High Dimensional and Banded Vector Autoregressions

Shaojun Guo† Yazhen Wang† Qiwei Yao*

†Academy of Mathematics and Systems Science, CAS, Beijing, 100190, China
‡Department of Statistics, London School of Economics, London, WC2A 2AE, UK
‡Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
*Guanghua School of Management, Peking University, Beijing, 100871, China

guoshaoj@amss.ac.cn yzwang@stat.wisc.edu q.yao@lse.ac.uk

2 March 2015

Abstract

We consider a class of vector autoregressive (VAR) models with banded coefficient matrices. The setting represents a type of sparse structure for high-dimensional time series, though the implied autocovariance matrices are not banded. The structure is also practically meaningful when the order of component time series is arranged appropriately. The convergence rates for the estimated banded autoregressive coefficient matrices are established. We also propose a Bayesian information criterion (BIC) for determining the width of the bands in the coefficient matrices, which is proved to be consistent. By exploring some approximate banded structure for the autocovariance functions of banded VAR processes, some consistent estimators for the autocovariance matrices are constructed. Illustration with both simulated and real data sets is reported.

Some key words: Banded auto-coefficient matrices; BIC; Convergence in Frobenius norm; Large autocovariance matrices; Vector autoregressive model.

*Partially supported by Key Laboratory of RCSDS (S. Guo), NSF Grants DMS-10-5635 and DMS-12-65203 (Y. Wang) and an EPSRC research grant (S. Guo and Q. Yao).
1 Introduction

In this modern information age, the availability of large or vast time series data brings the opportunities with challenges to time series analysts. The demand of modelling and forecasting high-dimensional time series arises from various practical problems such as panel study of economic, social and natural (such as weather) phenomena, financial market analysis, communication engineering. When the dimension of time series is large or merely moderately large, the statistical modelling is challenging as the vector autoregressive and moving average models suffer from the lack of identification, over-parameterization and flat likelihood functions. While pure vector autoregressive (VAR) models are perfectly identifiable, their usefulness is often compounded by a proper means of reducing the number of parameters.

In many practical situations it is often enough to collect the information from ‘neighbour’ variables. (The definition of neighbourhoods is case-dependent.) For example, sales, prices, weather indices or electricity consumptions (influenced by temperature) depend more on those at close range locations. The information from further locations may become redundant given that from neighbours. See, for example, se Can and Megbolugbe (1997) for a house price data example which exhibits such a dependence structure. In this paper, we propose a class of VAR models to cater for such dynamic structures. We assume that the autoregressive coefficient matrices are banded, i.e. non-zero coefficients form a narrow band along the main diagonal in each autoregressive coefficient matrix. The setting specifies explicit autoregression over neighbour component series only. Note that the non-zero cross correlations among all component series may still present, as the implied auto-covariance matrices are not banded. This is an effective way to impose sparse structure for high-dimensional VAR models, as the number of parameters in each autoregressive coefficient matrix is reduced from $p^2$ to $O(p)$, where $p$ denotes the dimension of time series. In practice, a banded structure may be employed by arranging the order of component series appropriately. Based on the imposed banded structure, we propose the least squares estimators for the autoregressive coefficient matrices, which attain the convergence rate $\sqrt{p/n}$ under the Frobenius norm and $\sqrt{\log p/n}$ under the spectral norm when the dimension of time series $p$ diverges together with the length of time series (i.e. the sample size) $n$.

In practice the maximum width of the non-zero coefficient bands in the coefficient matrices is unknown, which is called the bandwidth. We propose a marginal Bayesian information criterion
to identify the true bandwidth. It is shown that this criterion leads to consistent bandwidth
determination when both $n$ and $p$ tend to infinity.

We also address the estimation for the autocovariance functions for high-dimensional banded
VAR models. Although the autocovariance matrices of a banded VAR process are unlikely to be
banded, they admit some asymptotic banded approximations when the covariance of innovations
is also banded. Based on this property, the band-truncated sample autocovariance matrices are
consistent estimators with the convergence rate $\log(n/\log p)^{1/2}/\sqrt{\log(p)/n}$, which is faster than the
standard banding covariance estimators (Bickel and Levina, 2008). See also Wu and Pourahmadi
(2009), and Bickel and Gel (2011) for the estimation for the banded covariance matrices for time
series.

Most existing literature on high-dimensional VAR models draw inspiration and energy from the
recent developments in ‘large $p$ small $n$’ regression paradigm. For example, Hsu et al. (2008) proposed
the Lasso penalization for subset autoregression. Haufe et al. (2009) introduced the group sparsity
for coefficient matrices and advocated to use the group lasso penalization. A truncated weighted
lasso penalization approach was proposed by Shojaie and Michailidis (2010) to exploring graphical
granger causality in a VAR model. Song and Bickel (2011) proposed a lasso penalty for select the
variables and lags simultaneously. Bolstad et al. (2011) inferred sparse causal networks through
VAR processes and proposed a group lasso procedure. Kock and Callot (2012) established oracle
inequalities for high-dimensional VAR. Han and Liu (2013) proposed an alternative Dantzig-type
penalization and formulated the estimation problem into a linear program. Different from the above
approaches, Davis et al. (2012) proposed a two-stage method based on partial coherence measure
to identify sparse structures in autoregressive coefficient matrices. Their method belongs to the
more traditional spectral approach for time series. Chen et al. (2013) studied sparse covariance and
precision matrix in high dimensional time series under a general dependence structure.

The rest of this paper is organized as follows. Section 2 presents the banded VAR model and
the associated estimation methods, including a BIC rule for determining the bandwidth parameter.
The asymptotic properties are established in Section 3. Section 4 deals with the estimation for auto-
covariance functions of banded VAR processes. Numerical illustration with both simulated and real
data is reported in Section 5. All technical proofs are relegated to the Appendix.
2 Methodology

2.1 Banded VAR models

Let \( y_t \) be a \( p \times 1 \) time series process defined by

\[
y_t = A_1 y_{t-1} + \cdots + A_d y_{t-d} + \varepsilon_t,
\]

where \( \varepsilon_t \) is the innovation at time \( t \), \( E \varepsilon_t = 0 \) and \( \text{Var}(\varepsilon_t) = \Sigma_\varepsilon \), and \( \varepsilon_t \) is independent of \( y_{t-1}, y_{t-2}, \cdots \). Furthermore, all the coefficient matrices \( A_1, \cdots, A_d \) are banded matrices in the sense that

\[
a^{(\ell)}_{ij} = 0 \quad \text{for all} \quad |i - j| > k_0 \quad \text{and} \quad 1 \leq \ell \leq d,
\]

where \( a^{(\ell)}_{ij} \) denotes the \((i, j)\)-th element of \( A_\ell \). Thus the maximum number of non-zero elements in each row of \( A_\ell \) is \( 2k_0 + 1 \) which is the bandwidth, and \( k_0 \) is called the bandwidth parameter. We always assume that \( k_0 \geq 0 \) and \( d \geq 1 \) are fixed integers, and \( p \) is much greater than both \( k_0 \) and \( d \). Our goal is to determine the bandwidth parameter \( k_0 \) and to estimate the banded coefficient matrices \( A_1, \cdots, A_d \). For simplicity, we assume that the autoregressive order \( d \) is known, as the order determination problem has already been thoroughly studied and well-documented. See, e.g., Chapter 4 of Lütkepohl (2007).

Under the condition \( \det(I_p - A_1 z - \cdots - A_d z^d) \neq 0 \) for any \( |z| \leq 1 \), model (2.1) admits a (weakly) stationary solution \( \{y_t\} \), where \( I_p \) denotes the \( p \times p \) identity matrix. Throughout this paper, \( y_t \) is referred to this stationary process. If, in addition, \( \varepsilon_t \) is i.i.d., \( y_t \) is also strictly stationary.

In model (2.1), we do not require \( \text{Var}(\varepsilon_t) = \Sigma_\varepsilon \) to be banded. But even it is, the autocovariance matrices are not necessarily banded; see (4.1) below. Therefore, the proposed banded model is applicable when the linear dynamics of each component series depends predominately on its neighbour series, though any pair components of \( y_t \) may still be correlated at some time lags.

2.2 Estimating banded autoregressive coefficient matrices

Since each row of \( A_\ell \) has maximum \( 2k_0 + 1 \) non-zero elements, there are at most \((2k_0 + 1) d\) regressors in each row on the RHS of (2.1). For \( i = 1, \cdots, p \), let \( \beta_i \) be the column vector obtained by stacking the non-zero elements in the \( i \)-th rows of \( A_1, \cdots, A_d \) together. Let \( \tau_i \) denote the length of \( \beta_i \). Then

\[
\tau_i \equiv \tau_i(k_0) = \begin{cases} (2k_0 + 1)d & i = k_0 + 1, k_0 + 2, \cdots, p - k_0, \\ (2k_0 + 1 - j)d & i = k_0 + 1 - j \text{ or } p - k_0 + j, \quad j = 1, \cdots, k_0. \end{cases}
\]
Now (2.1) can be written as

\[ y_{i,t} = x_{i,t}^T \beta_i + \varepsilon_{i,t}, \quad i = 1, \cdots, p, \]  

(2.4)

where \( y_{i,t}, \varepsilon_{i,t} \) are, respectively, the \( i \)-th component of \( y_t \) and \( \varepsilon_t \), and \( x_{i,t} \) is the \( \tau_i \times 1 \) vector consisting of the corresponding components of \( y_{t-1}, \cdots, y_{t-d} \). Consequently, the least squares estimator of \( \beta_i \) based on (2.4) admits the form

\[ \hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T y(i), \]  

(2.5)

where \( y(i) = (y_{i,d+1}, \cdots, y_{i,n})^T \), and \( X_i \) is an \( (n-d) \times \tau_i \) matrix with \( x_{i,d+j}^T \) as its \( j \)-th row.

By estimating \( \beta_i \) for \( i = 1, \cdots, p \) separately based on (2.5), we obtain the least squares estimators \( \hat{A}_1, \cdots, \hat{A}_d \) for the coefficient matrices in (2.1). Furthermore, the residual sum of squares resulting from estimating the non-zero elements in the \( i \)-th rows of \( A_1, \cdots, A_d \) is

\[ \text{RSS}_i \equiv \text{RSS}_i(k_0) = y_i^T \{ I_{n-d} - X_i(X_i^T X_i)^{-1} X_i^T \} y(i). \]  

(2.6)

We write RSS\(_i\) as the function of \( k_0 \) to reflect the fact that the above estimation is based on the assumption that the bandwidth is \((2k_0 + 1)\) in the sense of (2.2).

### 2.3 Determination of bandwidth

In practice the bandwidth is unknown and we need to estimate the bandwidth parameter \( k_0 \). We propose to determine \( k_0 \) based on the marginal BIC defined as

\[ \text{BIC}_i(k) = \log \{ \text{RSS}_i(k) \} + \frac{1}{n} d \tau_i(k) \log(p \lor n) C_n, \quad i = 1, \cdots, p, \]  

(2.7)

where \( \text{RSS}_i(k) \) and \( \tau_i(k) \) are defined, respectively, in (2.6) and (2.3), and \( C_n > 0 \) is a constant which diverges together with \( n \); see condition A2 below. We often take \( C_n \) to be \( \log \log n \). The estimator for \( k_0 \) is now defined as

\[ \hat{k} = \max_{1 \leq i \leq p} \{ \arg \min_{1 \leq k \leq K} \text{BIC}_i(k) \}, \]  

(2.8)

where \( K \geq 1 \) is a prescribed integer.
3 Asymptotic properties

3.1 Regularity conditions

For any vector \( v = (v_1, \cdots, v_j) \) and matrix \( B = (b_{ij}) \), let

\[
\|v\|_q = \left( \sum_{j=1}^{p} |v_j|^q \right)^{1/q}, \quad \|v\|_\infty = \max_{1 \leq j \leq p} |v_j|,
\]

\[
\|B\|_q = \max_{\|v\|=1} \|Bv\|_q, \quad \|B\|_F = \left( \sum_{i,j} b_{ij}^2 \right)^{1/2},
\]

i.e. \( \| \cdot \|_q \) denotes the \( L_q \) norm of a vector or matrix, and \( \| \cdot \|_F \) is the Frobenius norm for a matrix.

First we note that the VAR model (2.1) can be casted into the following VAR(1) form

\[
\tilde{y}_t = \tilde{A} \tilde{y}_{t-1} + \tilde{\varepsilon}_t,
\]

where

\[
\tilde{y}_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-d+1} \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_d \\ I_p & 0_p & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & I_p & 0 \end{pmatrix}, \quad \tilde{\varepsilon}_t = \begin{pmatrix} \varepsilon_t \\ 0_{p \times 1} \\ \vdots \\ 0_{p \times 1} \end{pmatrix}.
\] (3.1)

Now some regularity conditions are in order.

A1. For \( \tilde{A} \) defined in (3.1), \( \|\tilde{A}\|_2 \leq C \) and \( \|\tilde{A}^{j_0}\|_2 \leq \delta^{j_0} \), where \( C > 0, \delta \in (0,1) \) and \( j_0 \geq 1 \) are constants free of \( n \) and \( p \), and \( j_0 \) is an integer.

A2. With \( a_{ij}^{(\ell)} \) be the \((i,j)\)-th element of \( A_\ell \), we assume that \( |a_{i,i+j_0}^{(\ell)}| \) or \( |a_{i,i-k_0}^{(\ell)}| \) is greater than \( C_n \sqrt{\log(p \vee n)/n} \) for some \( 1 \leq \ell \leq d, 1 \leq i \leq p \), where \( C_n \to \infty \) as \( n \to \infty \).

A3. Let \( \Sigma_0 = \text{Var}(y_t) \) and \( \Sigma_{ii} \) be the \(i\)-th diagonal element of \( \Sigma_0 \). Assume that \( \lambda_{\min}(\Sigma_0) \geq \kappa_1 \) and \( \max_{1 \leq i \leq p} |\Sigma_{ii}| \leq \kappa_2 \) for some positive constants \( \kappa_1 \) and \( \kappa_2 \) free of \( p \), where \( \lambda_{\min}(\cdot) \) denotes the minimum eigenvalue.

A4. The innovation process \( \{\varepsilon_t, t = 0, \pm 1, \pm 2, \cdots\} \) are i.i.d. with zero mean and covariance \( \Sigma_\varepsilon \).

Furthermore, one of the two assertions below holds.

(i) \( \max_{1 \leq j \leq p} E|\varepsilon_{j,t}|^{2q} \leq C \) and \( p = O(n^\beta) \), where \( q > 2, \beta \in (0, (q-2)/4) \) and \( C > 0 \) are some constants free of \( n \) and \( p \).
(ii) $\max_{1 \leq j \leq p} E \exp(t|\varepsilon_{j,t}|^{2\alpha}) \leq C$ for all $t \in (0,t_0]$, and $\log p = o(n^{\alpha/2})$, where $t_0 > 0$, $\alpha \in (0,1]$ and $C > 0$ are some constants free of $n$ and $p$.

Condition A1 implies $y_t$ to be strictly stationary provided $\varepsilon_t$ is i.i.d.. Condition A1 also implies that for any $j \geq 1$, $\|\tilde{A}_j\|_2 \leq C\delta^j$ with some constant $C > 0$ and $\delta \in (0,1)$. The i.i.d assumption is not essential and it is imposed to simplify the proofs. Condition A2 ensures that the bandwidth $(2k_0 + 1)$ are asymptotically identifiable as $\sqrt{\log(p \vee n)/n}$ is the minimum order of a non-zero coefficient to be identifiable; see, e.g., Luo and Chen (2013). Condition A3 ensures that the covariance matrix $\Sigma_0$ is strictly positive definite. Condition A4 specifies the two asymptotic modes: (i) the high-dimension cases with $p = O(n^\beta)$, and (ii) the ultra high-dimension cases with $\log p = o(n^{\alpha - \alpha/2})$. Note that the larger $p$ is, the faster the distribution tails of $\varepsilon_t$ decay under condition A4.

### 3.2 Asymptotic theorems

We first state the consistency of the BIC selector $\hat{k}$, defined in (2.8), for determining the bandwidth parameter $k_0$.

**Theorem 1.** Under conditions A1 – A4, $P(\hat{k} = k_0) \to 1$ as $n \to \infty$.

Since $k_0$ is unknown, we let $k_0 = \hat{k}$ in Section 2.2 for constructing the estimators. Theorem 2 below addresses to $\hat{A}_1, \ldots, \hat{A}_d$ constructed with $k_0$ estimated by $\hat{k}$.

**Theorem 2.** Let conditions A1 – A4 hold. As $n \to \infty$, it holds for $j = 1, \ldots, d$ that

$$\|\hat{A}_j - A_j\|_F = O_P\left(\sqrt{p/n}\right) \quad \text{and} \quad \|\hat{A}_j - A_j\|_2 = O_P\left(\sqrt{(\log p)/n}\right).$$

Conditions A4(i) and A4(ii) impose a high moment condition and an exponential tail condition on the innovation, respectively. Of course, A4(ii) places a much stronger condition on the innovation than A4(i) in a sense that if the exponential condition in A4(ii) holds for a given $\alpha \in (0,1]$, the moment condition in A4(i) is also true for any $q$. As a result, the asymptotic theory may allow the dimension $p$ to grow polynomially or exponentially with sample size $n$ under the corresponding cases in A4(i) or A4(ii), which is reflected by the different requirements on $p$ in A4(i) and A4(ii).

Although the convergence rates in Theorem 2 have the same expressions in terms of $n$ and $p$, due to the difference on the allowable growth of dimension $p$ on sample size $n$, the actual asymptotic framework set-up and convergence rates may be quite different under these two cases. For example, under Condition A4(i) we may allow $p$ to grow like $n^\beta$ with $\beta$ specified in A4(i) and obtain the...
convergence rate $\sqrt{\log n/n}$ for $\hat{A}_j$ under the spectral norm, while under Condition A4(ii) we can take $p$ exponentially large as $\exp\{n^{\alpha/(2-\alpha)-2\epsilon}\}$ with $\alpha$ specified in A4(ii) and a small but fixed positive constant $\epsilon$ and get the convergence rate $n^{\alpha/(4-2\alpha)-\epsilon-1/2}$ for $\hat{A}_j$ under the spectral norm.

4 Estimation for auto-covariance functions of banded VAR

For the banded VAR process $y_t$ defined by (2.1), the auto-covariance function $\Sigma_j = \text{Cov}(y_t, y_{t+j})$ is unlikely to be banded. For example for a stationary banded VAR(1) process, it can be shown that $\Sigma_0 = \text{Var}(y_t) = \Sigma_\varepsilon + \sum_{i=1}^{\infty} A_i^T \Sigma_\varepsilon (A_i^T)^i$.

(4.1)

Note that for any banded matrices $B_1$ and $B_2$ with bandwidths $2k_1 + 1$ and $2k_2 + 1$, the product $B_1 B_2$ is a banded matrix with the enlarged bandwidth $2(k_1 + k_2) + 1$ in general. Thus $\Sigma_0$ presented in (4.1) is not a banded matrix. Nevertheless if $\text{Var}(\varepsilon_t) = \Sigma_\varepsilon$ is also banded (see condition A5 below), Theorem 3 below shows that $\Sigma_j$ can be approximated by some banded matrices.

A5. Matrix $\Sigma_\varepsilon$ is banded with bandwidth $2s_0 + 1$ and $\|\Sigma_\varepsilon\|_1 \leq C < \infty$, where $C, s_0 > 0$ are constants free of $p$, and $s_0$ is an integer.

Theorem 3. Let conditions A1 and A5 hold. For any integers $r, j \geq 1$, there exist banded matrices $\Sigma_0^{(r)}$ with bandwidth $2(2rk_0 + s_0) + 1$, and $\Sigma_j^{(r)}$ with bandwidth $2((2r + j)k_0 + s_0) + 1$ such that

$$
\|\Sigma_0^{(r)} - \Sigma_0\|_2 \leq C_1 \delta^{2(r+1)}, \quad \|\Sigma_0^{(r)} - \Sigma_0\|_1 \leq C_2 r \delta^{2(r+1)},
$$

$$
\|\Sigma_j^{(r)} - \Sigma_j\|_2 \leq C_3 \delta^{2(r+j)+1}, \quad \|\Sigma_j^{(r)} - \Sigma_j\|_1 \leq C_4 r \delta^{2(r+j)+1},
$$

where $C_1, C_2, C_3, C_4$ are positive constants independent of $r$ and $p$, and $\delta \in (0, 1)$ is specified in condition A1.

Theorem 3 can be intuitively illustrated with (4.1), i.e. $\Sigma_0$ for a banded VAR(1) process. Under condition A5, $\Sigma_0^{(r)} = \Sigma_\varepsilon + \sum_{1 \leq i \leq r} A_i^T \Sigma_\varepsilon (A_i^T)^i$ is a banded matrix with bandwidth $2(2rk_0 + s_0) + 1$. Condition A1 ensures that the norms of the difference $\Sigma_0 - \Sigma_0^{(r)} = \sum_{i>r} A_i^T \Sigma_\varepsilon (A_i^T)^i$ admit the required upper bounds.

Theorem 3 also paves the way to estimate $\Sigma_j$ by using the banding method of Bickel and Levina (2008), as $\Sigma_j$ can be treated as a banded matrix with a bounded error. To this end, we define the
banding operator as follows: for any matrix $H = (h_{ij})$, $B_r(H) = (h_{ij}I(|i - j| \leq r))$. Then the banding estimator for $\Sigma_j$ is defined as

$$\hat{\Sigma}_j^{(r_n)} = B_{r_n}(\tilde{\Sigma}_j),$$

where $\tilde{\Sigma}_j = \frac{1}{n} \sum_{t=1}^{n-j} (y_t - \bar{y})(y_{t+j} - \bar{y})^T$, $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$.

where $r_n = C \log(n/\log p)$, and $C > 0$ is a constant with $C \geq 4^{-1}(\log \delta^{-1})^{-1}$. Theorem 4 below presents the convergence rates for $\hat{\Sigma}_0^{(r_n)}$ and $\hat{\Sigma}_j^{(r_n)}$. The rates are faster than those presented in Bickel and Levina (2008), due to the approximate banded structure presented in Theorem 3.

**Theorem 4.** Assume that the conditions A1 – A5 hold. Then as $n \to \infty$ and $p \to \infty$,

$$\|\hat{\Sigma}_0^{(r_n)} - \Sigma_0\|_2 = O_P\left(n\sqrt{\frac{\log p}{n} + \delta^{2(r_n+1)}}\right) = O_P\left(\log(n/\log p)\sqrt{\frac{\log p}{n}}\right),$$

$$\|\hat{\Sigma}_j^{(r_n)} - \Sigma_j\|_2 = O_P\left(n\sqrt{\frac{\log p}{n} + \delta^{2(r_n+j)+1}}\right) = O_P\left(\log(n/\log p)\sqrt{\frac{\log p}{n}}\right),$$

and

$$\|\hat{\Sigma}_0^{(r_n)} - \Sigma_0\|_1 = O_P\left(\log(n/\log p)\sqrt{\frac{\log p}{n}}\right) = \|\hat{\Sigma}_j^{(r_n)} - \Sigma_j\|_1.$$

From a practical point of view, we need to select $r_n$ in a data-driven way. A standard way to choose $r_n$ is to minimize the risk

$$R_j(r) = E\|\hat{\Sigma}_j^{(r)} - \Sigma_j\|_1$$

and the oracle bandwidth is given by $r_{nj} = \arg\min_r R_j(r)$. In practice $R_j(r)$ is unavailable due to unknown $\Sigma_j$. We replace it by an estimator obtained via a version of wild bootstrap. To this end, let $u_1, \ldots, u_n$ be i.i.d. with $E u_t = \text{Var}(u_t) = 1$. Then a bootstrap estimator for $\Sigma_j$ is defined as

$$\Sigma_j^* = \frac{1}{n} \sum_{t=1}^{n-j} u_t(y_t - \bar{y})(y_{t+j} - \bar{y})^T.$$

For example, we may draw $u_t$ from the standard exponential distribution. Consequently the bootstrap estimator for $R_j(r)$ is defined as

$$R_j^*(r) = E\{\|B_r(\Sigma_j^*) - \hat{\Sigma}_j\|_1 | y_1, \ldots, y_n\}.$$

We choose $r_{nj}$ to minimize $R_j^*(r)$. In practice we use the approximation

$$R_j^*(r) \approx \frac{1}{q} \sum_{k=1}^{q} \|B_r(\Sigma_j^*_{j,k}) - \hat{\Sigma}_j\|_1,$$

where $\Sigma_j^*_{j,1}, \ldots, \Sigma_j^*_{j,q}$ are $q$ bootstrap estimates for $\Sigma_j$, obtained by repeating the above wild bootstrap scheme $q$ times, and $q$ is a large integer.
5 Numerical properties

5.1 Simulations

In this section, we conduct simulation to evaluate the finite sample properties of the proposed methods in the VAR(1) model,

\[ y_t = A y_{t-1} + \varepsilon_t, \]

where \( \{\varepsilon_t\} \) are independent and \( N(0, I_p) \). We consider two settings for the banded coefficient matrix \( A = (a_{ij}) \) as follows:

(i) \( \{a_{ij}, |i - j| \leq k_0\} \) are generated independently from \( U[-1, 1] \). Since the spectral norm of \( A \) must be smaller than 1, we re-scale \( A \) by \( \eta \cdot A/\|A\|_2 \), where \( \eta \) is generated from \( U[0.3, 1.0] \).

(ii) \( \{a_{ij}, |i - j| < k_0\} \) are generated independently from the mixture distribution \( \xi \cdot 0 + (1 - \xi) \cdot N(0, 1) \) with \( P(\xi = 1) = 0.4 \). The elements \( \{a_{ij}, |i - j| = k_0\} \) are i.i.d. discrete variables taking values -4, 0 and 4 with probability 0.3, 0.4 and 0.3, respectively. \( A \) is then rescaled as in (i) above.

In Setting (ii), there are about \( 0.4(2k_0 + 1)p \) non-zero elements within the band, i.e., \( A \) is more sparse than that in Setting (i).

We set \( n = 200, p = 100, 200, 400, 800, \) and \( k_0 = 1, 2, 3, 4 \). We repeat each setting 500 times. Table 1 lists the relative frequencies of the occurrence of the events \( \{\hat{k} = k\}, \{\hat{k} > k_0\} \) and \( \{\hat{k} < k_0\} \) in the 500 replications. Overall there is tendency that \( \hat{k} \) under-estimated \( k_0 \), especially when \( k_0 = 3 \) or 4. (In fact when \( k_0 = 4 \), \( \hat{k} \) chose 3 most times.) Note that the constraint \( \|A\| < 1 \) makes most non-zero elements small or very small when \( p \) is large. Only the coefficients at least as large as \( \sqrt{\log(p \vee n)/n} \) are identifiable; see condition A2. It is also clear that the estimation performs better in Setting (ii) than in Setting (i), as condition A2 is more likely to hold at the boundaries of the band in Setting (ii).

The BIC (2.7) is defined for each row separately. One natural alternative would be

\[
\text{BIC}(k) = \sum_{i=1}^p \log(\text{RSS}_i(k)) + \frac{1}{n}|\tilde{\tau}(k)| \log(p \vee n) C_n,
\]

where \( \tilde{\tau}(k) = (2p + 1)k - k^2 - k \) is the total number of parameters in the model. This leads to the estimator for the bandwidth parameter as follows.

\[
\tilde{k} = \arg \min_{1 \leq k \leq K} \text{BIC}(k). \tag{5.1}
\]

10
Although this joint BIC approach can be shown to be consistent, its finite sample performance, reported in Table 2, is worse than that of the marginal BIC (2.7). See also Table 1.

We also calculate both $L_1$ and $L_2$ errors in estimating the banded coefficient matrix $A$. The means and the standard deviations of the errors for Setting (i) are reported in the left panel of Table 3. In the right panel of Table 3, we report the results from estimating $A$ using the true values for the bandwidth parameter $k_0$. The loss in the accuracy in estimating $A$ caused by unknown $k_0$ is almost negligible. The results for Setting (ii) are similar, and are therefore omitted.

To evaluate the performance of the estimation for the auto-covariance matrices $\Sigma_0$ and $\Sigma_1$, we set $k_0 = 3$, and the spectral norm of $A$ at 0.8. Furthermore, $\varepsilon_t$ are independent and follow a multi-normal distribution with mean zero and variance $\Sigma_\varepsilon$, where $\Sigma_\varepsilon = (\sigma_{ij},\varepsilon) = 0.8 I(|i - j| = 1) + I(i = j)$. Table 4 lists the average estimation errors and the standard deviations over 100 replications, measured by matrix $L_1$-norm. For the sake of comparison, we also report the results for a threshold estimator and the sample covariance estimator. For the banded estimator, we choose $r$ which minimizes the bootstrap loss defined in (4.2) with $q = 100$. For the threshold estimator, the thresholding parameter is selected in the same manner. Table 4 shows clearly that the proposed banding method performs much better than the threshold estimator since it is directly adaptive to the underlying structure, while the sample covariance performs much worse than both the banding and threshold methods.

5.2 Real data examples

We illustrate the proposed method with two real data sets in this section.

Example 1. Consider the weekly temperature data across the 71 cities in China in the period of 1 January 1990 – 17 December 17 2000 (i.e., $p = 71$ and $n = 572$). The 71 cities are arranged from north to south along latitudes. Fig 1 displays the weekly temperature of three selected cities (i.e. Ha’erbin, Shanghai and Hangzhou), showing strong seasonal behavior with period 52 (a year). Therefore, we set the seasonal period to be 52 and estimate the seasonal effects by taking averages of the same weeks across different years. The deseasonalized series, i.e. the original series subtracting estimated seasonal effects, are denoted as $\{y_t, t = 1, \cdots, 572\}$, and each $y_t$ has 71 components. Fig 2 displays the three component series of $y_t$ for cities Ha’erbin, Shanghai and Hangzhou.

We fit a banded VAR(1) to $y_t$, with the bandwidth parameter $k = 2$ determined by the proposed (marginal) BIC. This implies that for each of the 71 cities, the temperatures in the previous week
from the 4 neighbour cities are incorporated in the model. The estimated banded coefficient matrix \( \hat{A} \) is depicted in the left panel of Fig. 3.

We also examine the prediction performance of the banded VAR fitting. For each of the last 30 data points in the series, we use the data in the past to fit a banded VAR(1) model with the bandwidth parameter set at \( k_0 = 2 \). We calculate the one-step-ahead forecasts based on the fitted banded VAR models, and the two-step-ahead forecasts by plugging in the one-step-ahead forecasted values into the fitted models. The forecasts for the original temperatures are obtained by adding up the forecasted values of \( y_t \) and the corresponding (estimated) seasonal components. We calculate the mean absolute prediction error (MAPE) for the temperatures over the last 30 weeks for each of the 71 cities. The mean and standard deviation of those 71 MAPEs are listed in Table 5.

For the comparison purpose, we also fit the data with lasso VAR(1) obtained by minimizing

\[
\sum_{t=2}^{n} \|y_t - Ay_{t-1}\|^2 + \sum_{i,j=1}^{p} \lambda_i |a_{ij}|, 
\]

where \( \{\lambda_i, i = 1, \cdots, p\} \) are tuning parameters. Here the tuning parameters \( \{\lambda_i, i = 1, \cdots, p\} \) are estimated by 5-fold cross-validation (Bickel and Levina 2008). The estimated coefficient matrix \( \hat{A} \) is depicted in the right panel of Fig. 3. The post-sample forecasting errors are reported in Table 5 together with those based on the fitted banded VAR(1). The forecasting accuracies of the two models are comparable, though the banded VAR performs better. However the clear neighbourhood dependence of the banded VAR displayed in the panel on the left in Fig. 3 is attractive, in contrast to the lack of the structure in the sparse lasso fitting displayed in the other panel in Fig. 3.

**Example 2.** We consider the daily sales of a clothing brand in 21 provinces in China in the period of 1 January 2008 – 9 December 2012 (i.e., \( n = 1812, p = 21 \)). Fig. 4 is a map of those 21 provinces. The 21 provinces are arranged from north to south along latitudes. After subtracting the sample mean from the data, we fit the data with a banded VAR(1) and a lasso-VAR(1). For the banded VAR(1), the estimated bandwidth parameter is \( \hat{k} = 4 \). The estimated coefficient matrices are displayed in Fig. 5.

The post-sample forecasting for the last 100 data points in the series is conducted in the same manner as in Example 1. The forecasting errors by the banded VAR(1) and the lasso VAR(1) are summarized in Table 6. While the forecasting performances of the two methods are comparable, the banded structure of \( \hat{A} \) makes the interpretation of the fitted model much more plausible. While the
non-zero coefficients in the lasso estimator $\tilde{A}$ are scattered all over the places, it is difficult to argue why the sales in one province depends on some remote provinces more than on its close neighbours. See Fig.5

6 Conclusion

In this paper, we study a sparse vector autoregressive model with banded autoregressive coefficient matrices. The setting is practically relevant, as in many real-life scenarios it is often enough to include the information from ‘neighbour’ series. The model does not rule out the possible correlations between ‘remote’ series, as the implied autocovariance matrices are not necessarily banded. The fitted model facilitates easier interpretation than, for example, a sparse VAR model regularized by lasso. Our numerical examples show that it also offers competitive forecasting performance.

APPENDIX

In this section, we provide the detailed proofs of Theorems 1-4. We introduce some technical lemmas first in Appendix A.

Appendix A: Lemmas

To establish the limit theory of polynomial tail case (under condition A4(i)), we shall adopt the asymptotic theories using the functional dependent measure of Wu (2005, 2013). Assume that $z_i$ is a stationary process of the form $z_i = g(F_i)$, where $g(\cdot)$ is a measurable function and $F_i = (\cdots, e_{-1}, e_0, \cdots, e_i)$ with i.i.d random variables $\{e_i, i = 0, \pm 1, \cdots\}$. Wu(2005) defined the functional dependent measures in terms of how the outputs are affected by the inputs. To be specific, assume $\|z\|_q = (E|z|^q)^{1/q}$ with $q \geq 1$. The physical or functional dependent measure is defined as

$$\theta_{i,q} = \|z_i - z_i^*\|_q = \|g(F_i) - g(F_i^*)\|_q$$

where $z_i^* = g(F_i^*)$ is the coupled process of $z_i$, $F_i^* = (\cdots, e_{-1}, e_0^*, \cdots, e_i)$ with $\{e_0^*, e_0\}$ being i.i.d.

Intuitively, $\theta_{i,q}$ measures the dependency of $z_i$ on $e_0$ while keeping all other innovations unchanged.

Lemma 1 (Theorem 2 (ii) of Liu, Xiao and Wu (2013)). Denote that $S_n = n^{-1/2} \sum_{i=1}^n z_i$ and $\Theta_{m,q} = \sum_{i=m}^{\infty} \theta_{i,q}$. Assume that for each $m$, $\Theta_{m,q} = O(m^{-\alpha})$ with $\alpha > 1/2 - 1/q$ and $q > 2$. Then there exist positive constants $C_1, C_2$ and $C_3$ which only depend on $q$ such that for all $x > 0$,

$$P\{|S_n| \geq x\} \leq \frac{C_1 \Theta_{0,q}^{\alpha}}{(\sqrt{n}x)^q} + C_2 \exp \left(\frac{C_3 \Theta_{0,q}^{-1}x^2}{q}\right).$$
To prove the limit theory of sub-exponential tail case (under condition A4(ii)), we shall use the following Lemmas 2-4.

**Lemma 2.** Suppose that $X$ is a random variable. Then, $E \exp(t_0 |X|^v) < \infty$ for some $0 < v \leq 2$ and $t_0 > 0$ if and only if

$$\lim \sup_{q \to \infty} q^{-1/v} \|X\|_q < \infty.$$ 

**Proof of Lemma 2.** Assume that $\zeta = E \exp(t_0 |X|^v) < \infty$. Then, for any $q \geq 2$,

$$E|X|^q = q \int_0^\infty x^{q-1} P(|X| > x)dx \leq \zeta q^{v-1} t_0^{-q/v} \cdot \int_0^\infty x^{q/v-1} \exp\left(-x\right)dx = \zeta q^{v-1} t_0^{-q/v} \cdot \Gamma\left(\frac{q}{v}\right),$$

where $\Gamma(\cdot)$ is the Gamma-function. By Stirling’s formula,

$$\lim_{x \to \infty} \Gamma(x+1)\left\{\sqrt{2\pi x} \left(\frac{x}{e}\right)^x\right\}^{-1} = 1.$$

Then, for all sufficiently large $q$,

$$\|X\|_q \leq \left(\zeta q^{v-1} t_0^{-q/v}\right)^{1/q} \cdot q^{1/(2q)} \cdot \left(\frac{q}{v} - 1\right)^{1/v-1/q} \leq C q^{1/v},$$

where $C$ is a constant only depending on $\zeta$, $v$ and $t_0$. This implies that

$$\lim \sup_{q \to \infty} q^{-1/v} \|X\|_q < \infty.$$

Conversely, assume that $\lim \sup_{q \to \infty} q^{-1/v} \|X\|_q < \infty$. Then, there exists a positive constant $\phi_0 > 0$ such that, $\|X\|_q \leq \phi_0 \cdot q^{1/v}$ for all $q \geq 2$. Note that $\exp(x) = 1 + \sum_{k \geq 1} (k!)^{-1} x^k$. To prove that $E \exp(t_0 |X|^v) < \infty$ for some $t_0 > 0$, we only need to show that there exist positive constants $t_0$ and $k_0$ such that

$$\sum_{k \geq k_0} \frac{t_0^k \|X\|_{vk}^k}{k!} < \infty.$$

By Stirling’s formula, there exists a large integer $k_0$ such that for $k \geq k_0$,

$$\Gamma(k+1) = k! \geq \sqrt{\pi k} \left(\frac{k}{e}\right)^k.$$

With such $k_0$ and $t_0 = (2\phi_0^v v e)^{-1}$, we have that

$$\sum_{k \geq k_0} \frac{t_0^k \|X\|_{vk}^k}{k!} \leq \sum_{k \geq k_0} \frac{(t_0 \phi_0^v v e)^k}{\sqrt{\pi k} k^k} \leq \sum_{k \geq k_0} 2^{-k} < \infty.$$
Lemma 3. Suppose that \( \{X_1, X_2, \cdots, X_n\} \) are independent random variables and \( \sup_{i \leq n} E \exp(t_0 |X_i|^\alpha) \leq \zeta \) for some positive constants \( \alpha, t_0 \) and \( \zeta \) with \( 0 < \alpha \leq 1 \). Then, there exist positive constants \( C_j > 0 (j = 1, \cdots, 4) \) which depend only on \( \alpha, t_0 \) and \( \zeta \) such that for any \( x > 0 \) and all \( n \), the following concentration inequality holds:

\[
P \left\{ \left| \sum_{i=1}^n (X_i - E X_i) \right| > 3x \right\} \leq C_1 \exp \left( -\frac{x^2}{C_2 n + C_3 n^{1-\alpha} x} \right) + C_1 \exp \left( -\frac{x^{2\alpha}}{C_2 n^{\frac{2\alpha}{\alpha} - \alpha} + C_3 x^{\alpha}} \right) + nC_1 \exp (-C_4 x^\alpha). \tag{6.1} \]

In particular, if \( \alpha = 1 \), then

\[
P \left\{ \left| \sum_{i=1}^n (X_i - E X_i) \right| > 3x \right\} \leq C_1 \exp \left( -\frac{x^2}{C_2 n + C_3 x} \right) + C_1 n \exp (-C_4 x) \]

if any \( x > 0 \) and all \( n \). This is very close to Bernstein’s inequality except the last term.

**Proof of Lemma 3.** If \( \alpha = 1 \), it can be proved by Bernstein’s inequality directly. So we only consider the case of \( 0 < \alpha < 1 \) here. Let \( \xi_{n1} \) and \( \xi_{n2} \) be two constants with \( 0 < \xi_{n1} < \xi_{n2} \), which depends on \( n \) and will be defined belows. Denote \( \tilde{X}_{i1} = X_iI(|X_i| \leq \xi_{n1}) \), \( \tilde{X}_{i2} = X_iI(\xi_{n1} \leq |X_i| \leq \xi_{n2}) \) and \( \tilde{X}_{i3} = X_iI(|X_i| > \xi_{n2}) \). Then \( X_i = \tilde{X}_{i1} - E(\tilde{X}_{i1}) + \tilde{X}_{i2} - E(\tilde{X}_{i2}) + \tilde{X}_{i3} - E(\tilde{X}_{i3}) \), and hence

\[
P \left\{ \left| \sum_{i=1}^n (X_i - E(X_i)) \right| > 3x \right\} \leq \sum_{k=1}^3 P \left\{ \left| \sum_{i=1}^n (\tilde{X}_{ik} - E\tilde{X}_{ik}) \right| > x \right\}. \]

In the following, we will give an upper bound for each term separately.

Now consider the first term. Let \( \sigma^2 \) be a finite constant such that \( \sup_{i \leq n} E|X_i|^2 \leq \sigma^2 \). Note that \( |\tilde{X}_{i1}| \leq \xi_{n1} \) and \( E\tilde{X}_{i1}^2 \leq \sigma^2 \) for all \( i \). By Bernstein’s inequality for bounded variables, we get that

\[
P \left\{ \left| \sum_{i=1}^n (\tilde{X}_{i1} - E\tilde{X}_{i1}) \right| > x \right\} \leq 2 \exp \left( -\frac{x^2}{2n\sigma^2 + 2\xi_{n1} x/3} \right). \tag{6.2} \]

Let us handle the second term. To use Bernstein’s inequality, we only require an appropriate control of moments. Using integration by parts, we observe that

\[
E|\tilde{X}_{i2}|^q \leq q \int_{\xi_{n1}}^{\xi_{n2}} u^{q-1} P(|X_i| > u) du + \xi_{n1}^q P(|X_i| > \xi_{n1}).
\]
for each \( q \geq 2 \). For each integer \( q \geq 2 \),

\[
q \int_{\xi_{n1}}^{\xi_{n2}} u^{q-1} P(|X_i| > u) du \leq q \zeta \int_{\xi_{n1}}^{\xi_{n2}} u^{q-1} \exp(-t_0 u^\alpha) du
\]

\[
\leq q \alpha^{-1} \zeta (2t_0^{-1})^{q/\alpha} \int_{t_0 \xi_{n1}/2}^{t_0 \xi_{n2}/2} u^{q/\alpha - 1} \exp(-2u) du
\]

\[
\leq q \alpha^{-1} \zeta (2t_0^{-1} \xi_{n2}^{-\alpha})^q \exp(-2^{-1} t_0 \xi_{n1}) \int_{t_0 \xi_{n1}/2}^{t_0 \xi_{n2}/2} u^{q-1} \exp(-u) du
\]

\[
\leq q! \cdot 4 \alpha^{-1} \zeta (t_0^{-1} \xi_{n2}^{1-\alpha})^2 \exp(-2^{-1} t_0 \xi_{n1}) \cdot (2t_0^{-1} \xi_{n2}^{-\alpha})^{q-2}.
\]

Choose \( \xi_{n1} = (4t_0^{-1} \xi_{n2}^{-\alpha} \log(n))^{1/\alpha} \) and \( \xi_{n2} = n^{\frac{1}{2-\alpha}} \lor x \). Write \( \xi_n = n^{\frac{1}{2-\alpha}} \) and \( \nu = \max(16 \zeta \alpha^{-1} t_0^{-2}, \sigma^2) \). Then

\[
q \int_{\xi_{n1}}^{\xi_{n2}} u^{q-1} P(|X_i| > u) du \leq \frac{1}{2} q! \cdot \nu \cdot (1 \lor x^{2(1-\alpha)} \xi_n^{-2}) \cdot (2t_0^{-1} (\xi_n \lor x^{1-\alpha}))^{q-2}.
\]

We also have that \( \xi_{n1}^q P(|X_i| > \xi_{n1}) \leq \xi_{n1}^2 \exp(-t_0 \xi_{n1}) = \xi_{n1}^2 \exp(-t_0 \xi_{n1}) \cdot \xi_{n1}^{-q} \). A simple analysis yields that there exists a positive integer \( N_{\alpha,t_0} \) which depends only on \( \alpha \) and \( t_0 \) such that

\[
\xi_{n1} < \xi_{n2}, \xi_{n1}^2 \exp(-t_0 \xi_{n1}) \leq 4 \alpha^{-1} \zeta t_0^{-2}, 2t_0^{-1} \xi_{n2}^{-\alpha} \geq \xi_{n1}, \quad \text{and} \quad 4 \log(n) \leq t_0 \xi_{n2}^{-\alpha}
\]

if \( n > N_{\alpha,t_0} \). Then, if \( x \leq n^{\frac{1}{2-\alpha}} \),

\[
E|\tilde{X}_{i2}|^q \leq \frac{1}{2} q! \nu (2t_0^{-1} \xi_n)^{q-2}
\]

for each integer \( q \geq 2 \). Otherwise, if \( x > n^{\frac{1}{2-\alpha}} \),

\[
E|\tilde{X}_{i2}|^q \leq \frac{1}{2} q! \nu (x^{2(1-\alpha)} \xi_n^{-2}) \cdot (2t_0^{-1} x^{1-\alpha})^{q-2}
\]

for each integer \( q \geq 2 \). By Bernstein’s inequality, we obtain that

\[
P\left\{ \left| \sum_{i=1}^{n} (\tilde{X}_{i2} - E \tilde{X}_{i2}) \right| > x \right\} \leq 2 \exp \left( \frac{-x^2}{2n \nu + 4t_0^{-1} \xi_n^{1-\alpha} x} \right) + 2 \exp \left( \frac{-x^{2\alpha}}{2\nu n^{2-\alpha} + 4t_0^{-1} x^\alpha} \right). \quad (6.3)
\]

Consider the last term. Note that

\[
\left\{ \left( \sum_{i=1}^{n} (\tilde{X}_{i3} - E \tilde{X}_{i3}) \right) > x \right\} \subset \left\{ \sup_i |X_i| > \xi_{n2} \right\} \cup \left\{ \sup_i |X_i| \leq \xi_{n2}, \sum_{i=1}^{n} |E(X_i I(|X_i| > \xi_{n2}))| > x \right\}
\]

Therefore, we have

\[
P\left\{ \left( \sum_{i=1}^{n} (\tilde{X}_{i3} - E \tilde{X}_{i3}) \right) > x \right\} \leq P\left\{ \sup_i |X_i| > \xi_{n2} \right\} + P\left\{ \sup_i |X_i| \leq \xi_{n2}, \sum_{i=1}^{n} |E(X_i I(|X_i| > \xi_{n2}))| > x \right\}.
\]
Note that \( \zeta = \sup_{i \leq n} E \exp(t_0 |X_i|^\alpha) < \infty \). We observe that
\[
P \left\{ \sup_i |X_i| > \xi_{n2} \right\} \leq \zeta n \exp \left( -t_0 \xi_{n2}^\alpha \right) \leq \zeta n \exp \left( -t_0 x^\alpha \right).
\]
In a similar fashion, we obtain that
\[
\sum_{i=1}^{n} |E\{X_i | |X_i| > \xi_{n2}\}| \leq n \sigma \sqrt{P\{(|X_i| > \xi_{n2})\}} \leq n^{-1} \sigma \cdot \zeta \exp \left( 2 \log(n) - 2^{-1} t_0 \xi_{n2}^\alpha \right).
\]
As a result, for \( x > \sigma \zeta n^{-1} \) and \( n > N_{\alpha,t_0} \),
\[
P \left\{ \sum_{i=1}^{n} (\tilde{X}_{i3} - E \tilde{X}_{i3}) > x \right\} \leq \zeta n \exp \left( -t_0 x^\alpha \right). \tag{6.4}
\]
Combing the three inequalities \((6.2) - (6.4)\), we obtain that, for \( x > \sigma \zeta n^{-1} \) and \( n > N_{\alpha,t_0} \),
\[
P \left\{ \sum_{i=1}^{n} (X_i - EX_i) \right\} > 3x \right\} \leq 4 \exp \left( -\frac{x^2}{2n \nu + 4t_0^{-1} n^{\frac{1-\alpha}{2}} x} \right) + 2 \exp \left( -\frac{x^{2\alpha}}{2n \nu^{\frac{2-\alpha}{2}} + 4t_0^{-1} x^{\alpha}} \right) + n \zeta \exp \left( -t_0 x^\alpha \right).
\]
If \( x \leq \sigma \zeta n^{-1} \) or \( n \leq N_{\alpha,t_0} \), we can always multiply a large positive constant \( C \) in the right side to make the inequality hold. The proof is completed.

**Lemma 4.** Suppose that \( \{X_1 = (X_{1,1}, X_{1,2})^T, X_2 = (X_{2,1}, X_{2,2})^T, \cdots \} \) are independent random vectors and \( \sup_{i \leq n, j = 1,2} E \exp(t_0 |X_{i,j}|^{2\alpha}) \leq \zeta \) for some positive constants \( \alpha, t_0 \) and \( \zeta \) with \( 0 < \alpha \leq 1 \). Denote \( l_n \) be a sequence and may depend on \( n \) but \( 1 \leq l_n \leq O(n^\epsilon) \) with \( 0 \leq \epsilon < 1 \). Then, for each \( m \) and \( m' \) with \( m, m' = 1, 2 \), there exists a positive constants \( C_j (j = 1, \cdots, 4) \) such that for any \( x > 0 \), the following concentration inequality holds:
\[
P \left\{ \sum_{i=1}^{n} (X_{i,m} X_{i+l_n,m'} - EX_{i,m} X_{i+l_n,m'}) \right\} > 3(l_n+1)x \right\}
\[
\leq (l_n + 1)C_1 \exp \left( -\frac{x^2}{C_2 n + C_3 n^{\frac{1-\alpha}{2}} x} \right) + C_1(l_n + 1) \exp \left( -\frac{x^{2\alpha}}{C_2 n^{\frac{2-\alpha}{2}} + C_3 x^{\alpha}} \right) + C_1(l_n + 1) n \exp \left( -C_4 x^{\alpha} \right).
\]

**Proof of Lemma 4.** Without loss of generality, we assume that \( n/(l_n + 1) \) is a positive integer. Here we only prove the inequality for \( m = 1 \) and \( m' = 2 \). Similar techniques can be applied for other cases. Let \( Y_{ji} = X_{(i-1)(l_n+1)+j,1} X_{(i+1)(l_n+1)+j-1,2} \). Then, for each \( j \), \( \{Y_{ji}, i = 1, \cdots, n/(l_n + 1)\} \)
are independent with sup, \(E \exp(t_0|Y_{ji}|^\alpha) \leq \zeta < \infty\). With the help of \(Y_{ji}\), \(\sum_{i=1}^n (X_{i,1}X_{i+t_n,2} - E(X_{i,1}X_{i+t_n,2}))\) can be re-expressed as

\[
\sum_{i=1}^n (X_{i,1}X_{i+t_n,2} - E(X_{i,1}X_{i+t_n,2})) = \sum_{j=1}^{l_n+1} \sum_{i=1}^{n/(l_n+1)} (Y_{ji} - E(Y_{ji})).
\]

By Lemma 3, we obtain that there exist positive constants \(C\) such that \(x > \exp(1)\), consequently, this leads to the following uniform convergence rate:

\[
P\left( \left| \sum_{i=1}^n (X_{i,1}X_{i+t_n,2} - E(X_{i,1}X_{i+t_n,2})) \right| > 3x \right) \leq C_1 \exp \left( -\frac{x^2}{C_2 n + C_3 n^{\frac{1}{3-\alpha}} x} \right) + C_1 \exp \left( -\frac{x^{2\alpha}}{C_2 n^\alpha + C_3 x^\alpha} \right) + C_1 n \exp (-C_4 x^\alpha).
\]

Note that

\[
\left| \sum_{i=1}^n (X_{i,1}X_{i+t_n,2} - E(X_{i,1}X_{i+t_n,2})) \right| \leq (l_n + 1) \sup_{j \leq l_n+1} \left| \sum_{i=1}^{n/(l_n+1)} (Y_{ji} - E(Y_{ji})) \right|.
\]

Therefore,

\[
P\left( \left| \sum_{i=1}^n (X_{i,1}X_{i+t_n,2} - E(X_{i,1}X_{i+t_n,2})) \right| > 3(l_n + 1)x \right)
\leq (l_n + 1) \sup_{j \leq l_n+1} P\left( \left| \sum_{i=1}^{n/(l_n+1)} (Y_{ji} - E(Y_{ji})) \right| > 3x \right).
\]

This lemma is proved.

The following Lemmas 5, 6 and 7 are based on VAR(1) model with \(\|A_1\|_2 \leq \delta < 1\). Similar techniques can be applied to the general \(d\). Some notation are given. For each \(j, k = 1, \cdots, p\), define \(\hat{\Sigma}_{jk} = n^{-1} \sum_{t=1}^n y_{j,t} y_{k,t}\) and \(\Sigma_{jk} = E(\hat{\Sigma}_{jk})\). For \(i = 1, \cdots, p\), denote \(e_{(i)} = (\varepsilon_{i,2}, \cdots, \varepsilon_{i,n})^\top\) and \(x_{(i)} = (y_{i,1}, \cdots, y_{i,n-1})^\top\). We should note that Lemmas 5 and 6 have the same rate expressions but the actual rates are different, since they are under A4 (i) and (ii), respectively.

**Lemma 5.** Suppose that Conditions (A1) - (A3) and A4(i) hold.

(i) For each \(j, k = 1, \cdots, p\), there exist positive constants \(C_1, C_2\) and \(C_3\) not depending on \((j, k, n, p)\) such that

\[
P\left( \left| \hat{\Sigma}_{jk} - \Sigma_{jk} \right| > x \right) \leq \frac{C_1 n}{(nx)^q} + C_2 \exp \left( -C_3 n x^2 \right)
\]

holds for \(x > 0\); consequently, this leads to the following uniform convergence rate:

\[
\sup_{1 \leq j, k \leq p} \left| \hat{\Sigma}_{jk} - \Sigma_{jk} \right| = O_P\left( \sqrt{\frac{\log p}{n}} \right).
\]
(ii) For each \( j, k = 1, \cdots, p \), there exist positive constants \( C_1, C_2 \) and \( C_3 \) not depending on \( (j, k, n, p) \) such that
\[
P\{ |e_{(j)}^T x_{(k)}| \geq x \} \leq C_1 n \frac{1}{x^{2q}} + C_2 \exp \left( -C_3 x^2 \right)
\]
holds for \( x > 0 \); in particular, we have
\[
\sup_{1 \leq j, k \leq p} |e_{(j)}^T x_{(k)}| = O_P \left( \sqrt{n \log p} \right).
\]

**Proof of Lemma 5.** Here we only prove the part (i). The part (ii) can be proved analogously.

Denote \( \mu_q = \sup_{j \leq p} \| \varepsilon_{j0} \|_q \) for \( q \geq 2 \). To use the results of Lemma 1, we just need to bound the physical dependent measure of \( y_{j,i}y_{k,i} \) for each \( j \) and \( k \), denoted by \( \tilde{\theta}_{i,q,j,k} = \| y_{j,i}y_{k,i} - y^*_{j,i}y^*_{k,i} \|_q \) with \( y^*_{j,i} \) being the coupled process of \( y_{j,i} \). Denote the physical dependent measure of \( y_{j,i} \) by \( \theta_{i,2q,j} = \| y_{j,i} - y^*_{j,i} \|_2q \) with \( y^*_{j,i} \) being the coupled process of \( y_{j,i} \).

We will show (a) \( \sup_{j \leq p} \| y_{j,i} \|_2q \leq C \cdot \mu_{2q} \); (b) \( \sup_{j \leq p} \theta_{i,2q,j} \leq C \cdot \mu_{2q}(i + 1)^{\delta i} \), where \( C \) is some positive constant and depends only on the spectral norm of \( A_1 \) rather than \( q \). Observe that
\[
\| y_{j,i}y_{k,i} - y^*_{j,i}y^*_{k,i} \|_q \leq \| y_{j,i}y_{k,i} - y^*_{j,i}y_{k,i} \|_q + \| y_{j,i}y_{k,i} - y^*_{j,i}y_{k,i} \|_q \leq \sup_{j \leq p} \| y_{j,i} \|_2q (\theta_{i,2q,j} + \theta_{i,2q,k})
\]
If their bounds are obtained, then,
\[
\tilde{\Theta}_{m,q} = \sup_{j,k \leq p} \sum_{i=m}^{\infty} \tilde{\theta}_{i,2q,j,k} \leq C \cdot \mu_{2q}^2 \sum_{i=m}^{\infty} (i + 1)^{\delta i} \leq C \cdot \mu_{2q}^2 (1 - \delta)^{-2}(m + 1)^{\delta m} = o(m^{-\alpha})
\]
for any \( \alpha > 1 \). Apply Lemma 1 here and the proof of (i) is completed.

Let us turn to bound \( \sup_{j \leq p} \| y_{j,i} \|_2q \). Let \( A^l_1 \) be \( (a_{l,jk})_{j,k \leq p} \) with \( l \geq 1 \). Since \( A^l_1 \) is a banded matrix with the bandwidth \( \min(2lk_0 + 1, p) \), we can bound \( \| A^l_1 \|_{\infty} \) by
\[
\| A^l_1 \|_{\infty} = \max_{j \leq p} \sum_{k=1}^{p} |a_{l,jk}| \leq \sqrt{\min(2lk_0 + 1, p)} \| A^l_1 \|_2 \leq C(2lk_0 + 1)^{\delta l}, l \geq 1,
\]
which means that \( \| A^l_1 \|_{\infty} \leq C(2k_0 + 1)(l + 1)^{\delta l}, l \geq 0 \). Using the innovation representation \( y_t = \sum_{l=0}^{\infty} A^l_1 \varepsilon_{t-l} \), we get
\[
\| y_{j,i} \|_2q \leq \sum_{l=0}^{\infty} \| \sum_{k=1}^{p} a_{l,jk} \varepsilon_{k,i-l} \|_2q \leq \sum_{l=0}^{\infty} \sum_{k=1}^{p} |a_{l,jk}| \| \varepsilon_{k,i-l} \|_2q.
\]
As a result, \( \sup_{j \leq p} \| y_{j,i} \|_2q \leq C(2k_0 + 1)\mu_{2q} \sum_{l=0}^{\infty} (l + 1)^{\delta l} = C(2k_0 + 1)(1 - \delta)^{-2}\mu_{2q} < \infty \). Similarly, we can bound \( \sup_{j \leq p} \theta_{i,2q,j} \) above by \( C \cdot (i + 1)^{\delta i} \) with some positive constant \( C \) since we have a nice inequality
\[
\| y_{j,i} - y^*_{j,i} \|_2q = \| \sum_{k=1}^{p} a_{i,jk} (\varepsilon_{k,0} - \varepsilon^*_{k,0}) \|_2q \leq 2\mu_{2q} \| A^l_1 \|_{\infty}.
\]
Lemma 6. Suppose that Conditions (A1) - (A3) and A4(ii) hold.

(i) We have
\[ \sup_{1 \leq j, k \leq p} |\hat{\Sigma}_{jk} - \Sigma_{jk}| = O_P(\sqrt{\frac{\log p}{n}}). \]

(ii) We have
\[ \sup_{1 \leq j, k \leq p} |e_{(j)}^T x_{(k)}| = O_P(\sqrt{n \log p}). \]

Proof of Lemma 6. Here we only prove part (i). The proof of part (ii) can be derived similarly.

Note that \( y_t = A_1 y_{t-1} + \varepsilon_{t-l} \) and \( \|A_1\|_2 \leq \delta < 1 \). Let \( A_1' \) be \((a_{i,j,k})_{j,k \leq p} \). For each \( j \), \( y_{j,t} = \sum_{m=1}^{\infty} \sum_{m=1}^{p} a_{i,j,m} \varepsilon_{m,t-l} \) converges almost surely. Write \( \eta_{j,l} = \sum_{m=1}^{p} a_{i,j,m} \varepsilon_{m,t-l} \) for \( l \geq 0 \). We separate \( y_{j,t} \) to be two terms \( y_{j,t} = \sum_{l=0}^{N_n} \eta_{j,l} + \sum_{l=N_n+1}^{\infty} \eta_{j,l} \). Here we choose \( N_n \) to be \( N_\delta \log(n) \) with \( N_\delta > (1 + \alpha)\alpha^{-1}(\log \delta)^{-1} \). Hence, \( n\hat{\Sigma}_{jk} \) can be expressed as
\[ n\hat{\Sigma}_{jk} = \sum_{l,l'=0}^{N_n} \left( \sum_{t=0}^{n} \eta_{j,l,t} \eta_{j,l',t} \right) + \sum_{l,l'=N_n+1}^{\infty} \left( \sum_{t=1}^{n} \eta_{j,l,t} \eta_{j,l',t} \right) + \sum_{l=0}^{N_n} \sum_{l'=N_n+1}^{\infty} \left( \sum_{t=1}^{n} \eta_{j,l,t} \eta_{j,l',t} \right) + \sum_{l=N_n+1}^{\infty} \sum_{l'=0}^{N_n} \left( \sum_{t=1}^{n} \eta_{j,l,t} \eta_{j,l',t} \right) \]
\[ = S_{j,k,1} + S_{j,k,2} + S_{j,k,3} + S_{j,k,4} \]
and
\[ n(\hat{\Sigma}_{jk} - \Sigma_{jk}) = 4 \sum_{m=1}^{4} (S_{j,k,m} - ES_{j,k,m}). \]

Let us handle the first term \( S_{n1} - ES_{n1} \). Note that if \( \sup_{m,l} E \exp |t_0 \varepsilon_{m,l}|^{2\alpha} < \infty \),
\[ \zeta = \sup_{m,l,m',l'} E \exp \left( t_0 |\varepsilon_{m,l} \varepsilon_{m',l'}|^{\alpha} \right) < \infty. \]

By Lemma 4, we obtain the following equality:
\[ P(\left| \sum_{t=1}^{n} (\varepsilon_{m,t-l} \varepsilon_{m,t-l'} - E(\varepsilon_{m,t-l} \varepsilon_{m,t-l'})) \right| > 3(l + 1)x) \]
\[ \leq (l + 1)C_1 \exp \left( -\frac{x^2}{C_2 n + C_3 n^{1-\alpha} x} \right) + C_1 (l + 1) \exp \left( -\frac{x^{2\alpha}}{C_4 n^{2-\alpha} + C_3 x^{\alpha}} \right) \]
\[ + C_1 (l + 1) n \exp (-C_4 x^{\alpha}), \]
for some positive constants \( C_j (j = 1, \cdots , 4) \), where \( l_n = |l - l'| \). Take \( x = C \sqrt{n \log p} \) for some large constant \( C > 0 \). This leads to the following convergence rate:
\[ \bar{\eta}_n = \sup_{m,m' \leq p, l, l' \leq N_n} (l' + 1)^{-2} (l + 1)^{-2} \left| \sum_{t=1}^{n} (\varepsilon_{m,t-l} \varepsilon_{m,t-l'} - E(\varepsilon_{m,t-l} \varepsilon_{m,t-l'})) \right| = O_P(\sqrt{n \log p}). \]
Observe that
\[
\left| \sum_{t=1}^{n} (\eta_{jt,lt} - E(\eta_{jt,lt})) \right| \leq \sum_{m=1}^{p} |a_{t,jm}| \sum_{m'=1}^{p} |a_{t',km'}| \sum_{t=1}^{n} \left( \varepsilon_{m,t-1} \varepsilon_{m',t' - 1} - E(\varepsilon_{m,t-1} \varepsilon_{m',t'} - 1) \right) \leq C(2k_0 + 1)^2 (l + 1)^3 (l' + 1)^3 \delta^{l+l'} \eta_n
\]
and \(\sum_{l=0}^{N} (l + 1)^3 \delta^l < \infty\). Therefore,
\[
\sup_{j,k \leq p} \left| S_{jk,1} - ES_{jk,1} \right| \leq C \cdot (2k_0 + 1)^2 \eta_n = O_P(\sqrt{n \log p}).
\]

Consider the second term. Since \(\sup_{m,l} E \exp|t_0 \varepsilon_{m,l}|^{2\alpha} < \infty\), \(\tilde{\zeta}_{q,\varepsilon} = \sup_{m,l,m',l'} \left\| \varepsilon_{m,l} \varepsilon_{m',l'} \right\|_{q} \leq C \cdot q^{1/\alpha}\) for any \(q > 2\). Now we bound \(\left\| S_{jk,2} - ES_{jk,2} \right\|_{q}\). To be specific,
\[
\left\| \sum_{t=1}^{n} (\eta_{jt,lt} - E(\eta_{jt,lt})) \right\|_{q} \leq \sum_{m,m'=1}^{p} |a_{t,jm}| |a_{t',km'}| \sup_{m,m',l,l'} \left\| \varepsilon_{m,l} \varepsilon_{m',l'} \right\|_{q} \leq n(2k_0 + 1)^2 (l + 1)(l' + 1) \delta^{l+l'} \tilde{\zeta}_{q,\varepsilon}.
\]

Hence,
\[
\left\| S_{jk,2} - ES_{jk,2} \right\|_{q} \leq C n^{1/\alpha} \sum_{l,l' = N_n + 1}^{\infty} (l + 1)(l' + 1) \delta^{l+l'} \leq C \cdot nN_n^2 \delta^{2N_n} q^{1/\alpha}.
\]

Write \(\eta_{n2} = (nN_n^2 \delta^{2N_n})^{-1} (S_{jk,2} - ES_{jk,2})\). It follows from Lemma 2 that there exists a constant \(t > 0\) such that \(E \exp(t|\eta_{n2}|^{\alpha}) < \infty\). Consequently, for a large constant \(C > 0\), we have that
\[
P \left( \sup_{j,k \leq p} \left| S_{jk,2} - ES_{jk,2} \right| > C(\log n)^2 \right) \leq O(1) \cdot p^2 \exp \left( -tC^\alpha \cdot n(\log n)^{-2\alpha}(\log n)^{2\alpha} \right) \to 0
\]
as \(n \to \infty\), which means that \(\sup_{j,k \leq p} \left| S_{jk,2} - ES_{jk,2} \right| = O_P(\log n)^2 = o_P(\sqrt{n \log p})\). Similarly, we can obtain that \(\sup_{j,k \leq p} \left| S_{jk,m} - ES_{jk,m} \right| = o_P(\sqrt{n \log p})\), \(m = 3, 4\). Hence, it follows that
\[
\sup_{j,k \leq p} \left| \hat{\Sigma}_{jk} - \Sigma_{jk} \right| = O_P(\sqrt{\log p / n}).
\]
The proof is complete.

**Lemma 7.** Suppose that Conditions (A1) - (A3) and A4(i) or A4(ii) hold. Then, for each finite \(k\) with \(k \geq k_0\),
\[
\sup_{1 \leq i \leq p} \left| \frac{\text{RSS}_i(k)}{n \sigma_i^2} - 1 \right| = O_P(\sqrt{\log p / n})
\]
as \(n \to \infty\), where \(\text{RSS}_i(k)\) is defined in (2.6) and \(\sigma_i^2\) is the \((i, i)\)-th element of \(\Sigma_{\varepsilon}\).
Proof of Lemma 7. For $k > k_0$, the term $\text{RSS}_i(k)$ can be decomposed as

$$\text{RSS}_i(k) = e_i^T e_i - e_i^T X_i (X_i^T X_i)^{-1} X_i^T e_i = R_{i1} - R_{i2},$$

where $e_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n})^T$ and $X_i$ is a $(n - 1) \times \tau_i(k)$ matrix with $x_{i,1+j}$ as its $j$-th row. In the following, we will show that, under assumptions (A1-A3) and A4(i) or A4(ii),

(a) $\sup_{i \leq p} |R_{i1} - n\sigma_i^2| = O_P(\sqrt{n \log p})$; and (b) $\sup_{i \leq p} |R_{i2}| = O_P(\log p)$.

If they are proved, then it follows that $\sup_{i \leq p} \left| \frac{\text{RSS}_i(k)}{n\sigma_i^2} - 1 \right| \leq \sup_{i \leq p} \left| \frac{R_{i1}}{n\sigma_i^2} - 1 \right| + \sup_{i \leq p} \left| \frac{R_{i2}}{n\sigma_i^2} \right| = O_P(\sqrt{\log p / n}).$

Suppose first that condition A4(i) hold. Consider the term $R_{i1} - n\sigma_i^2$. Lemma 1 tells us that $\sup_{i \leq p} \left| e_i^T e_i - n\sigma_i^2 \right| = O_P(\sqrt{n \log p})$.

Let us handle the term $\sup_{i \leq p} |R_{i2}|$. Define

$$\mathcal{A}_n = \left\{ \inf_{i \leq p} \lambda_{\text{min}}(n^{-1}X_i^T X_i) > \kappa_1(1 - \eta) \right\}$$

with $0 < \eta < 1$. It follows from Lemma 5(i) and Condition (A3) that $P(\mathcal{A}_n) \to 1$ as $n \to \infty$. Over the event $\mathcal{A}_n$, the term $\sup_{i \leq p} |R_{i2}|$ can be bounded above by $(\kappa_1(1 - \eta))^{-1} k_0 \sup_{j,k \leq p} n^{-1} |e^T_{(j)} x_{(k)}|^2$. By Lemma 5 (ii), we obtain that $\sup_{j,k \leq p} |e^T_{(j)} x_{(k)}| = O_P(\sqrt{n \log p})$, which implies that (b) holds.

Suppose that conditions A4(ii) hold. Consider the term $R_{i1} - n\sigma_i^2$. By Lemma 3 and taking $x = C \sqrt{n \log p}$ with large constant $C > 0$, we have that

$$\sup_{i \leq p} \left| e_i^T e_i - n\sigma_i^2 \right| = O_P(\sqrt{n \log p}).$$

Similarly, Lemma 6 yields that (b) holds. The proof is completed.

Appendix B: Proof of Theorem 1

Without loss of generality, we consider the VAR(1) model with $\|A\|_1 \leq \delta < 1$. Our goal is to prove that $P(\hat{k} = k_0) \to 1$, i.e. $P(\hat{k} \neq k_0) \to 0$. If $\hat{k} \neq k_0$, then either $\hat{k} > k_0$ or $\hat{k} < k_0$ holds. Hence it suffices to show that

$$P(\hat{k} < k_0) \to 0 \ \text{and} \ \ P(\hat{k} > k_0) \to 0.$$
We follow the proof in Wang, Li and Leng (2009).

Consider the first case. Observe that $P\{\hat{k} < k_0\} \leq P\{\hat{k}_i < k_0\}$ for some $1 \leq i \leq p$ and the event $\{\hat{k}_i < k_0\}$ means $\{\min_{k < k_0} \text{BIC}_i(k) < \text{BIC}_i(k_0)\}$. To prove $P\{\hat{k} < k_0\} \to 0$, we only need to show that

$$P\{\min_{k < k_0} \text{BIC}_i(k) < \text{BIC}_i(k_0)\} \to 0,$$

for some $i$. Suppose that we have shown that there exists a constant $\eta > 0$ and an event $\mathcal{A}_n$ such that $P(\mathcal{A}_n) \to 1$ as $n \to \infty$ and on the event $\mathcal{A}_n$,

$$\text{RSS}_i(k) - \text{RSS}_i(k_0) \geq \eta \cdot \text{RSS}_i(k_0) \cdot (a_{i,i-k_0}^2 + a_{i,i+k_0}^2), \quad (6.5)$$

for sufficiently large $n$, where $a_{j,k}$ is the $(j,k)$-element of $A_1$. As a result, on the event $\mathcal{A}_n$ with large $n$, $\log(\text{RSS}_i(k)) - \log(\text{RSS}_i(k_0)) \geq \log(1 + \eta \cdot (a_{i,i-k_0}^2 + a_{i,i+k_0}^2))$. Note that $\log(1 + x) \geq \min\{0.5x, \log(2)\}$ for any $x > 0$. Consequently, with probability tending to one, $\log(\text{RSS}_i(k)) - \log(\text{RSS}_i(k_0))$ can be further bounded below by $\min\{0.5\eta(a_{i,i-k_0}^2 + a_{i,i+k_0}^2), \log(2)\}$. Condition (A3) implies that for some $1 \leq i^* \leq p$, $a_{i^*,i^*-k_0}^2 + a_{i^*,i^*+k_0}^2 \gg C_n \log(p/n)$ as $n \to \infty$. Hence, it follows that, with probability tending to 1,

$$\min_{k < k_0} \text{BIC}_i^*(k) - \text{BIC}_i^*(k_0) > \min\{0.5\eta(a_{i^*,i^*-k_0}^2 + a_{i^*,i^*+k_0}^2), \log(2)\} - k_0 \log(p \lor n)/n > 0.$$

Hence, $P\{\min_{k < k_0} \text{BIC}_i^*(k) < \text{BIC}_i^*(k_0)\} \to 0$ and consequently, $P\{\hat{k} < k_0\} \to 0$.

Let us turn to prove that (6.5). For $k < k_0$, denote $H_{i,k} = X_{i,k}(X_{i,k}^T X_{i,k})^{-1} X_{i,k}^T$, $X_{i,k_0} = (S_{i,k_0}^{(1)}, X_{i,k}, S_{i,k}^{(2)})$ and $\beta_{i,k_0} = (b_{i,1}, \beta_{i,k_0}^T, b_{i,2}^T)^T$, where $X_{i,k}$ is defined as (2.5) replaced $k_0$ by $k$. Then $\text{RSS}_i(k) = y_{(i)}^T (I_{n-1} - H_{i,k}) y_{(i)}$, and as a result, by Lemma 5 (ii) or Lemma 6 (ii),

$$\text{RSS}_i(k) - \text{RSS}_i(k_0) = (b_{i,1}^T, b_{i,2}^T)(S_{i,k_0}^{(1)}, S_{i,k_0}^{(2)})^T (I_{n-1} - H_{i,k})(S_{i,k}^{(1)}, S_{i,k}^{(2)}) \left( b_{i,1} \ b_{i,2} \right) + o_P(1).$$

From Lemma 5 (i) or Lemma 6 (i) and Lemma 7, there exists a small constant $\eta > 0$ such that, with probability tending to one,

$$\lambda_{\min}((S_{i,k}^{(1)}, S_{i,k}^{(2)})^T (I - H_{i,k})(S_{i,k}^{(1)}, S_{i,k}^{(2)})) > \eta(1 + \eta) \cdot n\sigma_i^2$$

and $\text{RSS}_i(k_0) \leq n\sigma_i^2(1 + \eta)$. Therefore, (6.5) follows.

Now we turn to the overfitting case, i.e. $P(\hat{k} > k_0) \to 0$. For $k > k_0$, denote $X_{i,k} = (S_{i,k_0}^{(1)}, X_{i,k_0}, S_{i,k_0}^{(2)})$ and $\beta_{i,k} = (0^T, \beta_{i,k_0}^T, 0^T)^T$. Denote also $S_{i,k} = (S_{i,k_0}^{(1)}, S_{i,k}^{(2)})$ and $\tilde{S}_{i,k} = (I_{n-1} - (I_{n-1} - H_{i,k})(S_{i,k}^{(1)}, S_{i,k}^{(2)}) \left( b_{i,1} \ b_{i,2} \right) + o_P(1).$
$\mathbf{H}_{i,k_0}\mathbf{S}_{i,k}$. Let $\eta$ be an arbitrary but fixed positive constant and define

$$B_n = \left\{ \inf_{k_0 \leq k \leq K} \inf_{1 \leq i \leq p} \frac{\text{RSS}_i(k)}{n\sigma_i^2} > (1 - \eta) \right\},$$

$$C_n = \bigcup_{1 \leq i \leq p} \bigcup_{k_0 \leq k \leq K} \lambda_{\text{min}}^{-1}(n^{-1}\mathbf{s}_{i,k}^T \mathbf{s}_{i,k}) < \kappa_1^{-1}(1 + \eta) \right\} \cap \left\{ \sup_{1 \leq j \leq k_0} \left| (n^{-1}\mathbf{s}_{i,k}^T \mathbf{s}_{i,k})_{jj} \right| < \kappa_2(1 + \eta) \right\}.$$

We first give a upper bound of $\text{RSS}_i(k_0) - \text{RSS}_i(k)$ with $k > k_0$. Note that for each $i$, $\text{RSS}_i(k)$ can be rewritten as

$$\text{RSS}_i(k) = \inf_{\mathbf{b}} \| \mathbf{y}(i) - \mathbf{X}_{i,k}\mathbf{b} \|^2 = \inf_{\mathbf{b}_1, \mathbf{b}_2} \| \mathbf{y}(i) - \mathbf{X}_{i,k_0}\mathbf{b}_1 - \mathbf{S}_i\mathbf{b}_2 \|^2.$$  

It can be verified that $\text{RSS}_i(k_0) = \| (\mathbf{I}_{n - 1} - \mathbf{H}_{i,k_0})\mathbf{y}(i) \|^2$ and

$$\text{RSS}_i(k) = \text{RSS}_i(k_0) - \| \mathbf{S}_i T \mathbf{b}_2 \|^2$$

where $T = (\mathbf{s}_{i,k}^T \mathbf{s}_{i,k})^{-1} \mathbf{s}_{i,k} e(i)$. Consequently, on the event $C_n$,

$$\text{RSS}_i(k_0) - \text{RSS}_i(k) = e_i^T (\mathbf{I}_{n - 1} - \mathbf{H}_{i,k_0})\mathbf{x}(k) \cdot \| \mathbf{S}_i T \mathbf{b}_2 \|^2 \leq \kappa_1^{-1}(1 + \eta) \cdot | \tau_i(k) - \tau_i(k_0) | \cdot \sup_{j, k \leq p} n^{-1/2} e_{j}^T (\mathbf{I}_{n - 1} - \mathbf{H}_{i,k_0})\mathbf{x}(k) |^2.$$  

Define

$$D_n = \left\{ \sup_{j, k \leq p} \left| n^{-1/2} e_{j}^T (\mathbf{I}_{n - 1} - \mathbf{H}_{i,k_0})\mathbf{x}(k) \right| \sigma_i^{-2} < \frac{\kappa_1(1 - \eta)}{(1 + \eta)} C_n \log(p \vee n) \right\}.$$  

On the set $B_n \cap C_n \cap D_n$, for all $k$ with $k_0 \leq k \leq K$,

$$\text{RSS}_i(k_0) - \text{RSS}_i(k) < \sigma_i^2(1 - \eta) | \tau_i(k) - \tau_i(k_0) | C_n \log(p \vee n) < \text{RSS}_i(k_0) C_n | \tau_i(k) - \tau_i(k_0) | n^{-1} \log(p \vee n).$$

Note that $\log(1 + x) \leq x$ for any $x > 0$. Hence, for all $k$ with $k_0 \leq k \leq K$, on the set $B_n \cap C_n \cap D_n$,

$$\text{BIC}_i(k) - \text{BIC}_i(k_0) = \log(\text{RSS}_i(k)) - \log(\text{RSS}_i(k_0)) + C_n | \tau_i(k) - \tau_i(k_0) | n^{-1} \log(p \vee n)$$

$$\geq \{- \{ \text{RSS}_i(k) - \text{RSS}_i(k_0) \} \} \{ \text{RSS}_i(k) \}^{-1} + C_n | \tau_i(k) - \tau_i(k_0) | n^{-1} \log(p \vee n) > 0.$$  

which means that over the set $B_n \cap C_n \cap D_n$, there must be $\tilde{k} \leq k_0$. To prove that $P\{ \tilde{k} > k_0 \} \to 0$, it suffices to show that $P\{ (B_n \cap C_n \cap D_n)^c \} \to 0$. In fact, it follows from Lemma 7 and Lemma 5 or 6 (i) that $P\{ B_n^c \} \to 0$ and $P\{ C_n^c \} \to 0$. It remains to show that $P\{ D_n^c \} \to 0$. Let $\Sigma_{i,k} = n^{-1}EX_{i,k}^T X_{i,k}$,
\( \hat{\Sigma}_{i,k} = n^{-1} \mathbf{X}_i^T \mathbf{X}_{i,k} \) and denote \( \tilde{H}_{i,k} = n^{-1} \mathbf{X}_{i,k} \Sigma_{i,k}^{-1} \mathbf{X}_i^T \) and
\( \tilde{x}(k) = (\mathbf{I}_{n-1} - \tilde{H}_{i,k}) \mathbf{x}(k) \). On the event \( \mathcal{C}_n \), we obtain that

\[
\sup_{j,k \leq p} |e_{(j)}^T (\mathbf{I}_{n-1} - \mathbf{H}_{i,k}) \mathbf{x}(k)|
\leq \sup_{j,k \leq p} |e_{(j)}^T \tilde{x}(k)| + \sup_{j,k \leq p} |e_{(j)}^T (\mathbf{H}_{i,k} - \tilde{H}_{i,k}) \mathbf{x}(k)|
\leq \sup_{j,k \leq p} |e_{(j)}^T \tilde{x}(k)| + n^{-1} \sup_{j,k \leq p} \|e_{(j)}^T \mathbf{X}_{i,k}\|_2 \|\Sigma_{i,k}^{-1}\|_2 \|\hat{\Sigma}_{i,k} - \Sigma_{i,k}\|_2 \|\mathbf{X}_i^T\| \|\mathbf{x}(k)\|_2
\leq \sup_{j,k \leq p} |e_{(j)}^T \tilde{x}(k)| + k_0 \kappa^2 \kappa_2 (1 + \eta)^2 \sup_{j,k \leq p} \|e_{(j)}^T \mathbf{x}(k)\| \cdot \|\hat{\Sigma}_{i,k} - \Sigma_{i,k}\|_2,
\]

where \( \sup_{1 \leq k \leq p} (n^{-1} \mathbf{x}(k) \mathbf{x}_k^T) \leq \kappa_2 (1 + \eta) \) is used in the above inequality. Hence, it follows from Lemmas 5 and 6, together with Condition A3, that \( P \{ \mathcal{D}_n \} \rightarrow 0 \) as \( n \rightarrow \infty \). This completes the proof.

**Appendix C: Proof of Theorem 2**

Without loss of generality, we consider the VAR(1) model. Since \( \hat{k} = k_0 \) with probability tending to 1, it suffices to consider the set \( \mathcal{A}_n = \{ \hat{k} = k_0 \} \). Over the set \( \mathcal{A}_n \), for each \( i \),

\[
\tilde{\beta}_i - \beta_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T e_{(i)}. \tag{6.6}
\]

For each \( i \), the law of large numbers for stationary process yields \( n^{-1} \mathbf{X}_i^T \mathbf{X}_i \) converges to a positive matrix almost surely, and furthermore, \( \lambda_{\min} (n^{-1} \mathbf{X}_i^T \mathbf{X}_i) \) is bounded away from zero with probability tending to one. As a matter of fact, if we define

\[
\mathcal{B}_n = \bigcap_{1 \leq i \leq p} \{ \lambda_{\min} (n^{-1} \mathbf{X}_i^T \mathbf{X}_i) > \kappa_1 (1 - \eta) \}
\]

with a small constant \( \eta \in (0,1) \), then it follows from by Lemma 5 or 6 under different moment conditions that \( P \{ \mathcal{B}_n \} \rightarrow 1 \) as \( n \rightarrow \infty \). Hence, over the event \( \mathcal{A}_n \cap \mathcal{B}_n \),

\[
\left\| \tilde{\beta}_i - \beta_i \right\|_2 \leq \kappa_1^2 (1 - \eta)^{-2} n^{-2} \left\| e_{(i)}^T \mathbf{X}_i \right\|_2^2 = C_1 n^{-2} \left\| e_{(i)}^T \mathbf{X}_i \right\|_2^2,
\]

where \( C_1 = \kappa_1^2 (1 - \eta)^{-2} > 0 \). It is not hard to see from Lemma 5(ii) or Lemma 6(ii) that, for all \( 1 \leq i \leq p \), \( n^{-1} E \| \mathbf{X}_i^T e_{(i)} \|_2^2 \leq C_2 \) with some constant \( C_2 > 0 \). Therefore, for a large positive constant
we obtain that
\[
P \left( \| \widehat{A}_1 - A_1 \|_F^2 > C \frac{p}{n} \right) = P \left( \| \widehat{A}_1 - A_1 \|_F^2 > C \frac{p}{n}, A_n \cap B_n \right) + P (A_n \cap B_n)
\]
\[
= \frac{n}{Cp n^2} \sum_{i=1}^{p} \| X_i^T e_{(i)} \|_2^2 + P (A_n \cap B_n)
\]
\[
= \frac{C_1 C_2}{C} + o(1).
\]

Take a sufficiently large C and the convergence rate of \( \| \widehat{A}_1 - A_1 \|_F \) is proved.

Now we prove the convergence rate of \( \| \widehat{A}_1 - A_1 \|_2 \). Note that for any matrix B, \( \| B \|_2^2 \leq \| B \|_1 \| B \|_\infty \). Hence, on the event \( A_n \),
\[
\| \widehat{A}_1 - A_1 \|_2 \leq \sqrt{\| \widehat{A}_1 - A_1 \|_1 \| \widehat{A}_1 - A_1 \|_\infty} \leq (2k_0 + 1) \sup_{i \leq p, j \leq \tau_i} | \beta_{ij} - \beta_{ij} |,
\]
where \( \beta_{ij} \) and \( \beta_{ij} \) is the \( j \)-th element of \( \beta_i \) and \( \beta_i \), respectively. Observe from \([6,6]\) that, for each \( i = 1, \ldots, p \),
\[
\sup_{i \leq p, j \leq \tau_i} | \beta_{ij} - \beta_{ij} | = \kappa_i^{-1} (1 - \eta)^{-1} (2k_0 + 1) \left( \sup_{i \leq p, j \leq \tau_i} | e_{(i)}^T x_{(j)} | \right).
\]
Hence, using Lemma 5(ii) or 6(ii), we have
\[
\sup_{i \leq p, j \leq \tau_i} | \beta_{ij} - \beta_{ij} | = O_P \left( \sqrt{\log \frac{p}{n}} \right).
\]
This also shows that
\[
\| \widehat{A}_1 - A_1 \|_2 = O_P \left( \sqrt{\log \frac{p}{n}} \right).
\]
The proof is completed.

### Appendix D: Proofs of Theorem 3 and 4

**Proof of Theorem 3.** The covariance matrix \( \Sigma_0 \) can be expressed as
\[
\Sigma_0 = \Sigma_e + \sum_{j=1}^{\infty} B_j \quad \text{with} \quad B_j = J \widetilde{A}^T J^T \Sigma_e J (\widetilde{A}^T)^j J^T, j \geq 1,
\]
where \( J = (I_{p \times p}, 0_{p \times (d-1)p}) \). Let \( \Phi_j = J \widetilde{A}^T J^T, j \geq 1 \). By the companion matrix \( \widetilde{A} \), it can be proven that \( \Phi_0 = I_p \) and \( \Phi_j = \sum_{k=1}^{\min(j,d)} \Phi_{j-k} A_k, j \geq 1 \). It is easy to see that for two banded matrix \( E \) and \( F \) with bandwidths \( 2r_1 + 1 \) and \( 2r_2 + 1 \), the product matrix \( EF \) is also banded and its bandwidth is at most \( 2(r_1 + r_2) + 1 \). Therefore, it can be verified that \( \Phi_j \) is banded with bandwidth
at most $2jk_0 + 1$ and then $B_j$ is also banded with its bandwidth at most $2(2jk_0 + s_0) + 1$ for $j \geq 1$.

Take $\Sigma_0^{(r)} = \Sigma_0 + \sum_{j=1}^r B_j$, which is banded with the bandwidth at most $2(2jk_0 + s_0) + 1$. Hence, $\Sigma_0 - \Sigma_0^{(r)} = \sum_{j=r+1}^\infty B_j$. Note that for any $j \geq 1$, $\|B_j\|_2 \leq \|\Sigma_0\|_2 \|A^{2j}\|_2 \leq C\delta^{2j}$ for some $C > 0$.

Write $C_1 = C\|\Sigma_0\|_2 (1 - \delta^2)^{-1}$. It follows that

$$\| \Sigma_0 - \Sigma_0^{(r)} \|_2 \leq \sum_{j=r+1}^\infty \|B_j\|_2 \leq C\|\Sigma_0\|_2 (1 - \delta^2)^{-1} \delta^{2(r+1)} = C_1 \delta^{2(r+1)}.$$

By using the inequality $\|B_j\|_1 \leq (2(2jk_0 + s_0) + 1)\|B_j\|_2 \leq C(2j + 1)\delta^{2j}$ for some $C > 0$, we can obtain

$$\| \Sigma_0 - \Sigma_0^{(r)} \|_1 \leq C_2 r \delta^{2(r+1)}.$$

Other inequalities can be proved analogously. The proof is complete.

**Proof of Theorem 4.** Now we prove the convergence rate of $\| \hat{\Sigma}_{n,0}^{(r_n)} - \Sigma_0 \|_2$. First, $\| \hat{\Sigma}_{n,0}^{(r_n)} - \Sigma_0 \|_2$ can be bounded above as

$$\| \hat{\Sigma}_{n,0}^{(r_n)} - \Sigma_0^{(r_n)} \|_2 + \| \Sigma_0^{(r_n)} - \Sigma_0 \|_2 = R_{n1} + R_{n2}.$$

Similar to Theorem 2, $R_{n1} \leq (4r_nk_0 + 2s_0 + 1)\sup_{j,k \leq p} |\hat{\Sigma}_{jk} - \Sigma_{jk}|$. From Lemma 5(i) or 6(i), we obtain that

$$R_{n1} = O_P\left( r_n \sqrt{\frac{\log p}{n}} \right).$$

From Theorem 3, $R_{n2} \leq O(\delta^{2(r_n+1)})$. Note that $r_n = C\log(n \log^{-1}(p))$ with $C > 4^{-1} (\log \delta^{-1})^{-1}$.

Combining these results, it follows that

$$\| \hat{\Sigma}_{n,0}^{(r_n)} - \Sigma_0 \|_2 = O_P\left( r_n \sqrt{\frac{\log p}{n}} + \delta^{2(r_n+1)} \right) = O_P\left( \log(n \log^{-1}(p)) \sqrt{\frac{\log p}{n}} \right).$$

The proofs of other results are similar and omitted.

**REFERENCES**

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36, 199-227.

Bickel, P.J. and Gel, Y.R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society, B*, 73, 549-92.
Bolstad, A., Van Veen, B. D. and Nowak, R. (2011). Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing, 59*, 2628-2640.

Can, A. and Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics, 14*, 203-222.

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics, 41*, 2994-3021.

Davis, R. A., Zhang, P. and Zheng, T. (2012). Sparse vector autoregressive modeling. [http://arxiv.org/abs/1207.0520](http://arxiv.org/abs/1207.0520).

Han, F. and Liu, H. (2013). A Direct Estimation of High Dimensional Stationary Vector Autoregressions. *Manuscript*.

Haufe, S., Nolte, G., Mueller, K. R., and Krämer, N. (2009). Sparse causal discovery in multivariate time series. [http://arxiv.org/abs/0901.2234](http://arxiv.org/abs/0901.2234).

Hsu, N. J., Hung, H. L., and Chang, Y. M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis, 52*, 3645-3657.

Kock, A. and Callot, L. (2012). Oracle inequalities for high dimensional vector autoregressions. Technical report, Aarhus University.

Liu, W., Xiao, H. and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica, 23*, 1257-1272.

Luo, S. and Chen, Z. (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference, 143*, 494-504.

Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer, New York.

Nbura, M. B. Giannone, D. and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics, 25*, 71-92.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics, 26*, 517-523.
Song, S. and Bickel, P. J. (2011). Large vector auto regressions. \url{http://arxiv.org/abs/1106.3915}

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, B, 71*, 671-683.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceeding of National Academy of Sciences, 102*, 14150-14154.

Wu, W.B. and Pourahmadi, M. (2009). Banding sample covariance matrices of stationary processes. *Statistica Sinica, 19*, 1755-68.
Table 1: Relative frequencies in simulation with 500 replications, where \( \hat{k} \) is defined in (2.8).

|       | \( k_0 = 1 \) | \( k_0 = 2 \) | \( k_0 = 3 \) | \( k_0 = 4 \) |
|-------|---------------|---------------|---------------|---------------|
| \( p = 100 \) | \( \hat{k} = k_0 \) | 0.826 | 0.870 | 0.730 | 0.554 |
|     | \( \hat{k} > k_0 \) | 0.168 | 0.084 | 0.060 | 0.144 |
|     | \( \hat{k} < k_0 \) | 0.008 | 0.046 | 0.210 | 0.312 |
| \( p = 200 \) | \( \hat{k} = k_0 \) | 0.996 | 0.930 | 0.804 | 0.680 |
|     | \( \hat{k} > k_0 \) | 0.034 | 0.048 | 0.024 | 0.028 |
|     | \( \hat{k} < k_0 \) | 0.0 | 0.022 | 0.172 | 0.292 |
| \( p = 400 \) | \( \hat{k} = k_0 \) | 0.990 | 0.948 | 0.826 | 0.680 |
|     | \( \hat{k} > k_0 \) | 0.010 | 0.004 | 0.012 | 0.004 |
|     | \( \hat{k} < k_0 \) | 0.0 | 0.048 | 0.162 | 0.374 |
| \( p = 800 \) | \( \hat{k} = k_0 \) | 0.990 | 0.932 | 0.720 | 0.560 |
|     | \( \hat{k} > k_0 \) | 0.010 | 0.004 | 0.008 | 0.004 |
|     | \( \hat{k} < k_0 \) | 0.0 | 0.064 | 0.272 | 0.436 |


Table 2: Relative frequencies in simulation with 500 replication, where $\tilde{k}$ is defined in (5.1).

|     | Setting (i) | Setting (ii) |
|-----|-------------|--------------|
|     | $\tilde{k} = k_0$ $\tilde{k} > k_0$ $\tilde{k} < k_0$ | $\tilde{k} = k_0$ $\tilde{k} > k_0$ $\tilde{k} < k_0$ |
| $p = 100$ | $k_0 = 1$ 0.644 0 0.356 | $k_0 = 1$ 0.676 0 0.324 |
|     | $k_0 = 2$ 0.420 0 0.580 | $k_0 = 2$ 0.350 0 0.650 |
| $p = 200$ | $k_0 = 1$ 0.562 0 0.438 | $k_0 = 1$ 0.606 0 0.394 |
|     | $k_0 = 2$ 0.318 0 0.682 | $k_0 = 2$ 0.282 0 0.718 |
| $p = 400$ | $k_0 = 1$ 0.478 0 0.522 | $k_0 = 1$ 0.532 0 0.468 |
|     | $k_0 = 2$ 0.234 0 0.766 | $k_0 = 2$ 0.200 0 0.800 |
| $p = 800$ | $k_0 = 1$ 0.438 0 0.562 | $k_0 = 1$ 0.466 0 0.534 |
|     | $k_0 = 2$ 0.112 0 0.888 | $k_0 = 2$ 0.108 0 0.892 |

Figure 1: Time series plots of the weekly temperature in January 1990 – December 2000 in, from top to bottom, Ha’erbin, Shanghai and Nanjing.
Table 3: Means and standard deviations (in parentheses) of the errors in estimating $\mathbf{A}$ under Setting (i) in simulation with 500 replications.

| $p$ | With estimated $k_0$ | With true $k_0$ |
|-----|----------------------|-----------------|
|     | $\|\widehat{\mathbf{A}} - \mathbf{A}\|_1$ | $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2$ | $\|\widehat{\mathbf{A}} - \mathbf{A}\|_1$ | $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2$ |
|     |               |                 |               |
| 100 | $k_0 = 1$ | 0.378 (0.058) | 0.274 (0.031) | 0.372 (0.049) | 0.272 (0.029) |
|     | $k_0 = 2$ | 0.535 (0.059) | 0.331 (0.028) | 0.529 (0.052) | 0.329 (0.026) |
|     | $k_0 = 3$ | 0.703 (0.082) | 0.387 (0.035) | 0.694 (0.069) | 0.383 (0.029) |
|     | $k_0 = 4$ | 0.851 (0.101) | 0.434 (0.047) | 0.847 (0.077) | 0.432 (0.028) |
| 200 | $k_0 = 1$ | 0.401 (0.059) | 0.285 (0.030) | 0.395 (0.047) | 0.283 (0.027) |
|     | $k_0 = 2$ | 0.579 (0.067) | 0.349 (0.030) | 0.578 (0.065) | 0.348 (0.028) |
|     | $k_0 = 3$ | 0.738 (0.081) | 0.403 (0.037) | 0.736 (0.063) | 0.402 (0.027) |
|     | $k_0 = 4$ | 0.904 (0.110) | 0.459 (0.053) | 0.886 (0.072) | 0.448 (0.025) |
| 400 | $k_0 = 1$ | 0.425 (0.046) | 0.300 (0.027) | 0.422 (0.043) | 0.299 (0.026) |
|     | $k_0 = 2$ | 0.604 (0.057) | 0.363 (0.027) | 0.602 (0.055) | 0.362 (0.026) |
|     | $k_0 = 3$ | 0.773 (0.081) | 0.421 (0.042) | 0.762 (0.060) | 0.415 (0.025) |
|     | $k_0 = 4$ | 0.952 (0.136) | 0.477 (0.067) | 0.931 (0.071) | 0.464 (0.027) |
| 800 | $k_0 = 1$ | 0.442 (0.040) | 0.312 (0.023) | 0.441 (0.038) | 0.311 (0.023) |
|     | $k_0 = 2$ | 0.625 (0.055) | 0.373 (0.026) | 0.622 (0.051) | 0.371 (0.023) |
|     | $k_0 = 3$ | 0.805 (0.090) | 0.434 (0.046) | 0.802 (0.062) | 0.429 (0.023) |
|     | $k_0 = 4$ | 0.983 (0.144) | 0.492 (0.072) | 0.966 (0.067) | 0.475 (0.024) |
Table 4: Means and standard deviations (in parentheses) of the errors in estimating autocovariance matrices in a simulation with 100 replications.

|                  | $\|\hat{\Sigma}_{n,0} - \Sigma_0\|_1$ |                  | $\|\hat{\Sigma}_{n,1} - \Sigma_1\|_1$ |
|------------------|----------------------------------|------------------|----------------------------------|
|                  | Banding | Thresholding | Sample | Banding | Thresholding | Sample |
| $p = 100$        | 2.13 (0.04) | 2.60 (0.02) | 14.4 (0.07) | 2.91 (0.03) | 3.46 (0.04) | 14.32 (0.07) |
| $p = 200$        | 2.69 (0.04) | 3.38 (0.03) | 29.3 (0.02) | 3.14 (0.03) | 4.23 (0.04) | 30.3 (0.02) |
| $p = 400$        | 2.34 (0.02) | 2.98 (0.02) | 55.1 (0.02) | 2.77 (0.03) | 3.71 (0.02) | 55.2 (0.02) |
| $p = 800$        | 2.68 (0.03) | 3.38 (0.02) | 112 (0.03)  | 2.91 (0.03) | 3.93 (0.03) | 110 (0.04) |

|                  | $\|\hat{\Sigma}_{n,0} - \Sigma_0\|_1$ |                  | $\|\hat{\Sigma}_{n,1} - \Sigma_1\|_1$ |
|------------------|----------------------------------|------------------|----------------------------------|
|                  | Spectral Norm |                    | Spectral Norm |                    |
| $p = 100$        | 1.06 (0.01) | 1.44 (0.02) | 4.00 (0.07) | 1.42 (0.01) | 1.76 (0.02) | 3.70 (0.02) |
| $p = 200$        | 1.33 (0.03) | 1.67 (0.02) | 6.49 (0.03) | 1.47 (0.01) | 1.91 (0.01) | 6.07 (0.02) |
| $p = 400$        | 1.18 (0.01) | 1.65 (0.01) | 10.7 (0.03) | 1.33 (0.01) | 1.93 (0.01) | 9.20 (0.02) |
| $p = 800$        | 1.38 (0.02) | 1.80 (0.01) | 17.0 (0.03) | 1.42 (0.01) | 2.26 (0.02) | 15.2 (0.03) |

Table 5: Means and standard deviations (in parentheses) of MAPE for one-step-ahead and two-step-ahead forecasts over 30 post-sample weekly temperatures of the 71 cities in China.

|                  | Banded VAR(1) | Lasso VAR(1) |
|------------------|--------------|-------------|
| One-step ahead   | 1.42 (1.18)  | 1.54 (1.17) |
| Two-step ahead   | 1.54 (1.28)  | 1.63 (1.25) |

Table 6: Means and standard deviations (in parentheses) of MAPE for one-step-ahead and two-step-ahead forecasts over 100 post-sample daily sales in 21 provinces in China.

|                  | Banded VAR(1) | Lasso VAR(1) |
|------------------|--------------|-------------|
| One-step ahead   | 0.305 (0.376)| 0.311 (0.360)|
| Two-step ahead   | 0.395 (0.387)| 0.390 (0.390)|

33
Figure 2: Time series plots of the deseasonalized weekly temperature in January 1990 – December 2000 in, from top to bottom, Ha’erbin, Shanghai and Nanjing.

Figure 3: The estimated Banded coefficient matrix $\hat{A}$ (on the left), and $\tilde{A}$ estimated by Lasso (on the right) for Example 1.
Figure 4: Plot of locations of 21 provinces in China used in Example 2.

Figure 5: The estimated banded coefficient matrix $\hat{A}$ (on the left), and $\tilde{A}$ estimated by Lasso (on the right) for Example 2.