Biometrika (2008), **95**, 2, *pp*. 365–379 doi: 10.1093/biomet/asn009

Printed in Great Britain

Modelling multiple time series via common factors

BY JIAZHU PAN

Department of Statistics and Modelling Science, University of Strathclyde, Livingstone Tower, Richmond Street, Glasgow G1 1XH, U.K.
jiazhu@stams.strath.ac.uk

AND QIWEI YAO

Department of Statistics, The London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.
q.yao@lse.ac.uk

SUMMARY

We propose a new method for estimating common factors of multiple time series. One distinctive feature of the new approach is that it is applicable to some nonstationary time series. The unobservable, nonstationary factors are identified by expanding the white noise space step by step, thereby solving a high-dimensional optimization problem by several low-dimensional subproblems. Asymptotic properties of the estimation are investigated. The proposed methodology is illustrated with both simulated and real datasets.

Some key words: Cross-correlation function; Dimension reduction; Factor model; Multivariate time series; Nonstationarity; Portmanteau test; White noise.

1. INTRODUCTION

An important problem in modelling multivariate time series is to reduce the number of parameters involved. For example, a vector autoregressive and moving average model (VARMA) with moderately large order (p, q) is viable in practice only if a parsimonious representation is identified, by imposing constraints on the coefficient matrices; see Tiao & Tsay (1989), Reinsel (1997) and references therein. An alternative strategy is to reduce the dimensionality. Attempts along this line include, among others, approaches based on principal component analysis (Priestley et al., 1974; Brillinger, 1981; Stock & Watson, 2002), canonical correlation analysis (Box & Tiao, 1977; Geweke, 1977; Geweke & Singleton, 1981; Tiao & Tsay, 1989; Anderson, 2002), reduced rank regression (Ahn, 1997; Reinsel & Velu, 1998) and factor models (Engle & Watson, 1981; Peña & Box, 1987; Forni et al., 2000; Bai & Ng, 2002).

In this paper, we revisit factor models. Although the form of the model concerned is the same as that in, for example, Peña & Box (1987), our approach is novel in three respects. First, we allow factors to be nonstationary and the nonstationarity is not necessarily driven by unit roots. The latter was investigated in the context of factor models by, for example, Ahn (1997) and Peña & Poncela (2006). Secondly, our estimation method is new and it identifies the unobserved factors by expanding the white noise space step by step, thereby solving a high-dimensional optimization problem by several low-dimensional subproblems. Finally, we allow dependence between the factors and the white noise in the model.

We do not impose distributional assumptions in the model. Instead we use the portmanteau test to identify the white noise space. The key assumption in the theoretical exploration is that the sample cross-covariance functions converge in probability to constant limits; see Condition 1 in § 3. This may be implied by the ergodicity of stationary processes, and may also be fulfilled for some nonstationary mixing processes, purely deterministic trends and random walks; see Remark 2 in § 3.

2. MODELS AND METHODOLOGY

2·1. *Factor models*

Let $\{Y_t\}$ be a $d \times 1$ time series satisfying

$$Y_t = AX_t + \varepsilon_t, \tag{1}$$

where X_t is an $r \times 1$ time series with finite second moments, $r \leqslant d$ is unknown, A is a $d \times r$ unknown constant matrix, and $\{\varepsilon_t\}$ is a sequence of vector white noise processes with mean μ_ε and covariance matrix Σ_ε; that is, ε_t and ε_s are uncorrelated for any $t \neq s$. Furthermore, we assume that there exists no linear combination of X_t, which is a white noise process; otherwise such a linear combination should be part of ε_t. We only observe $Y_1, \ldots, Y_n$ from model (1). To simplify the presentation, we assume that

$$S_0 \equiv n^{-1} \sum_{t=1}^{n} (Y_t - \bar{Y})(Y_t - \bar{Y})^{\mathrm{T}} = I_d, \tag{2}$$

where $\bar{Y} = n^{-1} \sum_{1 \leqslant t \leqslant n} Y_t$, and I_d denotes the $d \times d$ identity matrix. In practice this amounts to replacing Y_t by $S_0^{-1/2} Y_t$ before the analysis.

The component variables of the unobserved X_t are called the factors, and A is called the factor loading matrix. We may assume that the rank of A is r; otherwise (1) may be expressed equivalently in terms of a smaller number of factors. Model (1) is unchanged if we replace (A, X_t) by $(AH, H^{-1}X_t)$ for any invertible $r \times r$ matrix H, so we may assume that the column vectors of $A = (a_1, \ldots, a_r)$ are orthonormal:

$$A^{\mathrm{T}} A = I_r. \tag{3}$$

Even with the constraint (3), A and X_t are not uniquely determined in (1), as the aforementioned replacement is still applicable for any orthogonal H. However, the linear space spanned by the columns of A, denoted by $\mathcal{M}(A)$ and called the factor loading space, is a uniquely defined r-dimensional subspace in $\mathcal{R}^d$.

Model (1) has been studied by Peña & Box (1987) who assume that ε_t and X_{t+k} are uncorrelated for any integers t and k, and Y_t is stationary. Under those conditions, the number of factors r is the maximum rank of the autocovariance matrices of Y_t over all nonzero lags. Furthermore, both A and r may be estimated via standard eigenanalysis; see Peña & Box (1987).

2·2. *Estimation of A and r*

Our goal is to estimate $\mathcal{M}(A)$, or its orthogonal complement $\mathcal{M}(B)$, where $B = (b_1, \ldots, b_{d-r})$ is a $d \times (d-r)$ matrix for which (A, B) forms a $d \times d$ orthogonal matrix, i.e. $B^{\mathrm{T}} A = 0$ and $B^{\mathrm{T}} B = I_{d-r}$; see also (3). It follows from (1) that

$$B^{\mathrm{T}} Y_t = B^{\mathrm{T}} \varepsilon_t, \tag{4}$$

and hence $\{B^{\mathrm{T}}Y_t,\ t = 0, \pm 1, \ldots\}$ is a $(d-r) \times 1$ white noise process. Therefore,

$$\operatorname{corr}(b_i^{\mathrm{T}} Y_t, b_j^{\mathrm{T}} Y_{t-k}) = 0, \tag{5}$$

for any $i, j = 1, \ldots, d-r$ and $k = 1, \ldots, p$, where $p \geqslant 1$ is an arbitrary integer. Under assumption (2), $b_i^{\mathrm{T}} S_k b_j$ is the sample correlation coefficient between $b_i^{\mathrm{T}} Y_t$ and $b_j^{\mathrm{T}} Y_{t-k}$, where

$$S_k = n^{-1} \sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})^{\mathrm{T}}. \tag{6}$$

This suggests that we may estimate B by minimizing

$$\Psi_n(B) \equiv \sum_{k=1}^{p} ||B^{\mathrm{T}} S_k B||^2 = \sum_{k=1}^{p} \sum_{1 \leqslant i,j \leqslant d-r} \rho_k(b_i, b_j)^2, \tag{7}$$

where the matrix norm $||H||$ is defined as $\{\operatorname{tr}(H^{\mathrm{T}}H)\}^{1/2}$, and $\rho_k(b, a) = b^{\mathrm{T}} S_k a$.

Minimizing (7) leads to a constrained optimization problem with $d \times (d-r)$ variables. Furthermore, r is unknown. Below we present a stepwise expansion algorithm for estimating the columns of B as well as the the number of columns r. Put

$$\psi(b) = \sum_{k=1}^{p} \rho_k(b, b)^2, \quad \psi_m(b) = \sum_{k=1}^{p} \sum_{i=1}^{m-1} \{\rho_k(b, \hat{b}_i)^2 + \rho_k(\hat{b}_i, b)^2\},$$

and let $\alpha \in (0, 1)$ be the level of significance tests.

Step 1. Let $\hat{b}_1$ be a unit vector that minimizes $\psi(b)$. Compute the Ljung–Box–Pierce portmanteau test statistic:

$$L_{p,1} = n(n+2) \sum_{k=1}^{p} \frac{\rho_k(\hat{b}_1, \hat{b}_1)^2}{n-k}. \tag{8}$$

Terminate the algorithm with $\hat{r} = d$ and $\hat{B} = 0$ if $L_{p,1}$ is greater than the upper α-point of the χ_p^2 distribution. Otherwise proceed to Step 2.

Step 2. For $m = 2, \ldots, d$, let $\hat{b}_m$ minimize $\psi(b) + \psi_m(b)$ subject to the constraints

$$||b|| = 1, \quad b^{\mathrm{T}} \hat{b}_i = 0 \quad \text{for } i = 1, \ldots, m-1. \tag{9}$$

Terminate the algorithm with $\hat{r} = d - m + 1$ and $\hat{B} = (\hat{b}_1, \ldots, \hat{b}_{m-1})$ if

$$L_{p,m} \equiv n^2 \sum_{k=1}^{p} \frac{1}{n-k} \left[\rho_k(\hat{b}_m, \hat{b}_m)^2 + \sum_{j=1}^{m-1} \{\rho_k(\hat{b}_m, \hat{b}_j)^2 + \rho_k(\hat{b}_j, \hat{b}_m)^2\} \right] \tag{10}$$

is greater than the upper α-point of the $\chi^2_{p(2m-1)}$ distribution (Reinsel, 1997, pp. 149–50).

Step 3. If $L_{p,m}$ never exceeds the critical value for all $1 \leqslant m \leqslant d$, let $\hat{r} = 0$ and $\hat{B} = I_d$.

Remark 1. (i) The algorithm increases the dimension of $\mathcal{M}(B)$ by 1 each time until a newly selected direction $\hat{b}_m$ does not lead to a white noise process. Note that (9) ensures that all those $\hat{b}_j$ are orthogonal to each other.

(ii) The minimization problem in Step 2 is d-dimensional subject to constraint (9). It may be reduced to an unconstrained optimization problem with $d - m$ free variables. Note that the vector

b satisfying (9) is of the form

$$b = D_m u, \tag{11}$$

where u is any $(d-m+1)\times 1$ unit vector, D_m is a $d\times(d-m+1)$ matrix with columns given by the $(d-m+1)$ orthonormal eigenvectors of the matrix $I_d - B_{m-1}B_{m-1}^{\mathrm{T}}$ corresponding to the $(d-m+1)$-fold eigenvalue 1, where $B_m = (\hat{b}_1, \ldots, \hat{b}_m)$. Also note that any $k\times 1$ unit vector is of the form $u^{\mathrm{T}} = (u_1, \ldots, u_k)$, where

$$u_1 = \prod_{j=1}^{k-1}\cos\theta_j, \quad u_i = \sin\theta_{i-1}\prod_{j=i}^{k-1}\cos\theta_j \ \ (i=2,\ldots,k-1), \quad u_k = \sin\theta_{k-1}.$$

In the above expressions, $\theta_1, \ldots, \theta_{k-1}$ are $(k-1)$ free parameters.

(iii) Note that $\hat{B}^{\mathrm{T}}\hat{B} = I_{d-\hat{r}}$. We may let the columns of $\hat{A}$ be the $\hat{r}$ orthonormal eigenvectors of $I_d - \hat{B}\hat{B}^{\mathrm{T}}$ corresponding to the common eigenvalue 1. It holds that $\hat{A}^{\mathrm{T}}\hat{A} = I_{\hat{r}}$.

(iv) The multivariate portmanteau test statistic $L_{p,m}$ given in (10) has a normalized constant n^2, which is different from $n(n+2)$ used in the univariate case (8). For the univariate case, the modified constant $n(n+2)$ was suggested to improve the finite-sample accuracy; see Ljung & Box (1978). For multivariate cases, a different suggestion, proposed by Li & McLeod (1981), uses

$$L^*_{p,m} = L_{p,m} + \frac{p(p+1)(2m-1)}{2n} \tag{12}$$

instead of $L_{p,m}$ as the test statistic. Our numerical experiment indicates that both $L_{p,m}$ and $L^*_{p,m}$ work reasonably well with moderately large sample sizes, unless $d \gg r$. For the latter cases, both $L_{p,m}$ and $L^*_{p,m}$ may lead to substantially overestimated r. In our context, an obvious alternative is to use a more stable univariate version,

$$L'_{p,m} = n(n+2)\sum_{k=1}^{p}\frac{\rho_k(\hat{b}_m, \hat{b}_m)^2}{n-k}, \tag{13}$$

instead of $L_{p,m}$ in Step 2. Then the critical value of the test is the upper α-point of the χ^2_p distribution.

(v) Although we do not require the processes $\{Y_t\}$ and $\{X_t\}$ to be stationary, our method rests on the fact that there is no autocorrelation in the white noise process $\{\varepsilon_t\}$. Furthermore, the asymptotic χ^2 distributions of the portmanteau tests used in determining r typically rely on the assumption that $\{\varepsilon_t\}$ be independent and identically distributed. An early investigation of these tests in more general settings is given by Francq et al. (2005).

(vi) When Y_t is nonstationary, the sample cross-covariance function S_k is no longer a meaningful covariance measure. However, since ε_t is white noise and is stationary, $c_1^{\mathrm{T}}S_k c_2$ is the proper sample covariance between $c_1^{\mathrm{T}}Y_t$ and $c_2^{\mathrm{T}}Y_{t-k}$ for any vectors $c_1, c_2 \in \mathcal{M}(B)$. In fact our method relies on the fact that $c_1^{\mathrm{T}}S_k c_2$ is close to 0 for any $1 \leqslant k \leqslant p$. This also indicates that in practice we should not use large p as, for example, $c_1^{\mathrm{T}}S_p c_2$ is a poor estimator of $\mathrm{cov}(c_1^{\mathrm{T}}Y_t, c_2^{\mathrm{T}}Y_{t-p})$ when p is too large.

(vii) When the number of factors r is given, we may omit all the test steps, and stop the algorithm after obtaining $\hat{b}_1, \ldots, \hat{b}_r$ by solving the r optimization problems.

2·3. *Modelling with estimated factors*

Note that $\hat{A}\hat{A}^{\mathrm{T}} + \hat{B}\hat{B}^{\mathrm{T}} = I_d$. Once we have obtained $\hat{A}$, it follows from (1) that

$$Y_t = \hat{A}\xi_t + e_t, \tag{14}$$

where

$$\xi_t = \hat{A}^{\mathrm{T}} Y_t = \hat{A}^{\mathrm{T}} A X_t + \hat{A}^{\mathrm{T}} \varepsilon_t, \qquad e_t = \hat{B}\hat{B}^{\mathrm{T}} Y_t. \tag{15}$$

We treat e_t as a white noise process, and estimate $\mathrm{var}(e_t)$ by the sample variance of $\hat{B}\hat{B}^{\mathrm{T}} Y_t$.

We model the lower-dimensional process ξ_t by VARMA or state-space models. As we have pointed out, $\hat{A}$ may be replaced by $\hat{A}H$ for any orthogonal H. We may choose $\hat{A}$ appropriately such that ξ_t admits a simple model; see, for example, Tiao & Tsay (1989). Alternatively, we may apply principal components analysis to the factors; see Example 3 in § 4. Note that there is no need to update $\hat{B}$ now since $\mathcal{M}(\hat{A}H) = \mathcal{M}(\hat{A})$, which is the orthogonal complement of $\hat{B}$.

3. Theoretical properties

The factor loading matrix A is only identifiable up to $\mathcal{M}(A)$, a linear space spanned by its columns. We are effectively concerned with the estimation for the factor loading space $\mathcal{M}(A)$ rather than A itself. To make our statements clearer, we first introduce some notation.

For $r < d$, let $\mathcal{H}$ be the set consisting of all $d \times (d-r)$ matrices H satisfying the condition $H^{\mathrm{T}}H = I_{d-r}$. For $H_1, H_2 \in \mathcal{H}$, define

$$D(H_1, H_2) = \left\| \left(I_d - H_1 H_1^{\mathrm{T}}\right) H_2 \right\| = \left\{ d - r - \mathrm{tr}\left(H_1 H_1^{\mathrm{T}} H_2 H_2^{\mathrm{T}}\right) \right\}^{1/2}. \tag{16}$$

Note that $H_1 H_1^{\mathrm{T}}$ is the projection matrix on to the linear space $\mathcal{M}(H_1)$, and $D(H_1, H_2) = 0$ if and only if $\mathcal{M}(H_1) = \mathcal{M}(H_2)$. Therefore, $\mathcal{H}$ may be partitioned into the equivalent classes by D as follows: the D-distance between any two elements in each equivalent class is 0, and the D-distance between any two elements from two different classes is positive. Denote by $\mathcal{H}_D = \mathcal{H}/D$ the quotient space consisting of all those equivalent classes; that is, we treat H_1 and H_2 as the same element in $\mathcal{H}_D$ if and only if $D(H_1, H_2) = 0$. Then $(\mathcal{H}_D, D)$ forms a metric space in the sense that D is a well-defined distance measure on $\mathcal{H}_D$; see Lemma A1(i) in the Appendix. Furthermore, the functions $\Psi_n(\cdot)$, defined in (7), and

$$\Psi(H) \equiv \sum_{k=1}^{p} \| H^{\mathrm{T}} \Sigma_k H \|^2, \tag{17}$$

are well defined on $\mathcal{H}_D$; see Lemma A1(ii) in the Appendix. In the above expression, the Σ_k are given in Condition 1 below.

We only consider the asymptotic properties for the estimation of the factor loading space with r known; it remains open how to establish the theoretical properties when r is unknown. Then the estimator for B may be defined as

$$\hat{B} = \arg\min_{H \in \mathcal{H}} \Psi_n(H). \tag{18}$$

We need the following regularity conditions.

Condition 1. As $n \to \infty$, $S_k \to \Sigma_k$ in probability for $k = 0, 1, \ldots, p$, where Σ_k are nonnegative definite matrices, and $\Sigma_0 = I_d$.

Condition 2. The matrix B is the unique minimizer of $\Psi(\cdot)$ in the space $\mathcal{H}_D$; that is, $\Psi(\cdot)$ reaches its minimum value at B' if and only if $D(B', B) = 0$, where B is specified at the beginning of § 2·2.

Condition 3. There exist constants $a > 0, c > 0$ for which $\Psi(H) - \Psi(B) \geqslant a\{D(H, B)\}^c$ for any $H \in \mathcal{H}$.

Remark 2. (i) Condition 1 does not require that the process Y_t be stationary. In fact it may hold when $ES_k \to \Sigma_k$ and Y_t is φ-mixing in the sense that $\varphi(m) \to 0$ as $m \to \infty$, where

$$\varphi(m) = \sup_{k \geqslant 1} \sup_{U \in \mathcal{F}_{-\infty}^{k},\, V \in \mathcal{F}_{m+k}^{\infty},\, pr(U)>0} |\mathrm{pr}(V|U) - \mathrm{pr}(V)|, \tag{19}$$

and $\mathcal{F}_i^j = \sigma(Y_i, \ldots, Y_j)$; see Lemma A2 in the Appendix. It also gives a sufficient condition, which ensures that the convergence in Condition 1 is almost sure. Examples of nonstationary φ-mixing processes include, among others, stationary φ-mixing processes plus nonconstant treads, and standardized random walks such as $Y_t = Y_{t-1} + n^{-1/2}\varepsilon_t$, $t = 1, \ldots, n$, where $Y_0 \equiv 0$ and ε_t are independent and identically distributed with, for example, $E(\varepsilon_t^2) < \infty$. Condition 1 may also hold for some purely deterministic processes such as a linear trend $Y_t = t/n,\ t = 1, \ldots, n$.

(ii) Under model (1), $\Psi(B) = 0$. Condition 2 implies that $\Psi(C) \neq 0$ for any $C \in \mathcal{H}$ and $\mathcal{M}(C) \cap \mathcal{M}(A)$ is not an empty set.

THEOREM 1. *Under Conditions* 1 *and* 2, $D(\hat{B}, B) \to 0$ *in probability as* $n \to \infty$. *Furthermore,* $D(\hat{B}, B) \to 0$ *almost surely if the convergence in Condition* 1 *is also almost sure.*

THEOREM 2. *Let* $n^{1/2}(ES_k - \Sigma_k) = O(1)$, *and let* Y_t *be* φ*-mixing with* $\varphi(m) = O(m^{-\lambda})$ *for* $\lambda > p/(p-2)$ *and* $\sup_{t \geqslant 1} E\|Y_t\|^p < \infty$ *for some* $p > 2$. *Then*

$$\sup_{H \in \mathcal{H}} |\Psi_n(H) - \Psi(H)| = O_P\big(n^{-1/2}\big).$$

If, in addition, Condition 3 *also holds,* $D(\hat{B}, B) = O_P(n^{-1/(2c)})$.

Theorems 1 and 2 do not require Y_t to be a stationary process. Their proofs are given in the Appendix.

4. NUMERICAL PROPERTIES

We illustrate the methodology proposed in § 2 with two simulated examples, one stationary and one nonstationary, and one real dataset. The numerical optimization was solved using the downhill simplex method; see § 10·4 of Press et al. (1992). In the simulated examples, we set the significance level at 5% for the portmanteau tests used in our algorithm, and $p = 15$ in (8). The results with $p = 5$, 10 and 20 show similar patterns and, therefore, are not reported. We measure the errors in estimating the factor loading space $\mathcal{M}(A)$ by

$$D_1(A, \hat{A}) = ([\mathrm{tr}\{\hat{A}^{\mathrm{T}}(I_d - AA^{\mathrm{T}})\hat{A}\} + \mathrm{tr}(\hat{B}^{\mathrm{T}}AA^{\mathrm{T}}\hat{B})]/d)^{1/2}.$$

It may be shown that $D_1(A, \hat{A}) \in [0, 1]$, and it equals 0 if and only if $\mathcal{M}(A) = \mathcal{M}(\hat{A})$, and 1 if and only if $\mathcal{M}(A) = \mathcal{M}(\hat{B})$.

Example 1. Let $Y_{ti} = X_{ti} + \varepsilon_{ti}$ for $i = 1, 2, 3$, and $Y_{ti} = \varepsilon_{ti}$ for $i = 4, \ldots, d$, where

$$X_{t1} = 0{\cdot}8X_{t-1,1} + e_{t1}, \quad X_{t2} = e_{t2} + 0{\cdot}9e_{t-1,2} + 0{\cdot}3e_{t-2,2},$$

$$X_{t3} = -0{\cdot}5X_{t-1,3} - \varepsilon_{t3} + 0{\cdot}8\varepsilon_{t-1,3},$$

and all ε_{tj} and e_{tj} are independent and standard normal. Because of the presence of ε_{t3} in the equation for X_{t3}, X_t and ε_t are mutually dependent. In this setting, the number of true factors is $r = 3$, and the factor loading matrix may be taken as $A = (I_3, 0)^{\mathrm{T}}$, where 0 denotes the $3 \times (d-3)$ matrix with all elements equal to 0. We set the sample size at $n = 300$, 600 and 1000, and the dimension of Y_t at $d = 5$, 10 and 20. For each setting, we generated 1000 samples from this

Table 1. *Relative frequencies for* $\hat{r}$ *taking different values in Example* 1. *The true value of* r *is* 3

		$\hat{r}$						
d	n	0	1	2	3	4	5	$\geqslant 6$
5	300	0·000	0·209	0·444	0·345	0·002	0·000	
	600	0·000	0·071	0·286	0·633	0·010	0·000	
	1000	0·000	0·004	0·051	0·933	0·120	0·000	
10	300	0·000	0·219	0·524	0·255	0·002	0·000	0·000
	600	0·000	0·049	0·290	0·649	0·012	0·000	0·000
	1000	0·000	0·007	0·062	0·898	0·033	0·000	0·000
20	300	0·000	0·162	0·543	0·285	0·010	0·000	0·000
	600	0·000	0·033	0·305	0·609	0·053	0·000	0·000
	1000	0·000	0·004	0·066	0·822	0·103	0·005	0·000

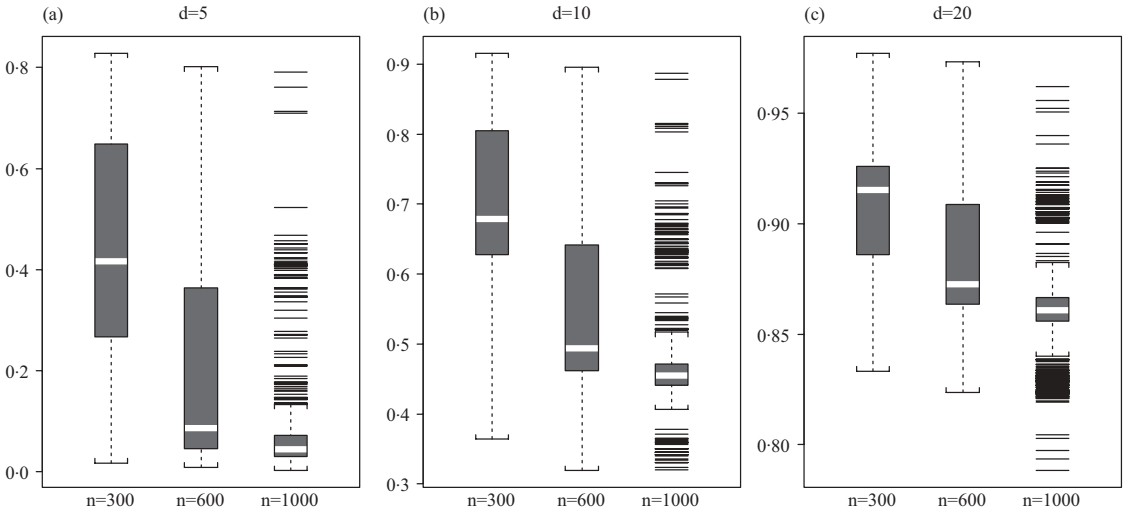

Fig. 1. Example 1. Boxplots of $D_1(A, \hat{A})$ for (a) $d = 5$, (b) $d = 10$, (c) $d = 20$.

model. The relative frequencies for $\hat{r}$ taking different values are reported in Table 1, which shows that, when the sample size n increases, the estimation of r becomes more accurate. We used $L'_{m,p}$ given in (13) in our simulation, since both $L_{m,p}$ and $L^*_{m,p}$ produced substantially overestimated r-values when $d = 10$ and 20. Figure 1 presents the boxplots of the errors $D_1(A, \hat{A})$. As the sample size increases, $D_1(A, \hat{A})$ decreases. Furthermore, the errors increase when d increases.

Example 2. We use the same setting as in Example 1 but with X_{t1}, X_{t2} and X_{t3} replaced by

$$\begin{aligned} X_{t1} - 2t/n &= 0{\cdot}8(X_{t-1,1} - 2t/n) + e_{t1}, \\ X_{t2} &= 3t/n, \\ X_{t3} &= X_{t-1,3} + (10/n)^{1/2} e_{t3} \quad \text{with} \quad X_{0,3} \sim N(0, 1). \end{aligned} \tag{20}$$

Thus X_{t1} is an AR(1) process with nonconstant mean, X_{t2} is a purely deterministic trend, and X_{t3} is a random walk. None of them is stationary. The relative frequencies for $\hat{r}$ taking different values are reported in Table 2. The boxplots of the estimation errors $D_1(A, \hat{A})$ are depicted in Fig. 2. The general pattern observed in the stationary example, Example 1, remains. The quality of our

Table 2. *Relative frequencies for $\hat{r}$ taking different values in Example 2. The true value of r is 3*

		$\hat{r}$						
d	n	0	1	2	3	4	5	$\geqslant 6$
5	300	0·000	0·000	0·255	0·743	0·002	0·000	
	600	0·000	0·000	0·083	0·907	0·010	0·000	
	1000	0·000	0·000	0·033	0·945	0·022	0·000	
10	300	0·000	0·000	0·283	0·695	0·022	0·000	0·000
	600	0·000	0·000	0·103	0·842	0·054	0·001	0·000
	1000	0·000	0·000	0·051	0·871	0·077	0·001	0·000
20	300	0·000	0·000	0·258	0·663	0·076	0·001	0·002
	600	0·000	0·000	0·035	0·673	0·278	0·012	0·002
	1000	0·000	0·000	0·099	0·733	0·162	0·006	0·000

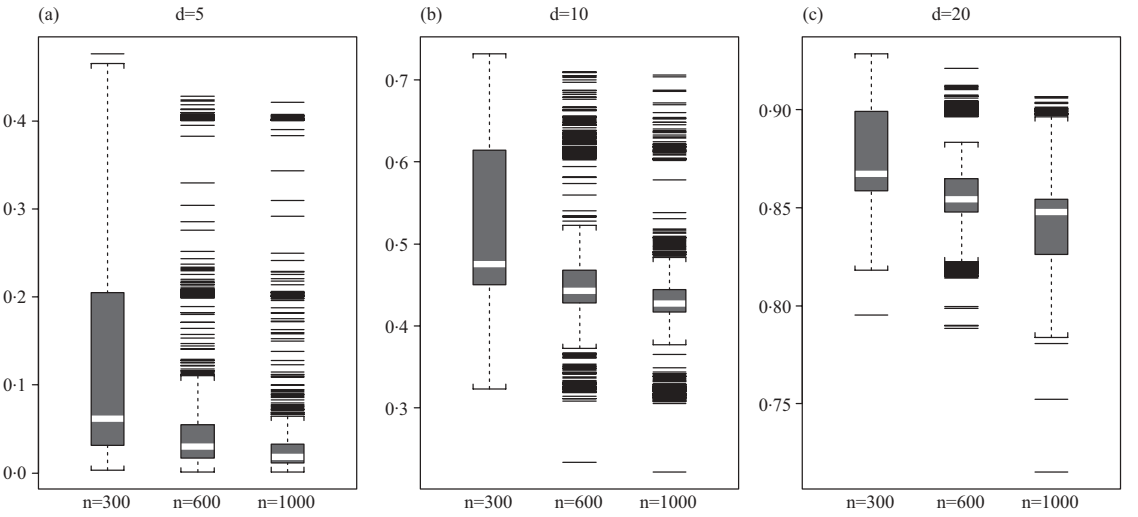

Fig. 2. Example 2. Boxplots of $D_1(A, \hat{A})$ for (a) $d = 5$, (b) $d = 10$, (c) $d = 20$.

estimation improves when sample sizes increase, because of the way in which the nonstationarity is specified in (20). For example, the sample $\{X_{t2}, t = 1, \ldots, n\}$ always consists of regular grid-points on the segment of the line $y = 3x$ between (0, 0) and (1, 3). Therefore, when n increases, we obtain more information from the same nonstationary system.

Our method rests on the simple fact that the quadratic forms of the sample cross-correlation function are close to zero along the directions perpendicular to the factor loading space, and are nonzero along the directions in the factor loading space; see Remarks 1(vi) and 2(ii). The departure from zero along the directions in the factor loading space in Example 2 is more pronounced than that in Example 1. This explains why the proposed method performs better in Example 2 than in Example 1, especially when $n = 300$ and 600.

Example 3. Figure 3 displays the monthly temperatures in cities of Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing and Hangzhou in Eastern China in January 1954–December 1986. The sample size is $n = 396$ and $d = 7$. As expected, the data show strong periodic behaviour with period 12. We fitted the data with factor models (1). With $p = 12$, the estimated number of factors is $\hat{r} = 4$. We applied principal components analysis to the estimated factors. The variances of the

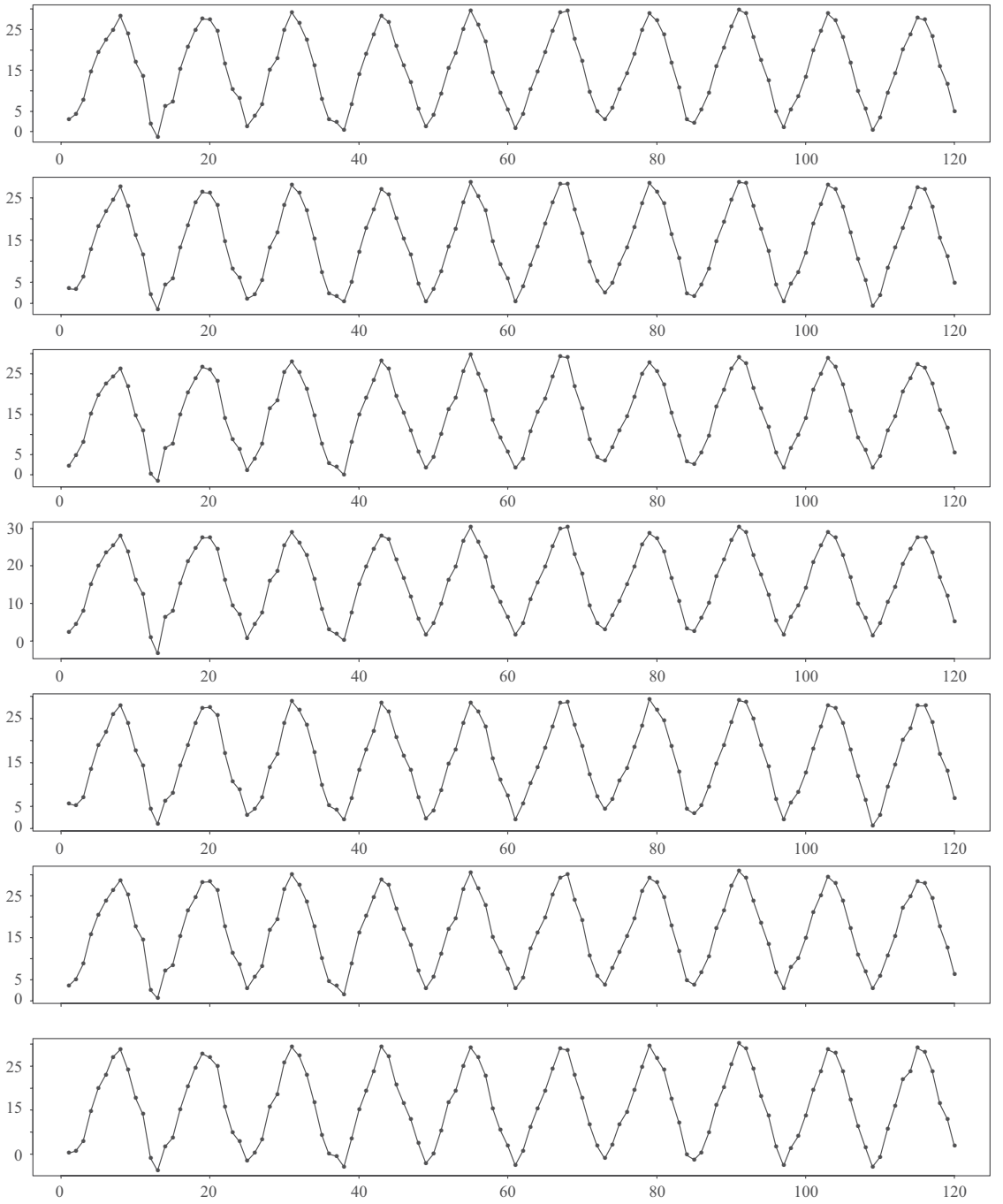

Fig. 3. Example 3. Time series plots of the monthly temperature in, from top to bottom, Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing and Hangzhou, corresponding to the first 10-year segments.

four principal component factors are, respectively, 542·08, 1·29, 0·07 and 0·06. The first factor accounts for over 99% of the total variation of the four factors, and 97·6% of the total variation of the original seven series. The first four principal component factors are plotted in Fig. 4, and their cross-correlation functions are displayed in Fig. 5. The periodic annual oscillation in the original

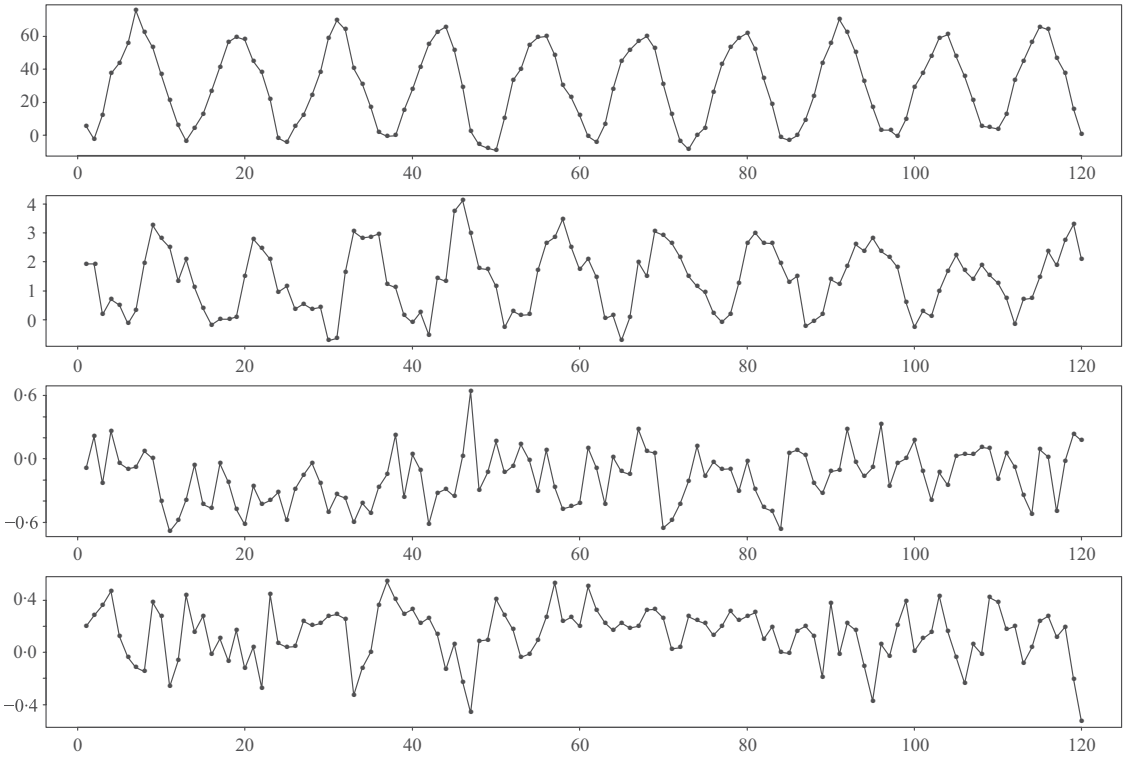

Fig. 4. Example 3. Time series plots of the first four principal component factors, corresponding to the first 10-year segments.

data is predominately reflected by the fluctuation of the first factor, and to a much less extent by that of the second factor; see Fig. 4. Furthermore, no periodic pattern is present in the third and the fourth principal component factors. This suggests that the annual temperature oscillation over this area may be seen as driven by one, or at most two, 'common factors'. The first two columns of the corresponding loading matrix $\hat{A}$ are

$$\begin{pmatrix} 0{\cdot}394 & 0{\cdot}386 & 0{\cdot}378 & 0{\cdot}387 & 0{\cdot}363 & 0{\cdot}376 & 0{\cdot}366 \\ -0{\cdot}086 & 0{\cdot}225 & -0{\cdot}640 & -0{\cdot}271 & 0{\cdot}658 & -0{\cdot}014 & 0{\cdot}164 \end{pmatrix}^{\mathrm{T}},$$

which indicates that the first principal component factor is effectively the average temperature over the seven cities. The residuals $\hat{B}^{\mathrm{T}} Y_t$ carry little dynamic information in the data; this is indicated by the cross-correlation functions, not shown. The sample mean and sample covariance matrix of e_t are, respectively,

$$\hat{\mu}_e = \begin{pmatrix} 3{\cdot}41 \\ 2{\cdot}32 \\ 4{\cdot}39 \\ 4{\cdot}30 \\ 3{\cdot}40 \\ 4{\cdot}91 \\ 4{\cdot}77 \end{pmatrix}, \qquad \hat{\Sigma}_e = \begin{pmatrix} 1{\cdot}56 & & & & & & \\ 1{\cdot}26 & 1{\cdot}05 & & & & & \\ 1{\cdot}71 & 1{\cdot}34 & 1{\cdot}91 & & & & \\ 1{\cdot}90 & 1{\cdot}49 & 2{\cdot}10 & 2{\cdot}33 & & & \\ 1{\cdot}37 & 1{\cdot}16 & 1{\cdot}46 & 1{\cdot}58 & 1{\cdot}37 & & \\ 1{\cdot}67 & 1{\cdot}26 & 1{\cdot}91 & 2{\cdot}09 & 1{\cdot}37 & 1{\cdot}97 & \\ 1{\cdot}41 & 1{\cdot}14 & 1{\cdot}58 & 1{\cdot}67 & 1{\cdot}39 & 1{\cdot}56 & 1{\cdot}53 \end{pmatrix}. \tag{21}$$

Figure 5 indicates that the first two factors are dominated by periodic components with period 12. We estimated those components simply by taking the averages of all values in each of the

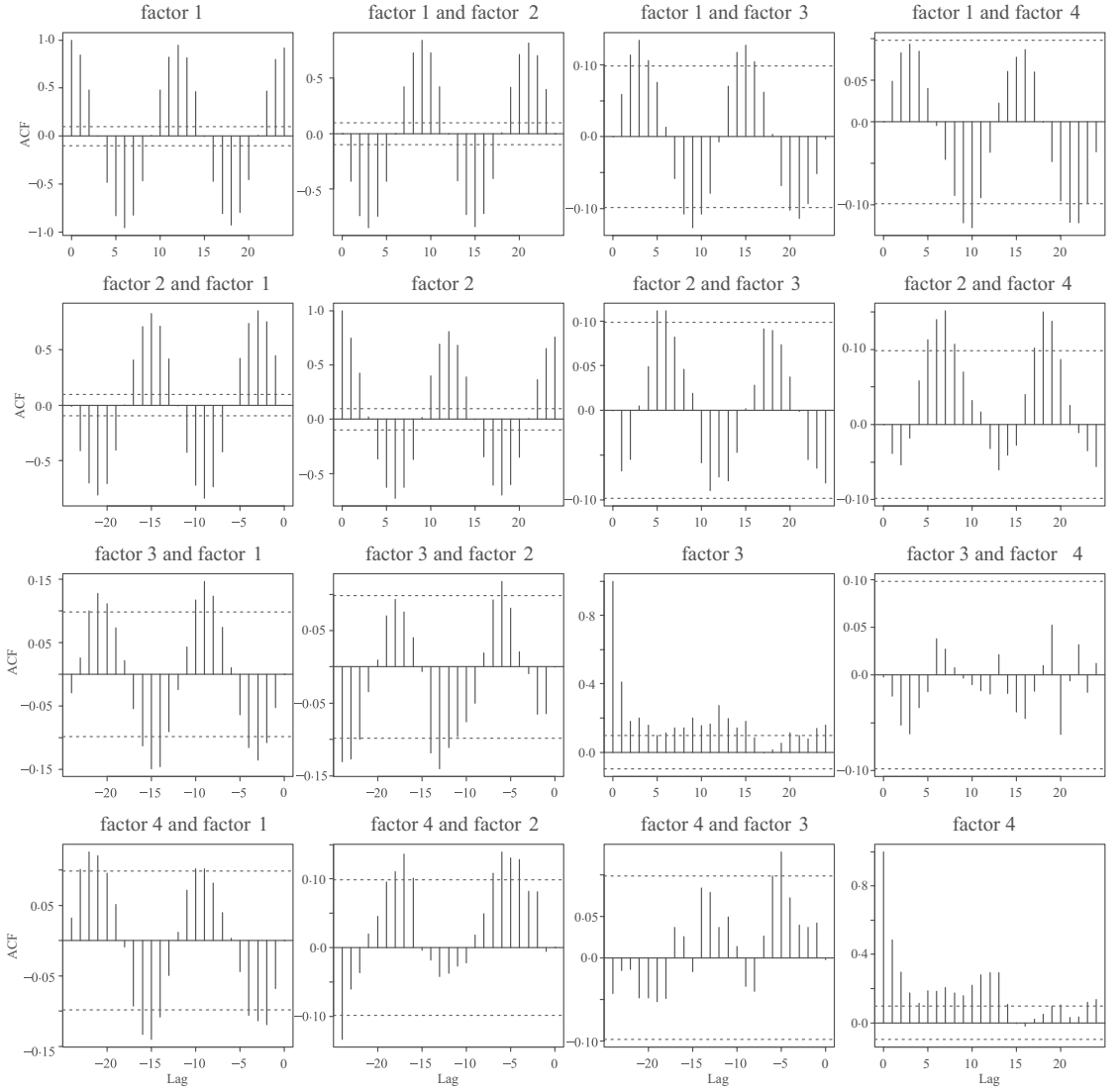

Fig. 5. Example 3. Sample cross-correlation functions of the first four principal component factors.

twelve months, leading to the estimated periodic components

$$(g_{1,1}, \ldots, g_{12,1}) = (-1{\cdot}61,\ 1{\cdot}33,\ 11{\cdot}74,\ 28{\cdot}06,\ 41{\cdot}88,\ 54{\cdot}51,\ 63{\cdot}77,\ 62{\cdot}14,\ 49{\cdot}48,\ 33{\cdot}74,\ 18{\cdot}29,\ 3{\cdot}50),$$
$$(g_{1,2}, \ldots, g_{12,2}) = (1{\cdot}67,\ 1{\cdot}21,\ 0{\cdot}47,\ 0{\cdot}17,\ 0{\cdot}41,\ 0{\cdot}48,\ 1{\cdot}37,\ 2{\cdot}13,\ 2{\cdot}98,\ 3{\cdot}05,\ 2{\cdot}78,\ 2{\cdot}22) \quad (22)$$

for, respectively, the first and the second factors. The cross-correlation functions of the four factors after removing the periodic components from the first two factors, not shown, indicate that the autocorrelation in each of those four series is not very strong, and that cross-correlations among those four series, at nonzero lag, are weak. We fitted a vector autoregressive model to those four series with the order 1 determined by the Akaike information criteria (AIC) (Brockwell

& Davis, 1991, p. 412) with the following estimated coefficients:

$$\hat{\varphi}_0 = \begin{pmatrix} 0{\cdot}07 \\ -0{\cdot}02 \\ -0{\cdot}11 \\ 0{\cdot}10 \end{pmatrix}, \quad \hat{\Phi}_1 = \begin{pmatrix} 0{\cdot}27 & -0{\cdot}31 & 0{\cdot}72 & 0{\cdot}40 \\ 0{\cdot}01 & 0{\cdot}36 & -0{\cdot}04 & 0{\cdot}04 \\ 0{\cdot}00 & -0{\cdot}01 & 0{\cdot}42 & -0{\cdot}02 \\ -0{\cdot}00 & 0{\cdot}03 & 0{\cdot}03 & 0{\cdot}48 \end{pmatrix}, \tag{23}$$

$$\hat{\Sigma}_u = \begin{pmatrix} 14{\cdot}24 & & & \\ -0{\cdot}17 & 0{\cdot}23 & & \\ -0{\cdot}02 & 0{\cdot}03 & 0{\cdot}05 & \\ 0{\cdot}04 & 0{\cdot}01 & -0{\cdot}00 & 0{\cdot}05 \end{pmatrix}. \tag{24}$$

The multivariate portmanteau tests, with lag $p = 12$, of both Li & Mcleod (1981) and Reinsel (1997, p. 149) for the residuals from the above fitted vector AR(1) model are nonsignificant at the 5% level. The univariate portmanteau test is nonsignificant at the 5% level for three out of the four component residual series, and is nonsignificant at the 1% level for the other component residual series. On the other hand, a vector AR(2) model was selected by the AIC for the 4-factor series with vector AR(1) as its closest competitor. In fact the AIC values are, respectively, 240·03, 0·11, 0·00, 6·38 and 18·76 for the autoregressive order 0, 1, 2, 3 and 4.

Overall, the fitted model for the monthly temperature vector Y_t is

$$Y_t = \hat{A}\xi_t + e_t,$$

where the white noise e_t has the mean and covariance matrix given in (21), and the 4×1 factor ξ_t follows the VAR(1) model

$$\xi_t - \alpha_t = \hat{\varphi}_0 + \hat{\Phi}_1(\xi_{t-1} - \alpha_{t-1}) + u_t,$$

in which the periodic component $\alpha_t^{\mathrm{T}} = (g_{m(t),1}, g_{m(t),2}, 0, 0)$, $g_{t,i}$ is given in (22),

$$m(t) = \{k \mid 1 \leqslant k \leqslant 12 \quad \text{and} \quad t = 12p + k \text{ for some integer } p \geqslant 0\},$$

and the white noise u_t has mean 0 and covariance matrix $\hat{\Sigma}_u$ given in (23).

ACKNOWLEDGEMENT

This project was partially supported by a grant from the U.K. Engineering and Physical Science Research Council. Jiazhu Pan was partially supported by the National Natural Science Foundation of China and the National Basic Research Program of China. The authors thank Professor Valdimir Spokoiny for helpful discussion, Mr Da Huang for making available the temperature data analysed in Example 3. Thanks also go to Professor D. M. Titterington and two referees for their helpful comments and suggestions.

APPENDIX

Proofs

We first introduce two lemmas.

LEMMA A1. (i) *It holds for any* $H_1, H_2, H_3 \in \mathcal{H}$ *that*

$$D(H_1, H_3) \leqslant D(H_1, H_2) + D(H_2, H_3).$$

(ii) *For any* $H_1, H_2 \in \mathcal{H}$, $\Psi(H_1) = \Psi(H_2)$ *and* $\Psi_n(H_1) = \Psi_n(H_2)$ *provided that* $D(H_1, H_2) = 0$.

Proof. (i) For any symmetric matrices M_1, M_2 and M_3, it follows from the standard triangle inequality for the matrix norm $\|\cdot\|$ that $\|M_1 - M_3\| \leqslant \|M_1 - M_2\| + \|M_2 - M_3\|$; that is

$$\{\mathrm{tr}(M_1^2 + M_3^2 - 2M_1M_3)\}^{1/2} \leqslant \{\mathrm{tr}(M_1^2 + M_2^2 - 2M_1M_2)\}^{1/2} + \{\mathrm{tr}(M_2^2 + M_3^2 - 2M_2M_3)\}^{1/2}. \quad \text{(A1)}$$

Let $M_1 = H_1H_1^{\mathrm{T}}$, $M_2 = H_2H_2^{\mathrm{T}}$ and $M_3 = H_3H_3^{\mathrm{T}}$. Since now $\mathrm{tr}(M_i^2) = \mathrm{tr}(M_i) = d - r$ for $i = 1, 2, 3$, the inequality required follows from (A1) and (16) directly.

(ii) Under the condition $D(H_1, H_2) = 0$, $H_1H_1^{\mathrm{T}} = H_2H_2^{\mathrm{T}}$ as it is the projection matrix onto the linear space $\mathcal{M}(H_1) = \mathcal{M}(H_2)$. Now

$$\|H_1^{\mathrm{T}}\Sigma_k H_1\|^2 = \mathrm{tr}\{(H_1^{\mathrm{T}}\Sigma_k H_1)^{\mathrm{T}} H_1^{\mathrm{T}}\Sigma_k H_1\} = \mathrm{tr}(\Sigma_k^{\mathrm{T}} H_1H_1^{\mathrm{T}}\Sigma_k H_1H_1^{\mathrm{T}}) = \|H_2^{\mathrm{T}}\Sigma_k H_2\|^2.$$

Hence $\Psi(H_1) = \Psi(H_2)$. The equality for Ψ_n may be proved in the same manner. □

LEMMA A2. *Let $\{Y_t\}$ be a φ-mixing process and let $E(S_k) \to \Sigma_k$. Suppose that Y_t can be represented as $Y_t = U_t + V_t$, where U_t and V_t are uncorrelated for each t, $\sup_{t\geqslant 1} E\|U_t\|^h < \infty$ for some constant $h > 2$, and*

$$\frac{1}{n}\sum_{t=1}^{n} V_t \to c \text{ in probability}, \qquad \frac{1}{n}\sum_{t=1}^{n} EV_t \to c, \quad \text{(A2)}$$

where c is a constant vector. It holds that

(i) $S_k \to \Sigma_k$ *in probability,*

(ii) $S_k \to \Sigma_k$ *almost surely provided that the mixing coefficients satisfy the condition*

$$\varphi(m) = \begin{cases} O\left(\frac{(m-b)}{(2b-2)} - \delta\right), & \text{if } 1 < b < 2, \\ O\left(\frac{(m-3)}{(b-\delta)}\right), & \text{if } b \geqslant 2, \end{cases} \quad \text{(A3)}$$

where $\delta > 0$ is a constant, and the convergence in condition (A2) is also almost sure.

Proof. Assertion (i) follows from the the law of large numbers for φ-mixing processes; see Theorem 8·1·1 of Lin & Lu (1997). Applying the result of Chen & Wu (1989) to the sequences $\{U_t\}$ and $\{U_tU_{t-i}^{\mathrm{T}}\}$, and using the almost sure version of Condition (A2), we may obtain (ii). □

Proof of Theorem 1. Applying the Cauchy–Schwarz inequality to the matrix norm, we have

$$\begin{aligned} |\Psi_n(H) - \Psi(H)| &\leqslant \sum_{k=1}^{p} |\|H^{\mathrm{T}}S_kH\|^2 - \|H^{\mathrm{T}}\Sigma_kH\|^2| \\ &\leqslant \sum_{k=1}^{p} \|H^{\mathrm{T}}(S_k - \Sigma_k)H\|(\|H^{\mathrm{T}}S_kH\| + \|H^{\mathrm{T}}\Sigma_kH\|) \\ &\leqslant \|H\|^4 \sum_{k=1}^{p} \|S_k - \Sigma_k\|(\|S_k\| + \|\Sigma_k\|). \end{aligned}$$

Note that $\|H\|^2 = d - r$ for any $H \in \mathcal{H}$, $\|S_k - \Sigma_k\| \to 0$ in probability, which is implied by Condition 1, and $\|S_k\| + \|\Sigma_k\| = O_P(1)$. Hence, in probability,

$$\sup_{H\in\mathcal{H}_D} |\Psi_n(H) - \Psi(H)| \to 0. \quad \text{(A4)}$$

LEMMA A1. (i) ensures that $(\mathcal{H}_D, D)$ is a well-defined metric space which is complete. Lemma A1(ii) guarantees that $\Psi_n(\cdot)$ is a well-defined stochastic process indexed by $H \in \mathcal{H}_D$, and $\Psi(\cdot)$ is a well-defined function on the metric space $(\mathcal{H}_D, D)$. It follows from the argmax theorem, see Theorem 3·2·2 and Corollary 3·2·3 of van der Vaart & Wellner (1996), that $D(\hat{B}, B) \to 0$ in probability.

To show the convergence with probability 1, note that the convergence in (A4) is with probability 1 provided that $S_k \to \Sigma_k$ with probability 1. Suppose by contradiction that there exists a δ such that $\mathrm{pr}\{\limsup_{n\to\infty} D(\hat{B}, B_0) > \delta\} > 0$. Let $\mathcal{H}'_D = \mathcal{H}_D \cap \{B : D(B, B_0) \geqslant \delta\}$. Then $\mathcal{H}'_D$ is a compact subset of $\mathcal{H}_D$. Note that, if $\sup_{H\in\mathcal{H}_D} |\Psi_n(H) - \Psi(H)| \to 0$ almost surely, then there exists a set of sample points Ω' satisfying $\Omega' \subset \{\limsup_{n\to\infty} D(\hat{B}, B_0) > \delta\}$ and $\mathrm{pr}(\Omega') > 0$ such that, for each $\omega \in \Omega'$, one can find a subsequence $\{\hat{B}_{n_k}(\omega)\} \subset \mathcal{H}'_D$ with $\hat{B}_{n_k}(\omega) \to B \in \mathcal{H}'_D$. Then, by the definition of $\hat{B}$,

$$\Psi(B) = \lim_{k\to\infty} \Psi_{n_k}\{\hat{B}_{n_k}(\omega)\} \leqslant \lim_{k\to\infty} \Psi(B_0) = \Psi(B_0)$$

holds for $\omega \in \Omega'$ and with positive probability. This is a contradiction to Condition 2. Therefore, it must hold that $D(\hat{B}, B_0) \to 0$ with probability 1. □

Proof of Theorem 2. Denote by $s_{(i,j),k}$ and $\sigma_{(i,j),k}$, respectively, the (i, j)th elements of S_k and Σ_k. By the Central Limit Theorem for φ-mixing processes, see Lin & Lu (1997) and an unpublished London School of Economics dissertation paper by J. Davidson, it holds that $n^{1/2}(s_{(i,j),k} - Es_{(i,j),k}) \to N_{(i,j),k}$ in distribution, where $N_{(i,j),k}$ denotes a Gaussian random variable, $i, j = 1, \ldots, d$. Hence, $\|n^{1/2}(S_k - ES_k)\| = O_P(1)$. It holds now that

$$\begin{aligned}
&\sup_{H\in\mathcal{H}_D} n^{1/2}|\Psi_n(H) - \Psi(H)| \\
&\quad\leqslant \sup_{H\in\mathcal{H}_D} n^{1/2} \sum_{k=1}^{p} |\|H^{\mathrm{T}} S_k H\|^2 - \|H^{\mathrm{T}} \Sigma_k H\|^2| \\
&\quad\leqslant \sup_{H\in\mathcal{H}_D} \sum_{k=1}^{p} \big\|H^{\mathrm{T}} n^{1/2}(S_k - ES_k)H\big\|(\|H^{\mathrm{T}} S_k H\| + \|H^{\mathrm{T}} \Sigma_k H\|) \\
&\qquad + \sup_{H\in\mathcal{H}_D} \sum_{k=1}^{p} \big\|H^{\mathrm{T}}\big\{n^{1/2}(ES_k - \Sigma_k)\big\}H\big\|(\|H^{\mathrm{T}} S_k H\| + \|H^{\mathrm{T}} \Sigma_k H\|) \\
&\quad\leqslant p \sup_{H\in\mathcal{H}_D, 1\leqslant k\leqslant p} \big\|H^{\mathrm{T}} n^{1/2}(S_k - ES_k)H\big\|(\|H^{\mathrm{T}} S_k H\| + \|H^{\mathrm{T}} \Sigma_k H\|) \\
&\qquad + p \sup_{H\in\mathcal{H}_D, 1\leqslant k\leqslant p} \big\|H^{\mathrm{T}}\big\{n^{1/2}(ES_k - \Sigma_k)\big\}H\big\|(\|H^{\mathrm{T}} S_k H\| + \|H^{\mathrm{T}} \Sigma_k H\|) \\
&\quad\leqslant p(d-r)^4 \Big\{ \sup_{1\leqslant k\leqslant p} \big\|n^{1/2}(S_k - ES_k)\big\|(\|S_k\| + \|\Sigma_k\|) \\
&\qquad + \sup_{1\leqslant k\leqslant p} \big\|n^{1/2}(ES_k - \Sigma_k)\big\|(\|S_k\| + \|\Sigma_k\|)\Big\} = O_P(1).
\end{aligned} \tag{A5}$$

By Condition 3, (A5) and the definitions of B and $\hat{B}$, we have that

$$\begin{aligned}
0 &\leqslant \Psi_n(B) - \Psi_n(\hat{B}) \\
&= \Psi(B) - \Psi(\hat{B}) + O_P\big(n^{-1/2}\big) \leqslant -a\{D(\hat{B}, B)\}^c + O_P\big(n^{-1/2}\big).
\end{aligned}$$

Now let $n \to \infty$ in the above expression. It must hold that $D(\hat{B}, B) = O_P(n^{-1/(2c)})$. □

REFERENCES

AHN, S. K. (1997). Inference of vector autoregressive models with cointegration and scalar components. *J. Am. Statist. Assoc.* **93**, 350–6.

ANDERSON, T. W. (2002). Canonical correlation analysis and reduced rank regression in autoregressive models. *Ann. Statist.* **30**, 1134–54.

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–222.
Brillinger, D. R. (1981). *Time Series Data Analysis and Theory*, extended ed. San Francisco: Holden-Day.
Box, G. & Tiao, G. (1977). A canonical analysis of multiple time series. *Biometrika* **64**, 355–65.
Brockwell, J. P. & Davis, R. A. (1991). *Time Series Theory and Methods*, 2nd ed. New York: Springer.
Chen, X. R. & Wu, Y. H. (1989). Strong law for a mixing sequence. *Acta Math. Appl. Sinica* **5**, 367–71.
Engle, R. & Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *J. Am. Statist. Assoc.* **76**, 774–81.
Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *Rev. Econ. Statist.* **82**, 540–54.
Francq, C., Roy, R. & Zakoïan, J.-M. (2005). Diagnostic checking in arma models with uncorrelated errors. *J. Am. Statist. Assoc.* **100**, 532–44.
Geweke, J. (1977). The dynamic factor analysis of economic time series models. In *Latent Variables in Socio-Economic Models*, Ed. D. J. Aigner and A. S. Goldberger, pp. 365–83 Amsterdam: North-Holland.
Geweke, J. & Singleton, K. (1981). Maximum likelihood confirmatory factor analysis of economic time series. *Int. Econ. Rev.* **22**, 37–54.
Li, W. K. & McLeod, A. I. (1981). Distribution of the residuals autocorrelations in multivariate arma time series models. *J. R. Statist. Soc.* B, **43**, 231–9.
Lin, Z. & Lu, C. (1997). *Limit Theory for Mixing Dependent Random Variables.* New York: Kluwer.
Ljung, G. M. & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.
Peña, D. & Box, G. E. P. (1987). Identifying a simplifying structure in time series. *J. Am. Statist. Assoc.* **82**, 836–43.
Peña, D. & Poncela, P. (2006). Nonstationary dynamic factor analysis. *J. Statist. Plan. Infer.* **136**, 1237–57.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.
Priestley, M. B., Subba Rao, T. & Tong, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariate systems. *IEEE Trans. Auto. Contr.* **19**, 703–4.
Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis*, 2nd ed. New York: Springer.
Reinsel, G. C. & Velu, R. P. (1998). *Multivariate Reduced Rank Regression*. New York: Springer.
Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Am. Statist. Assoc.* **97**, 1167–79.
Tiao, G. C. & Tsay, R. S. (1989). Model specification in multivariate time series (Discussion). *J. R. Statist. Soc.* B **51**, 157–213.
van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

[*Received September* 2006. *Revised September* 2007]