(\theta) = (p_\lambda'(\theta_1), \cdots, p_\lambda'(\theta_s))^\top$. By conditions in Theorem 3, we have

$$\|\overline{p}_{\lambda_N}(\widehat{\beta}_{\mathcal{S}})\|_2 \le \sqrt{s}p_{\lambda_N}'(d_N) = o_p(N^{-1/2}c_{2\lambda}^{-1}\lambda_{\min}(\Sigma_e)).$$

We then have

$$\widehat{\Gamma}_Z^{(\mathcal{S})}(\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}) = N^{-1}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}} + o_p(N^{-1/2}c_{2\lambda}^{-1}\lambda_{\min}(\Sigma_e)),$$

where $\widetilde{\mathcal{E}} = \mathrm{vec}(\mathcal{E})$ and $Z_{\mathcal{S}}$ denotes the submatrix of Z with column indexes in $\mathcal{S}$. Further it can be derived that

$$\sqrt{N}\Gamma_Z^{(\mathcal{S})}(\widehat{\beta}_{\mathcal{S}}-\beta_{\mathcal{S}}) = -\sqrt{N}(\widehat{\Gamma}_Z^{(\mathcal{S})}-\Gamma_Z^{(\mathcal{S})})(\widehat{\beta}_{\mathcal{S}}-\beta_{\mathcal{S}})+N^{-1/2}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}}+o_p(c_{2\lambda}^{-1}\lambda_{\min}(\Sigma_e)).$$

To prove the result, it suffices to show that for any $A_N \in \mathbb{R}^{m\times s}$ with $AA^\top \to G$, we have

$$\begin{aligned}&\|A_N(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})})(\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}})\| = o_p(\|\Gamma_Z^{(\mathcal{S})}(\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}})\|)\\ &N^{-1/2}A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}} \to_d N(0,G).\end{aligned}$$

For the first, using the Cauchy's inequality, it suffices to verify that

$$\max_i |\lambda_i(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})})| = o_p(\lambda_{\min}(\Gamma_Z^{(\mathcal{S})})) = o_p(c_{2\lambda}^{-1}),$$

which is directly implied by (A.3) of Lemma 1. Next, note that $\lambda_{\min}(\Gamma_{Ze}^{(\mathcal{S})}) \geq \lambda_{\min}(\Gamma_{\widetilde{X}}(0))\lambda_{\min}(\Sigma_e) \geq c_{2\lambda}^{-1}\lambda_{\min}(\Sigma_e)$. Consequently, the last term could be dominated by the second term. We then show $N^{-1/2}A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}} \to_d N(0,G)$.

Since m is finite, then it suffices to show that for any η with $\|\eta\| = 1$, that $N^{-1/2}\eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}} \to_d N(0,\eta^\top G\eta)$. First it can be derived $Z^\top\widetilde{\mathcal{E}} = \sum_t \mathrm{vec}(\widetilde{X}_{t-1}\mathcal{E}_t^\top)$. Define $J_{\mathcal{S}} = (I_s, \mathbf{0}_{s,dp^2-s})$. Then $Z_{\mathcal{S}}^\top\mathcal{E} = J_{\mathcal{S}}Z^\top\mathcal{E} = \sum_t J_{\mathcal{S}}(I_p\otimes \widetilde{X}_{t-1})\mathcal{E}_t$. Let $\xi_t = N^{-1/2}\eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}J_{\mathcal{S}}(I_p\otimes\widetilde{X}_{t-1})\mathcal{E}_t$. Then we have

$$N^{-1/2}\eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}Z_{\mathcal{S}}^\top\widetilde{\mathcal{E}} = \sum_t \xi_t.$$

Define the σ-field $\mathcal{F}_t = \sigma\{\varepsilon_{is}, 1 \leq i \leq N, -\infty < s \leq t\}$. As a result, the sequence $\{\sum_{s=1}^t \xi_s, \mathcal{F}_t\}$ constitutes a martingale array. To show the asymptotic normality of $\sum_{t=1}^T \xi_t$, we employ the central limit theorem of the martingale difference array. We then verify the two conditions in Corollary 3.1 of [15].

Define $\widetilde{\eta} = J_{\mathcal{S}}^\top(\Gamma_{Ze}^{(\mathcal{S})\top})^{-1/2}A_N^\top\eta \in \mathbb{R}^{dp^2}$. First, it can be derived that

$$\begin{aligned}S_{1T} &\overset{\text{def}}{=} \sum_{t=1}^T E\{\xi_t^2 I(|\xi_t| > \delta)|\mathcal{F}_{t-1}\} \leq \delta^{-2}\sum_{t=1}^T E(\xi_t^4|\mathcal{F}_{t-1})\\ &\leq \sum_t N^{-2}C\delta^{-2}\Big(\widetilde{\eta}^\top\Big[\Sigma_e\otimes\big\{(\widetilde{X}_{t-1}\widetilde{X}_{t-1}^\top\big\}\Big]\widetilde{\eta}\Big)^2\end{aligned}$$

where C is a finite constant since the errors take Gaussian distribution. To show $S_{1T} \to_p 0$, it suffices to prove that $E(S_{1T}) \to 0$. Let $U_t = \Sigma_e^{1/2}\otimes\widetilde{X}_t$ then we have $E(S_{1T}) = N^{-1}c\delta^{-2}E\{(U_t^\top\widetilde{\eta}\widetilde{\eta}^\top U_t)^2\}$. It can be derived U_t follows multivariate normal distribution $N(\mathbf{0}, \Sigma_e\otimes\Gamma_{\widetilde{X}}(0))$. Consequently, we have $U_t^\top\widetilde{\eta}\widetilde{\eta}^\top U_t = \widetilde{U}_t^\top\Sigma_u\widetilde{U}_t = \widetilde{U}_t^\top Q_u\Lambda_u Q_u^\top\widetilde{U}_t$, where $\widetilde{U}_t = (\Sigma_e^{-1/2}\otimes\Gamma_{\widetilde{X}}^{-1/2}(0))U_t$, $\Sigma_u = Q_u\Lambda_u Q_u^\top$ is the eigenvalue decomposition of Σ_u with $\Lambda_u = (\lambda_{u,1},\cdots,\lambda_{u,dp^2})$ being the diagonal matrix and Q_u being the orthogonal matrix. Consequently, $U_t^\top\widetilde{\eta}\widetilde{\eta}^\top U_t =$

$\sum_i \lambda_{u,i}\xi_{u,i}^2$, where $\xi_{u,i}^2$ independently follows $\chi^2(1)$ distribution. Then we have $E(U_t^\top \widetilde{\eta}\widetilde{\eta}^\top U_t)^2 = \{\mathrm{tr}(\Sigma_u)\}^2 + 2\mathrm{tr}(\Sigma_u^2)$. It can be calculated $\mathrm{tr}(\Sigma_u) = \widetilde{\eta}^\top(\Sigma_e \otimes \Gamma_{\widetilde{X}}(0))\widetilde{\eta} = \eta^\top A_N A_N^\top \eta \le \lambda_{\max}(G)$ as $N \to \infty$. Next, $\mathrm{tr}(\Sigma_u^2) = (\widetilde{\eta}^\top(\Sigma_e \otimes \Gamma_{\widetilde{X}}(0))\widetilde{\eta})^2 = (\eta^\top A_N A_N^\top \eta)^2 \le \lambda_{\max}^2(G)$ as $N \to \infty$. This implies $E(S_{1T}) \to 0$ immediately.

Next, define

$$S_{2T} \overset{\mathrm{def}}{=} \sum_{t=1}^T E(\xi_t^2|\mathcal{F}_{t-1}) = N^{-1}\sum_{t=1}^T \Big(\widetilde{\eta}^\top\Big[\Sigma_e \otimes \big\{\widetilde{X}_{t-1}\widetilde{X}_{t-1}^\top\big\}\Big]\widetilde{\eta}\Big)$$
$$= \eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}(\Sigma_e \otimes \widehat{\Gamma}_{\widetilde{X}}(0))(\Gamma_{Ze}^{(\mathcal{S})\top})^{-1/2}A_N^\top \eta.$$

and we then need to verify that $S_{2T} \to_p \eta^\top G\eta$. Define $\Delta_{\widetilde{X}} = \widehat{\Gamma}_{\widetilde{X}}(0) - \Gamma_{\widetilde{X}}(0))$ then we have $S_{2T} = \eta^\top G\eta + \eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}\Delta_{\widetilde{X}}(\Gamma_{Ze}^{(\mathcal{S})\top})^{-1/2}A_N^\top\eta$. It leaves to show that $\eta^\top A_N(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}\Delta_{\widetilde{X}}(\Gamma_{Ze}^{(\mathcal{S})\top})^{-1/2}A_N^\top\eta = o_p(1)$. To this end, it suffices to show

$$\max_i|\lambda_i\{(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}(\Sigma_e\otimes\Delta_{\widetilde{X}})(\Gamma_{Ze}^{(\mathcal{S})})^{-1/2}\}|\le\max_i|\lambda_i(\Delta_{\widetilde{X}})|)\lambda_{\min}^{-1}\{\Gamma_{\widetilde{X}}(0)\}=o_p(1).$$

Note that $\max_i|\lambda_i(\Delta_{\widetilde{X}})| = o_p(c_{2\lambda}^{-1})$ by (A.3) of Lemma 1. In addition, according to Proposition 2.3 of [2], we have $\lambda_{\min}\{\Gamma_{\widetilde{X}}(0)\} \ge c_{2\lambda}^{-1}$. Therefore we have $\max_i|\lambda_i(\Delta_{\widetilde{X}})|)\lambda_{\min}^{-1}\{\Gamma_{\widetilde{X}}(0)\} = o_p(1)$. Consequently, the desired results hold.

B.4. Proof of Theorem 4

It can be derived that

$$\widehat{\Sigma}_{\mathcal{S}} - \Sigma_{\mathcal{S}} = (\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1}(\widehat{\Gamma}_{Ze}^{(\mathcal{S})} - \Gamma_{Ze}^{(\mathcal{S})})(\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1} + \big\{(\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1} - (\Gamma_Z^{(\mathcal{S})})^{-1}\big\}\Gamma_{Ze}^{(\mathcal{S})}(\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1}$$
$$+ (\Gamma_Z^{(\mathcal{S})})^{-1}\Gamma_{Ze}^{(\mathcal{S})}\big\{(\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1} - (\Gamma_Z^{(\mathcal{S})})^{-1}\big\} \overset{\mathrm{def}}{=} \Delta_1 + \Delta_2 + \Delta_3.$$

Further note that

$$(\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1} - (\Gamma_Z^{(\mathcal{S})})^{-1} = (\widehat{\Gamma}_Z^{(\mathcal{S})})^{-1}\big(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})}\big)(\Gamma_Z^{(\mathcal{S})})^{-1}.$$

Define the events $\mathcal{H} = \{\lambda_{\min}^{-1}(\widehat{\Gamma}_Z^{(\mathcal{S})}) < 2c_{2\lambda}\}$. By (3.7), we have $P(\mathcal{H}) \to 1$ if $s = o(N\min(1,\omega^2))$. Then under $\mathcal{H}$, it suffices to show

$$\max_i|\lambda_i(\widehat{\Gamma}_Z^{(\mathcal{S})} - \Gamma_Z^{(\mathcal{S})})| = o_p\big\{c_{2\lambda}^{-3}c_{1\lambda}\lambda_{\max}^{-1}(\Sigma_e)\big\} \tag{B.10}$$
$$\max_i|\lambda_i(\widehat{\Gamma}_{Ze}^{(\mathcal{S})} - \Gamma_{Ze}^{(\mathcal{S})})| = o_p\big\{c_{2\lambda}^{-2}\big\}. \tag{B.11}$$

First by (A.2) and $s = o(N\min(1,\eta^2))$, where η is set to be $\eta = c_{1\lambda}^2 c_{2\lambda}^{-3}\lambda_{\max}^{-1}(\Sigma_e)$, (B.10) can be obtained. Next, (B.11) can be obtained from Lemma 3.

Acknowledgements

The author thanks the editor and reviewers for their constructive comments and suggestions which strengthened the contributions of the work.

References

[1] Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics.* **146** 304–317. MR2465175

[2] Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics.* **43** 1535–1567. MR3357870

[3] Basu, S., Shojaie, A., and Michailidis, G. (2015). Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research.* **16** 417–453. MR3335801

[4] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data. Methods, Theory and Applications*. Springer Science & Business Media. MR2807761

[5] Chen, L., Guo, B., Huang, J., He, J., Wang, H., Zhang, S., and Chen, S. X. (2018). Assessing air-quality in Beijing-Tianjin-Hebei region: The method and mixed tales of PM2. 5 and O3. *Atmospheric Environment*.

[6] De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics.* **146** 318–328. MR2465176

[7] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics.* **3** 521. MR2750671

[8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association.* **96** 1348–1360. MR1946581

[9] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B.* **70** 849–911. MR2530322

[10] Fan and Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory.* **57** 5467–5484. MR2849368

[11] Fan, J., Lv, J., and Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics.* **3** 291–317.

[12] Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics.* **32** 928–961. MR2065194

[13] Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics.* **42** 819–849. MR3210988

[14] Fan, J. and Yao, Q. (2008). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Science & Business Media. MR1964455

[15] Hall, P. and Heyde, C. C. (2014). *Martingale Limit Theory and Its Application*. Academic Press. MR0624435

[16] Han, F. and Liu, H. (2013). Transition matrix estimation in high dimensional time series. In *Proceedings of the 30th International Conference on Machine Learning.* 172–180.

[17] Härdle, W. K., Wang, W., and Yu, L. (2016). Tenet: Tail-event driven network risk. *Journal of Econometrics.* **192** 499–513. MR3488092

[18] He, K., Yang, F., Ma, Y., Zhang, Q., Yao, X., Chan, C. K., Cadle, S., Chan, T., and Mulawa, P. (2001). The characteristics of PM2.5 in Beijing, China. *Atmospheric Environment.* **35** 4959–4970.

[19] Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems.* 2726–2734. MR3015038

[20] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media. MR2172368

[21] Lv, J., Fan, Y., et al. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics.* **37** 3498–3528. MR2549567

[22] Priestley, M. B. (1981). *Spectral Analysis and Time Series*, vol. 1. Academic Press, London. MR0628735

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.* **58** 267–288. MR1379242

[24] Vershynin, R. (2011). Lectures in geometric functional analysis. Preprint.

[25] Wang, L., Kim, Y., and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Annals of Statistics.* **41** 2505–2536. MR3127873

[26] Zhang, C.-H., et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics.* **38** 894–942. MR2604701

[27] Zhang, X., Pan, R., Guan, G., Zhu, X., and Wang, H. (2018). Logistic regression with network structure. *Statistica Sinica*, to appear.

[28] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research.* **7** 2541–2563. MR2274449

[29] Zhu, X., Wang, W., Wang, H., and Härdle, W. K. (2019). Network quantile autoregression. *Journal of Econometrics.* **1** 345–358. MR3994021

[30] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association.* **101** 1418–1429. MR2279469

[31] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics.* **36** 1509–1533. MR2435447