Biometrika (2016), **103**, 4, *pp.* 889–903
Printed in Great Britain

doi: 10.1093/biomet/asw046
Advance Access publication 27 October 2016

High-dimensional and banded vector autoregressions

BY SHAOJUN GUO

Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China
sjguo@ruc.edu.cn

YAZHEN WANG

Department of Statistics, University of Wisconsin, Madison, Wisconsin 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. This setting represents a type of sparse structure for high-dimensional time series, although the implied auto-covariance matrices are not banded. The structure is also practically meaningful when the component time series are ordered appropriately. We establish the convergence rates of the estimated banded autoregressive coefficient matrices. 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 structures 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 matrix; Bayesian information criterion; Frobenius norm; Vector autoregressive model.

1. INTRODUCTION

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

In many practical situations it is enough to collect information from neighbouring variables, although the definition of neighbourhood is case-dependent. For example, sales, prices, weather indices and electricity consumption influenced by temperature depend on the values at nearby locations, in the sense that information from farther locations may become redundant given the information from neighbours; see se Can & Mebolugbe (1997) for a house price example which exhibits such a dependence structure. In this paper, we propose a class of vector autoregressive models to deal with such dynamic structures. We assume that the autoregressive coefficient matrices are banded, i.e., nonzero coefficients form a narrow band along the main diagonal. The setting specifies explicit autoregression over neighbouring component series only. Nevertheless, nonzero cross-correlations among all the component series may exist, as the implied auto-covariance matrices are not banded. This is an effective way to impose sparse structure, 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 obtained by ordering the component series appropriately. The ordering can be deduced from subject knowledge, aided by statistical tools such as the Bayesian information criterion; see § 5·2. With the imposed banded structure, we propose least-squares estimators for the autoregressive coefficient matrices, which attain a convergence rate of $(p/n)^{1/2}$ under the Frobenius norm and a rate of $(\log p/n)^{1/2}$ under the spectral norm, when p diverges together with the length n of the time series.

In practice, the maximum width of the nonzero coefficient bands in the coefficient matrices, called the bandwidth, is unknown. We propose a marginal Bayesian information criterion to identify the true bandwidth. We show that this criterion leads to consistent bandwidth determination as both n and p tend to infinity.

We also address estimation of the auto-covariance functions for high-dimensional banded autoregressive models. Although the auto-covariance matrices of a banded 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 auto-covariance matrices are consistent estimators with convergence rate $\log(n/\log p)(\log p/n)^{1/2}$, which is faster than that of the standard banding covariance estimators (Bickel & Levina, 2008). See also Wu & Pourahmadi (2009), Bickel & Gel (2011) and Leng & Li (2011) for estimation of the banded covariance matrices of time series.

Most existing work on high-dimensional autoregressive models draws inspiration from recent developments in high-dimensional regression. 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 use of group lasso penalization. A truncated weighted lasso and group lasso penalization approaches were proposed by Shojaie & Michailidis (2010) and Basu et al. (2015), respectively, to explore graphical Granger causality. Basu & Michailidis (2015) focused on stable Gaussian processes and investigated the theoretical properties of L_1-regularized estimates of transition matrices in sparse autoregressive models. Bolstad et al. (2011) inferred sparse causal networks through vector autoregressive processes and proposed a group lasso procedure. Kock & Callot (2015) established oracle inequalities for high-dimensional vector autoregressive models. Han & Liu (2015) proposed an alternative Dantzig-type penalization and formulated the estimation problem as a linear program. Chen et al. (2013) studied sparse covariance and precision matrices in high-dimensional time series under a general dependence structure.

2. Methods

2·1. *Banded vector autoregressive models*

Let y_t be a $p \times 1$ time series 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, with $E(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = E(\varepsilon_t \varepsilon_t^{\mathrm{T}}) = \Sigma_\varepsilon$, and ε_t is independent of $y_{t-1}, y_{t-2}, \ldots$. Furthermore, the coefficient matrices $A_1, \ldots, A_d$ are all banded in the sense that

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

where $a_{ij}^{(\ell)}$ denotes the (i,j)th element of A_ℓ. The maximum number of nonzero elements in each row of A_ℓ is the bandwidth $2k_0 + 1$, and k_0 is called the bandwidth parameter. We assume that $k_0 \geqslant 0$ and $d \geqslant 1$ are fixed integers, and that $p \gg k_0, d$. 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 thoroughly studied; see, for example, Lütkepohl (2007, Ch. 4).

Under the condition $\det(I_p - A_1 z - \cdots - A_d z^d) \neq 0$ for any $|z| \leqslant 1$, where I_p denotes the $p \times p$ identity matrix, model (1) admits a weakly stationary solution $\{y_t\}$. Throughout this paper, y_t refers 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 it is, the auto-covariance matrices are not necessarily banded; see (12) below. Therefore, the proposed banded model is applicable when the linear dynamics of each component series depends predominantly on its neighbouring series, though there may be nonzero correlations among all component series of y_t.

2·2. *Estimating banded autoregressive coefficient matrices*

Since each row of A_ℓ has at most $2k_0 + 1$ nonzero 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 nonzero elements in the ith 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

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

where $y_{i,t}$ and $\varepsilon_{i,t}$ are the ith components of y_t and ε_t, respectively, 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

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

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 jth row.

By estimating β_i $(i = 1, \ldots, p)$ separately based on (5), we obtain the least-squares estimators $\hat{A}_1, \ldots, \hat{A}_d$ 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 this as a function of k_0 to stress that the above estimation presupposes that the bandwidth is $2k_0 + 1$ in the sense of (2).

2·3. *Determination of the bandwidth*

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

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

where $\text{RSS}_i(k)$ 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 below. We often take C_n to be $\log \log n$. An estimator for k_0 is

$$\hat{k} = \max_{1 \leqslant i \leqslant p} \left\{ \arg\min_{1 \leqslant k \leqslant K} \text{BIC}_i(k) \right\}, \tag{8}$$

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

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

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

Let L be a prescribed integer upper bound on d, often taken to be 10 or $[n^{1/2}]$. Let

$$(\hat{k}_i, \hat{d}_i) = \arg\min_{1 \leqslant k \leqslant K,\, 1 \leqslant \ell \leqslant L} \widetilde{\text{BIC}}_i(k, \ell) \quad (i = 1, \ldots, p),$$

and let $\hat{k} = \max_{1 \leqslant i \leqslant p} \hat{k}_i$ and $\hat{d} = \max_{1 \leqslant i \leqslant p} \hat{d}_i$. Proposition S1 in the Supplementary Material shows that under Conditions 1–4 in § 3·1, $\text{pr}(\hat{k} = k_0, \hat{d} = d) \to 1$ as $n, p \to \infty$.

Remark 2. The banded structure of the coefficient matrices $A_1, \ldots, A_d$ depends 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 that minimizes the bandwidth, but such a procedure is computationally burdensome for large p. For most applications, meaningful orderings are suggested by practical considerations. We can then calculate

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

for each suggested ordering, and choose the ordering which minimizes (10). In expression (10), $\text{BIC}_i(\cdot)$ and $\hat{k}$ are defined as in (7) and (8). Two real-data examples in § 5·2 indicate that this scheme works well in applications.

3. Asymptotic properties

3·1. *Regularity conditions*

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

$$\|v\|_q = \left(\sum_{j=1}^{p} |v_j|^q\right)^{1/q}, \quad \|v\|_\infty = \max_{1\leqslant j\leqslant p} |v_j|, \quad \|B\|_q = \max_{\|v\|_q=1} \|Bv\|_q, \quad \|B\|_{\mathrm{F}} = \left(\sum_{i,j} b_{ij}^2\right)^{1/2};$$

that is, $\|\cdot\|_q$ denotes the ℓ_q-norm of a vector or matrix, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix.

First, note that the model (1) can be formulated as

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

where

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

Now we list the regularity conditions required for our asymptotic results.

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

Condition 1′. For $\tilde{A}$ defined in (11), $\|\tilde{A}^{j_0}\|_2 \leqslant \delta^{j_0}$, $\|\tilde{A}\|_\infty \leqslant C$ and $\|\tilde{A}^{j_0}\|_\infty \leqslant \delta^{j_0}$, where $C > 0$, $\delta \in (0,1)$ and $j_0 \geqslant 1$ 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 $1 \leqslant \ell \leqslant d$, where $C_n \to \infty$ as $n \to \infty$.

Condition 3. We have $\lambda_{\min}\{\mathrm{cov}(y_t)\} \geqslant \kappa_1$ and $\max_{1\leqslant i\leqslant p}|\sigma_{ii}| \leqslant \kappa_2$ for some positive constants κ_1 and κ_2 free of p, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue and σ_{ii} is the ith diagonal element of $\mathrm{cov}(y_t)$.

Condition 4. The innovation process $\{\varepsilon_t : t = 0, \pm 1, \pm 2, \ldots\}$ is independent and identically distributed with mean zero and covariance Σ_ε. Moreover, one of the following two properties holds:

(i) $\max_{1\leqslant i\leqslant p} E(|\varepsilon_{i,t}|^{2q}) \leqslant C$ and $p = O(n^\beta)$, where $q > 2$, $\beta \in (0, (q-2)/4)$ and $C > 0$ are constants free of n and p;
(ii) $\max_{1\leqslant i\leqslant p} E\{\exp(\lambda_0|\varepsilon_{i,t}|^{2\alpha})\} \leqslant C$ 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 that y_t is strictly stationary and that for any $j \geqslant 1$, $\|\tilde{A}^j\|_2 \leqslant C\delta^j$ with some $C > 0$ and $\delta \in (0,1)$. The independent and identically distributed assumption in Condition 4 is imposed to simplify the proofs but is not essential. Condition 2 ensures that the bandwidth $2k_0 + 1$ is asymptotically identifiable, as $\{n^{-1}\log(p \vee n)\}^{1/2}$ is the minimum order for a nonzero coefficient to be identifiable; see, e.g., Luo & 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) high-dimensional cases with $p = O(n^{\beta})$, and (ii) ultrahigh-dimensional cases with $\log p = o\{n^{\alpha/(2-\alpha)}\}$.

3·2. *Asymptotic theorems*

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

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

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

Since k_0 is unknown, we replace it by $\hat{k}$ in the estimation procedure for $A_1, \ldots, A_d$ described in § 2·2, and still denote the resulting estimators by $\hat{A}_1, \ldots, \hat{A}_d$. Theorem 2 addresses their convergence rates.

Theorem 2. *Suppose that Conditions 1–4 hold. As* $n \to \infty$, *for* $j = 1, \ldots, d$,

$$\|\hat{A}_j - A_j\|_{\mathrm{F}} = O_{\mathrm{p}}\{(p/n)^{1/2}\}, \quad \|\hat{A}_j - A_j\|_2 = O_{\mathrm{p}}\{(\log p/n)^{1/2}\}.$$

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

4. Estimation for auto-covariance functions

For the banded vector autoregressive 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 autoregressive process of order 1, 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 $2k_2 + 1$, in general the product $B_1 B_2$ is a banded matrix with the enlarged bandwidth $2(k_1 + k_2) + 1$. Hence Σ_0 in (12) is not a banded matrix. Nevertheless, if $\mathrm{var}(\varepsilon_t) = \Sigma_\varepsilon$ is also banded, Theorem 3 shows that Σ_j can be approximated by some banded matrices.

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

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

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

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, $\Sigma_0^{(r)} = \Sigma_\varepsilon + \sum_{1\leqslant i\leqslant r} A_1^i \Sigma_\varepsilon (A_1^{\mathrm{T}})^i$ 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 & Levina (2008), as Σ_j can be approximated by a banded matrix with a bounded error and may thus be effectively treated as a banded matrix. To this end, we define the banding operator as follows: for any matrix $H = (h_{ij})$, $B_r(H) = \{h_{ij} I(|i-j| \leqslant r)\}$. Then the banding estimator for Σ_j is defined as

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

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

THEOREM 4. *Suppose that Conditions 1–5 hold. Then, for any integer $j \geqslant 0$, as $n, p \to \infty$,*

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

and

$$\left\| \hat{\Sigma}_j^{(r_n)} - \Sigma_j \right\|_1 = O_{\mathrm{p}}\left\{ \log(n/\log p)(n^{-1}\log p)^{1/2} \right\}.$$

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

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

but in practice this is unavailable because Σ_j is unknown. Instead, we replace it with an estimator obtained by a 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

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

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

$$R_j^*(r) \approx \frac{1}{q}\sum_{k=1}^{q} \|B_r(\Sigma_{j,k}^*) - \hat{\Sigma}_j\|_1, \tag{13}$$

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, with q being a large integer.

5. Numerical properties

5·1. *Simulations*

In this section, we evaluate the finite-sample properties of the proposed methods for the model

$$y_t = Ay_{t-1} + \varepsilon_t,$$

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

(i) $\{a_{ij} : |i-j| \leqslant k_0\}$ are generated independently from Un$[-1, 1]$. Since the spectral norm of A must be smaller than 1, we rescale A by $\eta A/\|A\|_2$, where η is generated from Un$[0{\cdot}3, 1{\cdot}0)$;

(ii) $\{a_{ij} : |i-j| < k_0\}$ are generated independently from the mixture distribution $\xi \cdot 0 + (1 - \xi)N(0, 1)$ with $\mathrm{pr}(\xi = 1) = 0{\cdot}4$. The elements $\{a_{ij} : |i-j| = k_0\}$ are drawn independently from $\{-4, 4\}$ with probability 0·5 each; then A is rescaled as in (i) above.

In setting (ii), there are about $0{\cdot}4(2k_0 - 1)p$ zero elements within the band, i.e., A is sparser than 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. We report only the results with $K = 15$ in (8), as the results with other values of $K \geqslant k_0$ are similar. Table 1 lists the relative frequencies of occurrence of the events $\{\hat{k} = k\}$, $\{\hat{k} > k_0\}$ and $\{\hat{k} < k_0\}$ over the 500 replications. Overall, $\hat{k}$ underestimates k_0, especially when $k_0 = 3$ or 4. In fact, when $k_0 = 4$, $\hat{k}$ chose the value 3 most times. The constraint $\|A\| < 1$ makes most nonzero elements small or very small when p is large, and only those coefficients which are at least as large as $\surd\{\log(p \vee n)/n\}$ 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 Bayesian information criterion (7) is defined for each row separately. One natural alternative would be

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

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

		Setting (i)			Setting (ii)		
		$\{\hat{k} = k_0\}$	$\{\hat{k} > k_0\}$	$\{\hat{k} < k_0\}$	$\{\hat{k} = k_0\}$	$\{\hat{k} > k_0\}$	$\{\hat{k} < k_0\}$
$p = 100$	$k_0 = 1$	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 of occurrence* (%) *of the events* $\{\tilde{k} = k\}$, $\{\tilde{k} > k_0\}$ *and* $\{\tilde{k} < k_0\}$ *in a simulation study with* 500 *replications, where* $\tilde{k}$ *is as defined in* (14)

		Setting (i)			Setting (ii)		
		$\{\tilde{k} = k_0\}$	$\{\tilde{k} > k_0\}$	$\{\tilde{k} < k_0\}$	$\{\tilde{k} = k_0\}$	$\{\tilde{k} > k_0\}$	$\{\tilde{k} < k_0\}$
$p = 100$	$k_0 = 1$	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

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

$$\tilde{k} = \arg\min_{1 \leqslant k \leqslant K} \textsc{bic}(k). \tag{14}$$

Although this joint approach can be shown to be consistent, its finite-sample performance, reported in Table 2, is worse than that of the marginal Bayesian information criterion (7), presented in Table 1.

In addition, we calculate both L_1 and L_2 errors in estimating the banded coefficient matrix A. The means and standard deviations of the errors for setting (i) are reported in Table 3, which also shows results from estimating A using the true values of the bandwidth parameter k_0. The

Table 3. *Means* ($\times 10^2$) *with corresponding standard deviations* ($\times 10^2$*, 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		$\Vert\hat{A}-A\Vert_1$	$\Vert\hat{A}-A\Vert_2$	$\Vert\hat{A}-A\Vert_1$	$\Vert\hat{A}-A\Vert_2$
100	$k_0 = 1$	38 (6)	27 (3)	37 (5)	27 (3)
	$k_0 = 2$	54 (6)	33 (3)	53 (5)	33 (3)
	$k_0 = 3$	70 (8)	39 (4)	69 (7)	38 (3)
	$k_0 = 4$	85 (10)	43 (5)	85 (8)	43 (3)
200	$k_0 = 1$	40 (6)	28 (3)	40 (5)	28 (3)
	$k_0 = 2$	58 (7)	35 (3)	58 (6)	35 (3)
	$k_0 = 3$	74 (8)	40 (4)	74 (6)	40 (3)
	$k_0 = 4$	90 (11)	46 (5)	88 (7)	45 (3)
400	$k_0 = 1$	43 (5)	30 (3)	42 (4)	30 (3)
	$k_0 = 2$	60 (6)	36 (3)	60 (5)	36 (3)
	$k_0 = 3$	77 (8)	42 (4)	76 (6)	42 (3)
	$k_0 = 4$	95 (14)	48 (7)	93 (7)	46 (3)
800	$k_0 = 1$	44 (4)	31 (2)	44 (4)	31 (2)
	$k_0 = 2$	63 (5)	37 (3)	62 (5)	37 (2)
	$k_0 = 3$	81 (9)	43 (5)	80 (6)	43 (2)
	$k_0 = 4$	98 (14)	49 (7)	96 (7)	47 (2)

accuracy loss in estimating A due to 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 take the spectral norm of A to be 0·8. Furthermore, we now let ε_t be independent and $N(0, \Sigma_\varepsilon)$, where $\Sigma_\varepsilon = BB^{\mathrm{T}}$ and $B = (b_{ij})$, with $b_{11} = 1$ and $b_{ij} = 0{\cdot}8I(|i-j| = 1) + 0{\cdot}6I(i = j)$ for $i > 1$ or $j > 1$. Table 4 lists the average estimation errors and standard deviations over 100 replications, measured by the matrix L_1-norm. We also report Monte Carlo results for a thresholded estimator and the sample covariance estimator. For the banded estimator, we chose r to minimize the bootstrap loss defined in (13) 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 adapts directly to the underlying structure; the sample covariance estimator performs much worse than both the banding and the threshold methods.

5·2. *Real-data examples*

First we consider weekly temperature data across 71 cities in China collected from 1 January 1990 to 17 December 2000, a dataset with $p = 71$ and $n = 572$. Temperature time series exhibit strong seasonal behaviour with a period of 52 weeks. We therefore set the seasonal period to 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. Figure 1 displays the three component series of y_t for cities Ha'erbin, Shanghai and Nanjing.

We might naturally order the 71 cities according to their geographic locations. However, the method of ordering is not unique. For example, we could order the cities from north to south,

Table 4. *Means and corresponding standard deviations (in parentheses) of the errors in estimating auto-covariance matrices in a simulation study with* $n = 200$ *and* 100 *replications*

	Banding	$\Vert\hat{\Sigma}_{n,0} - \Sigma_0\Vert_1$ Thresholding	Sample	Banding	$\Vert\hat{\Sigma}_{n,1} - \Sigma_1\Vert_1$ 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)

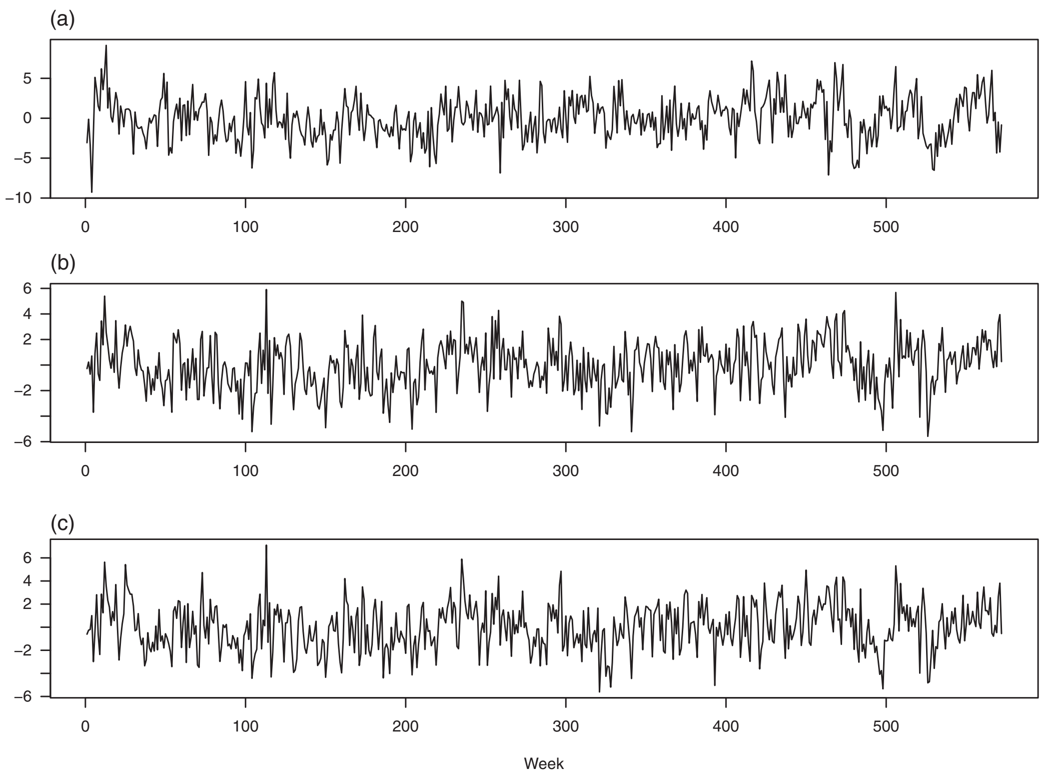

Fig. 1. Deseasonalized weekly temperature in degrees Celsius (°C) from January 1990 to December 2000, for (a) Ha'erbin, (b) Shanghai and (c) Nanjing.

from west to east, from northwest to southeast, or from southwest to northeast. Setting $d = 1$, each ordering would lead to a different banded autoregressive model of order 1. We compare these 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 optimal model, we compute (10) for these four orderings. The numerical results and the selected bandwidth parameters $\hat{k}$ are reported in Table 5. Three out

Table 5. *Results of the first real-data example: estimated bandwidth parameters, Bayesian information criterion values and average one-step-ahead and two-step-ahead post-sample predictive errors for the temperature data of* 71 *Chinese cities, with corresponding standard errors in parentheses*

Ordering	$\hat{k}$	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)

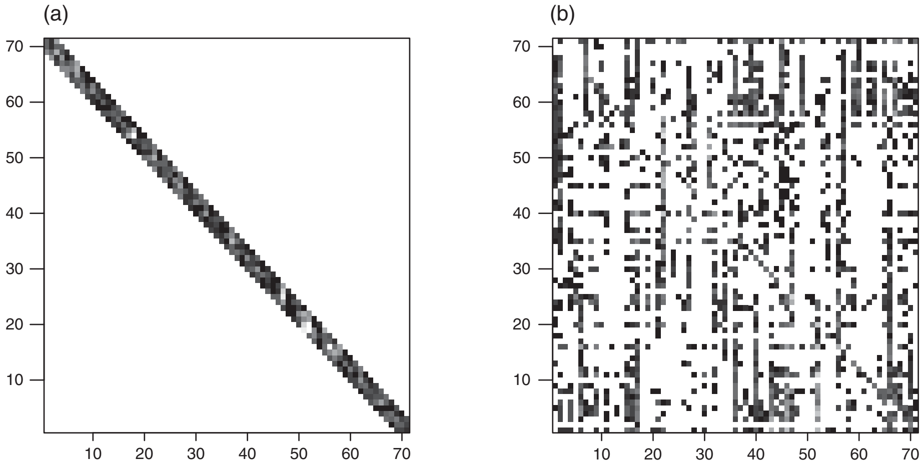

Fig. 2. Results of the first real-data example: (a) estimated banded coefficient matrix $\hat{A}$ for the model based on ordering from southwest to northeast; (b) estimated sparse coefficient matrix $\tilde{A}$ by lasso. White points represent zero entries, and grey or black points represent nonzero entries; the larger the absolute value of a coefficient, the darker the shade.

of the four models selected $\hat{k} = 2$, while the model based on ordering from west to east picks $\hat{k} = 4$. Overall, the model based on ordering from southwest to northeast is preferred; it also has the minimum one-step-ahead post-sample predictive errors. The performances of the four models in terms of prediction are very close.

Also included in Table 5 are the post-sample predictive errors of the sparse autoregressive model of order 1 obtained via lasso by minimizing

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

where $\lambda_1, \ldots, \lambda_p$ are tuning parameters estimated by five-fold crossvalidation as in Bickel & Levina (2008). The prediction accuracy of the sparse model via lasso is comparable to the predication accuracies of the banded autoregressive models, albeit slightly worse, especially for two-step-ahead prediction. However, the lack of any structure in the estimated sparse coefficient matrix $\tilde{A}$, displayed in Fig. 2(b), makes such fits difficult to interpret. In contrast, the banded coefficient matrix, depicted in Fig. 2(a), is attractive.

Fig. 3. Location plot of 21 provinces and some province-level municipalities in China; for example, Shanghai is a province-level municipality, and Ha'erbin, Hangzhou and Nanjing are the capitals of Heilongjiang, Zhejiang and Jiangsu provinces, respectively.

Table 6. *Results of the second real-data example: estimated bandwidth parameters, Bayesian information criterion values and average one-step-ahead and two-step-ahead post-sample predictive errors for clothing sales in* 21 *Chinese provinces, with corresponding standard errors in parentheses*

Ordering	$\hat{k}$	BIC	One-step ahead	Two-step ahead
North to south	4	114·9	0·314 (0·377)	0·407 (0·386)
West to east	7	115·2	0·323 (0·363)	0·409 (0·386)
Northwest to southeast	12	115·2	0·322 (0·361)	0·409 (0·395)
Southwest to northeast	5	115·1	0·316 (0·374)	0·407 (0·385)
Distance to Heilongjiang	3	114·7	0·313 (0·378)	0·407 (0·386)
Lasso	—	—	0·322 (0·362)	0·410 (0·393)

As a second example, we consider the daily sales of a clothing brand in 21 provinces of China from 1 January 2008 to 9 December 2012, a dataset with $n = 1812$ and $p = 21$. The map in Fig. 3 shows the relative geographical positions of the 21 provinces along with some province-level municipalities. We first subtract from each of the 21 series its mean. As in the previous example, we order the 21 provinces according to the four different geographic orientations, and fit a banded autoregressive model of order 1 to each ordering. The selected bandwidth parameters, the values according to (10) 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 from Heilongjiang, the most northeastern province; see Fig. 3. This results in a different ordering from the north-to-south ordering. Table 6 indicates that the minimum bandwidth parameter $\hat{k}$ is 3, attained by the ordering based on distances to Heilongjiang, followed by $\hat{k} = 4$, which is attained

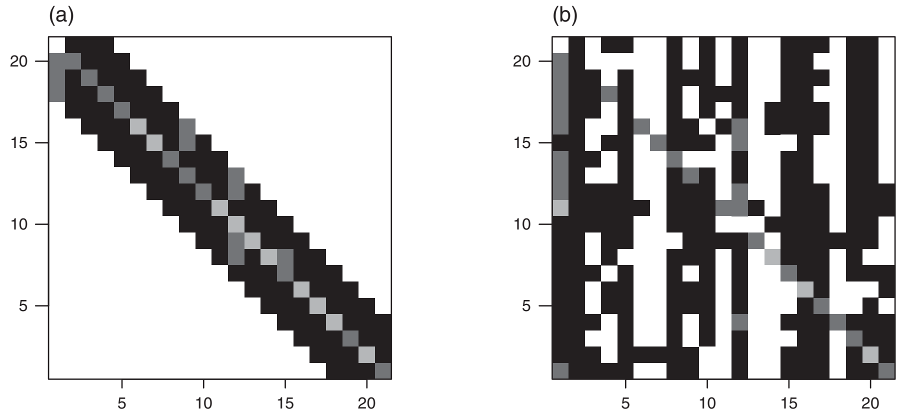

Fig. 4. Results of the second real-data example: (a) estimated banded coefficient matrix $\hat{A}$ for the model based on ordering by distances to Heilongjiang; (b) estimated sparse coefficient matrix $\tilde{A}$ by lasso. White points represent zero entries, and grey or black points represent nonzero entries; the larger the absolute value of a coefficient, the darker the shade.

by the north-to-south ordering. The post-sample prediction performances of these two models are almost the same, and are better than the performances of the other three banded models and the sparse autoregressive model.

The ordering based on the northwest-to-southeast direction leads to $\hat{k} = 12$. Therefore the corresponding banded model has 21 regressors for some components according to (3), i.e., no banded structure is observed in this case. Figure 3 indicates that the ordering from northwest to southeast groups some provinces which are quite far away from each other. Hence this is certainly a wrong ordering as far as the banded autoregressive structure is concerned.

The estimated coefficient matrix $\hat{A}$ for the banded vector autoregressive model of order 1 based on distances to Heilongjiang and the estimated $\tilde{A}$ by the lasso for the autoregressive model of order 1 are plotted in Fig. 4. The banded model facilitates interpretation; for example, sales in neighbouring provinces are closely associated with each other. Lasso fitting cannot reveal this phenomenon.

Acknowledgement

We are grateful to the editor, the associate editor and two referees for their insightful comments, which have led to significant improvement of our article. This research was supported in part by the Natural National Science Foundation of China, the U.S. National Science Foundation, and the U.K. Engineering and Physical Sciences Research Council. This paper was completed while Shaojun Guo was a Research Fellow at the London School of Economics and an Assistant Professor at the Chinese Academy of Sciences.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of Theorems 1–4, the consistency of the generalized Bayesian information criterion defined by (9) in § 2·3, and the consistency of the marginal Bayesian information criterion in the limit as $k_0 \to \infty$.

References

Basu, S. & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43**, 1535–67.

Basu, S., Shojaie, A. & Michailidis, G. (2015). Network Granger causality with inherent grouping structure. *J. Mach. Learn. Res.* **16**, 417–53.

Bickel, P. J. & 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**, 711–28.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Bolstad, A., Van Veen, B. D. & Nowak, R. (2011). Causal network inference via group sparse regularization. *IEEE Trans. Sig. Proces.* **59**, 2628–40.

Chen, X., Xu, M. & Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41**, 2994–3021.

Han, F. & Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.* **16**, 3115–50.

Haufe, S., Nolte, G., Müller, K. R. & Krämer, N. (2010). Sparse causal discovery in multivariate time series. *J. Mach. Learn. Res. Workshop Conf. Proc.* **6**, 97–106.

Hsu, N. J., Hung, H. L. & Chang, Y. M. (2008). Subset selection for vector autoregressive processes using lasso. *Comp. Statist. Data. Anal.* **52**, 3645–57.

Kock, A. & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econ.* **186**, 325–44.

Leng, C. & Li, B. (2011). Forward adaptive banding for estimating large covariance matrices. *Biometrika* **98**, 821–30.

Luo, S. & 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*. New York: Springer.

se Can, A. & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *J. Real Estate Finan. Econ.* **14**, 203–22.

Shojaie, A. & Michailidis, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **26**, 517–23.

Wu, W. B. & Pourahmadi, M. (2009). Banding sample covariance matrices of stationary processes. *Statist. Sinica* **19**, 1755–68.

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