Feature Screening for Network Autoregression Model

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

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

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

Supplementary Material

The supplementary material provides technical details. Section S1 provides useful lemmas to establish the theorems. Section S2 presents the detail to prove Proposition 1. Section S3–S5 establish Theorem 1, Corollary 1, and Theorem 2 respectively. Section S6 gives a discussion about how to select tuning parameter.

S1. Useful Lemmas

To prove the theoretical properties, three useful lemmas are established. The detailed technical proof of Lemma 1–3 are given in this subsection.

Lemma 1. *Assume X follows sub-Gaussian distribution with mean* 0 *and moment generating function satisfying $E\{\exp(tX)\} \leq \exp(\sigma^2 t^2/2)$. Then the random variable $Z = X^2 - E(X^2)$ follows sub-exponential distribution with mean 0, and the moment generating function satisfies $E\{\exp(tZ)\} \leq$*

$\exp(c_z^2 t^2)$ *for all* $|t| \le 1/c_z$ *where* c_z *is a positive constant.*

Proof: The proof can be found in Proposition 2.7.1 of Vershynin (2017).

Lemma 2. *Let* $X_i s$ *(*$1 \le i \le n$*) and* $Y_i s$ *(*$1 \le i \le n$*) be independent and identically distributed sub-Gaussian random variables with mean 0 and variances* σ_x^2 *and* σ_y^2 *respectively. In addition, assume* $Cov(X_i, Y_i) = \sigma_{xy}$*. Denote* $X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^n$ *and* $Y = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$*. Then we have*

$$E\{(X^\top MY)(X^\top WY)\} \le c_1\big\{tr(M)tr(W) + tr(\mathbb{W})\big\}, \tag{S1.1}$$

$$\mathrm{var}(X^\top MX) \le c_2\{tr(M^2) + tr(MM^\top)\}, \tag{S1.2}$$

where $M \in \mathbb{R}^{n\times n}$ *and* $W \in \mathbb{R}^{n\times n}$ *are arbitrary matrices,* $\mathbb{W} = MW + MW^\top + MM^\top + WW^\top$*, and* $c_1 = 2\{E(X_i^2 Y_i^2) + \sigma_x^2\sigma_y^2 + \sigma_{xy}^2\}$*,* $c_2 = 2\max\{\sigma_x^4, E(X_i^4) - \sigma_x^4\}$ *are finite positive constants.*

Proof: Let $M = (m_{ij}) \in \mathbb{R}^{n\times n}$ and $W = (w_{ij}) \in \mathbb{R}^{n\times n}$. Then we have $X^\top MY = \sum_{i,j} m_{ij} X_i Y_j$ and $(X^\top MY)^2 = \sum_{i_1,i_2,j_1,j_2} m_{i_1 i_2} m_{j_1 j_2} X_{i_1} X_{i_2} Y_{j_1} Y_{j_2}$. One could directly calculate that $E\{(X^\top MY)(X^\top WY)\} =$

$$\sum_{i\neq j} m_{ii} w_{jj} E(X_i^2 Y_j^2) + \sum_{i\neq j}\{m_{ij}w_{ij} + m_{ij}w_{ji}\}\{E(X_i Y_j)\}^2 + \sum_i m_{ii} w_{ii} E(X_i^2 Y_i^2)$$

$$= \sigma_{xy}^2\{\mathrm{tr}(MW) + \mathrm{tr}(MW^\top)\} + \sigma_x^2\sigma_y^2 \mathrm{tr}(M)\mathrm{tr}(W) + \sum_i m_{ii} w_{ii} c_{xy},$$

where $c_{xy} = E(X_i^2 Y_i^2) - \sigma_x^2\sigma_y^2 - 2\sigma_{xy}^2$. Since we have $\sum_i m_{ii}w_{ii} \leq \sum_i (m_{ii}^2 + w_{ii}^2) \leq \mathrm{tr}(MM^\top) + \mathrm{tr}(WW^\top)$, and X, Y are sub-Gaussian random vectors, we could have (S1.1) by letting $c_1 = 2\{E(X_i^2Y_i^2) + \sigma_x^2\sigma_y^2 + \sigma_{xy}^2\}$.

Next, we have $E(X^\top MX) = \sigma_x^2 \mathrm{tr}(M)$. Hence we have $\{E(X^\top MX)\}^2 = \sigma_x^4(\sum_i m_{ii})^2$. Therefore one could obtain $\mathrm{var}(X^\top MX) = E(X^\top MX)^2 - \{E(X^\top MX)\}^2 = \sigma_x^4\{\mathrm{tr}(M^2) + \mathrm{tr}(MM^\top)\} + \sum_i m_{ii}^2\{E(X_i^4) - \sigma_x^4\}$. Since $\sum_i m_{ii}^2 \leq \mathrm{tr}(MM^\top)$, by letting $c_2 = 2\max\{\sigma_x^4, E(X_i^4) - \sigma_x^4\}$ and by the result of Lemma 1, (S1.2) can be readily obtained.

Lemma 3. *Assume* $\mathbb{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n\times p}$, *where* $X_i = (X_{i1}, \cdots, X_{ip}) \in \mathbb{R}^p$ *independently follows sub-Gaussian distribution with* $E(X_i) = \mathbf{0}_p$ *with* $Cov(X_i) = \Sigma_x = (\sigma_{j_1j_2,x}) \in \mathbb{R}^{p\times p}$. *In addition, assume that* $Y \in \mathbb{R}^n$ *follows multivariate sub-Gaussian distribution with mean* $\mathbf{0}_n$, *and* $Cov(Y) = \Sigma_y \in \mathbb{R}^{n\times n}$. *Assume* $Cov(\mathbb{X}_j, Y) = \Sigma_{j,xy} \in \mathbb{R}^{n\times n}$, *where* $\mathbb{X}_j = (X_{1,j}, \cdots, X_{n,j})^\top \in \mathbb{R}^n$. *Moreover, assume* $\lambda_{\max}(\Sigma_x) \leq c_x < \infty$, $\lambda_{\max}(\Sigma_y) \leq c_y < \infty$, *where* c_x *and* c_y *are finite positive constants. Then we have*

$$P\Big\{\Big|n^{-1}(\mathbb{X}_j^\top Y) - \sigma_{j,xy}^{(n)}\Big| \geq \delta\Big\} \leq C_1\exp(-C_2n\delta^2), \tag{S1.3}$$

where $\sigma_{j,xy}^{(n)} = n^{-1}E(\mathbb{X}_j^\top Y)$, *and* C_1 *and* C_2 *are non-zero positive constants, which are only related to* c_x *and* c_y.

Proof: Let $Z_j = \mathbb{X}_j + Y$. We then have $\Sigma_{zj} \stackrel{\text{def}}{=} \text{Cov}(Z_j) = \sigma_{jj,x} I_n + (\Sigma_{j,xy} + \Sigma_{j,xy}^\top) + \Sigma_y$. One can directly derive that $\mathbb{X}_j^\top Y = 2^{-1}(Z_j^\top Z_j - \mathbb{X}_j^\top \mathbb{X}_j - Y^\top Y)$. Then we have

$$P\{|n^{-1}(\mathbb{X}_j^\top Y) - \sigma_{j,xy}^{(n)}| \geq \delta\} \leq P\{|n^{-1}(Z_j^\top Z_j) - (\sigma_{jj,x} + \sigma_y^{(n)} + 2\sigma_{j,xy}^{(n)})| \geq \delta_1\}$$
$$+ P\{|n^{-1}(\mathbb{X}_j^\top \mathbb{X}_j) - \sigma_{jj,x}| \geq \delta_1\} + P\{|n^{-1}(Y^\top Y) - \sigma_y^{(n)}| \geq \delta_1\}, \quad \text{(S1.4)}$$

where $\delta_1 = 2/3\delta$ and $\sigma_y^{(n)} = n^{-1}\text{tr}(\Sigma_y)$. We then derive an upper bound for the right hand side of (S1.4). It should be noted that $\mathbb{X}_j^\top \mathbb{X}_j$, $Y^\top Y$, and $Z_j^\top Z_j$ in the right hand side of (S1.4) are all in quadratic form and thus the proofs are similar. For the sake of simplicity, we take $Y^\top Y$ for an example and derive the upper bound for $P\{|n^{-1}(Y^\top Y) - \sigma_y^{(n)}| \geq \delta_1\}$. The same result could be proved similarly for the other two terms in the right hand side of (S1.4).

First we have $Y^\top Y = Y^\top \Sigma_y^{-1/2} \Sigma_y \Sigma_y^{-1/2} Y = \widetilde{Y}^\top \Sigma_y \widetilde{Y}$, where $\widetilde{Y} = \Sigma_y^{-1/2} Y$ follows sub-Gaussian distribution. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of Σ_y. Since Σ_y is a non-negative definite matrix, we could have the eigenvalue decomposition as $\Sigma_y = U^\top \Lambda U$, where $U = (U_1, \cdots, U_n)^\top \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_n\}$. As a consequence, we have $Y^\top Y = \sum_i \lambda_i \zeta_i^2$, where $\zeta_i = U_i^\top \widetilde{Y}$ and ζ_is are *i.i.d.*

from the standard sub-Gaussian distribution subG(1). It can be verified $\zeta_i^2 - 1$ satisfies sub-exponential distribution by Lemma 1. It can be easily verified that the sub-exponential distribution satisfies condition (P) on page 45 of Saulis and Statulevičius (2012). Thus we have $P\{|n^{-1}(Y^\top Y) - \sigma_y^{(n)}| \geq \delta_1\} = P\{\sum_i \lambda_i(\zeta_i^2 - 1)| \geq n\delta_1\} \geq c_1 \exp\{-c_2(\sum_i \lambda_i^2)^{-1} n^2 \delta_1^2\} = c_1 \exp\{-c_2 \mathrm{tr}^{-1} \lambda_{\max}^{-2}(\Sigma_y) n \delta_1^2\}$ by the Theorem 3.2, on Page 45 of Saulis and Statulevičius (2012). Similarly, there exists positive constants $c_1 > 0$ and $c_2 > 0$, such that $P\{|n^{-1}(\mathbb{X}_j^\top \mathbb{X}_j) - \sigma_{jj,x}| \geq \delta_1\} \leq c_1 \exp(-c_2 \sigma_{jj,x}^{-2} n \delta_1^2)$ and $P\{|n^{-1}(Z_j^\top Z_j) - (\sigma_{jj,x} + \sigma_y^{(n)} + 2\sigma_{j,xy}^{(n)})| \geq \delta_1\} \leq c_1 \exp\{-c_2 \mathrm{tr}^{-1}(\Sigma_{zj}^2) n^2 \delta_1^2\}$. It can be easily derived that $\sigma_{jj,x} \leq \lambda_{\max}(\Sigma_x) \leq c_x$ and $\mathrm{tr}(\Sigma_{zj}^2) \leq n\lambda_{\max}^2(\Sigma_{zj})$. Further we have $\lambda_{\max}(\Sigma_{zj}) \leq \lambda_{\max}(\Sigma_x) + 2\lambda_{\max}^{1/2}(\Sigma_x)\lambda_{\max}^{1/2}(\Sigma_y) + \lambda_{\max}(\Sigma_y) \leq \{\lambda_{\max}^{1/2}(\Sigma_x) + \lambda_{\max}^{1/2}(\Sigma_y)\}^2$ by the Cauchy's inequality. Lastly, by condition (C4), condition $\lambda_{\max}(\Sigma_x) \leq c_x < \infty$, and $\lambda_{\max}(\Sigma_y) \leq c_y < \infty$, the desired result (S1.3) can be obtained by using (S1.4).

S2. Proof of Proposition 1

In the proof of proposition 1, for convenience, we define $\widehat{R}_j^2 =$

$$n^{-3}\{(Y^\top W^\top WY)(\mathbb{X}_j^\top Y)^2 - 2(\mathbb{X}_j^\top Y)(\mathbb{X}_j^\top WY)(Y^\top WY) + (\mathbb{X}_j^\top WY)^2(Y^\top Y)\}.$$

Consequently $\widehat{\mathbf{R}}_j^2 = n^2\widehat{R}_j^2\big\{(Y^\top Y)(Y^\top W^\top WY) - (Y^\top W^\top Y)^2\big\}^{-1}$. In addition, define $R_j^2 = (\kappa_1^{(n)}\kappa_5^{(n)2} - 2\kappa_2^{(n)}\kappa_4^{(n)}\kappa_5^{(n)} + \kappa_3^{(n)}\kappa_4^{(n)2})\nu_0\nu_j^2$. It suffices to show $\max|n^{-2}\{(Y^\top Y)(Y^\top W^\top WY) - (Y^\top W^\top Y)^2\} - c_\kappa^{(n)}| = o_p(1)$. Due to the similarity of the proof, we only prove the first one.

To prove $\max_j|\widehat{R}_j^2 - R_j^2| \to_p 0$, it suffices to show that $P\{\max_j|\widehat{R}_j^2 - R_j^2| > \delta\} \to 0$ as $n \to \infty$. By the maximum inequality, it could be concluded that $P\{\max_j|\widehat{R}_j^2 - R_j^2| > \delta\} \le \sum_{j=1}^p P\{|\widehat{R}_j^2 - R_j^2| > \delta\}$. Next, we would prove that

$$P\big\{|\widehat{R}_j^2 - R_j^2| > \delta\big\} \le C_1\exp(-C_2 n\delta_0^2) \tag{S2.1}$$

for $1 \le j \le p$ and finite constants $C_1, C_2 > 0$, where $\delta_0 = (\delta/6)^{1/3}$. To achieve this, we first derive the inequality as $P\big\{|\widehat{R}_j^2 - R_j^2| > \delta\big\} \le$

$$\begin{aligned}
&P\big\{|n^{-1}(Y^\top W^\top WY) - \kappa_3^{(n)}\nu_0| > \delta_0\big\} + 2P\big\{|n^{-1}(\mathbb{X}_j^\top Y) - \kappa_4^{(n)}\nu_j| > \delta_0\big\}\\
&+ 2P\big\{|n^{-1}(\mathbb{X}_j^\top WY) - \kappa_5^{(n)}\nu_j| > \delta_0\big\} + P\big\{|n^{-1}(Y^\top WY) - \kappa_2^{(n)}\nu_0| > \delta_0\big\}\\
&+ P\big\{|n^{-1}(Y^\top Y) - \kappa_1^{(n)}\nu_0| > \delta_0\big\}.
\end{aligned} \tag{S2.2}$$

To derive the upper bound for each term of (S2.2), we apply Lemma 3. It can be derived $\mathrm{Cov}(WY) = W\Sigma_y W^\top$, $\lambda_{\max}(W\Sigma_y W^\top) \le \lambda_{\max}(\Sigma_y)\lambda_{\max}(WW^\top)$. By the conditions (C1)–(C4) and then applying Lemma 3, we could have $c_1\exp(-c_2 n\delta_0^2)$ as an upper bound for each term in (S2.2), where $c_1 > 0$

and $c_2 > 0$ are finite constants only related to κ_j $(1 \le j \le 5)$ and $\tau_{\max}$. Therefore, (S2.1) can be proved by letting $C_1 = 5c_1$ and $C_2 = c_2$. Consequently, the conclusion follows by the condition (C2) that $\log p = O(n^{\xi})$ with $0 \le \xi < 1$.

S3. Proof of Theorem 1

With the definition of $\widehat{R}_j^2$ and R_j^2 given in the Appendix S2, we know that the rank of $\widehat{\mathbf{R}}_j^2$s is exactly the same with that of $\widehat{R}_j^2$s across different js. To prove the screening consistency, we employ the following 5 steps.

Step 1. ($\|\beta\|^2 \le C_\beta < \infty$) Recall that $Y = \rho WY + \mathbb{X}\beta + \mathcal{E}$. Therefore we have $Y = (I_n - \rho W)^{-1}\mathbb{X}\beta + (I_n - \rho W)^{-1}\mathcal{E}$. One could further derive that $n^{-1}E(Y^\top Y) = n^{-1}\operatorname{var}(Y) = n^{-1}\operatorname{var}\{(I_n - \rho W)^{-1}\mathbb{X}\beta\} + n^{-1}\operatorname{var}\{(I_n - \rho W)^{-1}\mathcal{E}\} \ge n^{-1}\operatorname{var}\{(I_n - \rho W)^{-1}\mathbb{X}\beta\}$. By the condition that $n^{-1}E(Y^\top Y) = 1$, we have $n^{-1}\operatorname{var}\{(I_n - \rho W)^{-1}\mathbb{X}\beta\} \le 1$. This leads to

$$\beta^\top E\Big\{n^{-1}\mathbb{X}^\top (I_n - \rho W^\top)^{-1}(I_n - \rho W)^{-1}\mathbb{X}\Big\}\beta \le 1. \tag{S3.1}$$

By (C4), we know $\lambda_{\min}\{(I_n - \rho W^\top)^{-1}(I_n - \rho W)^{-1}\} \ge \tau_{\min}/(\beta^\top \Sigma_x \beta + \sigma^2) \ge \tau_{\min}/\sigma^2$. Thus $(\mathbb{X}\beta)^\top (I_n - \rho W^\top)^{-1}(I_n - \rho W)^{-1}(\mathbb{X}\beta) \ge \tau_{\min}\beta^\top \mathbb{X}^\top \mathbb{X}\beta/\sigma^2$. Then (S3.1) implies $\tau_{\min}\beta^\top E(n^{-1}\mathbb{X}^\top \mathbb{X})\beta/\sigma^2 \le 1$. Since we have $E(n^{-1}\mathbb{X}^\top \mathbb{X}) =$

Σ and $\lambda_{\min}(\Sigma) \geq \tau_{\min}$ by condition (C4), then it can be further derived that $\tau_{\min}^2/\sigma^2\|\beta\|^2 \leq 1$. Consequently, it can be concluded $\|\beta\|^2 \leq C_\beta$ by letting $C_\beta = \tau_{\min}^{-2}\sigma^2$.

STEP 2. ($\sum_{j=1}^p R_j^2 \leq C_r < \infty$) By the definition of R_j^2, we have $\sum_j R_j^2 = (\kappa_1^{(n)}\kappa_5^{(n)2} - 2\kappa_2^{(n)}\kappa_4^{(n)}\kappa_5^{(n)} + \kappa_3^{(n)}\kappa_4^{(n)2})\nu_0 \sum_{j=1}^p \nu_j^2$. By the convergence of $\kappa_j^{(n)}$ ($1 \leq j \leq 5$) in condition (C3), it can be conclude that $\kappa_j^{(n)} \leq C_\kappa$ for some positive constant C_κ. As a consequence, by STEP 1 and condition (C4), one can conclude that there exist a finite constant C_b such that

$$\beta^\top \Sigma \beta \leq \tau_{\max}\|\beta\|^2 < C_b,$$
$$\sum_{j=1}^p \nu_j^2 = \sum_{j=1}^p (\beta^\top \Sigma_{\cdot j})^2 = \beta^\top \Sigma^2 \beta < \tau_{\max}^2 \|\beta\|^2 \leq \beta_{\max}^2 \tau_{\max}^2 |\mathcal{M}_T|,$$

where $\beta_{\max} = \max_i |\beta_i|$. Consequently, by letting $C_r = \widetilde{c}_\beta \tau_{\max}^2 |\mathcal{M}_T| < \infty$ where $\widetilde{c}_\beta = 4C_\kappa^3 \beta_{\max}^2$, we then have $\sum_{j=1}^p R_j^2 \leq C_r$.

STEP 3. ($\max_j |\widehat{R}_j^2 - R_j^2| \to_p 0$) The result can be guaranteed by Proposition 1.

STEP 4. Recall $\mathbf{R}_j^2 = (c_\kappa^{(n)})^{-1} R_j^2$ and $\gamma_{\min}^* = \min_{j \in \mathcal{M}_T} \mathbf{R}_j^2$. Define $\mathcal{M}_T^* = \{j : \mathbf{R}_j^2 > \gamma_{\min}^*\}$. By definition, we have $\mathcal{M}_T \subset \mathcal{M}_T^*$. Equally, we have $\mathcal{M}_T^* = \{j : R_j^2 > \gamma_{\min}\}$, where $\gamma_{\min} = c_\kappa \gamma_{\min}^*$ and $c_\kappa = (\kappa_1\kappa_3 - \kappa_2)c_\beta^2$. By condition (C5), we have $\gamma_{\min} \geq c_\gamma c_\kappa > 0$ as $n \to \infty$. Recall that $\widehat{\mathcal{M}}^R = \{j : \widehat{\mathbf{R}}_j^2 >$

$\gamma_{\min}^*/2\} = \{j : \widehat{R}_j^2 > 2^{-1}\gamma_{\min} z_n\}$, where $z_n = c_\kappa^{-1} n^{-2}\{(Y^\top Y)(Y^\top W^\top WY) - (Y^\top WY)^2\}$. In this step, we want to show that $\widehat{\mathcal{M}}^R$ should uniformly cover $\mathcal{M}_T^*$ with probability tending to one. Otherwise, there must exist at least one $j^* \in \mathcal{M}_T^*$ which is not covered by $\widehat{\mathcal{M}}^R$. By the definition of $\widehat{\mathcal{M}}^R$, we know that we must have $\widehat{R}_{j^*}^2 \leq 2^{-1}\gamma_{\min} z_n$. However, due to the definition of $\mathcal{M}_T^*$, if $j^* \in \mathcal{M}_T^*$, $R_{j^*}^2 > \gamma_{\min}$. Both of these imply that $|\widehat{R}_j^2 - R_j^2| > 2^{-1}\gamma_{\min}|2 - z_n|$. As a consequence, if $\mathcal{M}_T^* \not\subset \widehat{\mathcal{M}}^R$, we must have, $\max_j |\widehat{R}_j^2 - R_j^2| > 2^{-1}\gamma_{\min}|2 - z_n|$. Therefore we have $P(\mathcal{M}_T^* \not\subset \widehat{\mathcal{M}}^R) \leq P(\max_j |\widehat{R}_j^2 - R_j^2||2 - z_n|^{-1} > \gamma_{\min}/2) \leq P(\max_j |\widehat{R}_j^2 - R_j^2||1 - |1 - z_n||^{-1} > \gamma_{\min}/2) =$

$$
\begin{aligned}
&P\Big(\max_j |\widehat{R}_j^2 - R_j^2||1 - |1 - z_n||^{-1} > \gamma_{\min}/2\Big||1 - z_n| \leq \epsilon\Big)P\Big(|1 - z_n| \leq \epsilon\Big) \\
&+ P\Big(\max_j |\widehat{R}_j^2 - R_j^2||1 - |1 - z_n||^{-1} > \gamma_{\min}/2\Big||1 - z_n| > \epsilon\Big)P\Big(|1 - z_n| > \epsilon\Big) \\
&\leq P(|1 - z_n| > \epsilon) + P(\max_j |\widehat{R}_j^2 - R_j^2| > 2^{-1}|1 - \epsilon|\gamma_{\min}).
\end{aligned}
$$

By similar technique as in Step 3, we have $P(|z_n - 1| > \epsilon) \leq P\big(|n^{-1}(Y^\top Y) - \kappa_1^{(n)}(\beta^\top\Sigma\beta + \sigma^2)| > \epsilon_0\big) + P\big(|n^{-1}(Y^\top W^\top WY) - \kappa_3^{(n)}(\beta^\top\Sigma\beta + \sigma^2)| > \epsilon_0\big) + P\big(|n^{-1}(Y^\top WY) - \kappa_2^{(n)}(\beta^\top\Sigma\beta + \sigma^2)| > \epsilon_0\big)$, where $\epsilon_0 = (2^{-1}c_\kappa\epsilon)^{1/2}$. Consequently, by Lemma 3, we have $P(|z_n - 1| > \epsilon) \leq 3c_1\exp(-c_2 n\epsilon_0^2) \to 0$. Next, by letting $\delta = 2^{-1}|1 - \epsilon|\gamma_{\min}$, we have $P(\max_j |\widehat{R}_j^2 - R_j^2| > 2^{-1}|1 - \epsilon|\gamma_{\min}) \to 0$ as $n \to \infty$. This suggests $P(\mathcal{M}_T^* \subset \widehat{\mathcal{M}}^R) \to_p 1$ as $n \to \infty$.

STEP 5. We next verify that the size of $\widehat{\mathcal{M}}^R$ could be uniformly bounded. First, by Step 2, we have $\sum_{j=1}^p R_j^2 \le C_r$. Define $\mathcal{M}^* = \{j : \mathbf{R}_j^2 > \gamma_{\min}^*/4\}$, which can be equivalent spelled as $\mathcal{M}^* = \{j : R_j^2 > \gamma_{\min}/4\}$. Then we have $C_r \ge \sum_{j\in\mathcal{M}^*} R_j^2 \ge |\mathcal{M}^*|\gamma_{\min}/4$. Then we have $|\mathcal{M}^*| \le 4C_r\gamma_{\min}^{-1} \doteq m_{\max}$, where $m_{\max} = c_\beta \tau_{\max}^2 \gamma_{\min}^{-1} |\mathcal{M}_T|$. By condition (C5) and Step 2 we have $m_{\max} < \infty$. If $|\widehat{\mathcal{M}}^R| > |\mathcal{M}^*|$, we must have $\widehat{\mathcal{M}}^R \not\subset \mathcal{M}^*$. This means there must exist at least one $j \in \widehat{\mathcal{M}}^R$ with $\widehat{R}_j^2 > \gamma_{\min} z_n/2$, but $j \notin \mathcal{M}^*$ with $R_j^2 \le \gamma_{\min}/4$. We immediately know that $\max_j |\widehat{R}_j^2 - R_j^2| \ge 4^{-1}\gamma_{\min}|2z_n - 1|$. Then we have, $P(|\widehat{\mathcal{M}}^R| > m_{\max}) \le P(\max_j |\widehat{R}_j^2 - R_j^2||2z_n - 1|^{-1} \ge \gamma_{\min}/4) \le P(\max_j |\widehat{R}_j^2 - R_j^2||1-2|z_n - 1||^{-1} \ge \gamma_{\min}/4) =$

$$P\Big(\max_j |\widehat{R}_j^2 - R_j^2||1-2|1-z_n||^{-1} \ge \gamma_{\min}/4 \Big| |1-z_n| \le \epsilon\Big) P\Big(|1-z_n| \le \epsilon\Big)$$
$$+ P\Big(\max_j |\widehat{R}_j^2 - R_j^2||1-2|1-z_n||^{-1} \ge \gamma_{\min}/4 \Big| |1-z_n| > \epsilon\Big) P\Big(|1-z_n| > \epsilon\Big)$$
$$\le P(|1-z_n| > \epsilon) + P(\max_j |\widehat{R}_j^2 - R_j^2| > |1-2\epsilon|\gamma_{\min}/4).$$

Consequently, by the similar technique in the previous step, $P(|z_n - 1| > \epsilon) \to 0$ and $P(\max_j |\widehat{R}_j^2 - R_j^2| > |1-2\epsilon|\gamma_{\min}/4) \to 0$ as $n \to \infty$. This suggest that $P(|\widehat{\mathcal{M}}^R| \le m_{\max}) \to 1$ as $n \to \infty$.

S4. Corollary from Theorem 1

Corollary 1. *Let* $\gamma_{\min}^{*(k)}$ *be the* k*th smallest element in* $\{\mathbf{R}_j^2 : j \in \mathcal{M}_T\}$ *and hence* $\gamma_{\min}^* = \gamma_{\min}^{*(1)}$. *Accordingly let* $\mathcal{M}_T^{(k)} = \{j \in \mathcal{M}_T : \mathbf{R}_j^2 \geq \gamma_{\min}^{*(k)}\}$ *and* $m_{\max}^{(k)} = c_\beta(\gamma_{\min}^{*(k)})^{-1}\tau_{\max}^2|\mathcal{M}_T^{(k)}|$. *Assume Conditions (C1)–(C4) and* $\gamma_{\min}^{*(k)} = 2c_\gamma$, *we then have*

$$P(\mathcal{M}_T^{(k)} \in \widehat{\mathcal{M}}^R) \to 1, \quad P(|\widehat{\mathcal{M}}^R| \leq m_{\max}^{(k)}) \to 1.$$

It implies that we are still able to have a compact model size $m_{\max}^{(k)}$ which detects $|\mathcal{M}_T| - k + 1$ important features consistently if these important features have relatively large signal.

The proof is similar to Theorem 1 but slightly different in Step 4 and 5. In Step 4 and 5, one could replace $\gamma_{\min}^*$, $\mathcal{M}_T$, and $m_{\max}$ by $\gamma_{\min}^{*(k)}$, $\mathcal{M}_T^{(k)}$, and $m_{\max}^{(k)}$ to obtain the result. The rest are the same with the proof of Theorem 1.

S5. Proof of Theorem 2

In this part we aim to establish the parameter consistency. For convenience we define $s = |\mathcal{M}|$ in the following. Following Fan and Li (2001), it is sufficient to show that for any $\varepsilon > 0$, there exists a constant $C > 0$ such that

$$\lim_{N\to\infty} P\Big\{ \sup_{|u|=C} \ell_1(\rho + N^{-1/2}u) < \ell_1(\rho) \Big\} > 1-\epsilon. \tag{S5.1}$$

Then, by (S5.1), with probability at least $1-\epsilon$, there exists a local optimizer $\widehat{\rho}$ in the ball $\{\rho + N^{-1/2}uC : |u| \le 1\}$. As a result, we have $|\widehat{\rho}-\rho| = O_p(n^{-1/2})$. To show (S5.1), we applies Taylor's expansion to obtain that $\sup_{\|u\|=1} \Big\{ \ell_1(\rho + n^{-1/2}uC) - \ell_1(\rho) \Big\} =$

$$\begin{aligned} &\sup_{\|u\|=1} \Big\{ Cn^{-1/2}\ell_1'(\rho)u + 2^{-1}C^2 n^{-1}\ell_1''(\rho)u^2 + o_p(1) \Big\} \\ &\quad \le C|n^{-1/2}\ell_1'(\rho)| - 2^{-1}C^2\Big\{ -n^{-1}\ell_1''(\rho)\Big\} + o_p(1). \end{aligned} \tag{S5.2}$$

We next show that (S5.2) is negative asymptotically with probability tending to 1. To this end, we consider $\ell_1'(\rho)$ and $\ell_1''(\rho)$ separately in the following two steps. For convenience, define $\alpha = E(\varepsilon_i^4) - \sigma^4$.

STEP 1. (PROOF OF $|n^{-1/2}\ell_1'(\rho)| = O_p(1)$). First it can be proved,

$$\ell_1'(\rho) = -\mathrm{tr}\big\{(I_n - \rho W)^{-1}W\big\} + \widehat{\sigma}^{-2}Y^\top(I_n - \rho W^\top)(I_n - P_X)WY, \tag{S5.3}$$

where $\widehat{\sigma}^2 = n^{-1}Y^\top(I_n - \rho W^\top)(I_n - P_X)(I_n - \rho W)Y$. Let $S_1 = \sigma^{-2}\{Y^\top(I_n - \rho W^\top)(I_n - P_X)WY\}$ and $s_1 = \mathrm{tr}\big\{(I_n - \rho W)^{-1}W\big\}$. We next show that $\widehat{\sigma}^2 \to_p \sigma^2$ and $n^{-1/2}(S_1 - s_1) = O_p(1)$.

STEP 1.1 ($\widehat{\sigma}^2 \to_p \sigma^2$) First it can be derived $\widehat{\sigma}^2 = \mathcal{E}^\top(I_n - P_X)\mathcal{E}$. One could verify that $E(\widehat{\sigma}^2) = (1 - s/n)\sigma^2 \to \sigma^2$ by the condition in Theorem 2. Next we have $\mathrm{var}(\widehat{\sigma}^2) \le n^{-2}\sigma^4 \,\mathrm{var}\{\mathrm{tr}(I_n - P_X)\} + n^{-2}\sigma^4 2c_2 E\{\mathrm{tr}(I_n - P_X)^2\} = 2\sigma^4 c_2 n^{-2}(n-s) \to 0$ by condition (C1) and (S1.2) of Lemma 2. This completes the proof of Step 1.1.

STEP 1.2 ($n^{-1/2}(S_1 - s_1) = O_p(1)$) It can be written that $S_1 = \sigma^{-2}\mathcal{E}^\top(I_n - P_X)W(I_n - \rho W)^{-1}\mathcal{E} = \sigma^{-2}\mathcal{E}^\top W(I_n - \rho W)^{-1}\mathcal{E} - \sigma^{-2}\mathcal{E}^\top P_X W(I_n - \rho W)^{-1}\mathcal{E}$. Define the first part to be S_{11} and the second to be S_{12}. Without loss of generality, we assume $\sigma^2 = 1$. Next we prove $n^{-1/2}(S_{11} - s_1) = O_p(1)$ and $n^{-1/2}S_{12} = o_p(1)$. For the first result, one could verify that $E(S_{11} - s_1) = 0$ and $n^{-1}\,\mathrm{var}(S_{11}) \le$

$$2^{-1}c_2\Big(n^{-1}\mathrm{tr}\big[\{W(I_n - \rho W)^{-1}\}^2\big] + n^{-1}\mathrm{tr}\big\{W(I_n - \rho W)^{-1}(I_n - \rho W^\top)^{-1}W^\top\big\}\Big)$$

$\to 2^{-1}c_2(\kappa_6 + \kappa_3)$ by (S1.2) of Lemma 2. Hence we have $n^{-1/2}(S_{11} - s_1) = O_p(1)$. Next, we have $S_{12} = \mathrm{tr}[(\mathbb{X}_\mathcal{M}^\top \mathbb{X}_\mathcal{M})^{-1}\{X_\mathcal{M}^\top W(I_n - \rho W)^{-1}\mathcal{E}\mathcal{E}^\top X_\mathcal{M}\}]$. By the trace inequality, we have,

$$|S_{12}| \le \lambda_{\min}^{-1}(\widehat{\Sigma}_\mathcal{M})|n^{-1}\mathrm{tr}\{\mathbb{X}_\mathcal{M}^\top W(I_n - \rho W)^{-1}\mathcal{E}\mathcal{E}^\top \mathbb{X}_\mathcal{M}\}|,$$

where $\widehat{\Sigma}_{\mathcal{M}} = n^{-1}\mathbb{X}_{\mathcal{M}}^\top\mathbb{X}_{\mathcal{M}}$. It can be concluded that $\lambda_{\min}(\widehat{\Sigma}_{\mathcal{M}}) \geq \tau_{\min} > 0$ with probability tending to 1 as $n \to \infty$ by condition (C4) and $s = o(n^{(1-\xi)/3})$, where the proof is similar to Wang (2009) and ignored here. Then it leads to show $n^{-1/2}[n^{-1}\{\mathcal{E}^\top\mathbb{X}_{\mathcal{M}}\mathbb{X}_{\mathcal{M}}^\top W(I_n-\rho W)^{-1}\mathcal{E}\}] = o_p(1)$. First $n^{-1}E\{\mathcal{E}^\top\mathbb{X}_{\mathcal{M}}\mathbb{X}_{\mathcal{M}}^\top W(I_n-\rho W)^{-1}\mathcal{E}\} = \sigma^2\kappa_5^{(n)}\mathrm{tr}(\Sigma_{\mathcal{M}}) \leq s\sigma^2\kappa_5^{(n)}\lambda_{\max}(\Sigma_{\mathcal{M}})$. Next, by Lemma (S1.1) in 2, var $\left\{\mathcal{E}^\top\mathbb{X}_{\mathcal{M}}\mathbb{X}_{\mathcal{M}}^\top W(I_n-\rho W)^{-1}\mathcal{E}\right\} \leq$

$$c_1E\Big[\mathrm{tr}\{(\mathbb{X}_{\mathcal{M}}^\top M\mathbb{X}_{\mathcal{M}})^2\}+\mathrm{tr}\{(\mathbb{X}_{\mathcal{M}}^\top MM^\top\mathbb{X}_{\mathcal{M}})(\mathbb{X}_{\mathcal{M}}^\top\mathbb{X}_{\mathcal{M}})\}\Big]+\mathrm{var}\Big[\mathrm{tr}\big(\mathbb{X}_{\mathcal{M}}^\top M\mathbb{X}_{\mathcal{M}}\big)\Big],$$

$\stackrel{\mathrm{def}}{=} V_1+V_2+V_3$, where $M = W(I_n-\rho W)^{-1}$. Note $\mathrm{tr}\{(\mathbb{X}_{\mathcal{M}}^\top M\mathbb{X}_{\mathcal{M}})^2\} = \sum_{j,k\in\mathcal{M}}(\mathbb{X}_j^\top M\mathbb{X}_k)(\mathbb{X}_j^\top M^\top\mathbb{X}_k)$. Then we have $V_1 \leq c_1s^2\{\mathrm{tr}(M)^2+\mathrm{tr}(M^2)+\mathrm{tr}(MM^\top)\}$ by condition (C1) and (S1.1) of Lemma 2. Further we have $n^{-2}\mathrm{tr}(M)^2 \to \kappa_5$ by condition (C3) and $n^{-2}\{\mathrm{tr}(M^2)+\mathrm{tr}(MM^\top)\} \to 0$ by the (5.3) of Lemma 2 in Zhu et al. (2017). As a consequence, we have $n^{-3}V_1 \to 0$ by conditions in Theorem 2. By similar techniques, one could have $n^{-3}V_2 \to 0$. Next, it can be derived by Cauchy's inequality, $V_3 \leq$

$$\sum_{j,k\in\mathcal{M}}E\{(\mathbb{X}_j^\top M\mathbb{X}_j)(\mathbb{X}_k^\top M^\top\mathbb{X}_k)\} \leq \sum_{j,k\in\mathcal{M}}[\{E(\mathbb{X}_j^\top M\mathbb{X}_j)^2E(\mathbb{X}_k^\top M^\top\mathbb{X}_k)^2\}]^{1/2}.$$

As a result, by (S1.1) of Lemma 2, we have $V_3 \leq c_1s^2\{\mathrm{tr}(M)^2+\mathrm{tr}(M^2)+\mathrm{tr}(MM^\top)$. Similarly, by condition (C4) and (5.3) of Lemma 2 in Zhu et al.

(2017), we have $n^{-3}V_3 \to 0$. Therefore, it can be concluded that $n^{-1/2}S_{12} \to_p 0$.

STEP 2. (PROOF OF $-n^{-1}\ell_1''(\rho) \to_p \sigma_{2\rho}^2$) It can be derived that

$$\begin{aligned}\ell''(\rho) &= -\mathrm{tr}\big\{(I_n-\rho W)^{-1}W(I_n-\rho W)^{-1}W\big\} \\ &\quad -\widehat{\sigma}^{-2}Y^\top W^\top(I_n-P_X)WY + 2(n\widehat{\sigma}^4)^{-1}\sigma^4 S_1^2.\end{aligned}$$

By the previous step, we have $n^{-1}S_1 - s_1 = o_p(1)$. Next, let the second term be $S_2 = \sigma^{-2}\{Y^\top W^\top(I_n-P_X)WY\}$ and $s_2 = n^{-1}\mathrm{tr}\{(I_n-\rho W)^{-1}WW^\top(I_n-\rho W^\top)^{-1}\}$. We then show that $n^{-1}S_2 - s_2 = o_p(1)$. Let $M = (I_n-\rho W)^{-1}W$. Then we have $S_2 =$

$$\begin{aligned}&\sigma^{-2}[\mathrm{tr}(\mathcal{E}^\top M^\top M\mathcal{E}) - \mathrm{tr}(\mathcal{E}^\top M^\top P_X M\mathcal{E}) + 2\mathrm{tr}\{\mathcal{E}^\top M^\top(I_n-P_X)M(\mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}})\} \\ &\quad +\mathrm{tr}\{(\mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}})^\top M^\top(I_n-P_X)M(\mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}})\}] \stackrel{\mathrm{def}}{=} \sigma^{-2}(S_{21}-S_{22}+S_{23}+S_{24}).\end{aligned}$$

By similar techniques in Step 1, one could verify that $n^{-1}\sigma^{-2}S_{21} - s_2 = o_p(1)$ and $n^{-1}\sigma^{-2}S_{2j} = o_p(1)$ for $2 \le j \le 4$. Let $M_1 = M + M^\top$, then one could verify that $-n^{-1}\ell_1''(\rho) - \sigma_{2\rho}^2 = o_p(1)$, where $\sigma_{2\rho}^2 = 2^{-1}\lim_{n\to\infty} n^{-1}\mathrm{tr}[\{M - n^{-1}\mathrm{tr}(M)I_n\}^2] = \kappa_3 + \kappa_6 - 2\kappa_5^2$.

By the results of Step 1 and Step 2, it can be concluded that the quadratic term will dominate the linear term in (S5.2) as long as a sufficiently large C is chosen. Then with probability tending to 1, we have $\ell_1(\rho + n^{-1/2}u) < \ell_1(\rho)$ as $n \to \infty$. This completes the proof of (S5.2).

S6. Selection of Tuning Parameter

For practical implementation, the selection of the tuning parameter c_γ is important. Different c_γ may lead to different selected model. Under a classical regression setup with $p < n$, this problem has been extensively studied. A number of selection criterions, such as AIC (Akaike, 1973), BIC (Schwarz, 1978), and EBIC (Chen and Chen, 2008; Wang, 2009), are proposed and carefully investigated. Practically, we could set the maximum number of features to be selected as p', with $p' < n$. For example, $p' = [n/\log(n)]$, where $[m]$ is the maximum integer, which is no larger than m.

Thus, in this case, the tuning parameter could be selected in the following steps. First, the features are sorted according to the value of $\widehat{\mathbf{R}}_j^2$. Second, $\widehat{\mathcal{M}}_j$ could be defined containing the first j features with the largest $\widehat{\mathbf{R}}_j^2$s. Third, the model could be selected via AIC, BIC, or EBIC methods. For example, for EBIC method, we define for $1 \le j \le p'$,

$$\text{EBIC}_\tau^j = -2\ell_j(\widehat{\theta}_j) + j\log(n) + 2\tau\log\{P(\widehat{\mathcal{M}}_j)\}, \tag{S6.1}$$

where $\ell_j(\theta)$ is the log likelihood of the model $\widehat{\mathcal{M}}_j$, $\widehat{\theta}_j$ is the maximum likelihood estimator of $\theta_j = (\rho, \beta_{\widehat{\mathcal{M}}_j}^\top)^\top$, $P(\widehat{\mathcal{M}}_j) = 1/p'$ and τ is a constant between 0 and 1. When $\tau = 0$, the method is the same with the original BIC. As a result, the model with the smallest EBIC_τ^j could be selected.

Practically, the computation of log likelihood $\ell_j(\widehat{\theta}_j)$ is intensive, since the determinant of a high dimensional matrix $(I - \rho W)$ is involved. Alternatively, we use another method to save computational cost here. It is shown that it works well in numerical studies. Define $\text{RSS}_{\widehat{\mathcal{M}}_j} = Y^\top(I_n - H_j)Y$ and $\sigma^2_{\widehat{\mathcal{M}}_j} = n^{-1}\text{RSS}_{\widehat{\mathcal{M}}_j}$, where $H_j = \mathbb{X}_{\rho,j}^\top(\mathbb{X}_{\rho,j}^\top\mathbb{X}_{\rho,j})^{-1}\mathbb{X}_{\rho,j}$ and $\mathbb{X}_{\rho,j} = (WY, \mathbb{X}_{\widehat{\mathcal{M}}_j}) \in \mathbb{R}^{n\times(j+1)}$. Thus $-2\ell_j(\widehat{\theta}_j)$ in (S6.1) could be replaced by $n\log(\sigma^2_{\widehat{\mathcal{M}}_j})$ as an approximation, which leads to

$$\widetilde{\text{EBIC}}_\tau^j = n\log(\sigma^2_{\widehat{\mathcal{M}}_j}) + j\log(n) + 2\tau\log\{P(\widehat{\mathcal{M}}_j)\}. \tag{S6.2}$$

Then c_γ and $\widehat{\mathcal{M}}^R$ could be selected based on the value of $\widetilde{\text{EBIC}}_\tau^j(1 \le j \le p')$ similarly. In this way, we do not need to obtain the maximum likelihood estimator for the SAR model with different js. We illustrate the performance of the method by numerical studies.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory, Ed. B. N. Petrov & F. Csaki*, 267?81. Budapest: Akademia Kiado.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika*. 95, 759?71.

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.

Saulis, L. and Statulevičius, V. (2012). Limit Theorems for Large Deviations. *Springer Science & Business Media*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*. 6, 461?64.

Vershynin, R. (2017). High-Dimensional Probability: An Introduction with Applications. *Cambridge University Press*.

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

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