Biometrika (2016), **00**, 0, *pp*. 1–12
Printed in Great Britain

High Dimensional and Banded Vector Autoregressions

BY SHAOJUN GUO
Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, P.R.China
sjguo@ruc.edu.cn

YAZHEN WANG
Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.
yzwang@stat.wisc.edu

AND QIWEI YAO
Department of Statistics, London School of Economics, London WC2A 2AE, U.K.
q.yao@lse.ac.uk

SUMMARY

We consider a class of vector autoregressive 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 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 auto-covariance functions of banded vector autoregressive processes, consistent estimators for the auto-covariance matrices are constructed.

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

1. INTRODUCTION

The demand for modelling and forecasting high-dimensional time series arises from panel study of economic, social and natural phenomena, financial market analysis, communication engineering and other domains. When the dimension of time series is large or even moderately large, statistical modelling is challenging as vector autoregressive and moving average models suffer from lack of identification, over-parameterization and flat likelihood functions. While pure vector autoregressive models are perfectly identifiable, their usefulness is often hampered by the lack of proper means of reducing the number of parameters.

In many practical situations it is often enough to collect the information from neighbour variables, though the definition of neighbourhoods is case-dependent. For example, sales, prices, weather indices or electricity consumptions influenced by temperature depend on those at close range locations, in the sense that the information from farther locations may become redundant given that from neighbours. See, for example, Can and Mebolugbe (1997) for a house price data example which exhibits such a dependence structure. In this paper, we propose a class of vector

autoregressive (hereafter 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. Nevertheless, non-zero cross correlations among all component series may still be existent, 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 number of time series. In practice, a banded structure may be employed by arranging the order of component series appropriately. The ordering can be deduced from subject knowledge aided by statistical tools such as Bayesian Information Criterion (hereafter BIC); see two real data examples in Section 5·2. With the imposed banded structure, we propose least squares estimators for the autoregressive coefficient matrices which attain the convergence rate $(p/n)^{1/2}$ under the Frobenius norm and $(\log p/n)^{1/2}$ under the spectral norm when p diverges together with the length n of time series.

In practice the maximum width of the non-zero coefficient bands in the coefficient matrices, which is called the bandwidth, is unknown. We propose a marginal BIC 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 of 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 banded. Because of this property, the band-truncated sample autocovariance matrices are consistent estimators with the convergence rate $\log(n/\log p)(\log p/n)^{1/2}$, which is faster than the standard banding covariance estimators (Bickel and Levina , 2008). See also Wu and Pourahmadi (2009), Bickel and Gel (2011) and Leng and Li (2011) for the estimation of the banded covariance matrices of time series.

Most existing work on high-dimensional VAR models draws inspiration from recent developments in the 'large p small n' regression paradigm. For example, Hsu et al. (2008) proposed lasso penalization for subset autoregression. Haufe et al. (2010) introduced the group sparsity for coefficient matrices and advocated to use group lasso penalization. A truncated weighted lasso and group lasso penalization approaches were proposed by Shojaie and Michailidis (2010) and Basu et al. (2015), respectively, to explore graphical granger causality in a VAR model. Basu and Michailidis (2015) focused on stable Gaussian processes and investigated the theoretical properties of L_1-regularized estimates of transition matrix in sparse VAR models. Bolstad et al. (2011) inferred sparse causal networks through VAR processes and proposed a group lasso procedure. Kock and Callot (2015) established oracle inequalities for high-dimensional VAR models. Han and Liu (2015) proposed an alternative Dantzig-type penalization and formulated the estimation problem into a linear program. Chen et al. (2013) studied sparse covariance and precision matrix in high dimensional time series under a general dependence structure.

2. METHODOLOGY

2·1. *Banded VAR models*

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

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

where ε_t is the innovation at time t, $E(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) = E(\varepsilon_t \varepsilon_t^{\mathrm{T}}) = \Sigma_\varepsilon$, and ε_t is independent of $y_{t-1}, y_{t-2}, \ldots$. Furthermore, all the coefficient matrices $A_1, \ldots, A_d$ are banded matrices

in the sense that

$$a_{ij}^{(\ell)} = 0, |i-j| > k_0, \ell = 1, \ldots, d, \tag{2}$$

where $a_{ij}^{(\ell)}$ denotes the (i,j)-th element of A_ℓ. Thus the maximum number of non-zero elements in each row of A_ℓ is the bandwidth $2k_0+1$, and k_0 is called the bandwidth parameter. We 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 k_0 and to estimate the banded coefficient matrices $A_1, \ldots, A_d$. For simplicity, we assume that the autoregressive order d is known, as the order determination problem has already been throughly studied; 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 (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, ε_t is independent and identically distributed, y_t is also strictly stationary.

In model (1), we do not require $\mathrm{var}(\varepsilon_t) = \Sigma_\varepsilon$ to be banded. But even if $\mathrm{var}(\varepsilon_t)$ is banded, the autocovariance matrices are not necessarily banded; see (12) below. Therefore, the proposed banded model is applicable when the linear dynamics of each component series depends predominately on its neighbour series, though non-zero corrlations may still be existent among all component series of y_t.

2·2. *Estimating banded autoregressive coefficient matrices*

Since each row of A_ℓ has maximum $2k_0+1$ non-zero elements, there are at most $(2k_0+1)d$ regressors in each row on the right-hand side of (1). For $i = 1, \ldots, p$, let β_i be the column vector obtained by stacking the non-zero elements in the i-th rows of $A_1, \ldots, A_d$ together. Let τ_i denote the length of β_i. Then

$$\tau_i \equiv \tau_i(k_0) = \begin{cases} (2k_0+1)d, & i = k_0+1,\, k_0+2, \ldots, p-k_0, \\ (2k_0+1-j)d, & i = k_0+1-j \ \text{ or } \ p-k_0+j, \quad j = 1, \ldots, k_0. \end{cases} \tag{3}$$

Now (1) can be written as

$$y_{i,t} = x_{i,t}^{\mathrm{T}} \beta_i + \varepsilon_{i,t}, \qquad i = 1, \ldots, p, \tag{4}$$

where $y_{i,t}$, $\varepsilon_{i,t}$ are, respectively, the i-th component of y_t and ε_t, and $x_{i,t}$ is the $\tau_i \times 1$ vector consisting of the corresponding components of $y_{t-1}, \ldots, y_{t-d}$. Consequently, the least squares estimator of β_i based on (4) is

$$\widehat{\beta}_i = (X_i^{\mathrm{T}} X_i)^{-1} X_i^{\mathrm{T}} y_{(i)}, \tag{5}$$

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

By estimating β_i, $i = 1, \ldots, p$, separately based on (5), we obtain the least squares estimators $\widehat{A}_1, \ldots, \widehat{A}_d$ for the coefficient matrices in (1). Furthermore, the resulting residual sum of squares is

$$\mathrm{RSS}_i \equiv \mathrm{RSS}_i(k_0) = y_{(i)}^{\mathrm{T}} \{I_{n-d} - X_i (X_i^{\mathrm{T}} X_i)^{-1} X_i^{\mathrm{T}}\} y_{(i)}. \tag{6}$$

We write RSS_i as a 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).

2·3. *Determination of bandwidth*

In practice the bandwidth is unknown and we need to estimate k_0. We propose to determine k_0 based on the marginal BIC,

$$\mathrm{BIC}_i(k) = \log \mathrm{RSS}_i(k) + \frac{1}{n} d\tau_i(k) C_n \log(p \vee n), \qquad i = 1, \ldots, p, \tag{7}$$

where $\mathrm{RSS}_i(k)$ and $\tau_i(k)$ are defined, respectively, in (6) and (3), $p \vee n = \max(p, n)$, and $C_n > 0$ is some constant which diverges together with n; see Condition 2. We often take C_n to be $\log \log n$. An estimator for k_0 is now defined as

$$\widehat{k} = \max_{1 \le i \le p} \big\{ \arg \min_{1 \le k \le K} \mathrm{BIC}_i(k) \big\}, \tag{8}$$

where $K \ge 1$ is a prescribed integer. Our numerical study shows that the procedure is insensitive to the choice of K as long as $K \ge k_0$. In practice, we often take K to be $[n^{1/2}]$ or choose K by checking the curvature of $\mathrm{BIC}_i(k)$ directly.

Remark 1. If the order d is unknown, we can modify the criterion in (8) as follows. Let $\mathrm{RSS}_i(k, \ell)$ and $\tau_i(k, \ell)$ be defined similarly as (6) and (3). The marginal BIC is defined as

$$\widetilde{\mathrm{BIC}}_i(k, \ell) = \log \mathrm{RSS}_i(k, \ell) + \frac{1}{n} \tau_i(k, \ell) C_n \log(p \vee n), \qquad i = 1, \ldots, p. \tag{9}$$

Let L be a prescribed integer upper bound on d. In applications, L is often taken to be 10 or $[n^{1/2}]$. For each $i = 1, \ldots, p$,

$$(\widehat{k}_i, \widehat{d}_i) = \arg \min_{1 \le k \le K, 1 \le \ell \le L} \widetilde{\mathrm{BIC}}_i(k, \ell),$$

and let $\widehat{k} = \max_{1 \le i \le p} \widehat{k}_i$ and $\widehat{d} = \max_{1 \le i \le p} \widehat{d}_i$. Proposition 1 in the supplementary material shows that under Conditions 1–4, $\mathrm{pr}(\widehat{k} = k_0, \widehat{d} = d) \to 1$ as n and $p \to \infty$.

Remark 2. The banded structure of the coefficient matrices $A_1, \ldots, A_d$ depends critically on the order of the component series of y_t. In principle it is possible to derive a complete data-driven method to deduce the optimal ordering which minimizes the bandwidth. However such a procedure is computationally burdensome especially for large p. For most applications meaningful orderings are suggested by practical consideration. We can then calculate the collective BIC value

$$\mathrm{BIC} = \sum_{i=1}^{p} \mathrm{BIC}_i(\widehat{k}) \tag{10}$$

for each suggested ordering, and choose one which minimizes (10). In the expression (10), $\mathrm{BIC}_i(\cdot)$ and $\widehat{k}$ are defined as, respectively, in (7) and (8). Examples 1 and 2 indicate that this pragmatic scheme works well in applications.

3. ASYMPTOTIC PROPERTIES

3·1. *Regularity conditions*

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

$$\|v\|_q = \Big(\sum_{j=1}^{p} |v_j|^q \Big)^{1/q}, \quad \|v\|_\infty = \max_{1 \le j \le p} |v_j|, \quad \|B\|_q = \max_{\|v\|_q = 1} \|Bv\|_q, \quad \|B\|_F = \Big(\sum_{i,j} b_{ij}^2 \Big)^{1/2},$$

i.e., $\|\cdot\|_q$ denotes the ℓ_q norm of a vector or matrix, and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

First we note that VAR(d) model (1) can be formulated into the following VAR(1) form,

$$\widetilde{y}_t = \widetilde{A}\widetilde{y}_{t-1} + \widetilde{\varepsilon}_t,$$

where

$$\widetilde{y}_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-d+1} \end{pmatrix}, \quad \widetilde{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_d \\ I_p & 0_p & \cdots & \cdots \\ \vdots & \cdots & \vdots & \cdots \\ 0 & \cdots & I_p & 0 \end{pmatrix}, \quad \widetilde{\varepsilon}_t = \begin{pmatrix} \varepsilon_t \\ 0_{p\times 1} \\ \vdots \\ 0_{p\times 1} \end{pmatrix}. \tag{11}$$

Some regularity conditions are stated as follows.

Condition 1. For $\widetilde{A}$ defined in (11), $\|\widetilde{A}\|_2 \le C$ and $\|\widetilde{A}^{j_0}\|_2 \le \delta^{j_0}$, where $C > 0$, $\delta \in (0,1)$ and $j_0 \ge 1$ are constants free of n and p, and j_0 is an integer.

Condition 1'. For $\widetilde{A}$ defined in (11), $\|\widetilde{A}^{j_0}\|_2 \le \delta^{j_0}$, $\|\widetilde{A}\|_\infty \le C$ and $\|\widetilde{A}^{j_0}\|_\infty \le \delta^{j_0}$, where $C > 0$, $\delta \in (0,1)$ and $j_0 \ge 1$ are constants free of n and p, and j_0 is an integer.

Condition 2. Let $a_{ij}^{(\ell)}$ be the (i,j)-th element of A_ℓ. For each $i = 1, \ldots, p$, $|a_{i,i+k_0}^{(\ell)}|$ or $|a_{i,i-k_0}^{(\ell)}|$ is greater than $\{C_n k_0 n^{-1}\log(p \vee n)\}^{1/2}$ for some $1 \le \ell \le d$, where $C_n \to \infty$ as $n \to \infty$.

Condition 3. The minimal eigenvalue $\lambda_{\min}\{\mathrm{cov}(y_t)\} \ge \kappa_1$ and $\max_{1\le i\le p}|\sigma_{ii}| \le \kappa_2$ for some positive constants κ_1 and κ_2 free of p, where σ_{ii} is the i-th diagonal element of $\mathrm{cov}(y_t)$, and $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.

Condition 4. The innovation process $\{\varepsilon_t; t = 0, \pm1, \pm2, \ldots\}$ is independent and identically distributed with zero mean and covariance Σ_ε. Furthermore, one of the two assertions below holds:

(i) $\max_{1\le i\le p} E(|\varepsilon_{i,t}|^{2q}) \le C$ 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) $\max_{1\le i\le p} E\{\exp(\lambda_0|\varepsilon_{i,t}|^{2\alpha})\} \le C$ and $\log p = o\{n^{\alpha/(2-\alpha)}\}$, where $\lambda_0 > 0$, $\alpha \in (0,1]$ and $C > 0$ are constants free of n and p.

Provided $\{\varepsilon_t\}$ is independent and identically distributed, Condition 1 implies y_t to be strictly stationary. It also implies that for any $j \ge 1$, $\|\widetilde{A}^j\|_2 \le C\delta^j$ with some constant $C > 0$ and $\delta \in (0,1)$. The independent and identically distributed assumption in Condition 4 is imposed to simplify the proofs but not essential. Condition 2 ensures that the bandwidth $(2k_0+1)$ are asymptotically identifiable as $\{n^{-1}\log(p \vee n)\}^{1/2}$ is the minimum order of a non-zero coefficient to be identifiable; see, e.g., Luo and Chen (2013). Condition 3 guarantees that the covariance matrix $\mathrm{var}(y_t)$ is strictly positive definite. Condition 4 specifies the two asymptotic modes: (i) the high-dimension cases with $p = O(n^\beta)$, and (ii) the ultra high-dimension cases with $\log p = o\{n^{\alpha/(2-\alpha)}\}$. The larger p is, the faster tail decay rates on the distribution of $\varepsilon_{i,t}$ is required under Condition 4.

3·2. *Asymptotic theorems*

We first state the consistency of the selector $\widehat{k}$, defined in (8), for determining the bandwidth parameter k_0.

THEOREM 1. *Under Conditions 1–4,* $pr(\widehat{k} = k_0) \to 1$ *as* $n \to \infty$.

Remark 3. In Theorem 1 above, k_0 is assumed to be fixed, but as in applications small k_0 is of particular interest. We can allow the bandwidth parameter k_0 to diverge as $n, p \to \infty$. To show its consistency, conditions would need to be strengthened. To be specific, if $k_0 \ll C_n^{-1} n / \log(p \vee n)$, $\mathrm{pr}(\widehat{k} = k_0) \to 1$ as $n \to \infty$ under Conditions 1' and 2–4 in Section 3.1; see Proposition 2 in the Supplementary Material.

Since k_0 is unknown, we replace k_0 by $\widehat{k}$ in the procedure of estimating $A_1, \ldots, A_d$ described in Section 2·2, and denote the resulted estimators still by $\widehat{A}_1, \ldots, \widehat{A}_d$. Theorem 2 below addresses their convergence rates.

THEOREM 2. *Let Conditions 1–4 hold. As $n \to \infty$, it holds for $j = 1, \ldots, d$ that*

$$\|\widehat{A}_j - A_j\|_F = O_P\Big\{(p/n)^{1/2}\Big\}, \ \ \|\widehat{A}_j - A_j\|_2 = O_P\Big\{(\log p/n)^{1/2}\Big\}.$$

Conditions 4(i) and 4(ii) impose, respectively, a high moment condition and an exponential tail condition on the innovation distribution. Although the convergence rates in Theorem 2 have the same expressions in terms of n and p, due to the different conditions imposed on them in Conditions 4(i) and 4(ii), the actual convergence rates are different under the two settings. For example, Condition 4(i) allows p to grow in the order n^{β}, which implies the convergence rate $(\log n/n)^{1/2}$ for $\widehat{A}_j$ under the spectral norm. On the other hand, Condition 4(ii) may allow p to diverge at the rate $\exp\{n^{\alpha/(2-\alpha)-2\epsilon}\}$ for a small constant $\epsilon > 0$, and the implied convergence rate for $\widehat{A}_j$ under the spectral norm is $n^{1/2+\epsilon-\alpha/(4-2\alpha)}$.

4. ESTIMATION FOR AUTO-COVARIANCE FUNCTIONS

For the banded VAR process y_t defined by (1), the auto-covariance function $\Sigma_j = \mathrm{cov}(y_t, y_{t+j})$ is unlikely to be banded. For example for a stationary banded VAR(1) process, it can be shown that

$$\Sigma_0 \equiv \mathrm{var}(y_t) = \Sigma_\varepsilon + \sum_{i=1}^{\infty} A_1^i \Sigma_\varepsilon (A_1^{\mathrm{T}})^i. \tag{12}$$

For any banded matrices B_1 and B_2 with bandwidths $2k_1 + 1$ and $2k_2 + 1$, respectively, the product $B_1 B_2$ is a banded matrix with the enlarged bandwidth $2(k_1 + k_2) + 1$ in general. Thus Σ_0 presented in (12) is not a banded matrix. Nevertheless if $\mathrm{var}(\varepsilon_t) = \Sigma_\varepsilon$ is also banded, Theorem 3 below shows that Σ_j can be approximated by some banded matrices.

Condition 5. Matrix Σ_ε is banded with bandwidth $2s_0 + 1$ and $\|\Sigma_\varepsilon\|_1 \le C < \infty$, where $C, s_0 > 0$ are constants free of p, and s_0 is an integer.

THEOREM 3. *Let Conditions 1 and 5 hold. For any integers $r, j \ge 0$, there exists a banded matrix $\Sigma_j^{(r)}$ with bandwidth $2\{(2r + j)k_0 + s_0\} + 1$ such that*

$$\|\Sigma_j^{(r)} - \Sigma_j\|_2 \le C_1\, \delta^{2(r+j)+1}, \ \ \|\Sigma_j^{(r)} - \Sigma_j\|_1 \le C_2\, r\, \delta^{2(r+j)+1},$$

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

Under Condition 5, $\Sigma_0^{(r)} = \Sigma_\varepsilon + \sum_{1 \le i \le r} A_1^i \Sigma_\varepsilon (A_1^{\mathrm{T}})^i$ is a banded matrix with bandwidth $2(2rk_0 + s_0) + 1$. Theorem 3 ensures that the norms of the difference $\Sigma_0 - \Sigma_0^{(r)} =$

$\sum_{i>r} A_1^i \Sigma_\varepsilon (A_1^{\mathrm{T}})^i$ admit the required upper bounds. Theorem 3 also paves the way for estimating Σ_j using the banding method of Bickel and Levina (2008), as Σ_j can be approximated by a banded matrix with a bounded error and thus may be effectively treated as a banded matrix. To this end, we define the banding operator as follows: for any matrix $H = (h_{ij})$, $B_r(H) = \big(h_{ij} I(|i-j| \le r)\big)$. Then the banding estimator for Σ_j is defined as

$$\widehat{\Sigma}_j^{(r_n)} = B_{r_n}(\widehat{\Sigma}_j), \;\; \widehat{\Sigma}_j = \frac{1}{n}\sum_{t=1}^{n-j}(y_t - \bar{y})(y_{t+j} - \bar{y})^{\mathrm{T}}, \;\; \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t, \tag{13}$$

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

THEOREM 4. *Assume that Conditions (1-5) hold. Then for any integer $j \ge 0$, as $n, p \to \infty$,*

$$\|\widehat{\Sigma}_j^{(r_n)} - \Sigma_j\|_2 = O_P\Big\{ r_n\big(n^{-1}\log p\big)^{1/2} + \delta^{2(r_n+j)+1}\Big\} = O_P\Big\{ \log(n/\log p)\big(n^{-1}\log p\big)^{1/2}\Big\},$$

and

$$\|\widehat{\Sigma}_j^{(r_n)} - \Sigma_j\|_1 = O_P\Big\{ \log(n/\log p)\,\big(n^{-1}\log p\big)^{1/2}\,\Big\}.$$

In practice we need to specify r_n. An ideal selection would be $r_n = \arg\min_r R_j(r)$, where

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

but in practice this is unavailable because Σ_j is unknown. We replace it by an estimator obtained via a version of wild bootstrap. To this end, let $u_1, \ldots, u_n$ be independent and identically distributed with $E(u_t) = \mathrm{var}(u_t) = 1$. A bootstrap estimator for Σ_j is defined as

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

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

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

We choose r_n to minimize $R_j^*(r)$. In practice we use the approximation

$$R_j^*(r) \approx \frac{1}{q}\sum_{k=1}^{q} \|B_r(\Sigma_{j,k}^*) - \widehat{\Sigma}_j\|_1, \tag{14}$$

where $\Sigma_{j,1}^*, \ldots, \Sigma_{j,q}^*$ are q bootstrap estimates for Σ_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 evaluate the finite-sample properties of the proposed methods for 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) $\{a_{ij};\ |i-j| \le k_0\}$ are generated independently from $U[-1, 1]$. Since the spectral norm of A must be smaller than 1, we re-scale A by $\eta A/\|A\|_2$, where η 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 $\text{pr}(\xi = 1) = 0.4$. The elements $\{a_{ij}; |i-j| = k_0\}$ are drawn independently from -4 and 4 with probability 0.5 each. Then A is rescaled as in (i) above.

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

We set $n = 200$, $p = 100, 200, 400, 800$, and $k_0 = 1, 2, 3, 4$. We repeat each setting 500 times. We only report the results with $K = 15$ in (8), as the results with other values of $K \ge k_0$ are similar. Table 1 lists the relative frequencies of the occurrence of the events $\{\widehat{k} = k\}$, $\{\widehat{k} > k_0\}$ and $\{\widehat{k} < k_0\}$ over the 500 replications. Overall $\widehat{k}$ under-estimates k_0, especially when $k_0 = 3$ or 4. In fact when $k_0 = 4$, $\widehat{k}$ chose 3 most times. The constraint $\|A\| < 1$ makes most non-zero elements small or very small when p is large, and that only the coefficients at least as large as $\sqrt{\log(p \vee n)/n}$ are identifiable; see Condition 2. Estimation performs better in setting (ii) than in setting (i), as Condition 2 is more likely to hold at the boundaries of the band in setting (ii).

The BIC (7) is defined for each row separately. One natural alternative would be

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

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

$$\widetilde{k} = \arg \min_{1 \le k \le K} \text{BIC}(k). \tag{15}$$

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 (7), presented in 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) is reported in Table 3. Table 3 also reports results from estimating A using the true values for the bandwidth parameter k_0. The accuracy loss 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 estimation performance for the auto-covariance matrices Σ_0 and Σ_1, we set $k_0 = 3$, and the spectral norm of A at 0.8. Furthermore, we let ε_t be independent and $N(0, \Sigma_\varepsilon)$ now, where $\Sigma_\varepsilon = BB^{\mathrm{T}}$ and $B = (b_{ij}), b_{11} = 1, b_{ij} = 0.8I(|i-j| = 1) + 0.6I(i = j), i > 1$ or $j > 1$. Table 4 lists the average estimation errors and the standard deviations over 100 replications, measured by matrix L_1-norm. We also report Monte Carlo results for a thresholded estimator and the sample covariance estimator. For the banded estimator, we choose r to minimize the bootstrap loss defined in (14) with $q = 100$. For the thresholded estimator, the thresholding parameter is selected in the same manner. Table 4 shows that the proposed banding method performs much better than the thresholded 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 from 1 January 1990 to 17 December 17 2000, i.e., $p = 71$ and $n = 572$. Fig.1 displays the weekly temperature of Ha'erbin, Shanghai and Hangzhou, showing strong seasonal behavior with period 52 weeks. 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 by $\{\, y_t;\ t = 1, \ldots, 572 \,\}$, and each y_t has 71 components.

Naturally we would order the 71 cities according to their geographic locations. However the choice is not unique. For example, we may order the cities from north to south, from west to east, from northwest to southeast, or from southwest to northeast. By setting $d = 1$, each ordering leads to a different banded VAR(1) model. We compare those four models by one-step ahead, and two-step ahead post-sample prediction for the last 30 data points in the series. To select an optimum model, we compute the BIC value according to (10). These numerical results and the selected bandwidth parameters $\widehat{k}$ are reported in Table 5. Three out of those four models select $\widehat{k} = 2$, while the model based on the ordering from west to east picks $\widehat{k} = 4$. Overall the model based on the ordering from southwest to northeast is preferred by the BIC, which also has the minimum one-step ahead post-sample predictive errors. The performances of the four models in terms of both the BIC and the prediction are very close.

Also included in Table 5 are the post-sample predictive errors of the lasso VAR(1) model obtained by minimizing

$$\sum_{t=2}^{n} \|y_t - A y_{t-1}\|^2 + \sum_{i,j=1}^{p} \lambda_i |a_{ij}|,$$

where $\{\lambda_i; i = 1, \ldots, p\}$ are tuning parameters estimated by a five-fold cross-validation as in Bickel and Levina (2008). The prediction accuracy of the lasso VAR(1) model is comparable to those of the banded VAR(1) models, though slightly worse especially for the two-step ahead prediction. However the lack of any structure in the estimated sparse coefficient matrix $\widetilde{A}$, displayed in Fig.2(b), makes such fit difficult to interpret. In contrast, the banded coefficient matrix, depicted in Fig.2(a), is attractive. It also makes data collection and administration easier.

Example 2. Now we consider the daily sales of a clothing brand in 21 provinces in China from 1 January 2008 to 9 December 2012, i.e., $n = 1812,\ p = 21$. Fig.4 plots the relative geographical positions of 21 provinces and province-level municipalities. For convenient analysis, we first subtract each of the 21 series by its mean. Similar to Example 1 above, we order the 21 provinces according to the four different geographic orientations, and fit a banded VAR(1) model for each ordering. The selected bandwidth parameters, the BIC values and the post sample prediction errors for the last 30 data points in the series are reported in Table 6. We also rank the series according to their geographic distances to Heilong Jiang, the most northwestern province; see Fig.4. This results in a different ordering to that from north to south. Table 6 indicates that the minimum bandwidth parameter $\widehat{k}$ is 3, attained by the ordering based on the distances to Heilong Jiang, followed by $\widehat{k} = 4$ attained by the north-to-south ordering. The post-sample prediction performances of those two models are almost the same, and are better than those of the other three banded VAR(1) models and the lasso VAR(1) model.

The ordering based on the direction from northwest to southeast leads to $\widehat{k} = 12$. Therefore the corresponding banded VAR model has 21 regressors for some components according to (3), i.e., no banded structure is observed in this case. Fig.4 indicates that the ordering from northwest

to southeast puts together some provinces which are distance away from each other. Hence this is certainly a wrong ordering as far as the banded VAR struture is concerned.

The estimated coefficient matrix $\widehat{A}$ for the banded VAR(1) model based on the distances to Heilong Jiang and the estimated $\widetilde{A}$ for the lasso VAR(1) model are plotted in Fig.3. The banded VAR(1) model facilitates an easy interpretation, i.e., the sales in the neighbour provinces are closely associated with each other. The lasso fitting cannot reveal this phenomenon.

ACKNOWLEDGEMENTS

We are grateful to the Editor, the Associate Editor and two referees for their insightful comments and valuable suggestions, which lead to significant improvement of our article. This research was supported in part by Natural NSF of China (S. Guo), NSF grants of U.S.A (Y. Wang) and EPSRC research grant (S. Guo and Q. Yao). This paper was completed when S. Guo was Research Fellow at London School of Economics and Assistant Professor at Chinese Academy of Sciences.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Theorems 1-4, the consistency of generalized BIC defined by (9) in Section 2.3 and the consistency of BIC selector in the setting $k_0 \to \infty$, as well as the detailed proofs of all the lemmas in this paper.

REFERENCES

BASU, S. AND MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43**, 1535–1567.

BASU, S., SHOJAIE, A. AND MICHAILIDIS, G. (2015). Network Granger Causality with Inherent Grouping Structure. *J. Mach. Learn. Res.* **16**, 417–453.

BICKEL, P. J. AND LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **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. *J. R. Statist. Soc. B* **73**, 711–728.

BOLSTAD, A., VAN VEEN, B. D. AND NOWAK, R. (2011). Causal network inference via group sparse regularization. *IEEE Trans. Sig. Proc.* **59**, 2628–2640.

SE CAN, A. AND MEGBOLUGBE, I. (1997). Spatial dependence and house price index construction. *J. Real. Est. Fin. Econ.* **14**, 203–222.

CHEN, X., XU, M. AND WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41**, 2994–3021.

HAN, F. AND LIU, H. (2015). A Direct Estimation of High Dimensional Stationary Vector Autoregressions. *J. Mach. Learn. Res.* **16**, 3115–3150.

HAUFE, S., NOLTE, G., MUELLER, K. R., AND KRäMER, N. (2010). Sparse causal discovery in multivariate time series. *J. Mach. Learn. Res.* W&CP **6**, 97–106.

HSU, N. J., HUNG, H. L., AND CHANG, Y. M. (2008). Subset selection for vector autoregressive processes using lasso. *Comp. Statist. Data. Ana.* **52**, 3645–3657.

KOCK, A. AND CALLOT, L.(2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econ.* **186**, 325-344.

LENG, C. AND LI, B. (2011). Forward adaptive banding for estimating large covariance matrices. *Biometrika* **98**, 821–830.

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. *J. Statist. Plan. Infer.* **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 autoregressions. *J. App. Econ.* **25**, 71-92.

SHOJAIE, A. AND MICHAILIDIS, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **26**, 517–523.

Table 1. *Relative frequencies multiplied by* 100 *for the occurrence of the events* $\{\widehat{k} = k\}$, $\{\widehat{k} > k_0\}$ *and* $\{\widehat{k} < k_0\}$ *in a simulation study with* 500 *replications, where* $\widehat{k}$ *is defined in (*8*).*

		Setting (i)			Setting (ii)		
		$\{\widehat{k} = k_0\}$	$\{\widehat{k} > k_0\}$	$\{\widehat{k} < k_0\}$	$\{\widehat{k} = k_0\}$	$\{\widehat{k} > k_0\}$	$\{\widehat{k} < k_0\}$
$p = 100$	$k_0 = 1$	82	17	1	98	2	0
	$k_0 = 2$	87	8	5	95	3	2
	$k_0 = 3$	73	6	21	83	2	15
	$k_0 = 4$	55	14	31	64	2	34
$p = 200$	$k_0 = 1$	91	9	0	97	3	0
	$k_0 = 2$	89	4	7	93	2	5
	$k_0 = 3$	65	3	32	83	0	17
	$k_0 = 4$	54	1	45	63	2	35
$p = 400$	$k_0 = 1$	95	5	0	99	1	0
	$k_0 = 2$	87	2	11	90	1	9
	$k_0 = 3$	66	2	32	76	1	23
	$k_0 = 4$	45	1	54	60	0	40
$p = 800$	$k_0 = 1$	97	3	0	100	0	0
	$k_0 = 2$	86	1	13	91	1	8
	$k_0 = 3$	59	1	40	67	1	32
	$k_0 = 4$	40	0	60	52	0	48

Table 2. *Relative frequencies multiplied by* 100 *for the occurrence of the events* $\{\widetilde{k} = k\}$, $\{\widetilde{k} > k_0\}$ *and* $\{\widetilde{k} < k_0\}$ *in a simulation study with* 500 *replications, where* $\widetilde{k}$ *is defined in (*15*).*

		Setting (i)			Setting (ii)		
		$\{\widetilde{k} = k_0\}$	$\{\widetilde{k} > k_0\}$	$\{\widetilde{k} < k_0\}$	$\{\widetilde{k} = k_0\}$	$\{\widetilde{k} > k_0\}$	$\{\widetilde{k} < k_0\}$
$p = 100$	$k_0 = 1$	64	0	36	88	0	12
	$k_0 = 2$	42	0	58	63	0	37
$p = 200$	$k_0 = 1$	56	0	44	84	0	16
	$k_0 = 2$	32	0	68	55	0	45
$p = 400$	$k_0 = 1$	48	0	52	83	0	17
	$k_0 = 2$	23	0	77	45	0	55
$p = 800$	$k_0 = 1$	44	0	56	76	0	24
	$k_0 = 2$	11	0	89	41	0	59

WANG, H., LI, B. AND LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B* **71**, 671–683.

WU, W. B. AND POURAHMADI, M. (2009). Banding sample covariance matrices of stationary processes. *Statist. Sin.* **19**, 1755–68.

Table 3. *Means with their corresponding standard deviations in parentheses of the errors in estimating A under setting (i) in a simulation study with $n = 200$ and 500 replications.*

		With estimated k_0		With true k_0	
p		$\|\widehat{A} - A\|_1$	$\|\widehat{A} - A\|_2$	$\|\widehat{A} - A\|_1$	$\|\widehat{A} - A\|_2$
$p = 100$	$k_0 = 1$	0.38 (0.06)	0.27 (0.03)	0.37 (0.05)	0.27 (0.03)
	$k_0 = 2$	0.54 (0.06)	0.33 (0.03)	0.53 (0.05)	0.33 (0.03)
	$k_0 = 3$	0.70 (0.08)	0.39 (0.04)	0.69 (0.07)	0.38 (0.03)
	$k_0 = 4$	0.85 (0.10)	0.43 (0.05)	0.85 (0.08)	0.43 (0.03)
$p = 200$	$k_0 = 1$	0.40 (0.06)	0.28 (0.03)	0.40 (0.05)	0.28 (0.03)
	$k_0 = 2$	0.58 (0.07)	0.35 (0.03)	0.58 (0.06)	0.35 (0.03)
	$k_0 = 3$	0.74 (0.08)	0.40 (0.04)	0.74 (0.06)	0.40 (0.03)
	$k_0 = 4$	0.90 (0.11)	0.46 (0.05)	0.88 (0.07)	0.45 (0.03)
$p = 400$	$k_0 = 1$	0.43 (0.05)	0.30 (0.03)	0.42 (0.04)	0.30 (0.03)
	$k_0 = 2$	0.60 (0.06)	0.36 (0.03)	0.60 (0.05)	0.36 (0.03)
	$k_0 = 3$	0.77 (0.08)	0.42 (0.04)	0.76 (0.06)	0.42 (0.03)
	$k_0 = 4$	0.95 (0.14)	0.48 (0.07)	0.93 (0.07)	0.46 (0.03)
$p = 800$	$k_0 = 1$	0.44 (0.04)	0.31 (0.02)	0.44 (0.04)	0.31 (0.02)
	$k_0 = 2$	0.63 (0.05)	0.37 (0.03)	0.62 (0.05)	0.37 (0.02)
	$k_0 = 3$	0.81 (0.09)	0.43 (0.05)	0.80 (0.06)	0.43 (0.02)
	$k_0 = 4$	0.98 (0.14)	0.49 (0.07)	0.96 (0.07)	0.47 (0.02)

Table 4. *Means with their corresponding standard deviations in parentheses of the errors in estimating autocovariance matrices in a simulation study with $n = 200$ and 100 replications.*

	$\|\widehat{\Sigma}_{n,0} - \Sigma_0\|_1$			$\|\widehat{\Sigma}_{n,1} - \Sigma_1\|_1$		
	Banding	Thresholding	Sample	Banding	Thresholding	Sample
	Matrix L_1-Norm			Matrix L_1-Norm		
$p = 100$	2.1 (0.04)	2.6 (0.02)	14 (0.07)	2.9 (0.03)	3.5 (0.04)	14 (0.07)
$p = 200$	2.7 (0.04)	3.4 (0.03)	29 (0.02)	3.1 (0.03)	4.2 (0.04)	30 (0.02)
$p = 400$	2.3 (0.02)	2.9 (0.02)	55 (0.02)	2.8 (0.03)	3.7 (0.02)	55 (0.02)
$p = 800$	2.7 (0.03)	3.4 (0.02)	112 (0.03)	2.9 (0.03)	3.9 (0.03)	110 (0.04)
	Spectral Norm			Spectral Norm		
$p = 100$	1.1 (0.01)	1.4 (0.02)	4.0 (0.07)	1.4 (0.01)	1.7 (0.02)	3.7 (0.02)
$p = 200$	1.3 (0.03)	1.7 (0.02)	6.5 (0.03)	1.5 (0.01)	1.9 (0.01)	6.1 (0.02)
$p = 400$	1.2 (0.01)	1.6 (0.01)	10 (0.03)	1.3 (0.01)	1.9 (0.01)	9.2 (0.02)
$p = 800$	1.4 (0.02)	1.8 (0.01)	17 (0.03)	1.4 (0.01)	2.3 (0.02)	15 (0.03)

[*Received February* 2015. *Revised August* 2016]

Table 5. *Results of Example 1: Estimated bandwidth parameters, BIC and average one-step-ahead and two-step-ahead post-sample predictive errors over 71 cities with their corresponding standard errors in parentheses.*

Ordering	$\widehat{k}$	BIC	One-step ahead	Two-step ahead
north to south	2	552.5	1.543 (1.170)	1.622 (1.245)
west to east	4	555.9	1.545 (1.152)	1.602 (1.247)
northwest to southeast	2	552.4	1.552 (1.167)	1.624 (1.249)
southwest to northeast	2	551.9	1.538 (1.160)	1.617 (1.253)
Lasso	-	-	1.545 (1.172)	1.632 (1.250)

Table 6. *Results of Example 2: Estimated bandwidth parameters, BIC and average one-step-ahead and two-step-ahead post-sample predictive errors over 21 provinces with their corresponding standard errors in parentheses.*

Ordering	$\widehat{k}$	BIC	One-step ahead	Two-step ahead
north to south	4	114.92	0.314 (0.377)	0.407 (0.386)
west to east	7	115.18	0.323 (0.363)	0.409 (0.386)
northwest to southeast	12	115.21	0.322 (0.361)	0.409 (0.395)
southwest to northeast	5	115.12	0.316 (0.374)	0.407 (0.385)
distance to Heilongjiang	3	114.68	0.313 (0.378)	0.407 (0.386)
Lasso	-	-	0.322 (0.362)	0.410 (0.393)

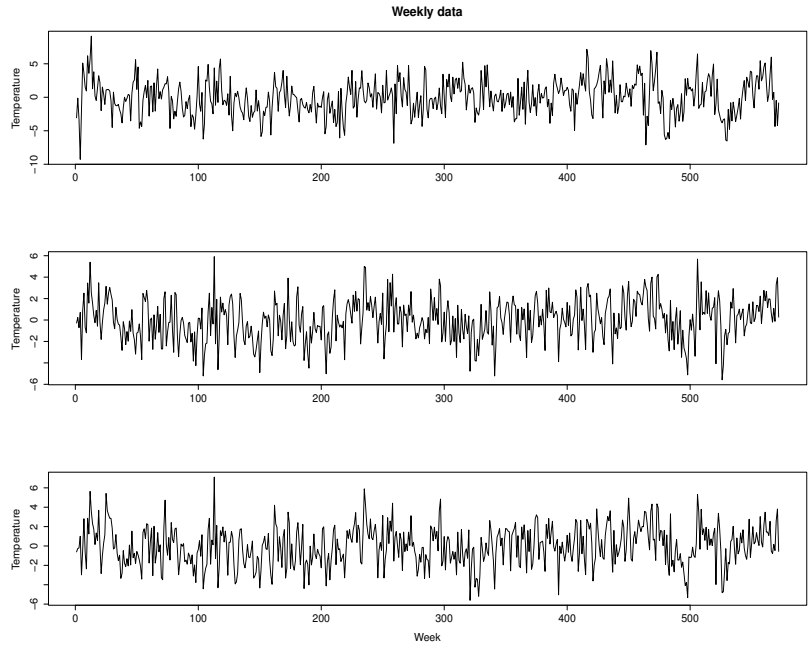

Fig. 1. Time series plots of the deseasonalized weekly temperature from January 1990 to December 2000, where Ha'erbin, Shanghai and Nanjing correspond to the plots from top to bottom.

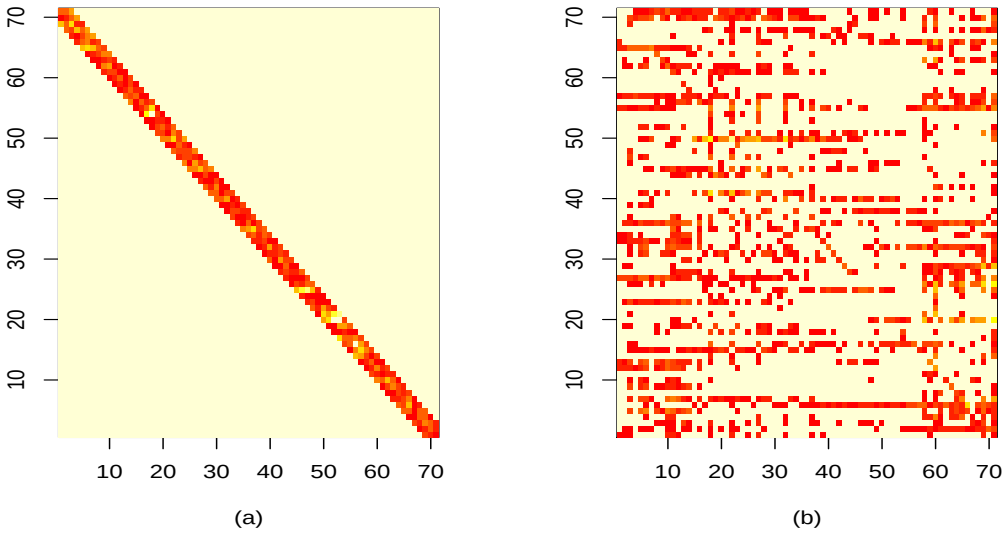

Fig. 2. Example 1: (a) Estimated banded coefficient matrix $\widehat{A}$ for the model based on the ordering from southwest to northeast, and (b) estimated sparse coefficient matrix $\widetilde{A}$ by lasso. The larger the absolute value of a coeffcient is, the darker the colour is.

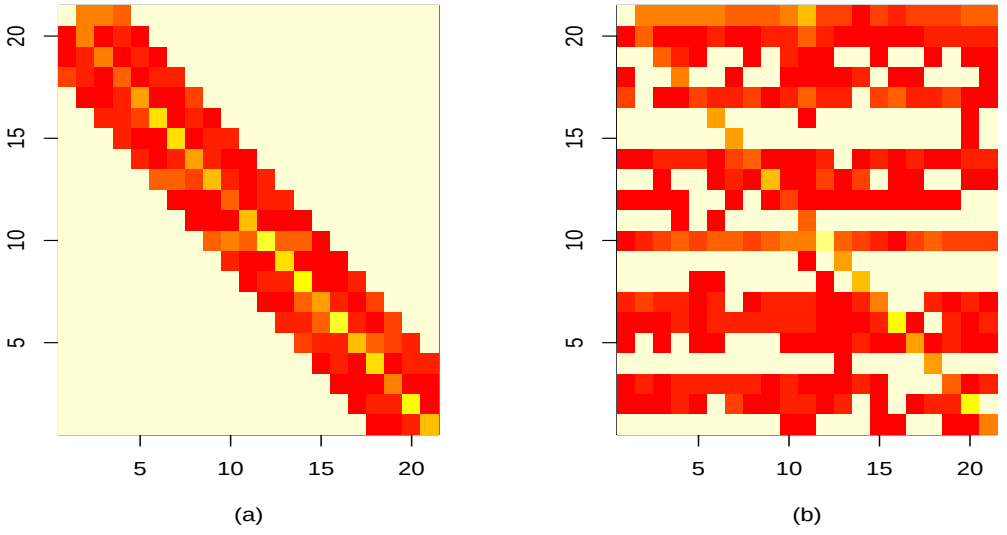

Fig. 3. Example 2: (a) Estimated banded coefficient matrix $\widehat{A}$ for the model based on the ordering using distances to Heilongjiang, and (b) estimated sparse coefficient matrix $\widetilde{A}$ by lasso. The larger the absolute value of a coeffcient is, the darker the colour is.

Fig. 4. Location plot of 21 provinces and province-level municipalities in China, where Shanghai is a province-level municipality, and Ha'erbin, Hangzhou and Nanjing are the capitals of Heilongjiang, Zhejiang, and Jiangsu provinces, respectively.