Pan, Jiazhu and Yao, Qiwei (2008) *Modelling multiple time series via common factors.* Biometrika, 95 (2). pp. 365-379. ISSN 1464-3510

Modelling multiple time series via common factors

JIAZHU PAN[1,2] AND QIWEI YAO[1,3]

[1]*Department of Statistics, London School of Economics, London, WC2A 2AE, UK*

[2]*School of Mathematical Sciences, Peking University, Beijing 100871, China*

[3]*Guanghua School of Management, Peking University, Beijing 100871, China*

j.pan@lse.ac.uk 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 via expanding the white noise space step by step; therefore solving a high-dimensional optimization problem by several low-dimensional sub-problems. Asymptotic properties of the estimation were investigated. The proposed methodology was illustrated with both simulated and real data sets.

Some key words: Factor models; Cross-correlation functions; Dimension reduction; Multivariate time series; Nonstationarity; Portmanteau tests; 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(p, q) with moderately large order (p, q) is practically viable only if a parsimonious representation is identified, resulted from imposing constraints on the coefficient matrices; see Tiao & Tsay (1989), Reinsel (1997) and the references within. An alternative strategy is to reduce the dimensionality. Attempts along this line include, among others, principal components analysis based approaches

(Priestley, Subba Rao & Tong 1974, Brillinger 1981, Stock & Watson 2002), canonical correlation analysis based methods (Box & Tiao 1977, Geweke 1977, Geweke & Singleton 1981, Tiao & Tsay 1989, and Anderson 2002), reduced rank regression methods (Ahn 1997, and Reinsel & Velu 1998), and factor models (Engle & Watson 1981, Peña & Box 1987, Forni et al 2000, Bai & Ng 2002).

In this paper, we revisit the factor models for multiple time series. Although the form of the model concerned is the same as that in, for example, Peña & Box (1987), our approach differs from those in the literature in following three aspects. 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 via expanding the white noise space step by step; therefore solving a high-dimensional optimization problem by several low-dimensional sub-problems. Finally, we allow the dependence between the factors and the white noise in the model. Therefore this overcomes the restriction that the rank of the autocovariance matrix at non-zero lag must not be beyond the number of factors; see Peña & Box (1987).

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 C1 in section 3 below. 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 section 3 below.

The rest of the paper is organized as follows. Section 2 presents the model, the new estimation method and the associated algorithm. The theoretical results for the estimation of the factor loading space is presented in section 3. Numerical illustration with both simulated and real data sets are reported in section 4. All technical arguments are relegated to the Appendix.

2. Models and Methodology

2·1. *Factor models*

Let $\{\mathbf{Y}_t\}$ be a $d \times 1$ time series generated admitting the decomposition

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\mathbf{X}_t$ is a $r \times 1$ time series with finite second moments, $r \leq d$ is unknown, $\mathbf{A}$ is a $d \times r$ unknown constant matrix, and $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of vector white noise process with mean $\boldsymbol{\mu}_\varepsilon$ and covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, *i.e.* $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_s$ are uncorrelated for any $t \neq s$. Furthermore we assume that there exists no linear combination of $\mathbf{X}_t$ which is a white noise process. (Otherwise such a linear combination should be part of $\boldsymbol{\varepsilon}_t$.) We only observe $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ from model (2.1). To simplify the presentation, we assume that

$$\mathbf{S}_0 \equiv \frac{1}{n}\sum_{t=1}^{n}(\mathbf{Y}_t - \bar{\mathbf{Y}})(\mathbf{Y}_t - \bar{\mathbf{Y}})^\tau = \mathbf{I}_d, \tag{2.2}$$

where $\bar{\mathbf{Y}} = n^{-1}\sum_{1\leq t\leq n}\mathbf{Y}_t$. This in practice amounts to replace $\mathbf{Y}_t$ by $\mathbf{S}_0^{-1/2}\mathbf{Y}_t$ before the analysis.

The component variables of the unobserved $\mathbf{X}_t$ are called the factors, $\mathbf{A}$ is called the factor loading matrix. We may assume that the rank of $\mathbf{A}$ is r. (Otherwise (2.1) may be expressed equivalently in terms of a smaller number of factors.) Note model (2.1) is unchanged if we replace $(\mathbf{A}, \mathbf{X}_t)$ by $(\mathbf{AH}, \mathbf{H}^{-1}\mathbf{X}_t)$ for any invertible $r \times r$ matrix $\mathbf{H}$. Therefore, we may assume that the column vectors of $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ are orthonormal, i.e.,

$$\mathbf{A}^\tau\mathbf{A} = \mathbf{I}_r, \tag{2.3}$$

where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. Note that even with constraint (2.3), $\mathbf{A}$ and $\mathbf{X}_t$ are not uniquely determined in (2.1), as the aforementioned replacement is still applicable for any orthogonal $\mathbf{H}$. However the linear space spanned by the columns of $\mathbf{A}$, denoted by $\mathcal{M}(\mathbf{A})$ and called the factor loading space, is a uniquely defined r-dimensional subspace in $\mathcal{R}^d$.

Model (2.1) has been studied by Peña & Box (1987) which assumes that $\boldsymbol{\varepsilon}_t$ and $\mathbf{X}_{t+k}$ are uncorrelated for any integers t and k, and $\mathbf{Y}_t$ is stationary. Under those conditions, the number of factors r is the maximum rank of the autocovariance matrices of $\mathbf{Y}_t$ over all non-zero lags. Further, both $\mathbf{A}$ and r may be estimated via standard eigenanalysis; see Peña & Box (1987).

Our approach is different. For example, we do not require stationarity conditions on the auto-dependence structures of $\mathbf{Y}_t$ and $\mathbf{X}_t$ in model (2.1). Furthermore, the capacity of model (2.1) is substantially enlarged since we allow the autocovariance matrices of $\mathbf{Y}_t$ to be full-ranked.

2·2. *Estimation of* $\mathbf{A}$ *(and* r*)*

Our goal is to estimate $\mathcal{M}(\mathbf{A})$, or its orthogonal complement $\mathcal{M}(\mathbf{B})$, where $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_{d-r})$ is a $d \times (d-r)$ matrix for which $(\mathbf{A}, \mathbf{B})$ forms a $d \times d$ orthogonal matrix, *i.e.* $\mathbf{B}^\tau \mathbf{A} = 0$ and $\mathbf{B}^\tau \mathbf{B} = \mathbf{I}_{d-r}$ (see also (2.3)). Now it follows from (2.1) that

$$\mathbf{B}^\tau \mathbf{Y}_t = \mathbf{B}^\tau \boldsymbol{\varepsilon}_t. \tag{2.4}$$

Hence $\{\mathbf{B}^\tau \mathbf{Y}_t,\ t = 0, \pm 1, \cdots\}$ is a $(d-r) \times 1$ white noise process. Therefore,

$$\mathrm{Corr}(\mathbf{b}_i^\tau \mathbf{Y}_t, \mathbf{b}_j^\tau \mathbf{Y}_{t-k}) = 0 \quad \text{for any } \ 1 \le i, j \le d-r \ \text{ and } \ 1 \le k \le p, \tag{2.5}$$

where $p \ge 1$ is an arbitrary integer. Note that under assumption (2.2), $\mathbf{b}_i^\tau \mathbf{S}_k \mathbf{b}_j$ is the sample correlation coefficient between of $\mathbf{b}_i^\tau \mathbf{Y}_t$ and $\mathbf{b}_j^\tau \mathbf{Y}_{t-k}$, where

$$\mathbf{S}_k = \frac{1}{n} \sum_{t=k+1}^{n} (\mathbf{Y}_t - \bar{\mathbf{Y}})(\mathbf{Y}_{t-k} - \bar{\mathbf{Y}})^\tau. \tag{2.6}$$

This suggests that we may estimate $\mathbf{B}$ by minimizing

$$\Psi_n(\mathbf{B}) \equiv \sum_{k=1}^{p} ||\mathbf{B}^\tau \mathbf{S}_k \mathbf{B}||^2 = \sum_{k=1}^{p} \sum_{1 \le i,j \le d-r} \rho_k(\mathbf{b}_i, \mathbf{b}_j)^2, \tag{2.7}$$

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

Minimizing (2.7) leads to a constrained optimization problem with $d \times (d-r)$ variables. Furthermore r is unknown. Below we present a stepwise expansion algorithm to estimate the columns of $\mathbf{B}$ as well as the the number of columns r. Put

$$\psi(\mathbf{b}) = \sum_{k=1}^{p} \rho_k(\mathbf{b}, \mathbf{b})^2, \qquad \psi_m(\mathbf{b}) = \sum_{k=1}^{p} \sum_{i=1}^{m-1} \left\{ \rho_k(\mathbf{b}, \widehat{\mathbf{b}}_i)^2 + \rho_k(\widehat{\mathbf{b}}_i, \mathbf{b})^2 \right\}.$$

White Noise Space Expansion Algorithm: let $\alpha \in (0,1)$ be the level of significance tests.

Step 1. Let $\widehat{\mathbf{b}}_1$ be a unit vector which minimizes $\psi(\mathbf{b})$. Compute the Ljung-Box-Pierce portmanteau test statistic

$$L_{p,1} = n(n+2) \sum_{k=1}^{p} \frac{\rho_k(\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_1)^2}{n-k}. \tag{2.8}$$

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

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

$$||\mathbf{b}|| = 1, \quad \mathbf{b}^\tau \widehat{\mathbf{b}}_i = 0 \quad \text{for } i = 1, \cdots, m-1. \tag{2.9}$$

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

$$L_{p,m} \equiv n^2 \sum_{k=1}^{p} \frac{1}{n-k} \big[\rho_k(\widehat{\mathbf{b}}_m, \widehat{\mathbf{b}}_m)^2 + \sum_{j=1}^{m-1} \{\rho_k(\widehat{\mathbf{b}}_m, \widehat{\mathbf{b}}_j)^2 + \rho_k(\widehat{\mathbf{b}}_j, \widehat{\mathbf{b}}_m)^2\}\big] \tag{2.10}$$

is greater than the top α-point of the χ^2-distribution with $p(2m-1)$ degrees of freedom (see, e.g. p.149-150 of Reinsel 1997).

Step 3. In the event that $L_{p,m}$ never exceeds the critical value for for all $1 \le m \le d$, let $\widehat{r} = 0$ and $\widehat{\mathbf{B}} = \mathbf{I}_d$.

Remark 1. (i) The algorithm grows the dimension of $\mathcal{M}(\mathbf{B})$ by 1 each time until a newly selected direction $\widehat{\mathbf{b}}_m$ does not lead to a white noise process. Note condition (2.9) ensures that all those $\widehat{\mathbf{b}}_j$ are orthogonal with each other.

(ii) The minimization problem in Step 2 is d-dimensional subject to constraint (2.9). It may be reduced to an unconstrained optimization problem with $d - m$ free variables. Note that the vector $\mathbf{b}$ satisfying (2.9) is of the form

$$\mathbf{b} = \mathbf{D}_m \mathbf{u}, \tag{2.11}$$

where $\mathbf{u}$ is any $(d-m+1) \times 1$ unit vector, $\mathbf{D}_m$ is a $d \times (d-m+1)$ matrix with the columns being the $(d-m+1)$ orthonormal eigenvectors of the matrix $\mathbf{I}_d - \mathbf{B}_{m-1}\mathbf{B}_{m-1}^\tau$, corresponding to the $(d-m+1)$-fold eigenvalue 1, where $\mathbf{B}_m = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_m)$. Also note that any $k \times 1$ unit vector is of the form $\mathbf{u}^\tau = (u_1, \cdots, u_k)$, where

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

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

(iii) Note $\widehat{\mathbf{B}}^\tau \widehat{\mathbf{B}} = \mathbf{I}_{d-\widehat{r}}$. We may let the columns of $\widehat{\mathbf{A}}$ be the $\widehat{r}$ orthonormal eigenvectors of $\mathbf{I}_d - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau$, corresponding to the common eigenvalue 1. It holds that $\widehat{\mathbf{A}}^\tau \widehat{\mathbf{A}} = \mathbf{I}_{\widehat{r}}$.

(iv) The multivariate portmanteau test statistic $L_{p,m}$ given in (2.10) has a normalized constant n^2 which is different from $n(n+2)$ used in the univariate case (2.8). 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 radically different suggestion was proposed by Li & McLeod (1981) which uses

$$L^*_{p,m} = L_{p,m} + \frac{p(p+1)(2m-1)}{2n} \tag{2.12}$$

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 $d >> r$. For the latter cases, both $L_{p,m}$ and $L^*_{p,m}$ may lead to substantially over-estimated r. In our context, an obvious alternative is to use a more stable univariate version

$$L'_{p,m} = n(n+2)\sum_{k=1}^{p} \frac{\rho_k(\widehat{\mathbf{b}}_m, \widehat{\mathbf{b}}_m)^2}{n-k} \tag{2.13}$$

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

(v) Although we do not require the processes $\{\mathbf{Y}_t\}$ and $\{\mathbf{X}_t\}$ to be stationary, our method rests on the fact that there is no autocorrelation in the while noise process $\{\boldsymbol{\varepsilon}_t\}$. Furthermore, the χ^2-asymptotic distributions of the portmanteau tests used in determining r typically rely on the assumption that $\{\boldsymbol{\varepsilon}_t\}$ be i.i.d. Using those tests beyond i.i.d. settings is worth of further investigation. Early attempts include, for example, Francq, Roy & Zakoïan (2005).

(vi) When $\mathbf{Y}_t$ is nonstationary, the sample cross-covariance function $\mathbf{S}_k$ is no longer a meaningful covariance measure. However since $\boldsymbol{\varepsilon}_t$ is a white noise and is stationary, $\mathbf{c}_1^\tau \mathbf{S}_k \mathbf{c}_2$ is the *proper* sample covariance of $\mathbf{c}_1^\tau \mathbf{Y}_t$ and $\mathbf{c}_2^\tau \mathbf{Y}_{t-k}$ for any vectors $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{M}(\mathbf{B})$. In fact our method relies on the fact that $\mathbf{c}_1^\tau \mathbf{S}_k \mathbf{c}_2$ is close to 0 for any $1 \le k \le p$. This also indicates that in practice we should not use large p as, for example, $\mathbf{c}_1^\tau \mathbf{S}_k \mathbf{c}_2$ is a poor estimate for $\mathrm{Cov}(\mathbf{c}_1^\tau \mathbf{Y}_t, \mathbf{c}_2^\tau \mathbf{Y}_{t-k})$ when p is too large.

(vii) When the number of factors r is given, we may skip all the test steps, and stop the algorithm after obtaining $\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_r$ from solving the r optimization problems.

2·3. *Modelling with estimated factors*

Note $\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\tau + \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau = \mathbf{I}_d$. Once we have obtained $\widehat{\mathbf{A}}$, it follows from (2.1) that

$$\mathbf{Y}_t = \widehat{\mathbf{A}}\boldsymbol{\xi}_t + \mathbf{e}_t, \tag{2.14}$$

where

$$\boldsymbol{\xi}_t = \widehat{\mathbf{A}}^\tau \mathbf{Y}_t = \widehat{\mathbf{A}}^\tau \mathbf{A}\mathbf{X}_t + \widehat{\mathbf{A}}^\tau \boldsymbol{\varepsilon}_t, \qquad \mathbf{e}_t = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau \mathbf{Y}_t = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau \boldsymbol{\varepsilon}_t. \tag{2.15}$$

We treat $\mathbf{e}_t$ as a white noise process, and estimate $\mathrm{Var}(\mathbf{e}_t)$ by the sample variance of $\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau \mathbf{Y}_t$.

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

3. Theoretical properties

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

For $r < d$, let $\mathcal{H}$ be the set consisting of all $d \times (d-r)$ matrix $\mathbf{H}$ satisfying the condition $\mathbf{H}^\tau\mathbf{H} = \mathbf{I}_{d-r}$. For $\mathbf{H}_1, \mathbf{H}_2 \in \mathcal{H}$, define

$$D(\mathbf{H}_1, \mathbf{H}_2) = ||(\mathbf{I}_d - \mathbf{H}_1\mathbf{H}_1^\tau)\mathbf{H}_2|| = \sqrt{d - r - \mathrm{tr}(\mathbf{H}_1\mathbf{H}_1^\tau\mathbf{H}_2\mathbf{H}_2^\tau)}. \tag{3.1}$$

Note that $\mathbf{H}_1\mathbf{H}_1^\tau$ is the projection matrix into the linear space $\mathcal{M}(\mathbf{H}_1)$, and $D(\mathbf{H}_1, \mathbf{H}_2) = 0$ if and only if $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$. 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, *i.e.* we treat $\mathbf{H}_1$ and $\mathbf{H}_2$ as the same element in $\mathcal{H}_D$ if and only if $D(\mathbf{H}_1, \mathbf{H}_2) = 0$. Then $(\mathcal{H}_D,\ D)$ forms a metric space in the sense that D is a well-defined distance measure on $\mathcal{H}_D$ (Lemma 1(i) in the Appendix below). Furthermore, the functions $\Psi_n(\cdot)$, defined in (2.7), and

$$\Psi(\mathbf{H}) \equiv \sum_{k=1}^{p} \parallel \mathbf{H}^\tau \boldsymbol{\Sigma}_k \mathbf{H} \parallel^2 \tag{3.2}$$

are well-defined on $\mathcal{H}_D$; see Lemma 1(ii) in the Appendix. In the above expression, $\boldsymbol{\Sigma}_k$ are given in condition C1 below.

We only consider the asymptotic properties for the estimation of the factor loading space while the number of factors r is assumed to be known. (It remains open how to establish the theoretical properties when r is unknown.) Then the estimator for $\mathbf{B}$ may be defined as

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{H}\in\mathcal{H}} \Psi_n(\mathbf{H}) \tag{3.3}$$

Some regularity conditions are now in order.

C1. As $n \to \infty$, $\mathbf{S}_k \to \mathbf{\Sigma}_k$ in probability for $k = 0, 1, \cdots, p$, where $\mathbf{\Sigma}_k$ are non-negative definite matrices, and $\mathbf{\Sigma}_0 = \mathbf{I}_d$.

C2. $\mathbf{B}$ is the unique minimizer of $\Psi(\cdot)$ in the space $\mathcal{H}_D$. That is, $\Psi(\cdot)$ reaches its minimum value at $\mathbf{B}'$ if and only if $D(\mathbf{B}', \mathbf{B}) = 0$, where $\mathbf{B}$ is specified in the beginning of section 2.2.

C3. There exist constants $a > 0, c > 0$ for which $\Psi(\mathbf{H}) - \Psi(\mathbf{B}) \geq a[D(\mathbf{H}, \mathbf{B})]^c$ for any $\mathbf{H} \in \mathcal{H}$.

Remark 2. (i) Condition C1 does not require that the process $\mathbf{Y}_t$ is stationary. In fact it may hold when $E\mathbf{S}_k \to \mathbf{\Sigma}_k$ and $\mathbf{Y}_t$ is φ-mixing in the sense that $\varphi(m) \to 0$ as $m \to \infty$, where

$$\varphi(m) = \sup_{k\geq 1} \sup_{U\in\mathcal{F}_{-\infty}^{k}, V\in\mathcal{F}_{m+k}^{\infty}, P(U)>0} \big|P(V|U) - P(V)\big|, \tag{3.4}$$

and $\mathcal{F}_i^j = \sigma(\mathbf{Y}_i, \cdots, \mathbf{Y}_j)$; see Lemma 2 in the Appendix below. It also gives a sufficient condition which ensures that the convergence in C1 is almost surely. Examples of nonstationary φ-mixing processes include, among others, stationary (φ-mixing) processes plus non-constant treads, and the standardized random walks such as $Y_t = Y_{t-1} + \varepsilon_t/n$, $t = 1, \cdots, n$, where $Y_0 \equiv 0$ and ε_t are i.i.d. with, for example, $E(\varepsilon_t^2) < \infty$. Condition C1 may also hold for some purely deterministic processes such as a linear trend $Y_t = t/n,\ t = 1, \cdots, n$.

(ii) Under model (2.1), $\Psi(\mathbf{B}) = 0$. Condition C2 implies $\Psi(\mathbf{C}) \neq 0$ for any $\mathbf{C} \in \mathcal{H}$ and $\mathcal{M}(\mathbf{C}) \cap \mathcal{M}(\mathbf{A})$ is not an empty set.

THEOREM 1. *Under conditions C1 and C2, $D(\widehat{\mathbf{B}}, \mathbf{B}) \to 0$ in probability as $n \to \infty$. Furthermore, it holds that $D(\widehat{\mathbf{B}}, \mathbf{B}) \to 0$ almost surely if the convergence in C1 is also almost surely.*

THEOREM 2. *Let $\sqrt{n}(E\mathbf{S}_k - \mathbf{\Sigma}_k) = O(1)$, and $\mathbf{Y}_t$ be φ-mixing with $\varphi(m) = O(m^{-\lambda})$ for $\lambda > \frac{p}{p-2}$ and $\sup_{t\geq 1} E\|\mathbf{Y}_t\|^p < \infty$ for some $p > 2$. Then it holds that*

$$\sup_{\mathbf{H}\in\mathcal{H}} |\Psi_n(\mathbf{H}) - \Psi(\mathbf{H})| = O_P(\frac{1}{\sqrt{n}}).$$

If, in addition, C3 also holds, $D(\widehat{\mathbf{B}}, \mathbf{B}) = O_P(n^{-\frac{1}{2c}})$.

Both Theorems 1 and 2 do not require $\mathbf{Y}_t$ to be a stationary process. Their proofs are given in the Appendix.

4. Numerical properties

We illustrate the methodology proposed in section 2 with two simulated examples (one stationary and one nonstationary) and one real data set. The numerical optimization was solved using the downhill simplex method; see section 10.4 of Press et al (1992). In the first two simulated examples below, we set the significance level at 5% for the portmanteau tests used in our algorithm, and $p = 15$ in (2.8). The results with $p = 5, 10$ and 20 are of similar patterns and, therefore, are not reported. We measure the errors in estimating the factor loading space $\mathcal{M}(\mathbf{A})$ by

$$D_1(\mathbf{A}, \widehat{\mathbf{A}}) = \big([\mathrm{tr}\{\widehat{\mathbf{A}}^\tau (I_d - \mathbf{A}\mathbf{A}^\tau)\widehat{\mathbf{A}}\} + \mathrm{tr}(\widehat{\mathbf{B}}^\tau \mathbf{A}\mathbf{A}^\tau \widehat{\mathbf{B}})]/d\big)^{1/2}.$$

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

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

$$X_{t1} = 0.8X_{t-1,1} + e_{t1}, \quad X_{t2} = e_{t2} + 0.9e_{t-1,2} + 0.3e_{t-2,2}, \quad X_{t3} = -0.5X_{t-1,3} - \varepsilon_{t3} + 0.8\varepsilon_{t-1,3},$$

and all ε_{tj} and e_{tj} are independent and standard normal. Note that due to the presence of ε_{t3} in the equation of X_{t3}, $\mathbf{X}_t$ and $\boldsymbol{\varepsilon}_t$ are dependent with each other. In this setting, the number of true factors is $r = 3$, and the factor loading matrix may be taken as $\mathbf{A} = (\mathbf{I}_3, \mathbf{0})^\tau$, where $\mathbf{0}$ denotes the $3 \times (d-3)$ matrix with all elements equal to 0. We set sample size at $n = 300, 600$ and 1000, and the dimension of $\mathbf{Y}_t$ at $d = 5, 10$ and 20. For each setting, we generated 1000 samples from this model. The relative frequencies for $\widehat{r}$ taking different values are reported in Table 1. It shows that when the sample size n increases, the estimation for r becomes more accurate. For example, when $n = 1000$ the relative frequency for $\widehat{r} = 3$ is 0.822 even for d as large as 20. We used $L'_{m,p}$ given in (2.13) in our simulation, since both $L_{m,p}$ and $L^*_{m,p}$ produced substantially over estimated r-values when $d = 10$ and 20. Figure 1 presents the boxplots of errors $D_1(\mathbf{A}, \widehat{\mathbf{A}})$. As the sample size increases, $D_1(\mathbf{A}, \widehat{\mathbf{A}})$ decreases. Furthermore, the errors also increases when d increases.

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

$$\begin{aligned} X_{t1} - 2t/n &= 0.8(X_{t-1,1} - 2t/n) + e_{t1}, \\ X_{t2} &= 3t/n, \\ X_{t3} &= X_{t-1,3} + \sqrt{\frac{10}{n}}\, e_{t3} \quad \text{with} \quad X_{0,3} \sim N(0,1), \end{aligned} \tag{4.1}$$

i.e. X_{t1} is an AR(1) process with non-constant mean, X_{t2} is a purely deterministic trend, and X_{t3} is a random walk. None of them are stationary. The relative frequencies for $\widehat{r}$ taking different values are reported in Table 2. The boxplots of the estimation errors $D_1(\mathbf{A}, \widehat{\mathbf{A}})$ are depicted in Figure 2 . The general pattern observed from the above stationary example (i.e. Example 1) retains. The quality of our estimation improves when sample sizes increases. This is due to the way in which the nonstationarity is specified in (4.1). For example, the sample $\{X_{t2}, t = 1, \cdots, n\}$ 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.

Note that our method rests on the simple fact that the quadratic forms of the sample cross-correlation function are close to 0 along the directions perpendicular to the factor loading space, and are non-zero along the directions in the factor loading space. (See Remark 1(vi) and Remark 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 the 7 cities in Eastern China in January 1954 — December 1986. The cities concerned are Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing and Hangzhou. The sample size $n = 396$ and $d = 7$. As well expected, the data show strong periodic behaviour with period 12. We fit the data with factor models (2.1). By setting $p = 12$, the estimated number of factors is $\widehat{r} = 4$. We applied principal components analysis to the estimated factors. The resulting 4 factors, in the descending order in terms of their variances, are plotted in Figure 4, and their cross-correlation functions are displayed in Figure 5. In fact the variances of the 4 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 4 factors, and 97.6% of the total variation of the original 7 series. The periodic annual oscillation in the original data is predominately reflected by the fluctuation of the first factor (the top panel in Figure 4), and to a much less extent,

also by that of the second factor (the second panel in Figure 4). This suggests that the annual temperature oscillation over this area may be seen as driven by one or at most two 'common factors'. The corresponding loading matrix is

$$\widehat{\mathbf{A}} = \begin{pmatrix} .394 & .386 & .378 & .387 & .363 & .376 & .366 \\ -.086 & .225 & -.640 & -.271 & .658 & -.014 & .164 \\ .395 & .0638 & -.600 & .346 & -.494 & -.074 & .332 \\ .687 & -.585 & -.032 & -.306 & .173 & .206 & -.139 \end{pmatrix}^{\tau}, \tag{4.2}$$

which indicates that the first factor is effectively the average temperatures over the 7 cities. The residuals $\widehat{\mathbf{B}}^{\tau}\mathbf{Y}_t$ carries little dynamic information in the data; see the cross-correlation functions depicted in Figure 6. The sample mean and sample covariance of $\mathbf{e}_t$ are, respectively,

$$\widehat{\boldsymbol{\mu}}_e = \begin{pmatrix} 3.41 \\ 2.32 \\ 4.39 \\ 4.30 \\ 3.40 \\ 4.91 \\ 4.77 \end{pmatrix}, \qquad \widehat{\boldsymbol{\Sigma}}_e = \begin{pmatrix} 1.56 & & & & & & \\ 1.26 & 1.05 & & & & & \\ 1.71 & 1.34 & 1.91 & & & & \\ 1.90 & 1.49 & 2.10 & 2.33 & & & \\ 1.37 & 1.16 & 1.46 & 1.58 & 1.37 & & \\ 1.67 & 1.26 & 1.91 & 2.09 & 1.37 & 1.97 & \\ 1.41 & 1.14 & 1.58 & 1.67 & 1.39 & 1.56 & 1.53 \end{pmatrix}. \tag{4.3}$$

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 January, February, $\cdots$, December, leading to the estimated periodic components

$$\begin{aligned} (g_{1,1}, \cdots, g_{12,1}) &= (-1.61, 1.33, 11.74, 28.06, 41.88, 54.51, 63.77, 62.14, 49.48, 33.74, 18.29, 3.50), \\ (g_{1,2}, \cdots, g_{12,2}) &= (1.67, 1.21, 0.47, 0.17, 0.41, 0.48, 1.37, 2.13, 2.98, 3.05, 2.78, 2.22) \end{aligned} \tag{4.4}$$

for, respectively, the first and the second factors. Figure 7 displays the cross-correlation functions of the 4 factors after removing the periodic components from the first two factors. It shows that the autocorrelation in each of those 4 series is not very strong, and furthermore cross correlation among those 4 series (at non-zero lag) are weak. We fitted a vector autoregressive model to those 4 series with the order 1 determined by the information criterion AICC (see, e.g., p.412 of

Brockwell & Davis 1991) with the estimated coefficients:

$$\widehat{\boldsymbol{\varphi}}_0 = \begin{pmatrix} .07 \\ -.02 \\ -.11 \\ .10 \end{pmatrix}, \quad \widehat{\boldsymbol{\Phi}}_1 = \begin{pmatrix} .27 & -.31 & .72 & .40 \\ .01 & .36 & -.04 & .04 \\ .00 & -.01 & .42 & -.02 \\ -.00 & .03 & .03 & .48 \end{pmatrix}, \tag{4.5}$$

$$\widehat{\boldsymbol{\Sigma}}_u = \begin{pmatrix} 14.24 & & & \\ -.17 & .23 & & \\ -.02 & .03 & .05 & \\ .042 & .01 & -.00 & .05 \end{pmatrix}. \tag{4.6}$$

Both multivariate portmanteau tests (with the lag value $p = 12$) of Li & Mcleod (1981) and Reinsel (1997, p.149) for the residuals from the above fitted vector AR(1) model are insignificant at the 5% level. The univariate portmanteau test is insignificant at the level 5% for three (out of the four) component residual series, and is insignificant at the level 1% 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 AR-order 0, 1, 2, 3 and 4.

Overall the fitted model for the month temperature vector $\mathbf{Y}_t$ is

$$\mathbf{Y}_t = \widehat{\mathbf{A}}\boldsymbol{\xi}_t + \mathbf{e}_t,$$

where the factor loading matrix $\widehat{\mathbf{A}}$ is given in (4.2), the mean and covariance of white noise $\mathbf{e}_t$ are given in (4.3), and the 4×1 factor $\boldsymbol{\xi}_t$ follows VAR(1) model

$$\boldsymbol{\xi}_t - \boldsymbol{\alpha}_t = \widehat{\boldsymbol{\varphi}}_0 + \widehat{\boldsymbol{\Phi}}_1(\boldsymbol{\xi}_{t-1} - \boldsymbol{\alpha}_{t-1}) + \mathbf{u}_t,$$

where the periodic component $\boldsymbol{\alpha}_t^\tau = (g_{m(t),1}, g_{m(t),2}, 0, 0)$, $g_{t,i}$ is given in (4.4),

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

the white noise $\mathbf{u}_t$ has mean 0 and covariance $\widehat{\boldsymbol{\Sigma}}_u$ given in (4.5).

Acknowledgment

This project was partially supported by an EPSRC research grant. The authors thank Professor Valdimir Spokoiny for helpful discussion, Mr Da Huang for making available the temperature data analyzed in Example 3. Thanks also go to two referees for their helpful comments and suggestions.

Appendix

Proofs

We use the same notation as in section 3. We first introduce two lemmas concerning the D-distance defined in (3.1) and condition C1. We then proceed to the proofs for Theorems 1 and 2.

Lemma 1. *(i) It holds for any* $\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3 \in \mathcal{H}$ *that*

$$D(\mathbf{H}_1, \mathbf{H}_3) \leq D(\mathbf{H}_1, \mathbf{H}_2) + D(\mathbf{H}_2, \mathbf{H}_3).$$

(ii) For any $\mathbf{H}_1, \mathbf{H}_2$, $\Psi(\mathbf{H}_1) = \Psi(\mathbf{H}_2)$ *and* $\Psi_n(\mathbf{H}_1) = \Psi_n(\mathbf{H}_2)$ *provided* $D(\mathbf{H}_1, \mathbf{H}_2) = 0$.

Proof. (i) For any symmetric matrices $\mathbf{M}_1, \mathbf{M}_2$ and $\mathbf{M}_3$, it follows from the standard triangle inequality for the matrix norm $||\cdot||$ that $\|\mathbf{M}_1 - \mathbf{M}_3\| \leq \|\mathbf{M}_1 - \mathbf{M}_2\| + \|\mathbf{M}_2 - \mathbf{M}_3\|$, that is

$$\sqrt{\mathrm{tr}(\mathbf{M}_1^2 + \mathbf{M}_3^2 - 2\mathbf{M}_1\mathbf{M}_3)} \leq \sqrt{\mathrm{tr}(\mathbf{M}_1^2 + \mathbf{M}_2^2 - 2\mathbf{M}_1\mathbf{M}_2)} + \sqrt{\mathrm{tr}(\mathbf{M}_2^2 + \mathbf{M}_3^2 - 2\mathbf{M}_2\mathbf{M}_3)}. \quad \text{(A.1)}$$

Let $\mathbf{M}_1 = \mathbf{H}_1\mathbf{H}_1^\tau$, $\mathbf{M}_2 = \mathbf{H}_2\mathbf{H}_2^\tau$ and $\mathbf{M}_3 = \mathbf{H}_3\mathbf{H}_3^\tau$. Since now $\mathrm{tr}(\mathbf{M}_i^2) = \mathrm{tr}(\mathbf{M}_i) = d - r$ for $i = 1, 2, 3$. The inequality required follows from (A.1) and (3.1) directly.

(ii) Under the condition $D(\mathbf{H}_1, \mathbf{H}_2) = 0$, $\mathbf{H}_1\mathbf{H}_1^\tau = \mathbf{H}_2\mathbf{H}_2^\tau$ as it is the projection matrix into the linear space $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$. Now

$$\| \mathbf{H}_1^\tau\mathbf{\Sigma}_k\mathbf{H}_1 \|^2 = \mathrm{tr}\{(\mathbf{H}_1^\tau\mathbf{\Sigma}_k\mathbf{H}_1)^\tau\mathbf{H}_1^\tau\mathbf{\Sigma}_k\mathbf{H}_1\} = \mathrm{tr}(\mathbf{\Sigma}_k^\tau\mathbf{H}_1\mathbf{H}_1^\tau\mathbf{\Sigma}_k\mathbf{H}_1\mathbf{H}_1^\tau) = \| \mathbf{H}_2^\tau\mathbf{\Sigma}_k\mathbf{H}_2 \|^2 .$$

Hence $\Psi(\mathbf{H}_1) = \Psi(\mathbf{H}_2)$. The equality for Ψ_n may be proved in the same manner. ■

Lemma 2. *Let* $\{\mathbf{Y}_t\}$ *be a* φ*-mixing process and* $E\mathbf{S}_k \to \mathbf{\Sigma}_k$. *Suppose that* $\mathbf{Y}_t$ *can be represented as* $\mathbf{Y}_t = \mathbf{U}_t + \mathbf{V}_t$, *where* $\mathbf{U}_t$ *and* $\mathbf{V}_t$ *are uncorrelated for each* t, $\sup_{t\geq 1} E\|\mathbf{U}_t\|^h < \infty$ *for some constant* $h > 2$, *and*

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{V}_t \xrightarrow{P} \mathbf{c}, \qquad \frac{1}{n}\sum_{t=1}^{n}E\mathbf{V}_t \to \mathbf{c}, \quad \text{(A.2)}$$

where $\mathbf{c}$ *is a constant vector. It holds that*

(i) $\mathbf{S}_k \to \mathbf{\Sigma}_k$ *in probability, and*

(ii) $\mathbf{S}_k \to \mathbf{\Sigma}_k$ *almost surely provided that the mixing coefficients satisfy the condition*

$$\varphi(m) = \begin{cases} O(m^{-\frac{b}{2b-2}-\delta}), & \text{if } 1 < b < 2, \\ O(m^{-\frac{2}{b}-\delta}), & \text{if } b \geq 2, \end{cases} \tag{A.3}$$

where $\delta > 0$ *is a constant, and the convergence in condition (A.2) is also almost surely.*

Proof. Assertion (i) follows from the the law of large number for φ-mixing processes; see, *eg.* Theorem 8.1.1 of Lin & Lu (1997). Applying the result of Chen & Wu (1989) to the sequences $\{\mathbf{U}_t\}$ and $\{\mathbf{U}_t\mathbf{U}_{t-i}^\tau\}$, and using the condition (A.2), we may obtain (ii). ■

Proof of Theorem 1. Applying the Cauchy-Schwartz inequality to the matrix norm, we have

$$|\Psi_n(\mathbf{H}) - \Psi(\mathbf{H})| \leq \sum_{k=1}^{p} \big| \parallel \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \parallel^2 - \parallel \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \parallel^2 \big|$$

$$\leq \sum_{k=1}^{p} \parallel \mathbf{H}^\tau (\mathbf{S}_k - \mathbf{\Sigma}_k)\mathbf{H} \parallel [\parallel \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \parallel + \parallel \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \parallel] \leq \parallel \mathbf{H} \parallel^4 \sum_{k=1}^{p} \|\mathbf{S}_k - \mathbf{\Sigma}_k\| [\parallel \mathbf{S}_k \parallel + \parallel \mathbf{\Sigma}_k \parallel].$$

Note that $\parallel \mathbf{H} \parallel^2 = d - r$ for any $\mathbf{H} \in \mathcal{H}$, $\|\mathbf{S}_k - \mathbf{\Sigma}_k\| \to 0$ in probability, which is implied by condition C1, and $\parallel \mathbf{S}_k \parallel + \parallel \mathbf{\Sigma}_k \parallel = O_P(1)$. Hence,

$$\sup_{\mathbf{H} \in \mathcal{H}_D} |\Psi_n(\mathbf{H}) - \Psi(\mathbf{H})| \xrightarrow{P} 0. \tag{A.4}$$

Lemma 1(i) ensures that $(\mathcal{H}_D, D)$ is a well-defined metric space which is complete. Lemma 1(ii) guarantees that $\Psi_n(\cdot)$ is a well-defined stochastic process index by $\mathbf{H} \in \mathcal{H}_D$, and $\Psi(\cdot)$ is well-defined function on the metric space $(\mathcal{H}_D, D)$. Now it follows from the argmax theorem (Theorem 3.2.2 and Corollary 3.2.3 of van der Vaart & Wellner 1996) that $D(\widehat{\mathbf{B}}, \mathbf{B}) \to 0$ in probability.

To show the convergence with probability 1, note that the convergence in (A.4) is with probability 1 provided $\mathbf{S}_k \to \mathbf{\Sigma}_k$ with probability 1. Suppose by contradiction that there exists a δ such that $P\{\limsup_{n\to\infty} D(\widehat{\mathbf{B}}, \mathbf{B}_0) > \delta\} > 0$. Denote $\mathcal{H}'_D = \mathcal{H}_D \cap \{\mathbf{B} : D(\mathbf{B}, B_0) \geq \delta\}$. Then $\mathcal{H}'_D$ is a compact subset of $\mathcal{H}_D$. Note that $\sup_{\mathbf{H}\in\mathcal{H}_D} |\Psi_n(\mathbf{H}) - \Psi(\mathbf{H})| \xrightarrow{a.s.} 0$ implies that there exists a set of sample points Ω' satisfying $\Omega' \subset \{\limsup_{n\to\infty} D(\widehat{\mathbf{B}}, \mathbf{B}_0) > \delta\}$ and $P(\Omega') > 0$ such that for each $\omega \in \Omega'$ one can find a subsequence $\{\widehat{\mathbf{B}}_{n_k}(\omega)\} \subset \mathcal{H}'_D$ with $\widehat{\mathbf{B}}_{n_k}(\omega) \to \mathbf{B} \in \mathcal{H}'_D$. Then, by the definition of $\widehat{\mathbf{B}}$,

$$\Psi(\mathbf{B}) = \lim_{k\to\infty} \Psi_{n_k}(\widehat{\mathbf{B}}_{n_k}(\omega)) \leq \lim_{k\to\infty} \Psi(B_0) = \Psi(\mathbf{B}_0)$$

holds for $\omega \in \Omega'$ and with positive probability. This is a contradiction to Condition C_2. Therefore it must hold that $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \ \rightarrow \ 0$ with probability 1. ■

Proof of Theorem 2. Denote by $s_{(i,j),k}$ and $\sigma_{(i,j),k}$, respectively, the (i,j)-th element of $\mathbf{S}_k$ and $\mathbf{\Sigma}_k$. By the Central Limit Theorem for φ-mixing processes (see Lin & Lu 1997, Davidson 1990), it holds that $\sqrt{n}\{s_{(i,j),k} - Es_{(i,j),k}\} \rightarrow N_{(i,j),k}$ in distribution, where $N_{(i,j),k}$ denotes a Gaussian random variable, $i, j = 1, ..., d$. Hence, $\| \sqrt{n}(\mathbf{S}_k - E\mathbf{S}_k) \| = O_P(1)$. It holds now that

$$\begin{aligned}
&\sup_{\mathbf{H}\in\mathcal{H}_D} \sqrt{n}|\Psi_n(\mathbf{H}) - \Psi(\mathbf{H})| \ \leq \ \sup_{\mathbf{H}\in\mathcal{H}_D} \sqrt{n} \sum_{k=1}^{p} \Big| \ \| \ \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \ \|^2 - \| \ \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \ \|^2 \ \Big| \\
\leq \ &\sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}} \sum_{k=1}^{p} \| \ \mathbf{H}^\tau \sqrt{n}(\mathbf{S}_k - E\mathbf{S}_k)\mathbf{H} \ \| \cdot [\| \ \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \ \| + \| \ \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \ \|] \\
&+ \sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}} \sum_{k=1}^{p} \| \ \mathbf{H}^\tau \{\sqrt{n}(E\mathbf{S}_k - \mathbf{\Sigma}_k)\}\mathbf{H} \ \| \cdot [\| \ \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \ \| + \| \ \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \ \|] \\
\leq \ &p \sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}, 1\leq k\leq p} \| \ \mathbf{H}^\tau \sqrt{n}(\mathbf{S}_k - E\mathbf{S}_k)\mathbf{H} \ \| \cdot [\| \ \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \ \| + \| \ \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \ \|] \\
&+ p \sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}, 1\leq k\leq p} \| \ \mathbf{H}^\tau \{\sqrt{n}(E\mathbf{S}_k - \mathbf{\Sigma}_k)\}\mathbf{H} \ \| \cdot [\| \ \mathbf{H}^\tau \mathbf{S}_k \mathbf{H} \ \| + \| \ \mathbf{H}^\tau \mathbf{\Sigma}_k \mathbf{H} \ \|] \\
\leq \ &p(d-r)^4 \{ \sup_{1\leq k\leq p} \| \ \sqrt{n}(\mathbf{S}_k - E\mathbf{S}_k) \ \| \cdot [\| \ \mathbf{S}_k \ \| + \| \ \mathbf{\Sigma}_k \ \|] \\
&+ \sup_{1\leq k\leq p} \| \ \sqrt{n}(E\mathbf{S}_k - \mathbf{\Sigma}_k) \ \|] \cdot [\| \ \mathbf{S}_k \ \| + \| \ \mathbf{\Sigma}_k \ \|]\} \ = \ O_P(1). \qquad \text{(A.5)}
\end{aligned}$$

By condition C3, (A.5) and the definitions of $\mathbf{B}$ and $\widehat{\mathbf{B}}$, we have that

$$\begin{aligned}
0 \ &\leq \ \Psi_n(\mathbf{B}) - \Psi_n(\widehat{\mathbf{B}}) \\
&= \ \Psi(\mathbf{B}) - \Psi(\widehat{\mathbf{B}}) + O_P(1/\sqrt{n}) \leq -a[D(\widehat{\mathbf{B}}, \mathbf{B})]^c + O_P(1/\sqrt{n}).
\end{aligned}$$

Now let $n \rightarrow \infty$ in the above expression, it must hold that $D(\widehat{\mathbf{B}}, \mathbf{B}) = O_P(n^{-\frac{1}{2c}})$. ■

References

Ahn, S.K. (1997). Inference of vector autoregressive models with cointegration and scalar components. *Journal of the American Statistical Association*, **93**, 350-356.

Anderson, T.G. & Lund, J. (1997). Estimating continuous time stochastic volatility models of the short term interest rates. *Journal of Econometrics*, **77**, 343-377.

Anderson, T.W. (2002). Canonical correlation analysis and reduced rank regression in autoregressive models. *The Annals of Statistics*, **30**, 1134-1154.

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 edition, Holden-Day, San Francisco.

Box, G. & Tiao, G. (1977). A canonical analysis of multiple time series. *Biometrika*, **64**, 355-365.

Brockwell, J.P. & Davis, R.A. (1991). *Time Series Theory and Methods* (2nd Edition). Springer, New York.

Chen, X. R. & Wu, Y. H. (1989). Strong law for a mixing sequence, *Acta Math. Appl. Sinica.* **5**, 367-371.

Davidson, J. (1990). Central limit theorems for nonstationary mixing processes and near-epoch dependent functions. *Discussion Paper No.* **EM/90/216**, Suntory-Toyota International Centre for Economics and Related Disciplines, London School of Economics, Houghton street, London WC2A 2AE.

Engle, R. & Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, **76**, 774-781.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *Review of Econ. Statist.* **82**, 540-554.

Francq, C., Roy, R. & Zakoïan, J.-M. (2005). Diagnostic checking in ARMA models with uncorrelated errors. *Journal of the American Statistical Association*, **100**, 532-544.

Geweke, J. (1977). The dynamic factor analysis of economic time series models. In "*Latent Variables in Socio-Economic Models*", eds. D.J. Aigner & A.S. Goldberger. North-Holland, Amsterdam, pp.365-383.

Geweke, J. & Singleton, K. (1981). Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review*, **22**, 37-54.

Li, W.K. & Mcleod, A.I. (1981). Distribution of the residuals autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, **B**, **43**, 231-239.

Lin, Z. & Lu, C. (1997). *Limit Theory for Mixing Dependent Random Variables.* Science Press/Kluwer Academic Publishers, New York/Beijing.

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, E.P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836-843.

Peña, D. & Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, **136**, 1237-1257.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge.

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. Automt. Control*, **19**, 703-704.

Reinsel, G.C. (1997). *Elements of Multivariate Time Series Analysis*, 2nd Edition. Springer, New York.

Reinsel, G.C. & Velu, R.P. (1998). *Multivariate Reduced Rank Regression*. Springer, New York.

Stock, J.H. & Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167-1179.

Tiao, G.C. & Tsay, R.S. (1989). Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society*, **B**, **51**, 157-213.

Table 1: Relative frequencies for $\widehat{r}$ taking different values in Example 1. (The true value of r is 3.)

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

Table 2: Relative frequencies for $\widehat{r}$ taking different values in Example 2. (The true value of r is 3.)

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

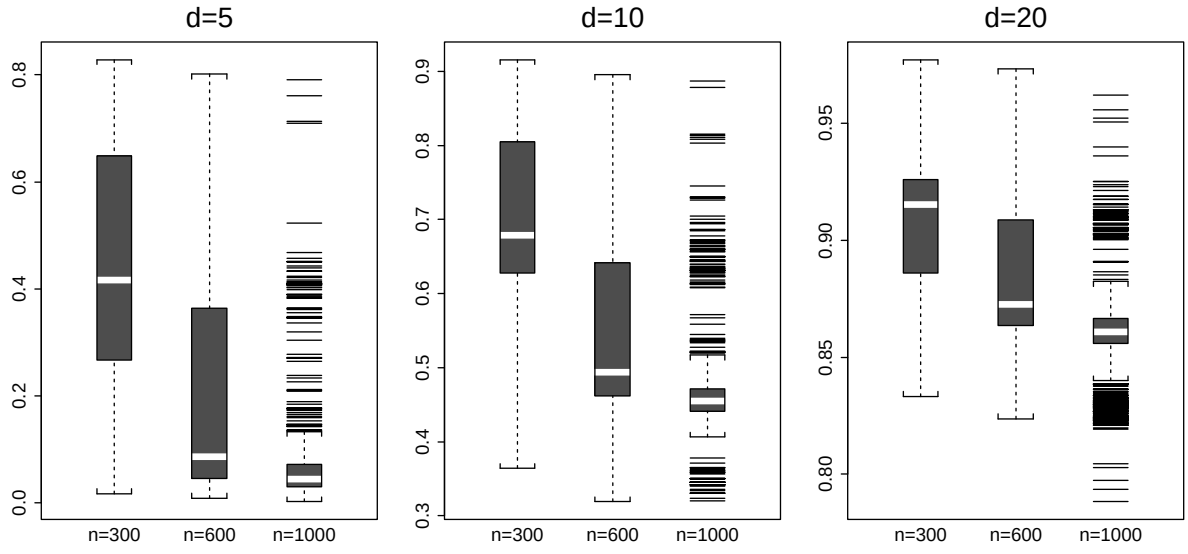

Figure 1: *Example 1 – Boxplots of* $D_1(\mathbf{A}, \widehat{\mathbf{A}})$.

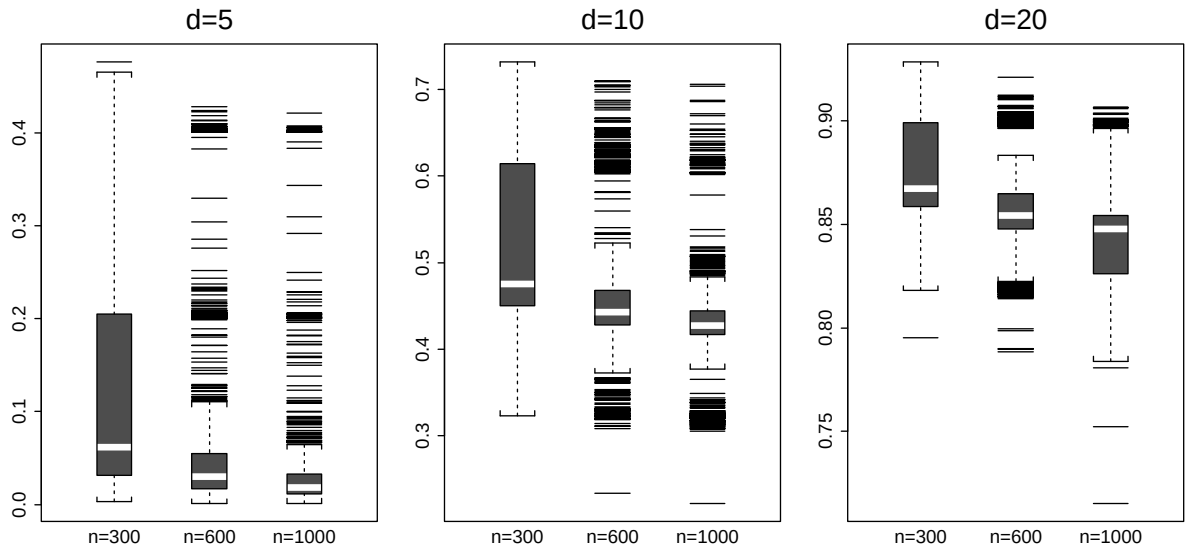

Figure 2: *Example 2 – Boxplots of* $D_1(\mathbf{A}, \widehat{\mathbf{A}})$.

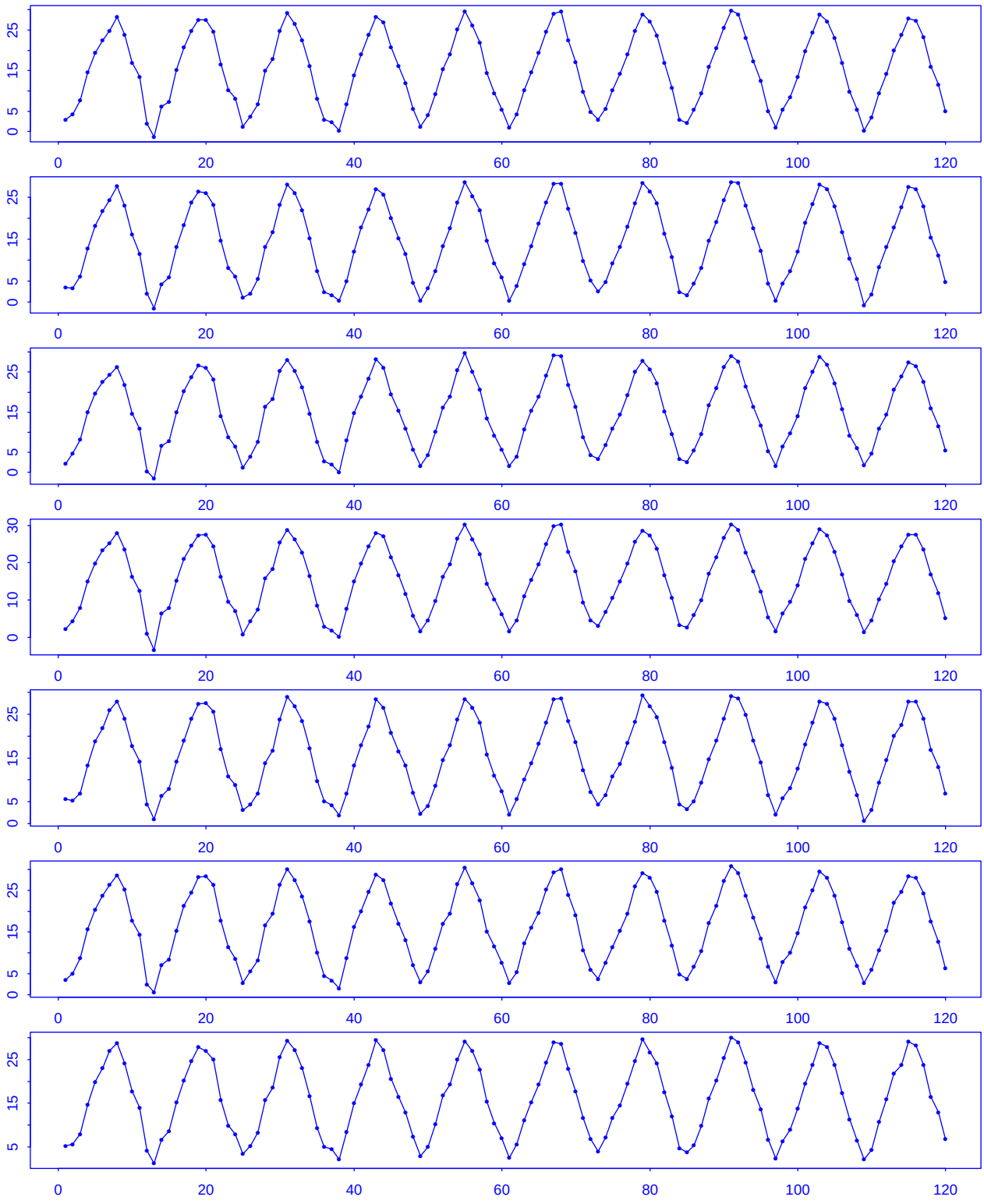

Figure 3: *Example 3 – Time series plots of the monthly temperature in (from top to bottom) Nanjing, Dongtai, Huoshan, Hefei, Shanghai, Anqing and Hangzhou (the first 10 year segments).*

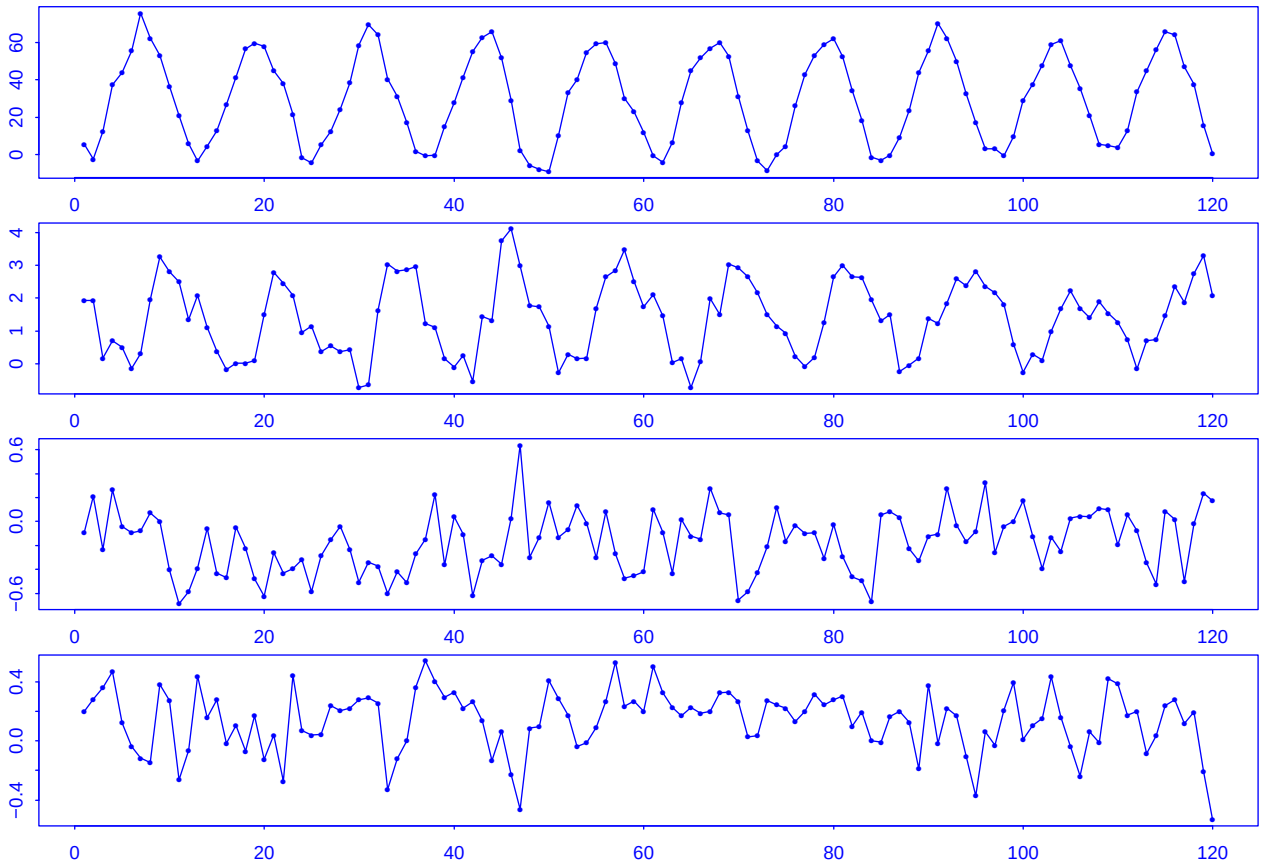

Figure 4: *Example 3 – Time series plots of the 4 estimated factors (the first 10 year segments)* .

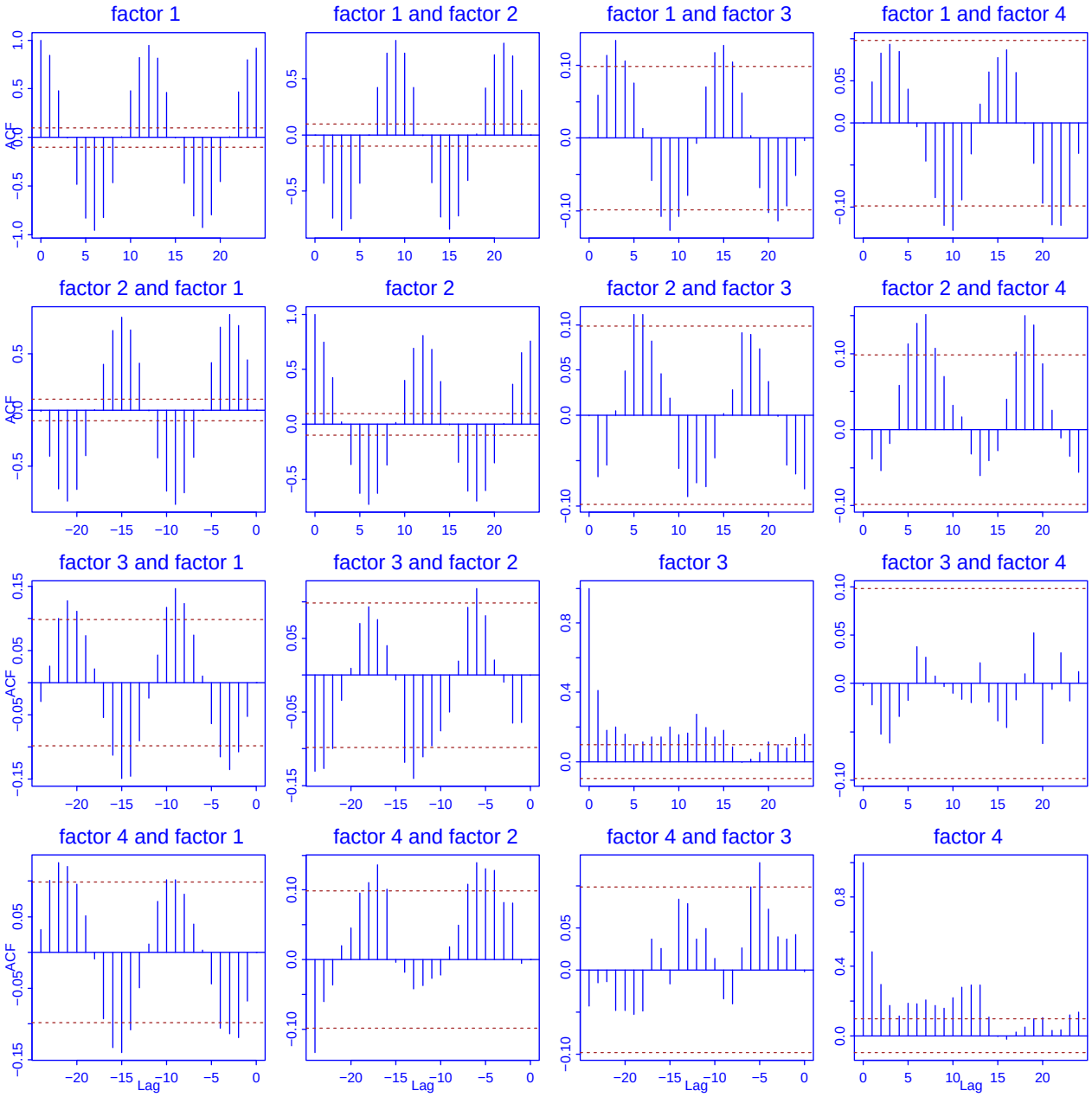

Figure 5: *Example 3 – Sample cross-correlation functions of the 4 estimated factors.*

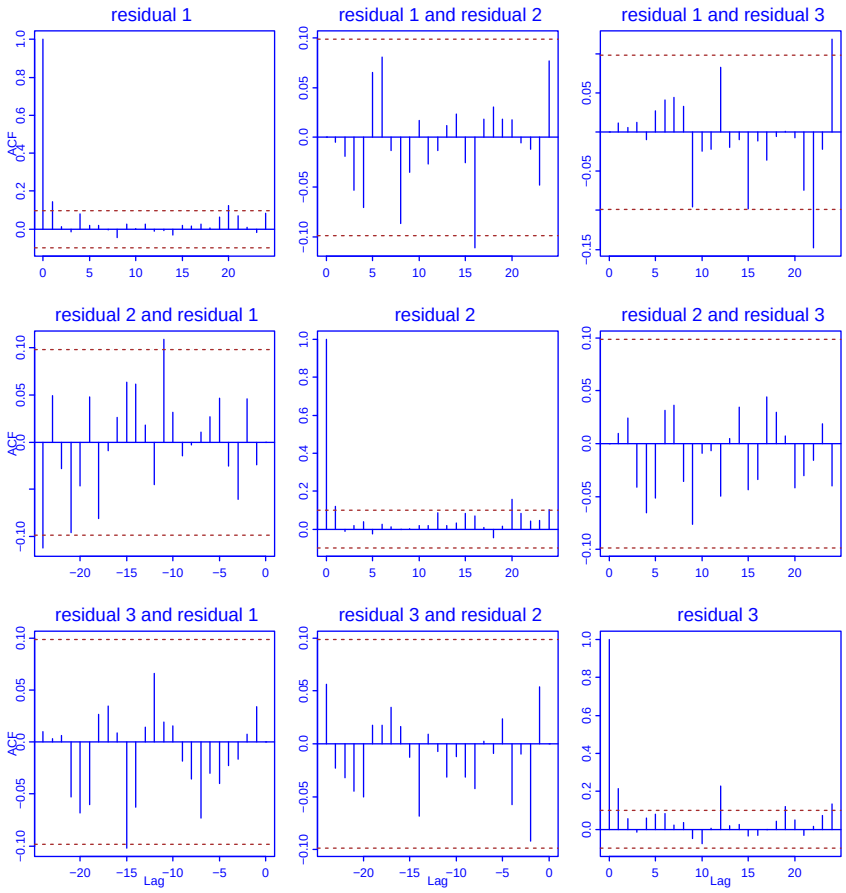

Figure 6: *Example 3 – Sample cross-correlation functions of the 3 residuals* $\widehat{\mathbf{B}}^{\tau}\mathbf{Y}_t$.

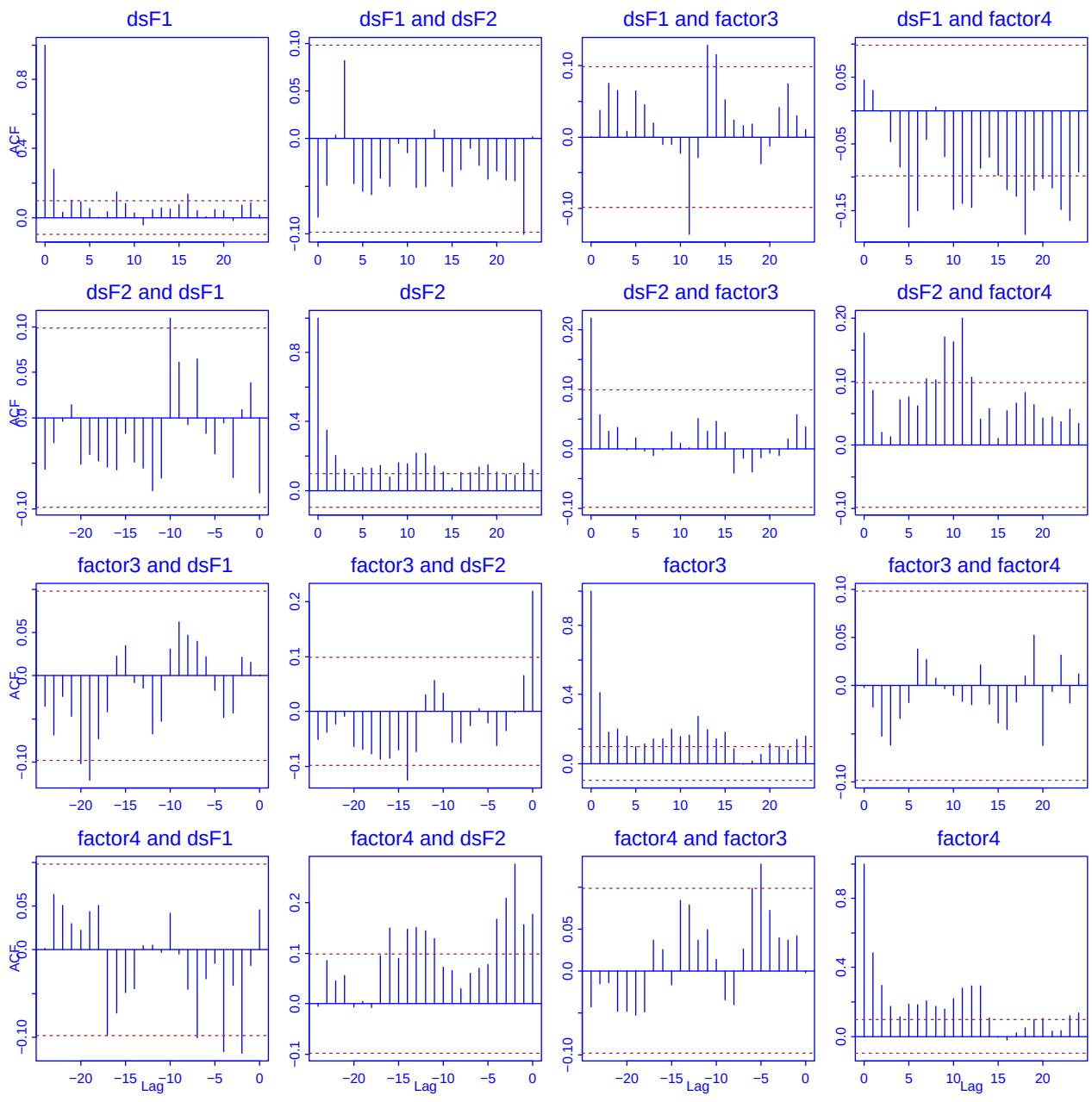

Figure 7: *Example 3 – Sample cross-correlation functions of the 4 factors, after removing the periodic components from the first two factors.*