Modelling Multivariate Volatilities by Common Factors: An Innovation Expansion Method

Jiazhu Pan[1] Daniel Peña[2] Wolfgang Polonik[3] Qiwei Yao[4]

[1]Department of Mathematics and Statistics, University of Strathclyde
26 Richmond Street, Glasgow, G1 1XH, UK

[2]Department of Statistics and Econometrics, University Carlos III
126 - 28903 Getafe, Madrid, Spain

[3]Division of Statistics, University of California at Davis
Davis, CA 95616, USA

[4]Department of Statistics, London School of Economics
Houghton Street, London WC2A 2AE, UK

Abstract

We consider a framework for modelling conditional variance (volatility) of a multivariate time series by common factors. We estimate the factor loading space and the number of factors by a stepwise algorithm of expanding the "innovation space". We develop the asymptotic theory on the proposed estimation method based on the empirical process theory. We further illustrate the method using both simulated and real data examples. Some novel asymptotic results on empirical processes constructed from nonstationary random sequences, which pave the way for the main result, are presented in the Appendix.

MSC2000 subject classifications: Primary 62M10; secondary 62H25, 62P20, 91B84.

Abbreviated title: Multivariate Volatility Factor Models

Key words and phrases: Dimension reduction, empirical processes, factor models, GARCH, multivariate volatility.

1 Introduction

In this modern information age, high dimensional data are available in various fields including finance, economics, psychometrics, biomedical signal processing and etc. For instance, macroeconomic series on output or employment are observed for a large number of countries, regions, or sectors, and time series of financial returns on many different assets are collected routinely. Practitioners frequently face the challenge from modelling high-dimensional time series, as it typically evolves a large number of parameters such that one runs into the so-called over-parameterization problem.

One of the effective ways to circumvent the aforementioned problem is to adopt a factor model, which provides a low-dimensional parsimonious representation for high-dimensional dynamics. There is a large body of literature on the factor modelling for time series. An incomplete list includes Sargent and Sims (1977), Geweke (1977), Chamberlain and Rothschild (1983), Forni, Hallin, Lippi and Reichlin (2002,2004), Bai and Ng (2002), Bai (2003), Hallin and Liška (2007), and Pan and Yao (2008), although all those papers deal with modeling the dynamics of the first moments or conditional first moments. The literature on modeling conditional second moments (i.e. volatilities) include, for example, Engle, Ng and Rothschild (1990), Lin (1992), and Hafner and Preminger (2009). They either assume that the factors are known, or search for factors using maximum likelihood methods.

In this paper, we consider a new method for factor-modelling for multivariate volatility processes. We introduce an innovation expansion method for the estimation of the factor loading space and the number of factors via expanding the "white noise space" (innovation space) step by step, which effectively decompose a a high-dimensional nonlinear optimization problem into several lower-dimensional sub-problems. Asymptotic theory on our approach is developed based the theory of empirical processes.

The rest of the paper is organized as follows. Section 2 introduces the methodology based on an innovation expansion algorithm. Section 3 develops asymptotic theory for the proposed method. Section 4 reports the illustration via both simulated and real data examples. Some novel results on the convergence of the empirical processes constructed from nonstationary processes are presented in the Appendix, which pave the way for establishing the main theoretical results in Section 3.

2 Models and Methodology

2.1 Factor models

Let $\{\mathbf{Y}_t\}$ be a $d \times 1$ time series, and $E(\mathbf{Y}_t|\mathcal{F}_{t-1}) = 0$, where $\mathcal{F}_t = \sigma(\mathbf{Y}_t, \mathbf{Y}_{t-1}, \cdots)$. Assume that $E(\mathbf{Y}_t\mathbf{Y}_t^\tau)$ exists, and we use the notation $\mathbf{\Sigma}_y(t) = \mathrm{Var}(\mathbf{Y}_t|\mathcal{F}_{t-1})$. The goal is to model $\mathbf{\Sigma}_y(t)$ via a common factor model

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

where $\mathbf{X}_t$ is a $r \times 1$ time series, $r < d$ is unknown, $\mathbf{A}$ is a $d \times r$ unknown constant matrix, $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of i.i.d. innovations with mean 0 and covariance matrix $\mathbf{\Sigma}_\varepsilon$, and $\boldsymbol{\varepsilon}_t$ is independent of $\mathbf{X}_t$ and $\mathcal{F}_{t-1}$.

Model (2.1) assumes that the volatility dynamics of $\mathbf{Y}_t$ is determined effectively by a lower dimensional volatility dynamics of $\mathbf{X}_t$ plus the static variation of $\boldsymbol{\varepsilon}_t$, as

$$\mathbf{\Sigma}_y(t) = \mathbf{A}\mathbf{\Sigma}_x(t)\mathbf{A}^\tau + \mathbf{\Sigma}_\varepsilon, \tag{2.2}$$

where $\mathbf{\Sigma}_x(t) = \mathrm{Var}(\mathbf{X}_t|\mathcal{F}_{t-1})$. The component variables of $\mathbf{X}_t$ are called the factors. There is no loss of generality in assuming $\mathrm{rk}(\mathbf{A}) = r$. (Otherwise (2.1) may be expressed equivalently in terms of a smaller number of factors.) Even so model (2.1) is still not completely identifiable, as $(\mathbf{A},\ \mathbf{X}_t)$ in the model may be replaced by $(\mathbf{A}\mathbf{\Gamma},\ \mathbf{\Gamma}^{-1}\mathbf{X}_t)$ for any $r \times r$ invertible matrix $\mathbf{\Gamma}$. However the factor loading space $\mathcal{M}(\mathbf{A})$, which is a linear space spanned by the columns of $\mathbf{A}$, is uniquely defined. We may impose a constraint

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

i.e. we require the column vectors of $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ to be orthonormal, where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. Note that such an orthonormal $\mathbf{A}$ is still not uniquely defined in (2.1). Nevertheless the sum on the right-hand side of (2.2) can be estimated coherently, although the terms involved may not be completely separable.

2.2 Estimation of A and r: an innovation expansion method

Note that the factor loading space $\mathcal{M}(\mathbf{A})$ is uniquely defined by the model. We are effectively concerned with the estimation for $\mathcal{M}(\mathbf{A})$ rather than the matrix $\mathbf{A}$ itself. This is equivalent to the estimation for orthogonal complement $\mathcal{M}(\mathbf{B})$, where $\mathbf{B}$ 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 $\mathbf{B}^\tau \mathbf{Y}_t$ are homoscedastic components since

$$E\{\mathbf{B}^\tau \mathbf{Y}_t \mathbf{Y}_t^\tau \mathbf{B} | \mathcal{F}_{t-1}\} = E\{\mathbf{B}^\tau \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\tau \mathbf{B} | \mathcal{F}_{t-1}\} = E\{\mathbf{B}^\tau \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\tau \mathbf{B}\} = E\{\mathbf{B}^\tau \mathbf{Y}_t \mathbf{Y}_t^\tau \mathbf{B}\} = \mathbf{B}^\tau \mathrm{Var}(\mathbf{Y}_t) \mathbf{B}.$$

This implies that

$$\mathbf{B}^\tau E[\{\mathbf{Y}_t \mathbf{Y}_t^\tau - \mathrm{Var}(\mathbf{Y}_t)\} I(\mathbf{Y}_{t-k} \in C)] \mathbf{B} = 0, \tag{2.5}$$

for any $t, k \geq 1$ and any measurable $C \subset \mathcal{R}^d$. For matrix $\mathbf{H} = (h_{ij})$, let $||\mathbf{H}|| = \{\mathrm{tr}(\mathbf{H}^\tau \mathbf{H})\}^{1/2}$ denote its norm. Then (2.5) implies that

$$\sum_{k=1}^{k_0} \sum_{C \in \mathcal{B}} w(C) \Big|\Big| \sum_{t=k_0+1}^{n} E[\mathbf{B}^\tau \{\mathbf{Y}_t \mathbf{Y}_t^\tau - \mathrm{Var}(\mathbf{Y}_t)\} \mathbf{B} I(\mathbf{Y}_{t-k} \in C)] \Big|\Big|^2 = 0 \tag{2.6}$$

where $k_0 \geq 1$ is a prescribed integer, $\mathcal{B}$ is a finite or countable collection of measurable sets, and the weight function $w(\cdot)$ ensures the sum on the right-hand side finite. In fact we may assume that $\sum_{C \in \mathcal{B}} w(C) = 1$. Even without the stationarity on $\mathbf{Y}_t$, $\mathrm{Var}(\mathbf{Y}_t)$ in (2.6) may be replaced by $\widehat{\Sigma}_y \equiv (n - k_0)^{-1} \sum_{k_0 < t \leq n} \mathbf{Y}_t \mathbf{Y}_t^\tau$. This is due to the fact $\mathbf{B}^\tau \mathrm{Var}(\mathbf{Y}_t) \mathbf{B} = \mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{B}$, and

$$\frac{1}{n - k_0} \sum_{t=k_0+1}^{n} \mathbf{B}^\tau \mathbf{Y}_t \mathbf{Y}_t^\tau \mathbf{B} \;=\; \frac{1}{n - k_0} \sum_{t=k_0+1}^{n} \mathbf{B}^\tau \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\tau \mathbf{B} \;\xrightarrow{a.s.}\; \mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{B},$$

see (2.4). Therefore $\mathbf{B}^\tau \widehat{\Sigma}_y \mathbf{B}$ is a consistent estimator for $\mathbf{B}^\tau \mathrm{Var}(\mathbf{Y}_t) \mathbf{B}$ for all t. Now (2.6) suggests to estimate $\mathbf{B} \equiv (\mathbf{b}_1, \cdots, \mathbf{b}_{d-r})$ by minimizing

$$\begin{aligned} \Phi_n(\mathbf{B}) &= \sum_{k=1}^{k_0} \sum_{C \in \mathcal{B}} w(C) \Big|\Big| \mathbf{B}^\tau \frac{1}{n - k_0} \sum_{t=k_0+1}^{n} (\mathbf{Y}_t \mathbf{Y}_t^\tau - \widehat{\boldsymbol{\Sigma}}_y) I(\mathbf{Y}_{t-k} \in C) \mathbf{B} \Big|\Big|^2 \qquad (2.7) \\ &= \sum_{k=1}^{k_0} \sum_{1 \leq i, j \leq d-r} \sum_{C \in \mathcal{B}} w(C) \Big\{ \mathbf{b}_i^\tau \frac{1}{n - k_0} \sum_{t=k_0+1}^{n} (\mathbf{Y}_t \mathbf{Y}_t^\tau - \widehat{\boldsymbol{\Sigma}}_y) I(\mathbf{Y}_{t-k} \in C) \mathbf{b}_j \Big\}^2 \end{aligned}$$

subject to the condition $\mathbf{B}^\tau \mathbf{B} = \mathbf{I}_{d-r}$. This is a high dimensional optimization problem. Further it does not explicitly address the issue how to determine the number of factors r. We present an algorithm which expands the innovation space step by step, and which also takes care of these two concerns. Note for any $\mathbf{b}^\tau \mathbf{A} = 0$,

$$Z_t \equiv \mathbf{b}^\tau \mathbf{Y}_t (= \mathbf{b}^\tau \boldsymbol{\varepsilon}_t)$$

is a sequence of independent random variables, and therefore, exhibits no conditional heteroskedasticity. The determination of the r is based on the likelihood ratio test for the null hypothesis that the conditional variance of Z_t given its lagged valued is a constant against the alternative that it follows a GARCH(1,1) model with normal innovations. See also Remark 1(vii) below.

Put

$$\Psi(\mathbf{b}) = \sum_{k=1}^{k_0} \Phi_k(\mathbf{b}),$$

$$\Phi_k(\mathbf{b}) = \sum_{C\in\mathcal{B}} w(C)[\mathbf{b}^\tau \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} (\mathbf{Y}_t\mathbf{Y}_t^\tau - \widehat{\mathbf{\Sigma}}_y) I(\mathbf{Y}_{t-k}\in C)\mathbf{b}]^2,$$

$$\Psi_m(\mathbf{b}) = \sum_{k=1}^{k_0}\Big\{2\sum_{i=1}^{m-1}\sum_{C\in\mathcal{B}} w(C)[\widehat{\mathbf{b}}_i^\tau \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} (\mathbf{Y}_t\mathbf{Y}_t^\tau - \widehat{\mathbf{\Sigma}}_y)I(\mathbf{Y}_{t-k}\in C)\,\mathbf{b}]^2 + \Phi_k(\mathbf{b})\Big\}.$$

We propose **An Innovation Expansion Algorithm** for estimating $\mathbf{B}$ and r as follows. Let $\alpha \in (0,1)$ specify the level of the significance tests involved.

Step 1. Compute $\widehat{\mathbf{b}}_1$ which minimises $\Psi(\mathbf{b})$ subject to the constraint $\mathbf{b}^\tau\mathbf{b} = 1$. Let $Z_t = \widehat{\mathbf{b}}_1^\tau \mathbf{Y}_t$. Compute the 2log-likelihood ratio test statistic

$$T = (n-k_0)\big\{1 + \log\big(\frac{1}{n-k_0}\sum_{t=k_0+1}^{n} Z_t^2\big)\big\} - \min\sum_{t=k_0+1}^{n}\big\{\frac{Z_t^2}{\sigma_t^2} + \log(\sigma_t^2)\big\}, \quad (2.8)$$

where $\sigma_t^2 = \alpha + \beta Z_{t-1}^2 + \gamma\sigma_{t-1}^2$, and the minimisation is taken over $\alpha > 0$, $\beta, \gamma \geq 0$ and $\beta + \gamma < 1$. Terminate the algorithm with $\widehat{r} = d$ and $\widehat{\mathbf{B}} = 0$ if T is greater than the top α-point of the χ_2^2-distribution. Otherwise proceed to Step 2.

Step 2. For $m = 2, \cdots, d$, compute $\widehat{\mathbf{b}}_m$ which minimizes $\Psi_m(\mathbf{b})$ subject to the constraint

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

Terminate the algorithm with $\widehat{r} = d - m + 1$ and $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_{m-1})$ if T, calculated as in (2.8) but with $Z_t = |\widehat{\mathbf{b}}_m^\tau \mathbf{Y}_t|$ now, is greater than the top α-point of the χ_2^2-distribution.

Step 3. In the event that T never exceeds the critical value (the top α-point of of the χ_2^2-distribution) for all $1 \leq m \leq d$, let $r = 0$ and $\widehat{\mathbf{B}} = \mathbf{I}_d$.

Remark 1. (i) The algorithm grows the dimension of $\mathcal{M}(\mathbf{B})$ by one each time until a newly selected direction $\widehat{\mathbf{b}}_m$ being relevant to the volatility dynamics of $\mathbf{Y}_t$. This effectively reduces the number of the factors in model (2.1) as much as possible without losing the significant information.

(ii) The minimization problem in Step 2 is a d-dimensional subject to constraint (2.9). It has only $(d-m+1)$ free variables. In fact, the vector $\mathbf{b}$ satisfying (2.9) is of the form

$$\mathbf{b} = \mathbf{A}_m \mathbf{u}, \tag{2.10}$$

where $\mathbf{u}$ is any $(d-m+1)\times 1$ unit vector, $\mathbf{A}_m$ is a $d\times(d-m+1)$ matrix with the columns being the $(d-m+1)$ unit eigenvectors, corresponding to the $(d-m+1)$-fold eigenvalue 1, of matrix $\mathbf{I}_d - \mathbf{B}_m\mathbf{B}_m^\tau$, and $\mathbf{B}_m = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_{m-1})$. Note that the other $(m-1)$ eigenvalues of $\mathbf{I}_d - \mathbf{B}_m\mathbf{B}_m^\tau$ are all 0.

(iii) We may let $\widehat{\mathbf{A}}$ consist of the $\widehat{r}$ (orthogonal) unit eigenvectors, corresponding to the common eigenvalue 1, of matrix $\mathbf{I}_d - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau$ (i.e. $\widehat{\mathbf{A}} = \mathbf{A}_{d-\widehat{r}+1}$). Note that $\widehat{\mathbf{A}}^\tau\widehat{\mathbf{A}} = \mathbf{I}_{\widehat{r}}$.

(iv) A general formal $d\times 1$ unit vector is of the form $\mathbf{b}^\tau = (b_1, \cdots, b_d)$, where

$$b_1 = \prod_{j=1}^{d-1}\cos\theta_j, \quad b_i = \sin\theta_{i-1}\prod_{j=i}^{d-1}\cos\theta_j \ \ (i=2,\cdots,d-1), \quad b_d = \sin\theta_{d-1},$$

where $\theta_1, \cdots, \theta_{d-1}$ are $(d-1)$ free parameters.

(v) We may choose $\mathcal{B}$ consisting of the balls centered at the origin in $\mathcal{R}^d$. Note that $E\mathbf{Y}_{t-k} = 0$. When the underlying distribution of $\mathbf{Y}_{t-k}$ is symmetric and unimodal, such a $\mathcal{B}$ is the collection of the minimum volume sets of the distribution of $\mathbf{Y}_{t-k}$, and this $\mathcal{B}$ determines the distribution of $\mathbf{Y}_{t-k}$ (Polonik 1997). In numerical implementation we simply use $w(C) = 1/K$, where K is the number the balls in $\mathcal{B}$.

(vi) Under the additional condition that

$$\mathbf{c}^\tau\mathbf{A}\{E(\mathbf{X}_t\mathbf{X}_t^\tau|\mathcal{F}_{t-1}) - E(\mathbf{X}_t\mathbf{X}_t^\tau)\}\mathbf{A}^\tau\mathbf{c} = 0 \tag{2.11}$$

if and only if $\mathbf{A}^\tau\mathbf{c} = 0$, (2.5) is equivalent to

$$E\{(\mathbf{b}_i^\tau\mathbf{Y}_t\mathbf{Y}_t^\tau\mathbf{b}_i - 1)I(Y_{t-k}\in C)\} = 0, \quad 1\le i\le d-r, \ \ k\ge 1 \text{ and } C\in\mathcal{B}.$$

See model (2.1). In this case, we may simply use $\Psi(\cdot)$ instead of $\Psi_m(\cdot)$ in Step 2 above. Note that for $\mathbf{b}$ satisfying constraint (2.9), (2.10) implies

$$\Psi(\mathbf{b}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\big\{\mathbf{u}^\tau\mathbf{A}_m^\tau\frac{1}{n-k_0}\sum_{t=k_0+1}^{n}(\mathbf{Y}\mathbf{Y}^\tau - \widehat{\boldsymbol{\Sigma}}_y)I(\mathbf{Y}_{t-k}\in C)\mathbf{A}_m\mathbf{u}\big\}^2. \tag{2.12}$$

Condition (2.11) means that all the linear combinations of $\mathbf{A}\mathbf{X}_t$ are genuinely (conditionally) heteroscadastic.

(vii) Note for any $\mathbf{b}^\tau \mathbf{A} = 0$, $\mathbf{b}^\tau \mathbf{Y}_t (= \mathbf{b}^\tau \boldsymbol{\varepsilon}_t)$ is a sequence of independent random variables, and therefore, $|\mathbf{b}^\tau \mathbf{Y}_t|$ (or $(\mathbf{b}^\tau \mathbf{Y}_t)^2$) is an uncorrelated time series. This suggests that we may replace the likelihood ratio test in Step 1 above by the standard Ljung-Box-Piece portmanteau test applying to $|\mathbf{b}^\tau \mathbf{Y}_t|$ (or $(\mathbf{b}^\tau \mathbf{Y}_t)^2$). However, the autocorrelations of, for example, the squared GARCH(1,1) processes are typically small or very small; see (4.30) of Fan and Yao (2003). This makes the Ljung-Box-Piece test almost powerless for detecting the dependence in the processes such as those specified in (4.2) below (unless the sample size is very large). On the other hand, simulation results in section 3 indicate that the potential of the Gaussian GARCH(1,1) based likelihood ratio test outlined in Step 1 is wide as it is powerful to detect various types of conditional heteroscedasticity even with heavy tailed innovations.

(viii) 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 optimisation problems.

2.3 Estimation for $\boldsymbol{\Sigma}_y(t)$

It is easy to see from (2.1) that

$$\mathbf{Z}_t \equiv \mathbf{A}^\tau \mathbf{Y}_t = \mathbf{X}_t + \mathbf{A}^\tau \boldsymbol{\varepsilon}_t,$$

where $\mathbf{Z}_t$ only has $r(< d)$ components. Note

$$\boldsymbol{\Sigma}_z(t) \equiv \mathrm{Var}(\mathbf{Z}_t | \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}_x(t) + \mathbf{A}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{A},$$

and $\mathbf{A}\mathbf{A}^\tau + \mathbf{B}\mathbf{B}^\tau = \mathbf{I}_d$. By (2.2), it holds that

$$\begin{aligned} \boldsymbol{\Sigma}_y(t) &= \mathbf{A}\boldsymbol{\Sigma}_z(t)\mathbf{A}^\tau + \mathbf{A}\mathbf{A}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{B}\mathbf{B}^\tau + \mathbf{B}\mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{A}\mathbf{A}^\tau + \mathbf{B}\mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{B}\mathbf{B}^\tau \\ &\equiv \mathbf{A}\boldsymbol{\Sigma}_z(t)\mathbf{A}^\tau + \mathbf{A}\mathbf{A}^\tau \boldsymbol{\Sigma}_\varepsilon \mathbf{B}\mathbf{B}^\tau + \mathbf{B}\mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon. \end{aligned} \tag{2.13}$$

By (2.1) and the fact $\mathbf{B}^\tau \mathbf{A} = 0$, $\mathbf{B}^\tau \varepsilon_t = \mathbf{B}^\tau \mathbf{Y}_t$. Hence a natural estimator for $\mathbf{B}^\tau \boldsymbol{\Sigma}_\varepsilon$ may be defined as

$$\frac{1}{n}\sum_{t=1}^n \widehat{\mathbf{B}}^\tau \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\tau = \frac{1}{n}\sum_{t=1}^n \widehat{\mathbf{B}}^\tau \boldsymbol{\varepsilon}_t \mathbf{Y}_t^\tau = \frac{1}{n}\sum_{t=1}^n \widehat{\mathbf{B}}^\tau \mathbf{Y}_t \mathbf{Y}_t^\tau.$$

This leads to a dynamic model for $\boldsymbol{\Sigma}_y(t)$ as follows

$$\widehat{\boldsymbol{\Sigma}}_y(t) = \widehat{\mathbf{A}}\widehat{\boldsymbol{\Sigma}}_z(t)\widehat{\mathbf{A}}^\tau + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\tau \widehat{\boldsymbol{\Sigma}}_y \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau + \widehat{\mathbf{B}}\widehat{\mathbf{B}}^\tau \widehat{\boldsymbol{\Sigma}}_y, \tag{2.14}$$

where $\widehat{\mathbf{\Sigma}}_y = n^{-1} \sum_{1 \le t \le n} \mathbf{Y}_t \mathbf{Y}_t^\tau$, and $\widehat{\mathbf{\Sigma}}_z(t)$ is obtained by fitting the data $\{\widehat{\mathbf{A}}^\tau \mathbf{Y}_t,\ 1 \le t \le n\}$ with, for example, either the dynamic correlation model of Engle (2002) or the CUC model of Fan, Wang and Yao (2008).

3 Theoretical properties

In this section we assume that the number of factors $r (< d)$ is known. Let $\mathcal{H}$ be the set consisting of all $d \times (d - r)$ matrices $\mathbf{H}$ satisfying the condition $\mathbf{H}^\tau \mathbf{H} = \mathbf{I}_{d-r}$. For $\mathbf{H}_1, \mathbf{H}_2 \in \mathcal{H}$, define

$$D(\mathbf{H}_1, \mathbf{H}_2) = ||(\mathbf{I}_d - \mathbf{H}_1 \mathbf{H}_1^\tau)\mathbf{H}_2|| = \{d - r - \mathrm{tr}(\mathbf{H}_1 \mathbf{H}_1^\tau \mathbf{H}_2 \mathbf{H}_2^\tau)\}^{1/2}. \tag{3.1}$$

Note that $\mathbf{H}_1 \mathbf{H}_1^\tau$ is the projection matrix on to 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; that is, 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 $\mathcal{H}_D$; see Lemma A.1(i) of Pan and Yao (2008). Furthermore, similar to the proof of Lemma A.1(ii) of Pan and Yao (2008), we may justify that $\Phi_n(\cdot)$ defined in (2.7), and $\Phi(\cdot)$ defined in (3.2) below are well-defined on $\mathcal{H}_D$. In fact Φ_n is a stochastic process indexed by the metric space $\mathcal{H}_\mathcal{D}$, and Φ is a deterministic function defined on $\mathcal{H}_D$.

Denote the indicator function of a set C by $I(C)$. To include nonstationary cases in our asymptotic theory, we introduce the following assumption which holds for a fairly general class of nonstationary processes; see Escanciano (2007).

Assumption 1. As $n \to \infty$, there exist limits of

$$(n - k_0)^{-1} \sum_{t=k_0}^{n} E\{I(\mathbf{Y}_{t-k} \le x)\}$$

and

$$(n - k_0)^{-1} \sum_{t=k_0}^{n} E\{\mathbf{Y}_t \mathbf{Y}_t^\tau I(\mathbf{Y}_{t-k} \le x)\}$$

for $x \in R^d$ and $k = 1, \ldots, k_0$.

Remark 2. It is easy to see that Assumption 1 holds for any stationary processes. Furthermore, it can be shown that it implies

1. $E\widehat{\mathbf{\Sigma}}_y \to \mathbf{\Sigma}$, where Σ is a nonnegative matrix,
2. $(n-k_0)^{-1}\sum_{t=k_0}^n E\{I(\mathbf{Y}_{t-k} \in C)\} \to a_k(C)$ uniformly for Borel measurable sets C, where $a_k(C)$ is a measure, $k = 1, \ldots, k_0$,
3. $(n-k_0)^{-1}\sum_{t=k_0}^n E\{\mathbf{Y}_t\mathbf{Y}_t^\tau I(\mathbf{Y}_{t-k} \in C)\} \to \mathbf{\Sigma}_k(C)$ uniformly for Borel sets C, where $\mathbf{\Sigma}_k(C)$ are nonnegative definite matrices depending on C, for $k = 1, \ldots, k_0$.

Under Assumption 1, we define

$$\Phi(\mathbf{B}) = \sum_{k=1}^{k_0} \sum_{C \in \mathcal{B}} w(C) \parallel \mathbf{B}^\tau\{\mathbf{\Sigma}_k(C) - a_k(C)\mathbf{\Sigma}\}\mathbf{B} \parallel . \tag{3.2}$$

We will use this denotation in our asymptotic theory for the proposed estimation. Our objective function is $\Phi_n(\mathbf{B})$, which is defined as (2.7). Then our estimator is the minimizer of $\Phi_n(\mathbf{B})$, i.e.

$$\widehat{\mathbf{B}} = \arg \min_{\mathbf{B} \in \mathcal{H}_\mathcal{D}} \Phi_n(\mathbf{B}).$$

Under the assumptions listed below, the estimator $\widehat{\mathbf{B}}$ is consistent.

Assumption 2. $\{\mathbf{Y}_t\}$ is φ-mixing in the sense that $\varphi(m) \to 0$ as $m \to \infty$, where

$$\varphi(m) = \sup_{k \geq 1} \sup_{U \in \mathcal{F}_{-\infty}^k,\, V \in \mathcal{F}_{m+k}^\infty,\, Pr(U) > 0} \big|P(V|U) - P(V)\big|, \tag{3.3}$$

and $\mathcal{F}_i^j = \sigma(Y_i, \ldots, Y_j)$. Furthermore, $\sup_t E \parallel \mathbf{Y}_t \parallel^{2+\delta} < \infty$ for a $\delta > 0$.

Assumption 3. There exists a $d \times (d-r)$ $(d \geq r)$ partial orthonormal matrix $\mathbf{B}_0$ which minimizes $\Phi(\mathbf{B})$, and $\Phi(\mathbf{B})$ reach its minimum value at a partial orthonormal matrix $\mathbf{B}$ if and only if $D(\mathbf{B}, \mathbf{B}_0) = 0$.

Assumption 4. Denote the distribution of Y_t by F_t. Let $F_{(n)}^* = \frac{1}{n-k}\sum_{t=k+1}^n F_{t-k}$. There exists a metric M such that

$$F_{(n)}^*(C_1 \triangle C_2) \leq M(C_1, C_2).$$

for any convex sets C_1, C_2, where $C_1 \triangle C_2 = (C_1 \cup C_2) \setminus (C_1 \cap C_2)$. Furthermore, $F_{(n)}^*$ has a probability density function $f_{(n)}^*$ satisfying that there exist positive constants β_0, c_0 and N_0 such that on the set $\{x \in \mathcal{R}^d : \|x\| \geq N_0\}$ we have $\|x\|^{d+\beta_0} f_{(n)}^*(x) \leq c_0$ and $\sup_x |f_{(n)}^*(x)| < \infty$.

Remark 3. When $\{\mathbf{Y}_t\}$ are identically distributed (not necessarily stationary), Assumption 4 can be replaced by the following assumption.

Assumption 4'. $\{\mathbf{Y}_t\}$ have a common distribution F which has a density function f satisfying that there exist positive constants β_0, c_0 and N_0 such that on the set $\{x \in \mathcal{R}^d : \|x\| \geq N_0\}$ we have $\|x\|^{d+\beta_0} f(x) \leq c_0$ and $\sup_x |f(x)| < \infty$.

Here we give an example to show that there are many non-stationary time series which satisfy Assumption 4 or Assumption 4'.

Example 1. Assume that $\{\mathbf{Y}_t = (Y_{1,t}, Y_{2,t}, \cdots, Y_{d,t})^\tau\}$ is 2-dependent sequence of Gaussian random vectors. That is, for any t, $\mathbf{Y}_t$ and $\mathbf{Y}_{t+m}$ are independent when $m \geq 2$, and $\mathbf{Y}_t$ follows d-variate normal distribution $N(0, \Sigma)$. Furthermore, $(Y_{1,t}, Y_{1,t+1})$ follows a two-variate normal distribution with mean vector $(0,0)^\tau$ and covariance matrix $\begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix}$ where the correlation ρ_t is time-varying. It is easy to see that $\{\mathbf{Y}_t\}$ satisfies Assumptions 2 and 4.

The following theorem is the main theoretical result.

Theorem 1. Let $\mathcal{C}$ denote the class of closed convex sets in $\mathcal{R}^d$. If the collection $\mathcal{B}$ is a countable subclass of $\mathcal{C}$, and Assumptions 1-4 hold, then $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \xrightarrow{P} 0$. Furthermore, $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \xrightarrow{a.s.} 0$ provided, in addition, the mixing coefficients in Assumption 2 satisfy Condition (A.7).

Proof. Denote

$$l_n^*(C) = \frac{1}{n-k_0} \sum_{t=k_0+1}^{n} I(\mathbf{Y}_{t-k} \in C), \tag{3.4}$$

and

$$L_n^*(C) = \frac{1}{n-k_0} \sum_{t=k_0+1}^{n} (\mathbf{Y}_t \mathbf{Y}_t^\tau) I(\mathbf{Y}_{t-k} \in C). \tag{3.5}$$

Note that

$$|\Phi_n(\mathbf{H}) - \Phi(\mathbf{H})| \leq \sum_{i=1}^{5} J_i(\mathbf{H}), \tag{3.6}$$

where

$$J_1(\mathbf{H}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\|\mathbf{H}^\tau(L_n^*(C) - EL_n^*(C))\mathbf{H}\|,$$

$$J_2(\mathbf{H}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\|\frac{1}{n-k_0}\sum_{t=k_0+1}^{n} E(\mathbf{H}^\tau\mathbf{Y}_t\mathbf{Y}_t^\tau\mathbf{H}I(\mathbf{Y}_{t-k}\in C)) - \mathbf{H}^\tau\mathbf{\Sigma}_k(C)\mathbf{H}\|,$$

$$J_3(\mathbf{H}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\|\mathbf{H}^\tau(\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma})\mathbf{H}\|\cdot|\frac{1}{n-k_0}\sum_{t=k_0+1}^{n} I(\mathbf{Y}_{t-k}\in C)|,$$

$$J_4(\mathbf{H}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\|\mathbf{H}^\tau\mathbf{\Sigma}\mathbf{H}\|\cdot|l_n^*(C) - El_n^*(C)|,$$

$$J_5(\mathbf{H}) = \sum_{k=1}^{k_0}\sum_{C\in\mathcal{B}} w(C)\|\mathbf{H}^\tau\mathbf{\Sigma}\mathbf{H}\||\frac{1}{n-k_0}\sum_{t=k_0+1}^{n} EI(\mathbf{Y}_{t-k}\in C) - a_k(C)|.$$

From the definition of $\|\cdot\|$, we have $\| \mathbf{H}_1\mathbf{H}_2 \|\leq\| \mathbf{H}_1 \|\|\ \mathbf{H}_2 \|$ for two conformable matrices. Note that $\| \mathbf{H}^\tau \|=\| \mathbf{H} \|\leq r$. Then, from Lemma A.2 and Assumption 1,

$$\begin{aligned}\sup_{\mathbf{H}\in\mathcal{H}_D} J_1(\mathbf{H}) &\leq \sup_{\mathbf{H}\in\mathcal{H}_D} \| \mathbf{H}^\tau \|\| \mathbf{H} \|\cdot \sup_{1\leq k\leq k_0, C\in\mathcal{B}} \| L_n^*(C) - EL_n^*(C) \| \cdot k_0 \sum_{C\in\mathcal{B}} w(C)\\ &\leq r^2 k_0 \sup_{1\leq k\leq k_0, C\in\mathcal{B}} \| L_n^*(C) - EL_n^*(C) \| \xrightarrow{P} 0,\end{aligned}$$

and

$$\begin{aligned}\sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}} J_4(\mathbf{H}) &\leq r^2 k_0 \sup_{1\leq k\leq k_0, C\in\mathcal{H}} \| \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} I(\mathbf{Y}_{t-k}\in C) - \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} EI(\mathbf{Y}_{t-k}\in C) \|\\ &\xrightarrow{P} 0.\end{aligned}$$

Note that, from Assumption 1 and Remark 2, we have

$$\sup_{\mathbf{H}\in\mathcal{H}_D} J_3(\mathbf{H}) \leq r^2 k_0\{\| \widehat{\mathbf{\Sigma}}_y - E\widehat{\mathbf{\Sigma}} \| + \| E\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma} \|\} \xrightarrow{P} 0,$$

and

$$\sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}} J_i(\mathbf{H}) \xrightarrow{P} 0$$

for $i = 2, 5$. Hence

$$\sup_{\mathbf{H}\in\mathcal{H}_D} |\Phi_n(\mathbf{H}) - \Phi(\mathbf{H})| \xrightarrow{P} 0. \tag{3.7}$$

By the argmax theorem (Theorem 3.2.2 and Corollary 3.2.3 of van der Vaart and Wellner (1996)), we have $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \xrightarrow{P} 0$.

For the strong consistency, if the additional condition (A.7) is satisfied, by Lemma A.2, all the convergence results above hold almost sure, and therefore

$$\sup_{\mathbf{H}\in\mathcal{H}_D} |\Phi_n(\mathbf{H}) - \Phi(\mathbf{H})| \overset{a.s.}{\to} 0. \tag{3.8}$$

This can imply that $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \overset{a.s.}{\to} 0$. In fact, suppose by contradiction that $D(\widehat{\mathbf{B}}, \mathbf{B}_0) \xrightarrow{a.s.} 0$ does not hold. Then there exists a δ such that $Pr\{\limsup_{n\to\infty} D(\widehat{\mathbf{B}}, \mathbf{B}_0) > \delta\} > 0$. Let $\mathcal{H}'_D = \mathcal{H}_D \cap \{\mathbf{B} : D(\mathbf{B}, \mathbf{B}_0) \geq \delta\}$. Thus $\mathcal{H}'_D$ is a compact subset of $\mathcal{H}_D$. Note that, if (3.8) holds, then there exists a set of sample points Ω' satisfying $\Omega' \subset \{\limsup_{n\to\infty} D(\widehat{\mathbf{B}}, \mathbf{B}_0) > \delta\}$ and $Pr(\Omega') > 0$ 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 $\widehat{\mathbf{B}}$,

$$\Phi(\mathbf{B}) = \lim_{k\to\infty} \Phi_{n_k}\{\widehat{\mathbf{B}}_{n_k}(\omega)\} \leq \lim_{k\to\infty} \Phi(\mathbf{B}_0) = \Phi(\mathbf{B}_0)$$

holds for $\omega \in \Omega'$ and with a positive probability. Hence, by Assumption 3, $D(\mathbf{B}, \mathbf{B}_0) = 0$. But $D(\mathbf{B}, \mathbf{B}_0) > \delta > 0$ because $\mathbf{B} \in \mathcal{H}'_D$. This is a contradiction.

In the case that $\{\mathbf{Y}_t\}$ is stationary, if we make an additional assumption, we can get a result on the rate of convergence of our estimator.

Assumption 5. There exist positive constants a and c such that $\Phi(\mathbf{B}) - \Phi(\mathbf{B}_0) \geq a[D(\mathbf{B}, \mathbf{B}_0)]^c$ for any $d \times r$ partial orthonormal matrix $\mathbf{B}$.

Theorem 2. Suppose that the series $\{\mathbf{Y}_t\}$ is strictly stationary with $E\{\parallel \mathbf{Y}_t \parallel\}^{2p} < \infty$ for some $p > 2$. Let $\mathcal{B}$ be a countable Vapnik-Chervonenkis (V-C) class[*] consisting of closed convex sets. If Assumptions 2-4 hold and the φ-mixing coefficients in Assumption 2 satisfy $\varphi_k = O(k^{-b})$ for some $b > \frac{p}{p-2}$. Then

$$\sup_{\mathbf{H}\in\mathcal{H}_\mathcal{D}} |\Phi_n(\mathbf{H}) - \Phi(\mathbf{H})| = O_P(n^{-1/2}).$$

If, in addition, Assumption 5 also holds,

$$D(\widehat{\mathbf{B}}, \mathbf{B}_0) = O_P(n^{-\frac{1}{2c}}).$$

Proof. Because we assume that $\{\mathbf{Y}_t\}$ is stationary in this theorem, $E\widehat{\mathbf{\Sigma}}_y = \mathbf{\Sigma}$, and $J_2(\mathbf{H})$ and $J_5(\mathbf{H})$ in inequality (3.6) are equal to zero. Denote the (i, j)th element of $\widehat{\mathbf{\Sigma}}_y$ and $\mathbf{\Sigma}$ by $\widehat{\sigma}_{(i,j)}$

[*] see van der Vaart and Wellner (1996)

and $\sigma_{(i,j)}$ respectively. From the Central Limit Theorem for $\beta-$mixing sequences, we have

$$\sqrt{n}(\widehat{\sigma}_{(i,j)} - \sigma_{(i,j)}) \xrightarrow{d} N_{i,j}$$

where $N_{i,j}$ is a random variable with Gaussian distribution, $i, j = 1, ..., d$. Then,

$$\| n^{1/2}(\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}) \| = O_P(1).$$

This implies that $\sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_3(\mathbf{H}) = O_P(1)$.

But, by Theorem 1 of Arcones and Yu (1994), the process $\{n^{1/2}(L_n^{*(i,j)}(C) - EL_n^{*(i,j)}(C), C \in \mathcal{B}\}$ indexed by $C \in \mathcal{B}$ converges weakly to a Gaussian process $\{g^{(i,j)}(C), C \in \mathcal{B}\}$ which has uniformly bounded and uniformly continuous paths with respect to the norm $\| \cdot \|$, where $L_n^{*(i,j)}(C)$ denotes the $(i, j)-$th element of $L_n^*(C)$. Hence, $\sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_1(\mathbf{H}) = O_P(1)$. By the same way, we have $\sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_4(\mathbf{H}) = O_P(1)$.

Therefore,

$$\begin{aligned} & \sup_{\mathbf{H}\in\mathcal{H}_D} n^{1/2}|\Phi_n(\mathbf{H}) - \Phi(\mathbf{H})| \\ \leq & \sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_1(\mathbf{H}) + \sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_3(\mathbf{H}) + \sup_{\mathbf{H}\in\mathcal{H}_{\mathcal{D}}} n^{1/2}J_4(\mathbf{H}) \\ = & O_P(1). \end{aligned} \tag{3.9}$$

Note that, from Assumption 5, (3.9) and the definitions of $\mathbf{B}_0$ and $\widehat{\mathbf{B}}$,

$$\begin{aligned} 0 & \leq \Phi_n(\mathbf{B}_0) - \Phi_n(\widehat{\mathbf{B}}) \\ & = \Phi(\mathbf{B}_0) - \Phi(\widehat{\mathbf{B}}) + O_P(1/\sqrt{n}) \leq -a[D(\widehat{\mathbf{B}}, \mathbf{B}_0)]^c + O_P(1/\sqrt{n}). \end{aligned}$$

Then,

$$D(\widehat{\mathbf{B}}, \mathbf{B}_0) = O_P(n^{-\frac{1}{2c}})$$

must hold. Otherwise, there is a contradiction. This completes the proof of Theorem 2.

Remark 4. (i) The result of Theorem 2 could be extended to include nonstationary cases if we can extend the results on the rate of convergence for empirical processes constructed from stationary mixing sequences to those from nonstationary mixing sequences. This is our future work which will be presented elsewhere.

(ii) The objective function in (2.7) can be modified to

$$\Lambda_n(\mathbf{B}) = \sup_{1\leq k\leq k_0, C\in\mathcal{B}} \| \mathbf{B}^\tau \frac{1}{n-k_0} \sum_{t=k_0+1}^{n} (\mathbf{Y}_t\mathbf{Y}_t^\tau - \widehat{\mathbf{\Sigma}}_y) I(\mathbf{Y}_{t-k} \in C)\mathbf{B} \|^2, \tag{3.10}$$

and under Assumption 1, define

$$\Lambda(\mathbf{B}) = \sup_{1\le k\le k_0, C\in\mathcal{B}} \parallel \mathbf{B}^\tau\{\mathbf{\Sigma}_k(C) - a_k(C)\mathbf{\Sigma}\}\mathbf{B} \parallel \tag{3.11}$$

where $\mathcal{B}$ is a collection of closed convex sets (not necessary to be countable in this case). Then, the estimator of B_0 is

$$\widetilde{\mathbf{B}} = \arg\min_{\mathbf{B}\in\mathcal{H}_\mathcal{D}} \Lambda_n(\mathbf{B}). \tag{3.12}$$

If Assumptions 1-5 hold when Φ is replaced by Λ, $\widetilde{\mathbf{B}}$ has the same properties as those of $\widehat{\mathbf{B}}$ in Theorem 1. But, in this case, the condition " the collection $\mathcal{B}$ is a countable class of closed convex sets" can be replaced by a weaker one "the collection $\mathcal{B}$ is the class of all closed convex sets". Similarly, the condition on the class $\mathcal{B}$ in Theorem 2 can be weakened to "$\mathcal{B}$ be a V-C class of closed convex sets".

4 Numerical properties

We illustrate the proposed method with both simulated and real data sets. We always set $k_0 = 30$, $\alpha = 5\%$, and the weight function $C(\cdot) \equiv 1$. Let $\mathcal{B}$ consist of all the balls centered at the origin. The optimisation problems involved were solved by adopting the Downhill Simplex routine from the Numerical Recipes of Press *et al.* (1992).

4.1 Simulated examples

Consider model (2.1) with $r = 3$ factors, and $d\times 3$ matrix $\mathbf{A}$ with $(1, 0, 0), (0, 0.5, 0.866)$ $(0, -0.866, 0.5)$ as its first 3 rows, and $(0, 0, 0)$ as all the other $(d-3)$ rows. We consider 3 different settings for $\mathbf{X}_t = (X_{t1}, X_{t2}, X_{t3})^\tau$, namely, two sets of GARCH(1,1) factors $X_{ti} = \sigma_{ti}e_{ti}$ and $\sigma_{ti}^2 = \alpha_i + \beta_i X_{t-1,i}^2 + \gamma_i\sigma_{t-1,i}^2$, where $(\alpha_i, \beta_i, \gamma_i)$, for $i = 1, 2, 3$, are

$$(1,\ 0.45,\ 0.45), \qquad (0.9,\ 0.425,\ 0.425), \qquad (1.1,\ 0.4,\ 0.4), \tag{4.1}$$

or

$$(1,\ 0.1,\ 0.8), \qquad (0.9,\ 0.15,\ 0.7), \qquad (1.1,\ 0.2,\ 0.6), \tag{4.2}$$

and one mixing setting with two ARCH(2) factors and one stochastic volatility factor:

$$\begin{aligned} X_{t1} &= \sigma_{t1} e_{t1}, \qquad \sigma_{t1}^2 = 1 + 0.6 X_{t-1,1}^2 + 0.3 X_{t-2,1}^2, \\ X_{t2} &= \sigma_{t2} e_{t2}, \qquad \sigma_{t2}^2 = 0.9 + 0.5 X_{t-1,2}^2 + 0.35 X_{t-2,2}^2, \\ X_{t3} &= \exp(h_t/2) e_{t3}, \qquad h_t = 0.22 + 0.7 h_{t-1} + u_t. \end{aligned} \tag{4.3}$$

We let $\{\varepsilon_{ti}\}$, $\{e_{ti}\}$ and $\{u_t\}$ be sequences of independent $N(0,1)$ or standardised t_p-distributed ($p = 4$ or 6) random variables. Note that the (unconditional) variance of X_{ti}, for each i, remains unchanged under the above three different settings. We set $k_0 = 30$, the level of the likelihood ratio tests at 5%, and the sample size $n = 300, 600$ or 1000. For each setting we repeat simulation 500 times.

Table 1: Relative frequency estimates of r with $d = 5$ and normal innovations

				$\widehat{r}$			
Factors	n	0	1	2	**3**	4	5
GARCH(1,1) with coefficients (4.1)	300	.000	.046	.266	**.666**	.014	.008
	600	.000	.002	.022	**.926**	.032	.018
	1000	.000	.000	.000	**.950**	.004	.001
GARCH(1,1) with coefficients (4.2)	300	.272	.236	.270	**.200**	.022	.004
	600	.004	.118	.312	**.500**	.018	.012
	1000	.006	.022	.174	**.778**	.014	.006
Mixture (4.3)	300	.002	.030	.166	**.772**	.026	.004
	600	.000	.001	.022	**.928**	.034	.014
	1000	.000	.000	.000	**.942**	.046	.012

We conducted the simulation with $d = 5$ first. Table 1 lists for the relative frequency estimates for r in the 500 replications. When sample size n increases, the relative frequency for $\widehat{r} = 3$ (i.e. the true value) also increases. Even for $n = 600$, the estimation is already very accurate for GARCH(1,1) factors (4.1) and mixing factors (4.2), less so for the persistent GARCH(1,1) factors (4.2). For $n = 300$, the relative frequencies for $\widehat{r} = 2$ were non-negligible, indicating the tendency of underestimating of r, although this tendency disappears when n increases to 600 or 1000.

Note that $(\mathbf{A}, \mathbf{X}_t)$ in model (2.1) may be equivalently replaced by $(\mathbf{A}\mathbf{\Gamma}^\tau, \mathbf{\Gamma}\mathbf{X}_t)$, where $\mathbf{\Gamma}$ is any $r \times r$ orthogonal matrix. However the linear vector space $\mathcal{M}(\mathbf{A})$ spanned by the columns of $\mathbf{A}$ is uniquely determined. To measure the difference between $\mathcal{M}(\mathbf{A})$ and $\mathcal{M}(\widehat{\mathbf{A}})$, we define

$$D(\mathbf{A}, \widehat{\mathbf{A}}) = \{|(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\tau)\widehat{\mathbf{A}}|_1 + |\mathbf{A}\mathbf{A}^\tau\widehat{\mathbf{B}}|_1\}/d^2, \tag{4.4}$$

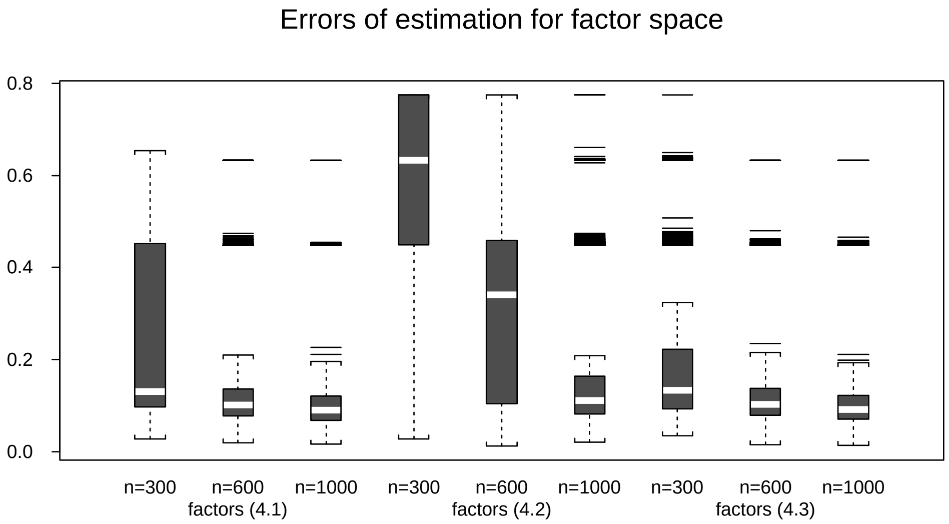

Figure 1: *Boxplots of* $D(\mathbf{A}, \widehat{\mathbf{A}})$ *with two sets of GARCH(1,1) factors specified, respectively, by (4.1) and (4.2), and mixing factors (4.3). Innovations are Gaussian and* $d = 5$.

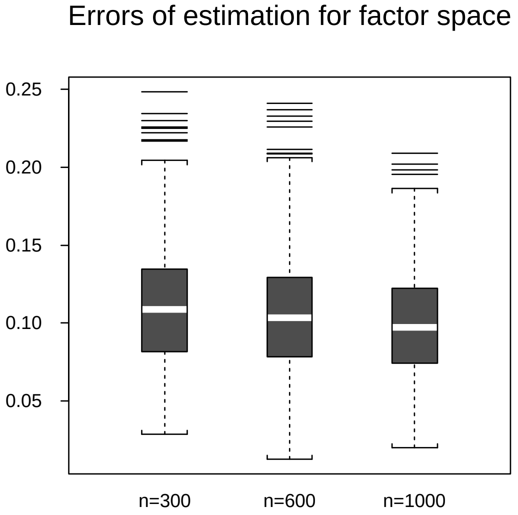

Figure 2: *Boxplots of* $D(\mathbf{A}, \widehat{\mathbf{A}})$ *with GARCH(1,1) factors (4.2), Gaussian innovations, and* $d = 5$. *The number of factors* $r = 3$ *is known.*

where $|\mathbf{A}|_1$ denotes the sum of the absolute values of all the elements in matrix $\mathbf{A}$. Note that $\mathcal{M}(\mathbf{A}) = \mathcal{M}(\widehat{\mathbf{A}})$ if and only if

$$(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\tau)\widehat{\mathbf{A}} = 0 \qquad \text{and} \qquad \mathbf{A}\mathbf{A}^\tau\widehat{\mathbf{B}} = 0.$$

Figure 1 displays the boxplots of $D(\mathbf{A}, \widehat{\mathbf{A}})$. The estimation was pretty accurate with GARCH factors (4.1) and mixing factors (4.3), especially with correctly estimated r. Note with $n = 600$ or 1000, those outliers (lying above the range connected by dashed lines) typically correspond to the estimates $\widehat{r} \neq 3$. The poor performance with GARCH factors (4.2) was largely due the misestimated r. To highlight this, Figure 2 plotted the boxplots of $D(\mathbf{A}, \widetilde{\mathbf{A}})$ with $r = 3$ given in the estimation. It shows that even with $n = 300$, the estimation for the factor space is accurate as long as we know the its dimension.

Table 2: Relative frequency estimates of r with $d = 5$, GARCH(1,1) factors (4.2) and t_p innovations

				$\widehat{r}$			
p	n	0	1	2	**3**	4	5
6	300	.154	.192	.346	**.240**	.044	.024
	600	.008	.086	.206	**.584**	.072	.044
	1000	.000	.001	.094	**.758**	.102	.036
4	300	.000	.014	.074	**.752**	.112	.048
	600	.000	.000	.000	**.724**	.182	.094
	1000	.000	.000	.000	**.706**	.180	.114

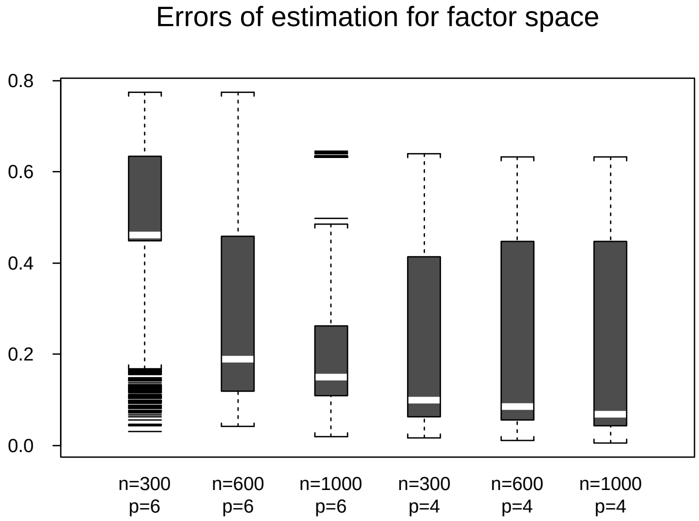

Figure 3: *Boxplots of $D(\mathbf{A}, \widehat{\mathbf{A}})$ with GARCH(1,1) factors (4.2), t_p innovations and $d = 5$.*

We repeated the above experiments with t_p-distributed ε_{ti} and e_{ti} for $p = 6$ and 4. Note $E|\varepsilon_{ti}|^p = E|e_{ti}|^p = \infty$ now. The results with GARCH(1,1) factors (4.2) are reported in Table 2 and Figure 3. One striking feature is that the estimation, especially with $n = 300$ or 600, is more accurate than that for the same model but with Gaussian innovations. Furthemore, the estimation for $\mathbf{A}$ becomes more accurate when n increases. However, with the very heavy tailed distribution t_4, the estimator for r may not be consistent, as the relative frequency for $\widehat{r} = 3$ when $n = 1000$ is smaller than that when $n = 600$ or 300.

Table 3: Relative frequency estimates of r with GARCH(1,1) factors, normal innovations and d=10 or 20

							$\widehat{r}$			
Coefficients	d	n	0	1	2	**3**	4	5	6	≥ 7
(4.1)	10	300	.002	.048	.226	**.674**	.014	.001	.004	.022
	10	600	.000	.000	.022	**.876**	.016	.012	.022	.052
	10	1000	.000	.000	.004	**.876**	.024	.022	.022	.052
	20	300	.000	.040	.196	**.626**	.012	.008	.010	.138
	20	600	.000	.000	.012	**.808**	.012	.001	.018	.149
	20	1000	.000	.000	.000	**.776**	.024	.012	.008	.180
(4.2)	10	300	.198	.212	.280	**.248**	.016	.008	.014	.015
	10	600	.032	.110	.292	**.464**	.018	.026	.012	.046
	10	1000	.006	.032	.128	**.726**	.032	.020	.016	.040
	20	300	.166	.266	.222	**.244**	.012	.004	.001	.107
	20	600	.022	.092	.220	**.472**	.001	.001	.012	.180
	20	1000	.006	.016	.092	**.666**	.018	.016	.014	.172

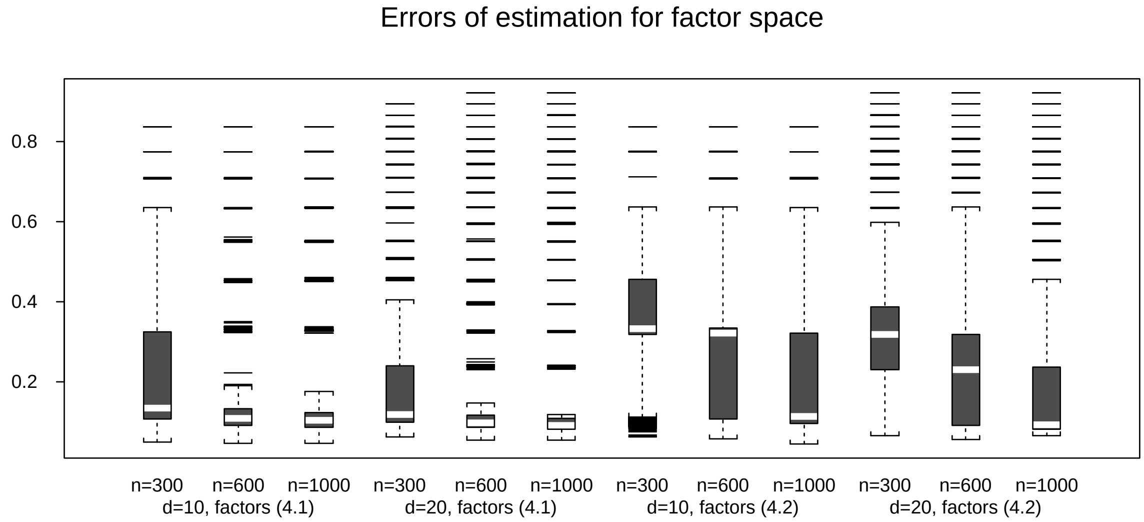

Figure 4: *Boxplots of* $D(\mathbf{A}, \widehat{\mathbf{A}})$ *with two sets of GARCH(1,1) factors specified in (4.1) and (4.2), normal innovations and* $d = 10$ *or 20.*

Finally we report the results with $d = 10$ and 20 in Table 3 and Figure 4. To save the space, we only report the results with the two sets GARCH(1,1) factors and Gaussian innovations. Comparing with Table 1, the estimation of r is only marginally worse than that with $d = 5$. Indeed the difference with $d = 10$ and 20 is not big either. Note the D-measures for different d are not comparable; see (4.4). Nevertheless, Figure 4 shows that the estimation for $\mathbf{A}$ becomes more accurate when n increases, and the estimation with the persistent factors (4.2) is less accurate than that with (4.1).

Further experiements with with $k_0 = 20, 40$ or 50 indicate that the procedure is robust with respect to the values of k_0 between 20 and 50. We also did some experiements with the likelihood ratio test replaced by Ljung-Box-Piece test. While the results with GARCH(1,1) factors (4.1) and mixing factors (4.2) are comparable with, though not as good as, Table 1, it is less satisfactory for GARCH (1,1) factors (4.2). For example, the relative frequency for $\widehat{r} = 3$ with $n = 1000$ is merely 0.540.

4.2 Real data examples

Figure 5 displays the daily log-returns of the S&P 500 index, the stock prices of Cisco System and Intel Corporation in 2 January 1997 – 31 December 1999. For this data set, $n = 758$ and $d = 3$. With $k_0 = 30$ and $\alpha = 5\%$, the estimated number of factors is $\widehat{r} = 1$ with $\widehat{\mathbf{A}}^\tau = (0.310,\ 0.687,\ 0.658)$. The time plots of the estimated factor $Z_t \equiv \widehat{\mathbf{A}}^\tau \mathbf{Y}_t$ and the two homoscedastic components $\widehat{\mathbf{B}}^\tau \mathbf{Y}_t$ are displayed in Figure 6. The P-value of the Gaussian-GARCH(1,1) based likelihood ratio test for the null hypothesis of the constant conditional variance for Z_t is 0.000. The correlograms of the squared and the absolute factor are depicted in Figure 7 which indicates the existence of heteroscedasticity in Z_t. The fitted GARCH(1,1) model for Z_t is

$$\widehat{\sigma}_t^2 = 2.5874 + 0.1416 Z_{t-1}^2 + 0.6509 \widehat{\sigma}_{t-1}^2. \tag{4.5}$$

In contrast, Figure 8 shows that there is little autocorrelation in squared or absolute components of $\widehat{\mathbf{B}}^\tau \mathbf{Y}_t$. The estimated constant covariance matrix in (2.14) is

$$\widehat{\boldsymbol{\Sigma}}_0 = \begin{pmatrix} 1.594 & & \\ 0.070 & 4.142 & \\ -1.008 & -0.561 & 4.885 \end{pmatrix}.$$

The overall fitted conditional variance process is as given in (2.14) with $\widehat{\boldsymbol{\Sigma}}_z(t) = \widehat{\sigma}_t^2$ defined in (4.5) above.

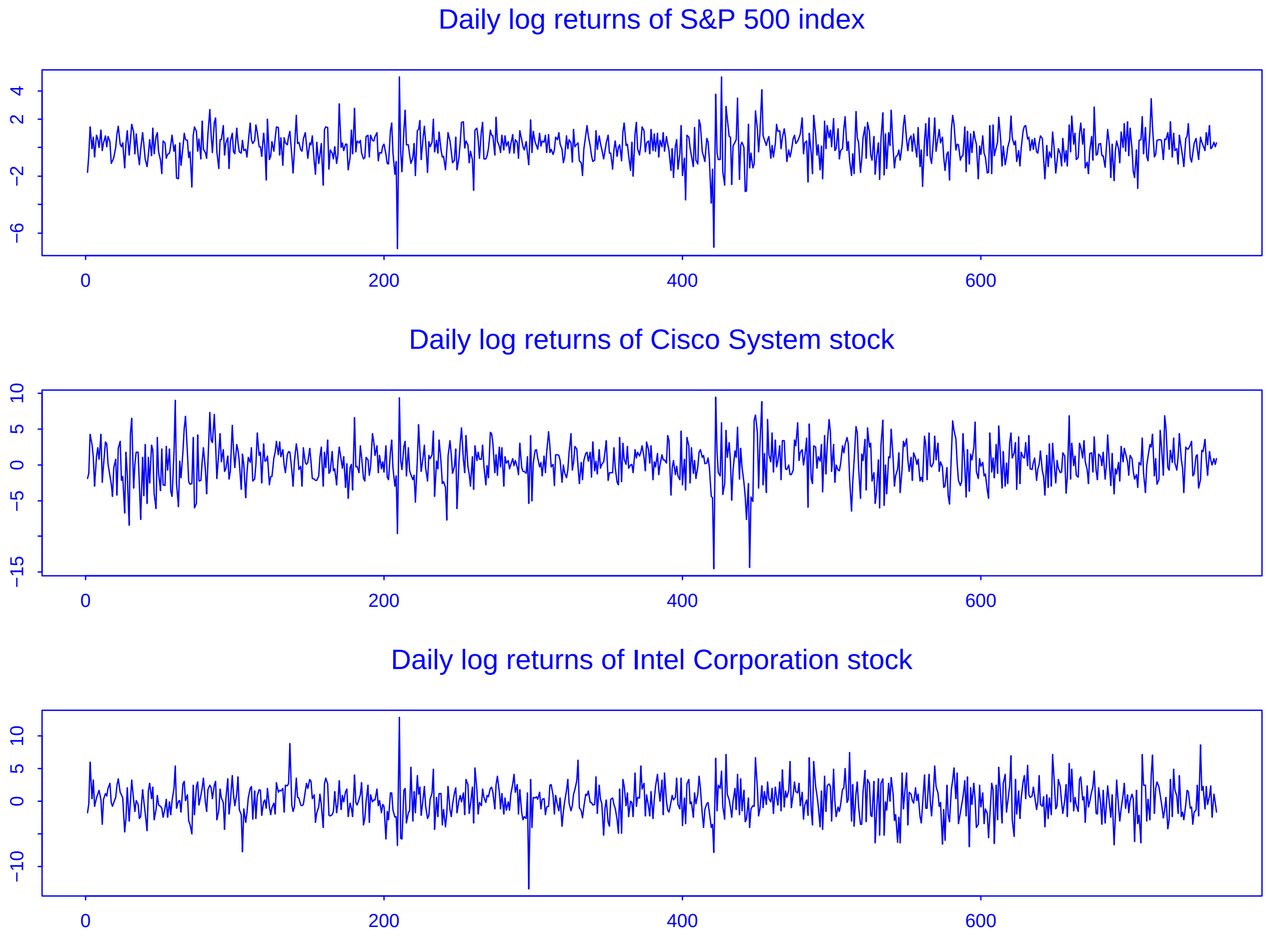

Figure 5: *Time plots of the daily log-returns of S&P 500 index, Cisco System and Intel Coprporation stock prices*

Appendix. Lemmas on Empirical Processes of Non-Stationary Sequences

Suppose that $\{L_n^*(g), g \in \mathcal{G}\}$ be a stochastic process indexed by a class of real-valued functions $\mathcal{G}$. Denote $L_n(g) = L_n^*(g) - EL_n^*(g)$. Suppose we know that

$$L_n(g) \to 0 \quad \text{in probability} \tag{A.1}$$

for any $g \in \mathcal{G}$. We are interested in proving a uniform convergence:

$$\sup_{f \in \mathcal{G}} \|L_n(f)\| \to 0 \quad \text{in probability}$$

as $n \to \infty$. We will use some concepts in the theory of empirical processes. Suppose that D is a metric on $\mathcal{G}$. For a pair of functions f, g with $f \le g$ and $D(f,g) \le \varepsilon$, $[f,g] := \{h \in \mathcal{G} : f \le h \le g\}$

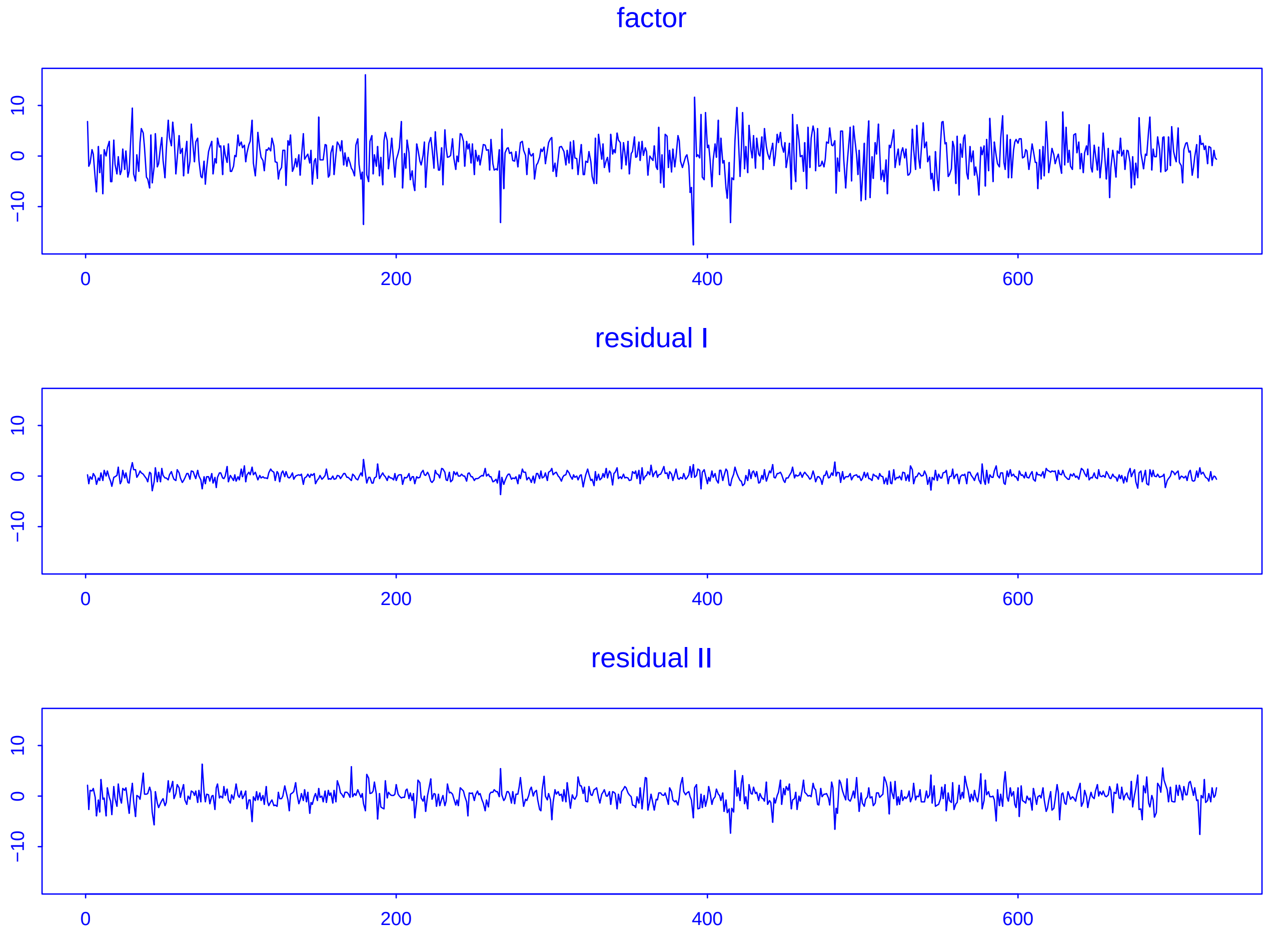

Figure 6: *Time plots of the estimated factor and two homoscedastic compoments for the S&P 500, Cisco and Intel data*

is called an ε-bracket, and

$$N_B(\varepsilon, \mathcal{G}, D) = \min\{k \geq 0 : \mathcal{G} \subset \bigcup_{j=1}^{k} B_j, B_1, \ldots, B_k \quad \text{are } \varepsilon- \text{ brackets}\}$$

is called bracketing numbers of $\mathcal{G}$ with respect to the metric D. In the other words, $N_B(\varepsilon, \mathcal{G}, D)$ denotes the minimal numbers of $\varepsilon-$brackets necessary to cover $\mathcal{G}$. Assume that

$$N_B(\varepsilon, \mathcal{G}, D) < \infty, \quad \forall \varepsilon > 0, \tag{A.2}$$

there exist some $C, \alpha > 0$ such that

$$\|EL_n^*(g_1) - EL_n^*(g_2)\| \leq \{D(g_1, g_2)\}^{\alpha}, \quad \forall g_1, g_2 \in \mathcal{G}, \tag{A.3}$$

and for each n, $L_n^*(g)$ is nondecreasing in g:

$$g_1 \leq g_2 \quad \Rightarrow L_n^*(g_1) \leq L_n^*(g_2). \tag{A.4}$$

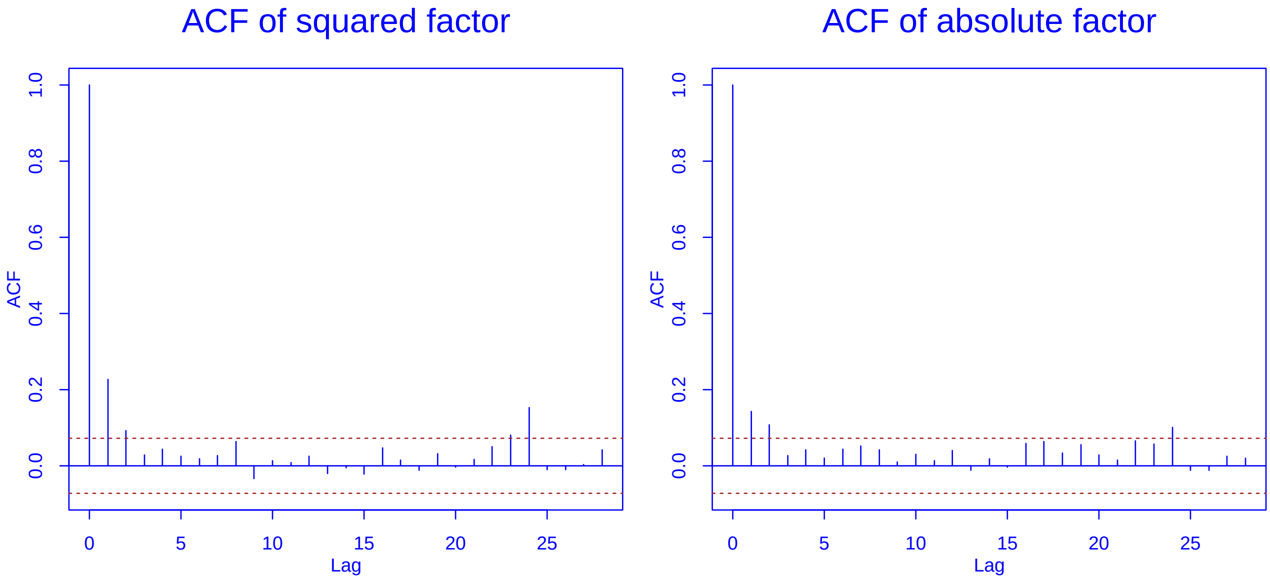

Figure 7: *The correlograms of squared and absulote factor for the the S&P 500, Cisco and Intel data*

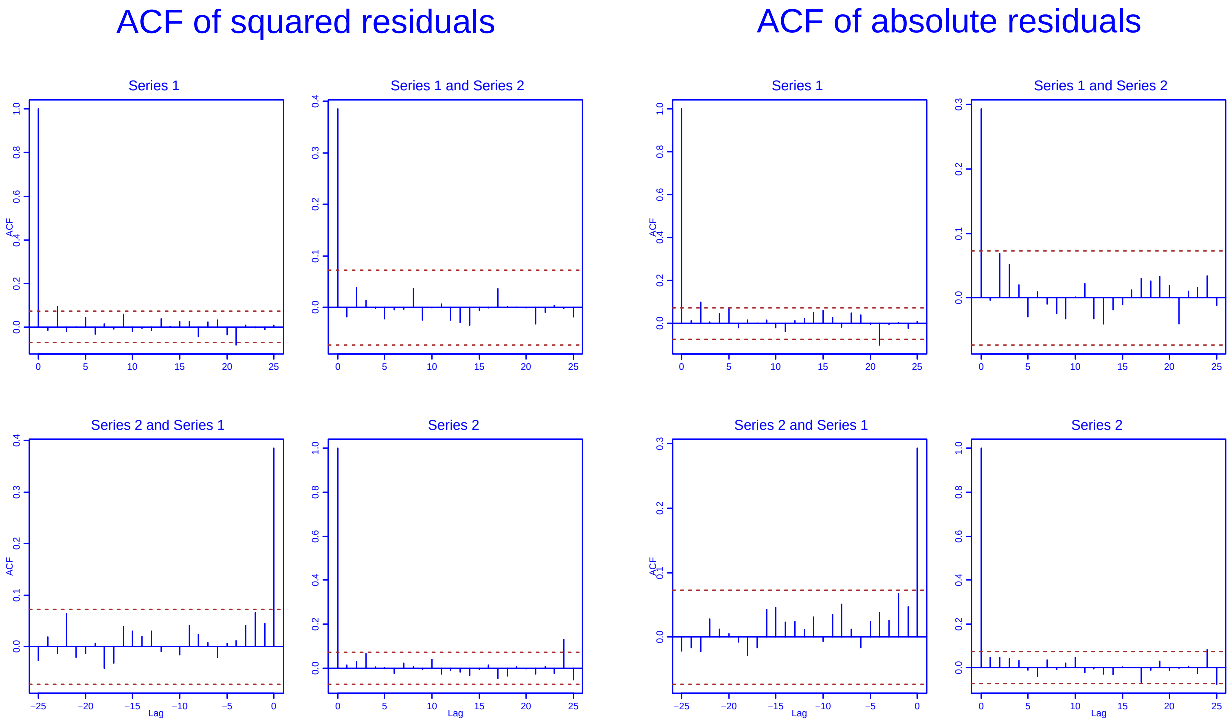

Figure 8: *The correlograms of squared and absulote homoscedastic compoments for the the S&P 500, Cisco and Intel data*

Lemma A.1 Suppose that Conditions (A.1)-(A.4) hold. Then, as $n \to \infty$,

$$\sup_{g \in \mathcal{G}} \|L_n(g)\| \to 0 \quad \text{in probability.}$$

Proof. The proof is a straightforward generalization of arguments laid out in Pollard (1984). For a given $\varepsilon > 0$, let $[l_1, u_1], \ldots, [l_N, u_N]$ denote the $N = N_B(\varepsilon, \mathcal{G}, D)$ brackets which cover $\mathcal{G}$. For a given $g \in \mathcal{G}$, let $l(g) \in \{l_1, \ldots, l_N\}$ and $u(g) \in \{u_1, \ldots, u_N\}$ be such that $l(g) \leq g \leq u(g)$. Using Conditions (A.3) and (A.4), we have

$$\begin{aligned}
Ł_n(g) = L_n^*(g) - EL_n^*(g) &\leq L_n^*(u(g)) - EL_n^*(g) \\
&\leq L_n^*(u(g)) - EL_n^*(u(g)) + [EL_n^*(u(g)) - EL_n^*(g)] \\
&\leq L_n^*(u(g)) - EL_n^*(u(g)) + C\varepsilon^\alpha \\
&= L_n(u(g)) + C\varepsilon^\alpha.
\end{aligned}$$

Similarly, we obtain the lower estimate

$$Ł_n(g) \geq L_n^*(l(g)) - EL_n^*(l(g)) - C\varepsilon^\alpha = L_n(l(g)) - C\varepsilon^\alpha.$$

Consequently, for every given ε, we have

$$\sup_{g \in \mathcal{G}} \|L_n(g)\| \leq \max\{\max_{j=1,\ldots,N} \|L_n(l_j)\|, \max_{j=1,\ldots,N} \|L_n(u_j)\|\} + C\varepsilon^\alpha.$$

Since N is a fixed finite number for a given ε by (A.1), by using (A.2), it follows that the right hand side in the above inequality is less than $2C\varepsilon^\alpha$ when n is large enough.

Remark A.1 If the convergence in (A.1) is almost sure, then the convergence in the result of Lemma A.1 is also almost sure.

Remark A.2 If Condition (A.4) is changed to "for each n, $L_n^*(g)$ is nonincreasing in g", the result of Lemma A.1 still holds.

Lemma A.2 Let $\{\mathbf{Y}_t\}$ be a time series and $\mathcal{C}$ be the class of (closed) convex sets in $\mathcal{R}^d$. Under Assumption 2 and Assumption 4, in probability,

$$\frac{1}{n-k_0} \sum_{t=k_0+1}^{n} \{\mathbf{Y}_t \mathbf{Y}_t^\tau I(\mathbf{Y}_{t-k} \in C) - E[\mathbf{Y}_t \mathbf{Y}_t^\tau I(\mathbf{Y}_{t-k} \in C)]\} \to 0, \tag{A.5}$$

$$\frac{1}{n-k_0} \sum_{t=k_0+1}^{n} \{I(\mathbf{Y}_{t-k} \in C) - EI(\mathbf{Y}_{t-k} \in C)]\} \to 0 \tag{A.6}$$

uniformly in $C \in \mathcal{C}$. Furthermore, if the mixing coefficients $\varphi(m)$ in Assumption 2 satisfy

$$\varphi(m) = \begin{cases} O(m^{-\frac{b}{2b-2}-\delta}), & \text{if } 1 < b < 2, \\ O(m^{-\frac{2}{b}-\delta}), & \text{if } b \geq 2, \end{cases} \tag{A.7}$$

where $\delta > 0$ is a constant, then the convergence in (A.5) and (A.6) is also almost sure.

Proof. We only prove (A.5). (A.6) can be proved in the same way.

Define a process $L_n^*(C)$ as in (3.5) in the proof of Theorem 1, i.e.

$$L_n^*(C) = \frac{1}{n-k_0}\sum_{t=k_0+1}^{n}(\mathbf{Y}_t\mathbf{Y}_t^\tau)I(\mathbf{Y}_{t-k}\in C).$$

Define

$$L_n^{\pm}(C) = \frac{1}{n-k_0}\sum_{t=k_0+1}^{n}(\mathbf{Y}_t\mathbf{Y}_t^\tau)^{\pm}I(\mathbf{Y}_{t-k}\in C).$$

Then, $L_n^*(C) = L_n^+(C) - L_n^-(C)$. Here, we use the conventional denotation: x^+ denotes $\max\{x,0\}$ and x^- denotes $\max\{0,-x\} = -\min\{0,x\}$ for a number x, and $M^{\pm}$ is defined in componentwise way for a matrix M. Note that $\{L_n^*(C), C \in \mathcal{C}\}$ is a process indexed by a class of indicator functions. We will apply Lemma A.1 to $L_n^+(C)$ and $L_n^-(C)$ respectively by checking conditions (A.1)-(A.4) one by one.

Obviously, (A.4) holds for $L_n^{\pm}(C)$.

To see that condition (A.3) holds, observing that with $p,q > 0$ such that $1/p + 1/q = 1$ and $p < 1 + \delta/2$ (with δ from Assumption 2), we have the following componentwise inequalities

$$\begin{aligned}
|EL_n^{\pm(i,j)}(C_1) - EL_n^{\pm(i,j)}(C_2)| &= |\frac{1}{n-k_0}\sum_{t=k_0+1}^{n} E|(\mathbf{Y}_t\mathbf{Y}_t^\tau)^{\pm(i,j)}(I(\mathbf{Y}_{t-k}\in C_1) - I(Y_{t-k}\in C_2)| \\
&\le \frac{1}{n-k_0}\sum_{t=k_0+1}^{n} E|(\mathbf{Y}_t\mathbf{Y}_t^\tau)^{\pm(i,j)}(I(\mathbf{Y}_{t-k}\in C_1\triangle C_2)| \\
&\le \frac{1}{n-k_0}\sum_{t=k_0+1}^{n}\big(E|(\mathbf{Y}_t\mathbf{Y}_t^\tau|^p)^{(i,j)}\big)^{1/p}\big(EI(\mathbf{Y}_{t-k}\in C_1\triangle C_2)\big)^{1/q} \\
&\le c\big(F(C_1\triangle C_2)\big)^{1/q}
\end{aligned}$$

where $L_n^{\pm(i,j)}$ denotes the (i,j)th element of $L_n^{\pm}$, and c denotes a constant which may be different at different line. That is, (A.3) holds with $\alpha = 1/q$.

For (A.2), see Polonik (1997).

(A.1) follows from the law of large number for $\varphi-$mixing processes; see Theorem 8.1.1 of Lin and Lu (1997).

Therefore, by Lemma A.1 and Remark A.2, we have

$$\sup_{C\in\mathcal{C}}\|L_n^*(C) - EL_n^*(C)\| \le \sup_{C\in\mathcal{C}}\|L_n^+(C) - EL_n^+(C)\| + \sup_{C\in\mathcal{C}}\|L_n^-(C) - EL_n^-(C)\| \to 0$$

in probability.

The proof of almost sure convergence part follows immediately from Lemma A.1 by noticing that the almost sure version of condition (A.1) holds under additional assumption (A.7) by the result of Chen and Wu (1989).

References

Arcones, M.A. and Yu, B. (1994). Central limit theorems for empirical processes and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, **7**, 47-71.

Boussama, F. (1998). Ergodicity, mixing and estimation in GARCH models. *PhD dissertation*, University Paris 7.

Comte, F. and Lieberman, O. (2003). Asymptotic theory for multivariate GARCH Processes. *Journal of Multivariate Analysis*, **84**, 61-84.

Chen, X. R. and Wu, Y. H. (1989). Strong law for a mixing sequence. *Acta Math. Appl. Sinica*, **5**, 367-71.

item Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135-171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191-221.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, **51**, 1281-1304.

Ding, Z., Engle, R. and Granger, C. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83-106.

Engle, R. F. (2002). Dynamic conditional correlation – a simple class of multivariate GARCH models. *Journal of Business and Economic Statistics*, **20**, 339-350.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalised ARCH. *Econometric Theory*, **11**, 122-150.

Engle, R. F. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. *NBER Working Papers 8554*, National Bureau of Economic Research, Inc.

Engle, R. F., Ng, V.K. and Rothschild, M. (1990). Asset pricing with a factor ARCH covariance structure: empirical estimates for Treasury bills. *Journal of Econometrics*, **45**, 213-238.

Escanciano, J. C. (2007). Weak convergence of non-stationary multivariate marked processes with applications to martingale testing. *Journal of Multivariate Analysis*, **98**, 1321-1336.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric methods*, Springer-Verlag, New York.

Fan, J., Wang, M. and Yao, Q. (2008). Modelling multivariate volatilities via conditionally uncorrelated components. *Journal of the Royal Statistical Society, Series B*, **70**, 679?702.

Forni, M., Hallin, M., Lippi, M. and Reichin, L. (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics*, **82**, 540-554.

Forni, M., Hallin, M., Lippi, M. and Reichin, L. (2004). The generalized dynamic factor model: Consistency and rates. *Journal of Econometrics*, **119**, 231-255.

Forni, M. and Lippi, M. (2001). The generalized factor model: Representation theory. *Econometric Theory*, **17**, 1113-1141.

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

Hafner, C. M. and Preminger, A. (2009). Asymptotic theory for a factor GARCH model. *Econometric Theory*, **25**, 336-363.

Hall, P. and Yao, Q. (2003). Inference for ARCH and GARCH models. *Econometrica*, **71**, 285-317.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, **102**, 603-617.

Lin, W.-L. (1992). Alternative estimators for factor GARCH models – a Monte Carlo comparison. *Journal of Applied Econometrics*, **7**, 259-279.

Lin, Z. Y. and Lu, C. R. (1996). *Limit Theory for Mixing Dependent Random Variables.* Science Press/Kluwer Academic Publishers.

Magnus, J.R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* (Revised edition) Wiley, New York.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95, 365-379.

Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika*, **90**, 967-975.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C* (2nd edition). Cambridge University Press, Cambridge.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and Their Applications.* **69**, 1-24.

Sargent, T. J. and Sims, C. A. (1977). Business cycle modelling without pretending to have too much a priori economic theory. In C. A. Sims (ed.), *New Methods in Business Cycle Research*, pp. 45-109. Minneapolis: Federal Reserve Bank of Minneapolis.

Tsay, R. (2001). *Analysis of Financial Time Series.* Wiley, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.* **22**, 94-116.