Modeling Multivariate Volatilities via Conditionally Uncorrelated Components

Jianqing Fan[1,2] Mingjin Wang[3] Qiwei Yao[2,3]

[1]Benheim Center of Finance and
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544, USA

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

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

Abstract

We propose to model multivariate volatility processes based on the newly defined conditionally uncorrelated components (CUCs). This model represents a parsimonious representation for matrix-valued processes. It is flexible in the sense that each CUC may be fitted separately with any appropriate univariate volatility model. Computationally it splits one high-dimensional optimization problem into several lower-dimensional subproblems. Consistency for the estimated CUCs has been established. A bootstrap method is proposed for testing the existence of CUCs. The proposed methodology is illustrated with both simulated and real data sets.

Key words: bootstrap test, causality in variance, dimension reduction, extended GARCH(1,1), financial returns, portfolio volatility, quasi-maximum likelihood estimator, time series.

1 Introduction

One of the most prolific areas of research in financial econometrics literature in the last two decades is to model time-varying volatility of financial returns. Many statistical models, most designed for univariate data, have been proposed for this purpose. From a practical point of view, there are at least two incentives to model several financial returns jointly. First, time-varying correlations among different securities are important and useful information for portfolio optimization, asset pricing and risk management. Secondly, models for a single security may be improved by incorporating the relevant information in other related ones. The quest for modeling multivariate processes, which are often represented by conditional covariance matrices, has motivated the attempts to extend univariate volatility models to multivariate cases, aiming for practical and/or statistical effectiveness. We list some of the endeavors below.

Let $\{\mathbf{X}_t\}$ be a vector-valued (return) time series with

$$E(\mathbf{X}_t|\mathcal{F}_{t-1}) = 0, \qquad \mathrm{Var}(\mathbf{X}_t|\mathcal{F}_{t-1}) = \mathbf{\Sigma}_t \equiv \big(\sigma_{t,ij}\big),$$

where $\mathcal{F}_t$ is the σ-algebra generated by $\{\mathbf{X}_t, \mathbf{X}_{t-1}, \cdots\}$, and $\mathbf{\Sigma}_t$ is an $\mathcal{F}_{t-1}$-measurable $d \times d$ semi-positive definite matrix. One of the most general multivariate GARCH(p, q) models is the BEKK representation (Engle and Kroner 1995)

$$\mathbf{\Sigma}_t = \mathbf{C} + \sum_{i=1}^{p}\sum_{j=1}^{m} \mathbf{A}_{ij}\mathbf{X}_{t-i}\mathbf{X}_{t-i}^{\tau}\mathbf{A}_{ij}^{\tau} + \sum_{i=1}^{q}\sum_{j=1}^{m} \mathbf{B}_{ij}\mathbf{\Sigma}_{t-i}\mathbf{B}_{ij}^{\tau}, \tag{1.1}$$

where $\mathbf{C}, \mathbf{A}_{ij}, \mathbf{B}_{ij}$ are $d \times d$ matrices, and $\mathbf{C}$ is positive definite. Although the form of the above model is quite general especially when m is reasonably large (Proposition 2.2 of Engle and Kroner 1995), it suffers from overparametrization. Similar to multivariate ARMA models, not all parameters in model (1.1) are necessarily identifiable even when $m = 1$. Overparametrization will also lead to a flat likelihood function, making statistical inference intrinsically difficult and computationally troublesome (Engle and Kroner 1995, and Jerez, Casals and Sotoca 2001).

To overcome the difficulties due to overparametrization, a dynamic conditional correlation (DCC) model (Engle 2002, Engle and Sheppard 2001) has been proposed. It is based on the decomposition

$$\mathbf{\Sigma}_t = \mathbf{D}_t\mathbf{R}_t\mathbf{D}_t, \tag{1.2}$$

where $\mathbf{D}_t = \mathrm{diag}(\sigma_{t,11}^{1/2}, \cdots, \sigma_{t,dd}^{1/2})$, $\sigma_{t,ii}$ is the conditional variance of the i-th component of $\mathbf{X}_t$, and $\mathbf{R}_t \equiv (\rho_{t,ij})$ is the conditional correlation matrix. A simple way to facilitate such a model is

to model each $\sigma_{t,ii}$ with a univariate volatility model and to model conditional correlation using a rolling exponential smoothing as follows

$$\rho_{t,ij} = \sum_{k=1}^{t-1} (\lambda_i \lambda_j)^{k/2} \varepsilon_{t-k,i} \varepsilon_{t-k,j} \Big/ \Big\{ \sum_{k=1}^{t-1} \lambda_i^k \varepsilon_{t-k,i}^2 \sum_{k=1}^{t-1} \lambda_j^k \varepsilon_{t-k,j}^2 \Big\}^{1/2},$$

where $\varepsilon_{ti} = X_{ti}/\sigma_{t,ii}^{1/2}$ and $\lambda_i, \lambda_j \in (0,1)$ are constants. Even with such a simple specification, the estimation typically involves solving a high-dimensional optimization problem as, for example, the Gaussian likelihood function cannot be factorized into several lower-dimensional functions. To overcome the computational difficulty, Engle (2002) proposes a two-step estimation procedure as follows: first fit each $\sigma_{t,ii}$ in (1.2) with a univariate GARCH(1,1) model, and then model the conditional correlation matrix $\mathbf{R}_t$ by a simple GARCH(1,1) form

$$\mathbf{R}_t = \mathbf{R}(1 - \theta_1 - \theta_2) + \theta_1(\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}') + \theta_2 \mathbf{R}_{t-1}, \tag{1.3}$$

and $\boldsymbol{\varepsilon}_t$ is a $d \times 1$ vector of the standardized residuals obtained in the separate GARCH(1,1) fittings for the d components of $\mathbf{X}_t$, and $\mathbf{R}$ is the sample correlation matrix of $\boldsymbol{\varepsilon}_t$. Note there are only two unknown parameters θ_1, θ_2 in model (1.3), so it can be easily implemented even for large or very large d. However it may not provide adequate fitting when the components of $\mathbf{X}_t$ exhibit different dynamic correlation structures; see the real data examples in section 4 below. Indeed, the conditional correlation matrix in (1.3) is a linear combination of the static sample correlation matrix $\mathbf{R}$ and the exponential smoothing of $\{\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}'\}$, which is a nonparametric estimator. When $\theta_1 + \theta_2 = 1$, it is a pure nonparametric (exponential smoothing) estimator. The biases are inevitable in such an estimation for the conditional correlation.

Alexander (2001) proposes an orthogonal GARCH model which fits each principal component (PC) with a univariate GARCH model separately, and treats all PCs as *conditionally* uncorrelated random variables. Since PCs are only unconditionally uncorrelated, such a misspecification may lead to non-negligible errors in the fitting; see the first example in section 4.

Other multivariate volatility models include, for example, vectorized multivariate GARCH models of Bollerslev, Engle and Wooldridge (1988), constant conditional correlation multivariate GARCH models of Bollerslev (1990), a multivariate stochastic volatility model of Harvey, Ruiz and Shephard (1994), a generalized orthogonal GARCH models of van der Weide (2002), an easy-to-fit ad hoc approach of Wang and Yao (2005), a hidden Markov switching model of Pelletier (2006); see also a survey in Bauwens, Laurent and Rombouts (2006) and the references therein.

In this paper, we propose a new alternative for modeling multivariate volatilities. The basic idea is to assume that $\mathbf{X}_t$ is a linear combination of a set of *conditionally uncorrelated components* (CUCs); see section 2.1 below. One fundamental difference from the orthogonal GARCH model is that we use CUCs, instead of PCs, which are genuinely conditionally uncorrelated. The advantages of the new approach include: (i) the CUC decomposition leads to a parsimonious and identifiable representation, and the number of parameters in the model is significantly reduced comparing to, for example, the BEKK representation or the vectorized multivariate GARCH models, (ii) it has the flexibility to model each CUC separately with any appropriate univariate volatility models, (iii) computationally it splits a high-dimensional optimization problem into several lower-dimensional subproblems, and (iv) it allows the volatility model for one CUC to depend on the lagged value of the other CUCs. However the estimation of CUCs involves solving a nonlinear optimization problem with $d(d-1)/2$ variables, where d is the dimension of $\mathbf{X}_t$. This poses some limitation on the dimensionality d with the available computing capacity. We view the CUC as a model capable to catch sophisticated dynamical correlation structures, but its potential may only be fully capitalized with further development in computing power and/or high-dimensional optimization algorithms.

The idea of using CUCs is similar to the so-called independent component analysis (Hyvärinen, Karhunen and Oja 2001). However instead of requiring all the component series are independent with each other, we only impose a weaker condition that the component series are conditionally uncorrelated; see (2.1) below. This relaxation is critical for the problem concerned in this paper. Of course like independent components, CUCs may not always exist. We propose a bootstrap test to assess the existence of CUCs. Our empirical experience indicates that for a large number of practical examples with small or moderately large d, there is no significant evidence to reject the hypothesis on the existence of CUCs.

Literature on applying independent components analysis to financial and economic time series includes, for example, Back and Weigend (1997), Kiviluoto and Oja (1998), Mălăroiu, Kiviluoto and Oja (2000), and van der Weide (2002). Although our basic idea is somehow similar to van der Weide (2002) which deals with Gaussian innovation models only, our approach is completely different; we separate the estimation for the CUCs from fitting the volatility models for the CUCs. In fact, fitting each CUC becomes a univariate volatility modeling problem.

The rest of the paper is organized as follows. Section 2 contains a detailed description of the

proposed new methodology and the associated theoretical results. Simulation results are reported in section 3. Illustration with two real data examples of dimension, respectively, $d = 4$ and $d = 10$ is presented in section 4. Applicability of the CUC method beyond its standard setting is discussed in section 5. Technical proofs are relegated to the Appendix.

2 Methodology

2.1 Basic setting

To simplify the matter concerned, we assume $\mathrm{Var}(\mathbf{X}_t) = \mathbf{I}_d$ — the $d \times d$ identity matrix. In practice, this amounts to replacing $\mathbf{X}_t$ by $\mathbf{S}^{-1/2}\mathbf{X}_t$, where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{X}_t$. We assume that each component of $\mathbf{X}_t$ is a linear combination of d conditionally uncorrelated components (CUCs) $Z_{t1}, \cdots, Z_{td}$ which satisfy the conditions $E(Z_{ti}|\mathcal{F}_{t-1}) = 0$, $\mathrm{Var}(Z_{ti}) = 1$, and

$$E(Z_{ti}Z_{tj}|\mathcal{F}_{t-1}) = 0, \quad \text{for all } i \neq j. \tag{2.1}$$

Put $\mathbf{Z}_t = (Z_{t1}, \cdots, Z_{td})^\tau$. The above setting implies that

$$\mathbf{X}_t = \mathbf{A}\mathbf{Z}_t, \tag{2.2}$$

for a constant matrix $\mathbf{A}$. Necessarily, $\mathrm{Var}(\mathbf{X}_t) = \mathbf{A}^\tau \mathrm{Var}(\mathbf{Z}_t)\mathbf{A} = \mathbf{A}\mathbf{A}^\tau = \mathbf{I}_d$. Hence, $\mathbf{A}$ is a $d \times d$ orthogonal matrix with $\frac{d}{2}(d-1)$ free elements and $\mathbf{Z}_t = \mathbf{A}^\tau \mathbf{X}_t$. Put

$$\mathrm{Var}(\mathbf{Z}_t|\mathcal{F}_{t-1}) = \mathrm{diag}(\sigma_{t1}^2, \cdots, \sigma_{td}^2), \tag{2.3}$$

i.e. $\sigma_{tj}^2 = \mathrm{Var}(Z_{tj}|\mathcal{F}_{t-1})$. It is easy to see that once we have specified σ_{tj}^2 – the volatility of the j-th CUC, for $j = 1, \cdots, d$, volatilities for any portfolios can be deduced accordingly. For example, for any portfolios $\xi_t = \mathbf{b}_1^\tau \mathbf{X}_t$ and $\eta_t = \mathbf{b}_2^\tau \mathbf{X}_t$ it holds that

$$\mathrm{Var}(\xi_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{d} b_{j1}^2 \, \sigma_{tj}^2, \qquad \mathrm{Cov}(\xi_t, \eta_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{d} b_{j1}b_{j2} \, \sigma_{tj}^2.$$

where $(b_{1j}, \cdots, b_{dj}) = \mathbf{b}_j^\tau \mathbf{A}$ $(j = 1, 2)$. Hence, the CUC decomposition (2.2) facilitates a parsimonious modeling for d-dimensional multivariate volatility process via d univariate volatility models. In this way, we reduce the number of parameters involved substantially.

The assumption that $\mathrm{Var}(\mathbf{X}_t) = \mathbf{I}_d$ is not essential. It is introduced to reduce the free parameters in $\mathbf{A}$ from d^2 to $d(d-1)/2$. This is similar to the independent component analysis which

performs a principal component analysis to reduce a d^2-dimensional optimization problem to a $d(d-1)/2$-dimensional one; see, for example, section 7.4 of Hyvärinen, Karhunen and Oja (2001), and also section 2.2.1 below.

2.2 Estimation of CUCs

2.2.1 Estimation for A

By (2.2), $Z_{tj} = \mathbf{a}_j^\tau \mathbf{X}_t$, and $\mathbf{a}_1, \cdots, \mathbf{a}_d$ are d orthogonal vectors. The goal is to estimate the orthogonal matrix $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_d)$. Condition (2.1) is equivalent to

$$\sum_{B \in \mathcal{B}_t} \big| E\{Z_{ti} Z_{tj} I(B)\} \big| = 0 \tag{2.4}$$

for any π-class $\mathcal{B}_t \subset \mathcal{F}_{t-1}$ such that the σ-algebra generated by $\mathcal{B}_t$ is equal to $\mathcal{F}_{t-1}$ (Theorem 7.1.1 of Chow and Teicher, 1997). In practice, we use some simple $\mathcal{B}_t$ for the sake of tractability. This leads to choosing an orthogonal matrix $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_d)^\tau$ which minimizes

$$\Psi_n(\mathbf{A}) \equiv \sum_{1 \le i < j \le d} \sum_{B \in \mathcal{B}} w(B) \sum_{k=1}^{k_0} \frac{1}{n-k} \Big| \mathbf{a}_i^\tau \Big\{ \sum_{t=k+1}^{n} \mathbf{X}_t \mathbf{X}_t^\tau I(\mathbf{X}_{t-k} \in B) \Big\} \mathbf{a}_j \Big|, \tag{2.5}$$

where $k_0 \ge 1$ is a prescribed integer, $\mathcal{B}$ consists of countable subsets in $\mathcal{R}^d$, and $w(\cdot)$ is a weight function such that $\sum_{B \in \mathcal{B}} w(B) < \infty$. We denote by $\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \cdots, \widehat{\mathbf{a}}_d)^\tau$ the resulting estimator.

Note that the order of $\mathbf{a}_1, \cdots, \mathbf{a}_d$ is arbitrary and $\mathbf{a}_i$ may be replaced by $-\mathbf{a}_i$. Therefore we measure the estimation error by

$$D(\widehat{\mathbf{A}},\ \mathbf{A}) = 1 - \frac{1}{d} \sum_{i=1}^{d} \max_{1 \le j \le d} |\mathbf{a}_i^\tau \widehat{\mathbf{a}}_j|. \tag{2.6}$$

Note that for any orthogonal matrices $\mathbf{A}$ and $\mathbf{B}$, $D(\mathbf{A}, \mathbf{B}) \ge 0$. Furthermore, if the columns of $\mathbf{A}$ are obtained from a permutation of the columns of $\mathbf{B}$ or their reflections, $D(\mathbf{A}, \mathbf{B}) = 0$.

In practice, we may let $\mathcal{B}$ be the collection of all the balls centered at the origin in $\mathcal{R}^d$. Note $E\mathbf{X}_t = 0$ and $\mathrm{Var}(\mathbf{X}_t) = \mathbf{I}_d$. When the distribution of $\mathbf{X}_t$ is spherically symmetric and unimodal, $\mathcal{B}$ is the collection of the minimum volume sets which determine the distribution of $\mathbf{X}_t$ (Polonik, 1997). With any given n observations, effectively such a $\mathcal{B}$ consists of $\{\mathbf{x} \in \mathcal{R}^d \,|\, ||\mathbf{x}|| \le ||\mathbf{X}_t||\}$ for $t = 1, \cdots, n$, and therefore has at most n different members. Hence we may let $w(B) = 1/n$.

To overcome the difficulties in handling the constraint $\mathbf{A}^\tau \mathbf{A} = \mathbf{I}_d$ in solving the above optimization problem, we parametrize $\mathbf{A}$ as follows:

$$\mathbf{A} = \prod_{1 \le i < j \le d} \mathbf{\Gamma}_{ij}(\varphi_{ij}), \tag{2.7}$$

where $\mathbf{\Gamma}_{ij}(\varphi_{ij})$ is obtained from the identity matrix $\mathbf{I}_d$ with the following replacements: both the (i,i)-th and the (j,j)-th elements are replaced by $\cos\varphi_{ij}$, the (i,j)-th and the (j,i)-th elements are replaced, respectively, by $\sin\varphi_{ij}$ and $-\sin\varphi_{ij}$ (Vilenkin 1968, van der Weide 2002). Obviously $\mathbf{\Gamma}_{ij}(\varphi_{ij})$ is an orthogonal matrix, so is $\mathbf{A}$ given in (2.7). Writing $\mathbf{A}$ in (2.2) in the form of (2.7), the constrained minimization of (2.5) over orthogonal $\mathbf{A}$ is transformed to an unconstrained minimization problem over a $\frac{d(d-1)}{2} \times 1$ vector $\boldsymbol{\varphi} = (\varphi_{12}, \varphi_{13}, \cdots, \varphi_{1d}, \varphi_{23}, \cdots, \varphi_{d-1,d})^\tau$. This minimization problem is typically solved by iterative algorithms. We stop the iteration when $D(\mathbf{A}_k, \mathbf{A}_{k+1})$ is smaller than a prescribed small constant, where $\mathbf{A}_k$ denotes the value of $\mathbf{A}$ in the k-th iteration. Note that $\Psi_n(\mathbf{A}) = \Psi_n(\mathbf{B})$ for any orthogonal $\mathbf{A}$ and $\mathbf{B}$ with $D(\mathbf{A}, \mathbf{B}) = 0$.

2.2.2 Asymptotic properties

Let

$$\Psi(\mathbf{A}) \equiv \sum_{1\le i<j\le d} \sum_{B\in\mathcal{B}} w(B) \sum_{k=1}^{k_0} \big|E\{\mathbf{a}_i^\tau \mathbf{X}_t \mathbf{X}_t^\tau \mathbf{a}_j I(\mathbf{X}_{t-k} \in B)\}\big|. \tag{2.8}$$

Theorem 1 below states that the estimator $\widehat{\mathbf{A}}$ is consistent under the regularity conditions (A1) – (A5) listed in the Appendix. Note that Theorem 1 does not require the condition that the CUCs exist. Instead condition (A3) only assumes that there exists an $\mathbf{A}_0$ which is a unique minimizer, under D-distance, of $\Psi(\mathbf{A})$. Since function Ψ_n cannot tell any difference between orthogonal $\mathbf{A}$ and $\mathbf{B}$ as long as $D(\mathbf{A}, \mathbf{B}) = 0$, we call $\widehat{\mathbf{A}}$ a consistent estimator of $\mathbf{A}_0$ if the D-distance between $\widehat{\mathbf{A}}$ and $\mathbf{A}_0$ converges to 0 in probability.

Theorem 1. Let $k_0 \ge 1$ be a fixed integer. Under conditions (A1)–(A3), $D(\widehat{\mathbf{A}}, \mathbf{A}_0) \to 0$ in probability as $n \to \infty$. If, in addition, condition (A4) holds, it holds that for any orthogonal $\mathbf{A}$

$$\Psi_n(\mathbf{A}) - \Psi(\mathbf{A}) = O_P(n^{-1/2}).$$

Furthermore, $n^{1/2} D(\widehat{\mathbf{A}}, \mathbf{A}_0) = O_P(1)$ provided that, in addition, condition (A5) also holds.

When the CUCs exist, $\Psi(\mathbf{A}_0) = 0$. On the other hand, when the CUCs do not exist, $\Psi(\mathbf{A}_0) \neq 0$ and $\mathbf{A}_0$ may now depend on the choice of $\mathcal{B}$. In this case, we naturally seek for an orthogonal transform such that the resulting components are the least conditionally correlated. Note that $\Psi(\cdot)$ defined in (2.8) may be written as

$$\Psi(\mathbf{A}) = \sum_{1\le i<j\le d} \sum_{B\in\mathcal{B}} w(B) \sum_{k=1}^{k_0} \big|\mathrm{Corr}(\mathbf{a}_i^\tau \mathbf{X}_t, \mathbf{a}_j^T \mathbf{X}_t | \mathbf{X}_{t-k} \in B)\big| P(\mathbf{X}_{t-k} \in B). \tag{2.9}$$

We view $\Psi(\mathbf{A})$ as a collective conditional correlation measure among the d direction $\mathbf{a}_1, \cdots, \mathbf{a}_d$. Thus, our criterion may be seen as to find an orthogonal transform $\mathbf{A}$ to minimizer $\Psi(\mathbf{A})$. (See also the discussion in section 5 below.) Theorem 2 below indicates that asymptotically the transformed components along any other orthogonal matrix $\widehat{\mathbf{B}}$ lead to a higher collective conditional correlation, in terms of $\Psi(\cdot)$, than that along $\widehat{\mathbf{A}}$.

Theorem 2. Let $k_0 \geq 1$ be a fixed integer, and conditions (A1) and (A2) hold. Then for any other orthogonal transform $\widehat{\mathbf{B}}$,

$$\limsup\{\Psi(\widehat{\mathbf{A}}) - \Psi(\widehat{\mathbf{B}})\} \leq 0.$$

The proof of Theorem 1 is more involved and is presented in the Appendix. We omit the proof of Theorem 2 to save space.

2.3 Modelling volatilities for CUCs

Once the CUCs have been identified, we may fit each σ_{tj}^2 with an appropriate univariate volatility model such as GARCH or stochastic volatility model; see the survey by Shephard (1996). As a simple illustration, we establish below an extended GARCH(1,1) model for each σ_{ti}^2 given in (2.3).

2.3.1 Extended GARCH(1,1) models

We assume, for the j-th CUC, $j = 1, \cdots, d$,

$$Z_{tj} = \sigma_{tj}\varepsilon_{tj}, \qquad \sigma_{tj}^2 = \gamma_j + \sum_{i=1}^{d} \alpha_{ji} Z_{t-1,i}^2 + \beta_j \sigma_{t-1,j}^2, \tag{2.10}$$

where $\{\varepsilon_{tj},\ -\infty < t < \infty\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1, ε_{tj} is independent of $\mathcal{F}_{t-1}$, $\gamma_j > 0$ and $\alpha_j, \alpha_{ji}, \beta_j \geq 0$. To ensure $\mathrm{Var}(Z_{tj}) = 1$, we set $\gamma_j = 1 - \beta_j - \sum_{1 \leq i \leq d} \alpha_{ji}$. This model contains extra $d-1$ terms $\sum_{i \neq j} \alpha_{ji} Z_{t-1,i}^2$ from the standard GARCH(1,1) model, which incorporates the possible association between the j-th CUC and the other CUCs, while the conditional zero-correlation condition (2.1) still holds. Such dependence is termed as that the i-th component (if $\alpha_{ji} \neq 0$) is causal in variance to the j-th component (Granger, Robins and Engle 1984). Note under the specification (2.10), the CUC becomes a restricted form of the BEKK representation.

In practice, we expect that σ_{tj}^2 may depend on $Z_{t-1,i}^2$ only for a small number of i's, including $i = j$, i.e. many coefficients α_{ji} (for $i \neq j$) may be 0. Section 2.3.3 below outlines a data-analytic approach for building such a component-dependent model.

Model (2.10) may be viewed as a special case of the vectorized ARMA-GARCH model of which the conditions for stationary and ergodicity may be found in, for example, Ling and McAleer (2003). When $\beta_j \in [0, 1)$, (2.10) admits the representation

$$\begin{aligned} \sigma_{tj}^2 &= \mathrm{Var}(Z_{tj}|\mathcal{F}_{t-1}) = \frac{\gamma_j}{1-\beta_j} + \sum_{i=1}^{d} \alpha_{ji} \sum_{k=1}^{\infty} \beta_j^{k-1} Z_{t-k,\,i}^2 \\ &= 1 - \frac{\sum_{i=1}^{d} \alpha_{ji}}{1-\beta_j} + \sum_{i=1}^{d} \alpha_{ji} \sum_{k=1}^{\infty} \beta_j^{k-1} Z_{t-k,i}^2. \end{aligned} \tag{2.11}$$

2.3.2 quasi-MLE

To facilitate a likelihood estimation, let us assume hypothetically that ε_{tj} in (2.10) is standard normal. The implied (negative) twice log-likelihood function for $\boldsymbol{\theta}_j \equiv (\alpha_{j1}, \cdots, \alpha_{jd}, \beta_j)^\tau$ is

$$l_j(\boldsymbol{\theta}_j) = \sum_{t=\nu+1}^{n} \left\{ \log \sigma_{tj}(\boldsymbol{\theta}_j)^2 + Z_{tj}^2 / \sigma_{tj}(\boldsymbol{\theta}_j)^2 \right\}, \tag{2.12}$$

for a given integer $\nu \geq 1$, where $\sigma_{tj}(\boldsymbol{\theta}_j)^2 = \mathrm{Var}(Z_{tj}|\mathcal{F}_{t-1})$ is given by (2.11). The quasi-maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}_j$ minimizes (2.12). In practice, we let $Z_{ti} \equiv 0$ for all $t \leq 0$ on the right hand side of (2.11). The sum in (2.12) is taken from $t = \nu + 1$ to alleviate the effect of this truncation.

2.3.3 Selection of causal components

To obtain a parsimonious representation for σ_{tj}^2, we may select only those significant $Z_{t-1,i}$ on the RHS of the second equation in (2.10). This is particularly important when the number of components d is large. It may be achieved by using the ideas for variable selection in regression analysis. Below we outline an algorithm based on a combination of the stepwise addition method and the BIC criterion, which is particularly computationally effective. An obvious alternative is to adopt a forward search algorithm based on the statistical tests for the causality in variance (Cheung and Ng 1996, and Hafner and Herwartz 2006).

We start with the standard GARCH(1,1) model (i.e. $\alpha_{jj} \neq 0$ and $\alpha_{ji} = 0$ for $j \neq i$). We then add one more $Z_{t-1,i}$ each time which maximizes the (quasi-)likelihood. More precisely, suppose the model contains $(k-1)$ terms $Z_{t-1,j_1}, \cdots, Z_{t-1,j_{k-1}}$ already. We choose an additional term

$Z_{t-1,\ell}$ among $\ell \notin \{j, j_1, \cdots, j_{k-1}\}$ which maximizes the quasi-likelihood function. Note that this is a two-step maximization problem: For each given $\ell \notin \{j, j_1, \cdots, j_{k-1}\}$, we compute the qMLE $\widetilde{\boldsymbol{\theta}}_j^{(k)}$ for $\boldsymbol{\theta}_j^{(k)} \equiv (\alpha_{jj}, \alpha_{jj_1}, \cdots, \alpha_{j\ell}, \beta_j)^\tau$ with the constraints $\alpha_{ji} = 0$, for $i \notin \{j, j_1, \cdots, j_{k-1}, \ell\}$. We then choose an $\ell \notin \{j, j_1, \cdots, j_{k-1}\}$ to minimize $l_j(\widetilde{\boldsymbol{\theta}}_j^{(k)})$, and denote by $l_j(k)$ the minimum value and the index of the selected variable j_k. Put

$$\mathrm{BIC}_j(k) = l_j(k) + (k+2)\log(n-\nu).$$

We choose r_j which minimizes $\mathrm{BIC}_j(k)$ over $0 \le k \le d$. Note that $k = 0$ corresponds to the standard GARCH(1,1) for Z_{tj}.

2.3.4 LADE

It is well documented that qMLE $\widetilde{\boldsymbol{\theta}}_j$ suffers from complicated asymptotic distributions and slow convergence rates if ε_{tj} is heavy-tailed in the sense that $E(|\varepsilon_{tj}|^4) = \infty$ (Hall and Yao 2003, and section 7.3 of Straumann 2005). On the other hand, a least absolute deviation estimator based on a log-transformation is always asymptotically normal with the standard root-n convergence rate provided $E(\varepsilon_{tj}^2) < \infty$; see Peng and Yao (2003).

To construct the LADE with the constraint $\mathrm{Var}(Z_{tj}) = 1$, we write $\varepsilon_{tj} = v_0 e_{tj}$ in the first equation in (2.10), where the median of e_{tj}^2 is equal to 1 and $v_0 = 1/\mathrm{STD}(e_{tj})$. With $\sigma_{tj}(\boldsymbol{\theta}_j)^2$ expressed in (2.11), parameters $\boldsymbol{\theta}_j$ and v_0 are (jointly) identifiable. Now

$$\log Z_{tj}^2 - \log\{\sigma_{tj}(\boldsymbol{\theta}_j)^2\} - \log v_0^2 = \log(e_{tj}^2).$$

Since the median of $\log(e_{tj}^2)$ is 0, the true values of the parameters minimize

$$E\big|\log Z_{tj}^2 - \log\{\sigma_{tj}(\boldsymbol{\theta}_j)^2\} - \log v_0^2\big|.$$

Therefore we may estimate the parameters by minimizing

$$\sum_{t=\nu+1}^{n} \big|\log Z_{tj}^2 - \log\{\sigma_{tj}(\boldsymbol{\theta}_j)^2\} - \log v_0^2\big|, \tag{2.13}$$

where $\sigma_{tj}(\boldsymbol{\theta}_j)^2$ is given in (2.11), with the part of $a_{ji} = 0$ for the non-causal component in the variance. So far $\boldsymbol{\theta}_j$ and v_0 are treated as free parameters. The estimators obtained are root-n consistent.

To make an explicit use of the condition that $\mathrm{Var}(\varepsilon_{tj}) = 1$, we may estimate parameters $\boldsymbol{\theta}_j$ as follows. With the initial estimate $\widehat{\boldsymbol{\theta}}_j^{(0)}$, let $\widehat{v}_0$ be the reciprocal of the sample standard deviation of the residuals $\{\widetilde{\varepsilon}_{tj}\}$, where $\widetilde{\varepsilon}_{tj} = Z_{tj}/\{\sigma_{tj}(\boldsymbol{\theta}_j^{(0)})\}$. With the given $\widehat{v}_0$ and $\widehat{\boldsymbol{\theta}}_j^{(0)}$, we can minimize

$$\sum_{t=\nu+1}^{n} w_t \big[\log Z_{tj}^2 - \log\{\sigma_{tj}(\boldsymbol{\theta}_j)^2\} - \log \widehat{v}_0^2\big]^2,$$

where $w_t = |\log Z_{tj}^2 - \log\{\sigma_{tj}(\widehat{\boldsymbol{\theta}}_j^{(0)})^2\} - \log \widehat{v}_0^2|^{-1}$. We may update $\widehat{v}_0$ and iterate further until the estimated $\boldsymbol{\theta}_j$ converges. Note that we have used a weighted L_2 loss function to approximate the L_1 loss to expedite the computation.

2.4 Inference based on bootstrapping

A natural question for the proposed approach is whether the CUCs $Z_{t1}, \cdots, Z_{td}$ exist or not, although the minimizer $\{\widehat{\mathbf{a}}_j\}$ of (2.5) always exists. To address this issue statistically, we may construct a test for the null hypothesis

$$H_0 : \ \mathbf{X}_t = \mathbf{A}\mathbf{Z}_t \quad \text{and} \quad \mathbf{Z}_t = \mathrm{diag}(\sigma_{t1}, \cdots, \sigma_{td})\boldsymbol{\varepsilon}_t, \tag{2.14}$$

where $\mathbf{A}^\tau\mathbf{A} = \mathbf{I}_d$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \cdots, \varepsilon_{td})^\tau$, $\{\varepsilon_{t1}\}, \cdots, \{\varepsilon_{td}\}$ are d independent series, and each of them is a sequence of i.i.d. r.v.s with mean 0 and variance 1. Note that the null hypothesis above is a sufficient but not necessary condition for the existence of CUCs. The independence condition is required to construct a bootstrap estimation of null distribution. Also note that $Z_{t1}, \cdots, Z_{td}$ may not be independent with each other.

When Z_{ti} and Z_{tj} are not conditionally uncorrelated, the left hand side of (2.4) is equal to positive constant instead of 0. Therefore, *large* values of $\Psi_n(\widehat{\mathbf{A}})$ indicate that the CUCs do not exist. We adopt a bootstrap method below to assess how large is large enough to reject H_0.

If the null hypothesis H_0 could not be rejected, we may also construct confidence sets for the coefficients $\mathbf{a}_j$ (i.e. the columns of $\mathbf{A}$) of the CUCs, and the parameters $\boldsymbol{\theta}_j$ based on the same bootstrap scheme. Formally confidence sets for $\boldsymbol{\theta}_j$ could also be constructed based on asymptotic distributions of, for example, the LADE $\widehat{\boldsymbol{\theta}}_j$, which may be derived in the similar manner of Peng and Yao (2003). However such an approach is based on the assumption that the CUCs are known (i.e. the vectors $\mathbf{a}_j$ are known), and, therefore, fails to take into account of the errors due to the estimation for $\mathbf{a}_j$.

Let $\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \cdots, \widehat{\mathbf{a}}_d)$ be the estimator derived from minimizing (2.5). Let $Z_{tj} = \widehat{\mathbf{a}}_j^\tau \mathbf{X}_t$ and $\widehat{\boldsymbol{\theta}}_j$ be an estimator for $\boldsymbol{\theta}_j$.

The bootstrap sampling scheme consists of the three steps below.

(i) For $j = 1, \cdots, d$, draw ε_{tj}^*, for $-\infty < t \le n$, by sampling randomly with replacement from the standardized residuals $\{\widehat{\varepsilon}_{\nu+1,j}, \cdots, \widehat{\varepsilon}_{nj}\}$ which are obtained from standardizing the raw residuals

$$Z_{tj}/\sigma_{tj}(\widehat{\boldsymbol{\theta}}_j), \qquad t = \nu + 1, \cdots, n.$$

(ii) For $j = 1, \cdots, d$, draw $Z_{tj}^* = \sigma_{tj}^* \varepsilon_{tj}^*$, for $-\infty < t \le n$, where

$$(\sigma_{tj}^*)^2 = 1 - \widehat{\beta}_j - \sum_{i=1}^{d} \widehat{\alpha}_{ji} + \sum_{i=1}^{d} \widehat{\alpha}_{ji} (Z_{t-1,i}^*)^2 + \widehat{\beta}_j (\sigma_{t-1,j}^*)^2.$$

(iii) Let $\mathbf{X}_t^* = \widehat{\mathbf{A}}(Z_{t1}^*, \cdots, Z_{td}^*)^\tau$ for $t = 1, \cdots, n$.

A test for the existence of the CUCs: Let $\Psi_n^*(\mathbf{A})$ be defined as in (2.5) with $\{\mathbf{X}_t\}$ replaced by $\{\mathbf{X}_t^*\}$, and the bootstrap estimator $\mathbf{A}^* = (\mathbf{a}_1^*, \cdots, \mathbf{a}_d^*)$ be computed in the same manner as $\widehat{\mathbf{A}}$ with Ψ_n replaced by Ψ_n^*. Note that the bootstrap sample $\{\mathbf{X}_t^*\}$ is drawn from the model with $\widehat{\mathbf{a}}_j^\tau \mathbf{X}_t^*$ as its *genuine* CUCs. Hence the conditional distribution of $\Psi_n^*(\mathbf{A}^*)$ (given the original sample $\{\mathbf{X}_t\}$) may be taken as an approximation for the distribution of $\Psi_n(\widehat{\mathbf{A}})$ under H_0. Thus we reject H_0 if $\Psi_n(\widehat{\mathbf{A}})$ is greater than the $[B\alpha]$-th largest value of $\Psi_n^*(\mathbf{A}^*)$ in a replication of the above bootstrap resampling for B times, where $\alpha \in (0, 1)$ is the size of the test and B is a large integer.

Confidence sets for $\mathbf{A}$: A bootstrap approximation for an $(1 - \alpha)$ confidence set of the transformation matrix $\mathbf{A}$ can be constructed as

$$\{\mathbf{A} \,\big|\, D(\mathbf{A}; \widehat{\mathbf{A}}) \le c_\alpha, \mathbf{A}^\tau \mathbf{A} = \mathbf{I}_d\}, \tag{2.15}$$

where c_α is the $[B\alpha]$-th largest value of $D(\mathbf{A}^*; \widehat{\mathbf{A}})$ in a replication of the bootstrap resampling for B times. Note that when $\mathbf{A}$ is in the confidence set, so is $\mathbf{B}$ if the columns of $\mathbf{B}$ form a permutation of the (reflected) columns of $\mathbf{A}$; see (2.6).

Interval estimators for the components of $\widehat{\boldsymbol{\theta}}_j$: A bootstrap confidence interval for any component, say, β_j of $\boldsymbol{\theta}_j$ may be obtained as follows. Repeat the above bootstrap sampling B times for some large integer B, resulting in bootstrap estimates $\beta_{j1}^*, \cdots, \beta_{jB}^*$. An approximate $(1-\alpha)$ confidence

interval for β_j is $(\beta^*_{j(b_1)},\ \beta^*_{j(b_2)})$, where $\beta^*_{j(i)}$ denotes the i-th smallest value among $\beta^*_{j1}, \cdots, \beta^*_{jB}$, and $b_1 = [B\alpha/2]$ and $b_2 = [B(1-\alpha/2)]$.

We have adopted the standard bootstrap procedure above. On the other hand, the wild bootstrap method has proved to be effective for the inference for mean functions in the presence of heteroscedastic noise (Wu 1986, Mammen 1993, and Hafner and Herwartz 2000). It is an interesting open question how to adapt the wild bootstrap idea for the inference on conditional second moments.

3 Simulation

We conduct a Monte Carlo experiment to illustrate the proposed CUC approach. In particular we check the accuracy of the estimation for the transformation matrix $\mathbf{A}$ in (2.2).

We consider a CUC extended GARCH(1,1) model with $d = 3$:

$$\mathbf{X}_t = \mathbf{A}\mathbf{Z}_t, \qquad \mathbf{Z}_t|\mathcal{F}_{t-1} \ \sim \ N(0,\ \mathrm{diag}\{\sigma^2_{t,1}, \sigma^2_{t,2}, \sigma^2_{t,3}\}), \tag{3.1}$$

where $\sigma^2_{t,i} = \gamma_i + \beta_i\sigma^2_{t-1,i} + \alpha_{i1}Z^2_{t-1,1} + \alpha_{i2}Z^2_{t-1,2} + \alpha_{i3}Z^2_{t-1,3}$, $i = 1, 2, 3$, and

$\mathbf{A}$			i	γ_i	β_i	α_{i1}	α_{i2}	α_{i3}
0	0.500	0.866	1	0.02	0.90	0.04	0	0.04
0	0.866	-0.500	2	0.10	0.80	0	0.10	0
-1	0	0	3	0.28	0.60	0	0	0.12

It is easy to see that $\mathbf{A}^\tau\mathbf{A} = \mathbf{I}_3$ and $\gamma_i = 1 - \alpha_{i1} - \alpha_{i2} - \alpha_{i3} - \beta_i$. Thus the variances of the CUCs are 1. Since $\alpha_{11} + \alpha_{12} + \alpha_{13} + \beta_1 = 0.98$, the volatility for the first CUC is highly persistent. On the contrary, the volatility persistence in the third component is less pronounced, as $\alpha_{31} + \alpha_{32} + \alpha_{33} + \beta_3 = 0.72$ only.

For each of 800 samples with size $n = 500$ or 1000 generated from the above model, we estimated $\mathbf{A}$ by minimizing $\Psi_n(\mathbf{A})$ defined in (2.5). Note that as far as the estimation of $\mathbf{A}$ is concerned, two orthogonal matrices are treated as identical if the D-distance between them is 0; see (2.6). The coefficients α_{ij}, β_i and γ_i were estimated using quasi-MLE based on Gaussian likelihood. The estimates are summarized in Table 1 and Figure 1. Estimation errors for $\alpha_{12}, \alpha_{21}, \alpha_{23}, \alpha_{31}$ and α_{32} are all very close to 0 and are not reported here to save space.

Table 1: Summary statistics of the estimation errors in simulation

n		$D(\widehat{\mathbf{A}}, \mathbf{A})$	$\widehat{\beta}_1$	$\widehat{\alpha}_{11}$	$\widehat{\alpha}_{13}$	$\widehat{\beta}_2$	$\widehat{\alpha}_{22}$	$\widehat{\beta}_3$	$\widehat{\alpha}_{33}$
500	mean	0.130	0.842	0.035	0.041	0.761	0.076	0.616	0.084
	median	0.128	0.884	0.030	0.036	0.803	0.072	0.668	0.076
	STD	0.080	0.147	0.030	0.029	0.175	0.045	0.257	0.058
	bias	-	-0.058	-0.005	0.001	-0.039	-0.024	0.016	-0.036
	RMSE	-	0.158	0.031	0.029	0.180	0.052	0.258	0.068
1000	mean	0.114	0.869	0.037	0.037	0.782	0.077	0.616	0.089
	median	0.102	0.885	0.036	0.035	0.804	0.076	0.641	0.087
	STD	0.077	0.078	0.019	0.019	0.119	0.033	0.214	0.043
	bias	-	-0.031	-0.003	-0.003	-0.018	-0.023	0.016	-0.032
	RMSE	-	0.084	0.020	0.019	0.120	0.041	0.215	0.054

Both the means and the standard deviations of $D(\widehat{\mathbf{A}}, \mathbf{A})$ are small. This indicates that the estimation for $\mathbf{A}$ seems to be reasonably accurate. The coefficients in each CUC models were also estimated accurately. The estimators are almost unbiased, as the biases are negligible in comparisons with the corresponding variances. The errors in estimation decrease as the sample size increases from 500 to 1000, roughly by a factor of $\sqrt{2}$.

Since most of the biases reported in Table 1 are negative (see also Figure 1), the coefficients in the GARCH models for CUCs were slightly underestimated. Also note that the estimation errors decrease when the volatility persistence (measured by $\alpha_{i1} + \alpha_{i2} + \alpha_{i3} + \beta_i$) increases; see Figure 1(a) with the sample size 1000. Figure 1(b) presents the estimation errors of the GARCH coefficients with $\mathbf{A}$ given. The difference between the estimation errors of the two cases is small.

4 Real data examples

In this section we illustrate the proposed method with two real data examples with, respectively, $d = 4$ and $d = 10$. First we analyze the 2527 daily log returns (in percentages) of the S&P 500 index, the stock prices of Cisco System, Intel Corporation and Sprint (SCIS for short) in the period from 2 January 1991 to 31 December 2000. This data set was downloaded from Yahoo!Finance. The close prices adjusted for dividends and splits were used to produce the return series plotted

in Figure 2. We use the first 2275 observations (i.e. the data up to the end of 1999) for estimating the parameters in the models, and leave the last 252 data points (i.e. the data in 2000) for checking the post-sample forecasting performance.

To account for the conditional mean of the return series, a vector AR(2) model, selected by both $M(i)$ (Tiao and Box, 1981) and AIC, was first fitted to the data. We denote by $\mathbf{Y}_t$, $t = 1, 2, \cdots, 2273$, the residuals resulted from this fitting. In the sequel, we focus on modeling the conditional covariance matrix process of $\mathbf{Y}_t$.

Let $\mathbf{S}$ be the sample covariance matrix of $\mathbf{Y}_t$, and $\mathbf{X}_t = \mathbf{S}^{-\frac{1}{2}}\mathbf{Y}_t$. The estimator $\widehat{\mathbf{A}}$ was obtained by minimizing $\Psi_n(\mathbf{A})$. For the sake of comparison, the estimator $\widetilde{\mathbf{A}}$ obtained by maximizing the likelihood function of the GO-GARCH(1,1) model (van der Weide, 2002) was also computed. We applied the bootstrap test described in section 2.4, with bootstrap sampling repeated 400 times, to test for the existence of the CUCs and obtained the P-value 0.34. This indicates that there exists no significant evidence against the hypothesis that the CUCs exist for this data set. The 95% bootstrap confidence set for the transformation matrix $\mathbf{A}$ is $\{\mathbf{A} \,|\, D(\mathbf{A}, \widehat{\mathbf{A}}) \leq 0.153,\ \mathbf{A}^\tau\mathbf{A} = \mathbf{I}_4\}$. Since $D(\mathbf{I}_4, \widehat{\mathbf{A}}) = 0.3035$, $\mathbf{I}_4$ is not contained in the confidence sets. Thus the principal components cannot be taken as the CUCs. Note also $D(\widetilde{\mathbf{A}}, \widehat{\mathbf{A}}) = 0.1616$, therefore $\widetilde{\mathbf{A}}$ is not contained in the confidence set either. This suggests that the MLE based on the GO-GARCH(1,1) model does not lead to CUCs and, therefore, it would be inappropriate to assume that the conditional covariance matrix of $\widetilde{\mathbf{A}}\mathbf{X}_t$ is diagonal, as implied by the GO-GARCH(1,1) approach.

Table 2 lists the estimated extended GARCH(1,1) modes for the estimated CUCs. The models were selected by the algorithm specified in section 2.3.3. There is a causality-in-variance relationship from the fourth CUC to the second CUC. Also the last two CUCs are highly persistent as the sum of all the GARCH and ARCH coefficients is close to 1 for both of them. Based on the fitted volatility $\widehat{\sigma}^2_{ti}$ $(i = 1, 2, 3, 4)$ for the CUCs, the conditional covariance matrix for the original residuals $\mathbf{Y}_t$ is of the form:

$$\widehat{\mathbf{H}}_t = \widehat{\mathbf{W}}\mathrm{diag}\{\widehat{\sigma}^2_{t1}, \widehat{\sigma}^2_{t2}, \widehat{\sigma}^2_{t3}, \widehat{\sigma}^2_{t4}\}\widehat{\mathbf{W}}^\tau, \tag{4.1}$$

where $\widehat{\mathbf{W}} = \mathbf{S}^{\frac{1}{2}}\widehat{\mathbf{A}}$.

For the comparison purpose, we also computed the estimated volatility processes for $\mathbf{Y}_t$ based on the O-GARCH(1,1) model of Alexander (2001), the DCC-GARCH(1,1) model of Engle (2002), and the GO-GARCH(1,1) model of van der Weide (2002). We also included the CUC-

GARCH(1,1) model in our comparison, *i.e.* we fitted for each CUC a standard GARCH(1,1) model without incorporating the lagged values from the other CUCs. As we have pointed out above, the GO-GARCH(1,1) model does not fit this data set well. In fact the estimated conditional correlation process between the S&P500 return and the Intel return based on the GO-GARCH(1,1) model is negatively correlated with its counterpart based on any other models mentioned above. Therefore we exclude the GO-GARCH(1,1) in the comparison below.

Figure 3 displays the time plots of the estimated conditional variance processes of the S&P500 return by the O-GARCH(1,1) model, the DCC-GARCH(1,1) model and the CUC-GARCH(1,1) model. While the estimated processes by the DCC-GARCH(1,1) and the CUC-GARCH(1,1) look similar, O-GARCH(1,1) certainly leads to a very different volatility profile. Comparing to the original return series in the Figure 2(a), the two peaks around $t = 850$ should not be there. It was caused by the extreme negative returns of the Cisco price in the same period; see Figure 2(b). Such a misleading phenomenon was resulted from treating the principal components as CUCs in the O-GARCH(1,1) model. The estimated conditional correlation processes between the S&P500 return and the Intel price return are plotted in Figure 4. The conditional correlation estimated by the CUC-GARCH(1,1) is more volatile than those estimated by the O-GARCH(1,1) and the DCC-GARCH(1,1). In particular the CUC-estimated conditional correlation is small in the middle period before it peaks up twice towards to the end. Those two peaks correspond to the two peaks in the volatility process of the S&P500 return. Note that the estimated correlations by the DCC and the CUC are quite different numerically from each other.

Table 2: Extended GARCH Model for CUCs of the SCIS data

j	j_i	$\sigma^2_{t,j}$
1		$\sigma^2_{t,1} = 0.1963 + 0.7603\sigma^2_{t-1,1} + 0.0434Z^2_{t-1,1}$
2	4	$\sigma^2_{t,2} = 0.6297 + 0.1148\sigma^2_{t-1,2} + 0.1622Z^2_{t-1,2} + 0.1015Z^2_{t-1,4}$
3		$\sigma^2_{t,3} = 0.0273 + 0.9031\sigma^2_{t-1,3} + 0.0697Z^2_{t-1,3}$
4		$\sigma^2_{t,4} = 0.0073 + 0.9421\sigma^2_{t-1,4} + 0.0506Z^2_{t-1,4}$

We now apply two diagnostic checking statistics to assess the different fitted models. Following the lead of Tse and Tsui (1999), we use the Box-Pierce statistic to check the cross-product of the standardized residuals. To this end, let $\widehat{u}_{ti} = Y_{ti}/\widehat{\sigma}^{1/2}_{t,ii}$ be the standardized residual for the i-th

component, where $\widehat{\sigma}_{t,ii}$ is the (i,i)-th element of the fitted conditional variance of $\mathbf{Y}_t$. Put

$$C_{t,ij} = \begin{cases} \widehat{u}_{ti}^2 - 1 & i = j \\ \widehat{u}_{ti}\widehat{u}_{tj} - \widehat{\rho}_{t,ij} & i \neq j, \end{cases} \tag{4.2}$$

where $\widehat{\rho}_{t,ij} = \widehat{\sigma}_{t,ij}/(\widehat{\sigma}_{t,ii}\widehat{\sigma}_{t,jj})^{1/2}$ is the estimated conditional correlation between Y_{ti} and Y_{tj}. If the model is correctly specified, there is no autocorrelation in $\{C_{t,ij}, t \geq 1\}$ for any fixed i and j. Define

$$Q(i,j;M) = n\sum_{k=1}^{M} r_{ij,k}^2, \tag{4.3}$$

where $r_{ij,k}$ is the sample autocorrelation of $C_{t,ij}$ at lag k. It is intuitively clear that large values of $Q(i,j;M)$ are indicative for the lack of fit for the conditional correlation between the i-th and the j-th components $\mathbf{Y}_t$ when $i \neq j$, and for the lack of fit for the conditional variance of the i-th component when $i = j$. We also employ a multivariate portmanteau statistic (section 5.5 of Reinsel, 1997) to test for the autocorrelation in the vectorized cross product of residuals $\boldsymbol{\xi}_t = \text{vech}(\widehat{\mathbf{e}}_t\widehat{\mathbf{e}}_t^{\tau})$, where $\widehat{\mathbf{e}}_t = \widehat{\mathbf{H}}_t^{-1/2}\mathbf{Y}_t$. Let $\widehat{\boldsymbol{\rho}}(\ell)$ be the autocovariance matrix of $\boldsymbol{\xi}_t$ at lag ℓ. The multivariate portmanteau statistic is defined as

$$P(k) = n^2\sum_{\ell=1}^{k}(n-\ell)^{-1}\text{tr}\{\widehat{\boldsymbol{\rho}}(\ell)\widehat{\boldsymbol{\rho}}(0)^{-1}\widehat{\boldsymbol{\rho}}(-\ell)\widehat{\boldsymbol{\rho}}(0)\}. \tag{4.4}$$

This may be seen as a multivariate extension of McLeod and Li (1983) which applied a univariate portmanteau test to squared residuals.

Table 3 lists the values of $Q(i,j;M)$, $1 \leq i < j \leq 4$ and $M = 5$, for five different models. Significant levels of $Q(i,j;M)$ were computed according to the χ_5^2-distribution; see Tse and Tsui (1999). Table 4 lists the values $P(k)$ for $1 \leq k \leq 5$. Although the asymptotic distribution of $P(k)$ is unavailable to conduct a formal testing, it is intuitively clear that large values of $P(k)$ would indicate the lack-of-fit of the model concerned.

Tables 3 and 4 indicate that O-GARCH(1,1) provided overall the poorest fit among the five models concerned according to both $Q(i,j;M)$ and $P(k)$, especially four of its Q-statistics are significant at the 0.05 level. On the other hand, the tests with $Q(1,2;5)$ and $Q(1,3;5)$ for both DCC-GARCH(1,1) and GO-GARCH(1,1) models are significant at least at level 10%, while both the CUC-GARCH(1,1) and CUC-extended GARCH(1,1) passed all the tests with the statistic $Q(i,j;M)$. Note that the values of $P(k)$ for the two CUC-based models are smaller than those for the O-GARCH(1,1) and the DCC-GARCH(1,1) models. Overall both the diagnostic statistics indicate that the CUC-extended GARCH(1,1) is the best model for this particular data set.

Table 3: $Q(i,j;M)$ with $M=5$ for the SCIS data

i,j	O-GARCH	DCC	GO-GARCH	CUC-GARCH	CUC-Ex GARCH
1,1	69.37***	5.00	5.23	5.19	5.19
2,2	10.38	9.05	8.91	8.22	8.12
3,3	2.11	4.67	5.88	1.55	1.59
4,4	1.29	1.08	0.97	0.46	0.41
1,2	48.11***	10.91*	10.31*	8.36	8.31
1,3	54.44***	15.79***	10.67*	4.73	4.55
1,4	18.69**	1.86	1.51	1.40	1.38
2,3	1.05	5.15	7.72	4.34	4.25
2,4	6.99	3.04	3.35	3.11	2.93
3,4	2.15	4.11	2.31	2.83	2.82

*Note: The tests significant at level 0.01, 0.05 and 0.1 are marked, respectively, by ***, ** and *.*

Table 4: $P(k)$ for the SCIS data

k	O-GARCH	DCC	GO-GARCH	CUC-GARCH	CUC-Ex GARCH
1	182.76	117.32	99.83	96.69	96.75
2	307.64	210.99	190.85	186.95	184.49
3	439.22	325.91	302.53	302.87	295.92
4	523.74	412.77	392.74	395.79	387.39
5	634.51	507.46	486.91	494.16	489.16

To make a post-sample comparison among these models, we need to construct proxies for unobserved conditional covariance matrices using the daily returns. Let $\widehat{\mathbf{H}}_{t+p|t} = (\widehat{\sigma}_{ij,t+p|t})$ be the p-days ahead forecast of the covariance matrix at t. Following the lead of Pelletier (2006) and Fan et al. (2007), we gauge the quality of forecasting based on the Adaptive Mean Absolute Deviations:

$$\text{AMAD}(p) = \frac{1}{d^2 n^*} \sum_t \sum_{i,j=1}^{d} \Big| \widehat{\sigma}_{ij,t+p|t} - \frac{1}{2v+1} \sum_{\ell=-v}^{v} Y_{i,t+p+\ell} Y_{j,t+p+\ell} \Big|, \tag{4.5}$$

where v is a nonnegative integer, the sum over t is over the n^* post-sample points. When $v=0$, AMAD reduces to the MAD used in Pelletier (2006) and the proxy for the covariance matrix at

time $t+p$ is just the cross product of the return vector at that day; When $v > 0$, the adjacent $2v+1$ days returns are used to average out the stochastic error in the proxy.

We use the last 252 observations in the SCIS data to compute the AMADs. Based on the p-step ahead forecast of the univariate GARCH model (see, e.g. pp.94 in Tsay 2002) for each component and the transformation matrix, $\widehat{\mathbf{H}}_{t+p|t}$ can be constructed in a straightforward way for O-GARCH, GO-GARCH and CUC-GARCH model. Forecast for DCC model follows the procedure of Pelletier (2006). The length of the samples used for parameter estimation are 500 and 1000, respectively, and the estimates are updated every five days, and no causal component is considered for the CUCs. Table 5 lists the results for $p = 1$ and 5.

Table 5: AMAD for the SCIS data

	p	O-GARCH	DCC	GO-GARCH	CUC-GARCH
$v = 0$					
$T = 500$	1	7.3095	7.2187	7.9227	7.2104
	5	7.2869	7.2383	7.7048	7.1909
$T = 1000$	1	7.2094	7.1636	7.8361	7.0859
	5	7.1781	7.1646	7.6988	7.1436
$v = 1$					
$T = 500$	1	4.5650	4.5897	4.8360	4.6894
	5	4.9543	4.9634	5.4663	4.8978
$T = 1000$	1	4.5856	4.5921	4.7972	4.6747
	5	4.9158	4.9343	5.7790	4.9142

The AMAD of the CUC-GARCH model is always the smallest when $v = 0$. When $v = 1$, this is still true for the five-day ahead forecast, but for the case of one-day ahead forecast, the AMADs of O-GARCH and DCC are both smaller than those of the CUC-GARCH model. On the other hand, the GO-GARCH model provides the worst forecasts for this data set. Overall the CUC-GARCH model outperforms the other three models in this forecasting comparison.

Our second example concerns the daily log-return of the exchange rates of the 10 European currencies to US dollar in 2 January 1990 – 31 December 1998, immediately before the introduction of the Euros. The currencies concerned are from Austria, Belgium, Finland, France, Germany,

Ireland, Italy, Netherlands, Portugal and Spain. For this data set, $n = 2263$ and $d = 10$. The diagnostic checks similar to those in the first example were carried out. To save space, we only list in Table 6 the multivariate portmanteau statistics for the different models. It is not surprising that $P(k)$ of the extended CUC-GARCH model is the smallest one in each row and the values of $P(k)$ for the CUC-GARCH model are smaller than those of the O-GARCH, DCC, and GO-GARCH model as $k > 3$. Note that the DCC model may be too simple to catch the dynamical structure of a 10-dimensional volatility process. The extension to incorporate more flexibility into the DCC structure would present an interesting line for further development.

Table 6: $P(k)$ for the Exchange Rates Data

k	O-GARCH	DCC	GO-GARCH	CUC-GARCH	CUC-Ex GARCH
1	6783	10271	6352	7297	6316
2	11224	16792	10918	11562	9871
3	15736	23530	15043	15288	12706
4	20538	30701	19871	19448	16205
5	23655	37077	23313	22839	19022
10	41631	62943	41770	41197	35928

Again the comparison based on post-sample forecasting was also in favour of the CUC approach. In fact we reserve the whole year data in 1998 (with 252 observations) for checking the post-sample forecasting performance. Both one-day ahead and five-day ahead forecasts are made based on the fitted models using 500 observations in the immediate past. Table 7 lists the AMAD-values (see (4.5)) for the forecasts based on the four different models. Except one case with $v = 1$ and $p = 1$, the CUC-GARCH model provides the best forecasts among the models concerned. Based on Tables 6 & 7, we would conclude that the CUC provides an alternative parsimonious representation for the dynamics of conditional covariance processes which is more accommodating than, for example, the simple DCC model when the dimension of the underlying process is large.

5 Conclusional remark

It is extremely effective for analysing multivariate time series to find an appropriate linear transformation such that the components of the transformed series exhibits certain "un-relatedness".

Table 7: AMAD for the exchange rates data

	p	O-GARCH	DCC	GO-GARCH	CUC-GARCH
	1	0.3097	0.3098	0.4074	0.2978
$v = 0$	2	0.3112	0.3109	0.4034	0.2987
	3	0.3089	0.3083	0.3885	0.2958
$T = 500$	4	0.3097	0.3089	0.3502	0.2974
	5	0.3097	0.3087	0.3752	0.2991
	1	0.2061	0.2038	0.2475	0.2088
$v = 1$	2	0.2127	0.2115	0.3439	0.2108
	3	0.2117	0.2102	0.3193	0.2087
$T = 500$	4	0.2138	0.2121	0.3145	0.2086
	5	0.2150	0.2133	0.3043	0.2094

There are at least three types of "un-relatedness". For modeling conditional covariance processes, the conditional uncorrelatedness is the correct measure which serves the purpose adequately, while the unconditional uncorrelatedness required in the orthogonal GARCH model (Alexander, 2001) is too weak and the independence in the independent component analysis is too strong.

Modeling multivariate volatility processes is a practically important and methodologically challenging problem. The CUC-based method proposed in this paper attempts to catch sophisticating conditional heteroscedasticity while maintaining a parsimonious representation for matrix processes. One natural question arises: do the CUCs so-defined exist? Empirical experiments with various real data sets indicate the P-value of the bootstrap test described in section 2.4 tends to decrease as d increases. However with small or moderately large d the hypothesis of the existence of CUCs have been rarely rejected in our empirical experiments.

In the event that the CUCs do not exist, we argue that it is very natural to find the linear transformation such that the resulting components are the least conditionally correlated, especially if we take the viewpoint that any statistical model is merely an approximation to the reality. In this sense, our CUC estimation leads to the least conditionally correlated directions (LCCD) and we build up an (approximate) volatility model by assuming the conditional correlations among those LCCD are 0. The LCCD are the directions which minimize $\Psi(\cdot)$ defined in (2.8); also see

(2.9). Theorem 1 indicates that the columns of $\widehat{\mathbf{A}}$ are the consistent estimators for the LCCD. Note that both Theorems 1 and 2 still apply when the CUCs do not exist (*i.e.* $\Psi(\mathbf{A}_0) \neq 0$); see condition A3 in the Appendix. Even if the CUCs do not exist, a CUC-GARCH(1,1) model, for example, still provides a more relevant fitting than O-GARCH(1,1) and GO-GARCH(1,1) models.

Finally we point out that for any multivariate time series $\mathbf{X}_t$, there always exists an $\mathcal{F}_{t-1}$-measurable orthogonal matrix $\mathbf{A}_{t-1}$ for which the components in $\mathbf{A}_{t-1}^{\tau}\mathbf{X}_t$ are conditionally uncorrelated. The CUC requires a further constraint $\mathbf{A}_{t-1} \equiv \mathbf{A}$. It is reasonable to assume that $\mathbf{A}_{t-1}$ varies smoothly in t (see also (1.3)). Therefore we may assume that the CUCs exist for a short time period in which $\mathbf{A}_{t-1} \approx \mathbf{A}$. This further extends the scope of the applicability of our method.

Acknowledgments

We thank the reviewers for very helpful comments and suggestions. Jianqing Fan was supported partially by NSF grant DMS-0355179, DMS-0704337, and Chinese NSF grant 10628104, Mingjin Wang was partially supported by EPSRC grant GR/R97436 and Chinese NSF grant 70201007, and Qiwei Yao was partially supported by EPSRC grants GR/R97436 and EP/C549058.

Appendix — Conditions and proof of Theorem 1

We first introduce two concepts: the mixing which measures the decaying speed of the auto-dependence for a time series over an increasing time span, and the Vapnik-Červonenkis (or VC) index which measures the complexity of a collection of sets.

Let $\mathcal{F}_i^j$ be the σ-algebra generated by $\{\mathbf{X}_t, i \leq t \leq j\}$. The β-mixing coefficients are defined as

$$\beta(n) = E\left\{\sup_{B \in \mathcal{F}_n^{\infty}} |P(B) - P(B|\mathcal{F}_{-\infty}^0)|\right\}.$$

(See §2.6.1 of Fan and Yao, 2003.)

For an arbitrary set of n points $\{x_1, \cdots, x_n\}$, there are 2^n possible subsets. Say that $\mathcal{B}$ picks out a certain subset from $\{x_1, \cdots, x_n\}$ if this can be formed as a set of the form $B \cap \{x_1, \cdots, x_n\}$ for a set B in $\mathcal{B}$. The collection $\mathcal{B}$ shatters $\{x_1, \cdots, x_n\}$ if each of its 2^n subsets can be picked out by $\mathcal{B}$. The VC-index of $\mathcal{B}$ refers to the smallest n for which no set of size n is shattered by

$\mathcal{B}$. A collection of sets $\mathcal{B}$ is called a VC-class if its VC-index is finite. The collections of sets of rectangles, balls and their unions are VC-classes. See Chapter 2.6 of van der Vaart and Wellner (1996) for further discussion on VC-classes.

Under the regularity conditions listed below, the estimator $\widehat{\mathbf{A}}$ is consistent; see Theorem 1 in section 2.2.2.

(A1) The collection $\mathcal{B}$ consists of countable subsets in $\mathcal{R}^d$, and is a VC-class. Furthermore $\sum_{B\subset\mathcal{B}} w(B) < \infty$.

(A2) The process $\{\mathbf{X}_t\}$ is strictly stationary with $E||\mathbf{X}_t||^2 < \infty$, where $||\cdot||$ denotes the Euclidean norm. Furthermore, the β-mixing coefficients $\{\mathbf{X}_t\}$ satisfy $\beta(n) = O(n^{-b})$ for some $b > 0$.

(A3) There exists a $d \times d$ orthogonal matrix $\mathbf{A}_0$ which minimizes $\Psi(\cdot)$ defined in (2.8). Furthermore the minimum value of Ψ is obtained at an orthogonal matrix $\mathbf{A}$ if and only if $D(\mathbf{A}, \mathbf{A}_0) = 0$.

(A4). $E\|\mathbf{X}_t\|^{2p} < \infty$ for some $p > 2$ and $b > p/(p-2)$, where b is given in (A2).

(A5) $\Psi(\mathbf{A}_0) - \Psi(\mathbf{A}) \leq -aD(\mathbf{A}, \mathbf{A}_0)$ for any orthogonal matrix $\mathbf{A}$ such that $D(\mathbf{A}, \mathbf{A}_0)$ is smaller than a small but fixed constant, where $a > 0$ is a constant.

Remark 1. Let $\mathcal{H}$ be the set consisting of all $d \times d$ orthogonal matrices. Then $\mathcal{H}$ may be partitioned into the equivalent classes defined by the distance D in (2.6) as follows: the D-distance between any two elements within an equivalent class is 0, and the D-distance between any two elements from different classes is greater than 0. Let $\mathcal{H}_D$ be the quotient space $\mathcal{H}/D$ consisting of those equivalent classes in $\mathcal{H}$, *i.e.* we treat $\mathbf{A}$ and $\mathbf{B}$ as the same element in $\mathcal{H}_D$ if and only if $D(\mathbf{A}, \mathbf{B}) = 0$. Condition (A3) ensures $\mathbf{A}_0$ is the unique minimizer of $\Psi(\mathbf{A})$ on $\mathcal{H}_D$.

We introduce some notation. Let

$$\mathbf{C}_{n,k}(B) = (n-k)^{-1} \sum_{t=k+1}^{n} \mathbf{X}_t \mathbf{X}_t^\tau I(\mathbf{X}_{t-k} \in B), \qquad \mathbf{C}_k(B) = E\{\mathbf{X}_t \mathbf{X}_t^\tau I(\mathbf{X}_{t-k} \in B)\}.$$

The lemma below shows that both $\Psi(\cdot)$ and $\Psi_n(\cdot)$ are Lipschitz continuous on $\mathcal{H}_D$ with D-distance.

Lemma 1. For any $\mathbf{U}, \mathbf{V} \in \mathcal{H}_D$, it holds that

$$|\Psi(\mathbf{U}) - \Psi(\mathbf{V})| \leq c \, \mathrm{tr} E(\mathbf{X}_t \mathbf{X}_t^T) \{D(\mathbf{U}, \mathbf{V})\}^{1/2},$$

and

$$|\Psi_n(\mathbf{U}) - \Psi_n(\mathbf{V})| \leq c\,\mathrm{tr}(n^{-1}\sum_{i=1}^{n}\mathbf{X}_t\mathbf{X}_t^T)\,\{D(\mathbf{U},\mathbf{V})\}^{1/2}$$

almost surely, where $c > 0$ is a constant and $\mathrm{tr}(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$.

Proof. We only prove the lemma for $\Psi(\cdot)$. The result for $\Psi_n(\cdot)$ may be shown in the same manner. Let $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_d)^\tau$, $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_d)^\tau$, $u_{ijk}(B) = E\{\mathbf{u}_i^\tau \mathbf{C}_k(B)\mathbf{u}_j\}$ and $v_{ijk}(B) = E\{\mathbf{v}_i^\tau \mathbf{C}_k(B)\mathbf{v}_j\}$. We assume that the orders and the directions of $\mathbf{u}_i$ and $\mathbf{v}_j$ are arranged such that $\mathbf{u}_i^\tau \mathbf{v}_i \in [0,1]$ for all i, and

$$D(\mathbf{U},\mathbf{V}) = 1 - \frac{1}{d}\sum_{i=1}^{d}\mathbf{u}_i^\tau\mathbf{v}_i = \frac{1}{d}\sum_{i=1}^{d}(1 - \mathbf{u}_i^\tau\mathbf{v}_i). \tag{6.1}$$

See (2.6). Put the spectral decomposition for $\mathbf{C}_k(B)$ as

$$\mathbf{C}_k(B) = \sum_{\ell=1}^{d}\mu_\ell(B,k)\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau,$$

where $\mu_1(B,k) \geq \cdots \geq \mu_d(B,k) \geq 0$ are the eigenvalues of $\mathbf{C}_k(B)$, and $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d$ are their corresponding (orthonormal) eigenvectors. It is easy to see that $\mu_\ell(B,k) \leq \mu_\ell$ for all k and B, where $\mu_1 \geq \cdots \geq \mu_d$ are the eigenvalues of the matrix $E\{\mathbf{X}_t\mathbf{X}^\tau\}$. Consequently, by noticing that $|\boldsymbol{\gamma}_\ell^\tau\mathbf{u}_j| \leq 1$ and $|\mathbf{v}_i^\tau\boldsymbol{\gamma}_\ell| \leq 1$, we have

$$\begin{aligned}
|u_{ijk}(B) - v_{ijk}(B)| \;&\leq\; \sum_{\ell=1}^{d}\mu_\ell|\mathbf{u}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{u}_j - \mathbf{v}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{v}_j| \\
&\leq\; \sum_{\ell=1}^{d}\mu_\ell\{|\mathbf{u}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{u}_j - \mathbf{v}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{u}_j| + |\mathbf{v}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{u}_j - \mathbf{v}_i^\tau\boldsymbol{\gamma}_\ell\boldsymbol{\gamma}_\ell^\tau\mathbf{v}_j|\} \\
&\leq\; \sum_{\ell=1}^{d}\mu_\ell\{|(\mathbf{u}_i - \mathbf{v}_i)^\tau\boldsymbol{\gamma}_\ell| + |\boldsymbol{\gamma}_\ell^\tau(\mathbf{u}_j - \mathbf{v}_j)|\}
\end{aligned}$$

By using the Cauchy-Schwartz's inequality, the above inequality is furthered bounded by

$$\begin{aligned}
&\sum_{\ell=1}^{d}\mu_\ell\{||\mathbf{u}_i - \mathbf{v}_i|| + ||\mathbf{u}_j - \mathbf{v}_j||\} \\
=\;& \sqrt{2}\{(1 - \mathbf{u}_i^\tau\mathbf{v}_i)^{1/2} + (1 - \mathbf{u}_j^\tau\mathbf{v}_j)^{1/2}\}\sum_{\ell=1}^{d}\mu_\ell.
\end{aligned} \tag{6.2}$$

Note that for $x \neq 0$, it holds that

$$|x+y| - |x| = y\,\mathrm{sgn}(x) + 2(x+y)\{I(-y < x < 0) - I(0 < x < -y)\}. \tag{6.3}$$

Hence,

$$\begin{aligned}\Psi(\mathbf{U}) &= \sum_{1\le i<j\le d}\sum_{B\in\mathcal{B}} w(B)\sum_{k=1}^{k_0}\big[|v_{ijk}(B)| + |v_{ijk}(B) + \{u_{ijk}(B) - v_{ijk}(B)\}| - |v_{ijk}(B)|\big] \\ &= \sum_{1\le i<j\le d}\sum_{B\in\mathcal{B}} w(B)\sum_{k=1}^{k_0}\big[|v_{ijk}(B)| + \{u_{ijk}(B) - v_{ijk}(B)\}\mathrm{sgn}\{v_{ijk}(B)\} \\ &+ 2u_{ijk}(B)\{I(B_1) - I(B_2)\}\big], \end{aligned} \tag{6.4}$$

where

$$B_1 = \{v_{ijk}(B) - u_{ijk}(B) < v_{ijk}(B) < 0\}, \quad B_2 = \{0 < v_{ijk}(B) < v_{ijk}(B) - u_{ijk}(B)\}.$$

On the set $B_1 \cup B_2$,

$$|u_{ijk}(B)| \le |u_{ijk}(B) - v_{ijk}(B)| + |v_{ijk}(B)| \le 2|u_{ijk}(B) - v_{ijk}(B)|.$$

This, combining with (6.2) and (6.4), implies that

$$\begin{aligned} & |\Psi(\mathbf{U}) - \Psi(\mathbf{V})| \\ \le\ & k_0 \sum_{1\le i<j\le d}\sum_{B\in\mathcal{B}} w(B)\big[\sqrt{2}\{(1-\mathbf{u}_i^\tau\mathbf{v}_i)^{1/2} + (1-\mathbf{u}_j^\tau\mathbf{v}_j)^{1/2}\}\sum_{\ell=1}^{d}\mu_\ell + 2|u_{ijk}(B)|I_1(B_1)\big] \\ \le\ & 5\sqrt{2}k_0\sum_{B\in\mathcal{B}} w(B)\sum_{1\le i<j\le d}\{(1-\mathbf{u}_i^\tau\mathbf{v}_i)^{1/2} + (1-\mathbf{u}_j^\tau\mathbf{v}_j)^{1/2}\}\sum_{\ell=1}^{d}\mu_\ell \\ \le\ & 10\sqrt{2}k_0 d\sum_{B\in\mathcal{B}} w(B)\sum_{\ell=1}^{d}\mu_\ell\sum_{i=1}^{d}(1-\mathbf{u}_i^\tau\mathbf{v}_i)^{1/2}. \end{aligned} \tag{6.5}$$

Now the lemma follows from (6.5) and the inequality

$$\sum_{i=1}^{d}(1-\mathbf{u}_i^\tau\mathbf{v}_i)^{1/2} \le d^{1/2}\Big\{\sum_{i=1}^{d}(1-\mathbf{u}_i^\tau\mathbf{v}_i)\Big\}^{1/2},$$

see also (6.1). This completes the proof.

Proof of Theorem 1. Since $\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)$ is a real symmetric matrix, it holds for any unit vectors $\mathbf{a}$ and $\mathbf{b}$ that

$$|\mathbf{a}^\tau\{\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)\}\mathbf{b}| \le ||\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)||,$$

where $||\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)||$ denotes the sum of the absolute values of the eigenvalues of $\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)$. This may be obtained by using the spectral decomposition of $\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)$. Conse-

quently it holds uniformly for any orthogonal matrix $\mathbf{A}$ that

$$\begin{aligned} |\Psi_n(\mathbf{A}) - \Psi(\mathbf{A})| &\leq c \sum_{1\leq i<j\leq d} \sup_{1\leq k\leq k_0, B\in\mathcal{B}} |\mathbf{a}_i^\tau \{\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)\}\mathbf{a}_j| \\ &\leq c\frac{d(d-1)}{2} \sup_{1\leq k\leq k_0, B\in\mathcal{B}} \|\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)\|, \end{aligned} \tag{6.6}$$

where $c > 0$ is a constant. Note the (i, j)-th element of $\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)$ is

$$\frac{1}{n-k}\sum_{t=k+1}^{n} X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B) - E\{X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B)\},$$

where X_{ti} denotes the i-th element of $\mathbf{X}_t$. Since $E|X_{ti}X_{tj}| < \infty$ and $\mathcal{B}$ is a VC-class, the covering number for the set of functions $\{X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B), B \in \mathcal{B}\}$ has a polynomial rate of growth for any underlying probability measure (Theorem 2.6.4, van der Vaart and Wellner 1996). Hence, it is a Glivenko-Cantelli class. It follows now from Theorem 3.4 of Yu (1994) that

$$\sup_{B\in\mathcal{B}} \Big| \frac{1}{n-k}\sum_{t=k+1}^{n} X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B) - E\{X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B)\} \Big| \xrightarrow{P} 0,$$

Consequently,

$$\sup_{B\in\mathcal{B}} |\lambda_{\max}(B, k)| \xrightarrow{P} 0, \qquad \sup_{B\in\mathcal{B}} |\lambda_{\min}(B, k)| \xrightarrow{P} 0,$$

where $\lambda_{\max}(B, k)$ and $\lambda_{\min}(B, k)$ denote, respectively, the maximum and the minimum eigenvalues of $\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)$. Thus

$$\sup_{B\in\mathcal{B}} \|\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)\| \xrightarrow{P} 0,$$

for $k = 1, \cdots, k_0$. Now it follows from (6.6) that

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

Combining this with Lemma 1 above and the continuity of the argmax mapping (Theorem 3.2.2 and Corollary 3.2.3, van der Vaart and Wellner, 1996), it holds that $D(\widehat{\mathbf{A}}, \mathbf{A}_0) \xrightarrow{P} 0$. This completes the proof of the first part of Theorem 1.

Under the additional condition $E|X_{ti}X_{tj}|^{2p} < \infty$ and the mixing condition given in Condition (A4), Theorem 1 of Arcones and Yu (1994) implies that the set of functions $\{X_{ti}X_{tj}I(\mathbf{X}_{t-k} \in B), B \in \mathcal{B}\}$ is a Donsker class, and hence the process $\{\mathbf{\Delta}_{n,k}(B), B \in \mathcal{B}\}$ indexed by $B \in \mathcal{B}$ converges weakly to a Gaussian process, where $\mathbf{\Delta}_{n,k}(B) = \sqrt{n}\{\mathbf{C}_{n,k}(B) - \mathbf{C}_k(B)\}$. It follows

from (6.3) that

$$\begin{aligned}\Psi_n(\mathbf{A}) &= \sum_{1\le i<j\le d}\sum_{B\in\mathcal{B}} w(B)\sum_{k=1}^{k_0}\big[|\mathbf{a}_i^T\mathbf{C}_k(B)\mathbf{a}_j| + n^{-1/2}\mathrm{sgn}\{\mathbf{a}_i^\tau\mathbf{C}_k(B)\mathbf{a}_j\}\mathbf{a}_i^\tau\mathbf{\Delta}_{n,k}(B)\mathbf{a}_j \\ &+ \mathbf{a}_i^\tau\mathbf{C}_{n,k}(B)\mathbf{a}_j\{I(B_3)-I(B_4)\}\big] \\ &= \Psi(\mathbf{A}) + O_P(n^{-1/2}), \end{aligned} \tag{6.7}$$

where

$$B_3 = \{n^{-1/2}\mathbf{a}_i^\tau\mathbf{\Delta}_{n,k}(B)\mathbf{a}_j < \mathbf{a}_i^\tau\mathbf{C}_k(B)\mathbf{a}_j < 0\}, \quad B_4 = \{0 < \mathbf{a}_i^\tau\mathbf{C}_k(B)\mathbf{a}_j < n^{-1/2}\mathbf{a}_i^\tau\mathbf{\Delta}_{n,k}(B)\mathbf{a}_j\}.$$

The last equality in (6.7) follows from the fact that on $B_3 \cup B_4$,

$$|\mathbf{a}_i^\tau\mathbf{C}_{n,k}(B)\mathbf{a}_j| \le |\mathbf{a}_i^\tau\mathbf{C}_k(B)\mathbf{a}_j| + n^{-1/2}|\mathbf{a}_i^\tau\mathbf{\Delta}_{n,k}(B)\mathbf{a}_j| \le 2n^{-1/2}|\mathbf{a}_i^\tau\mathbf{\Delta}_{n,k}(B)\mathbf{a}_j|.$$

It follows from (6.7) and condition (A5) that

$$\Psi_n(\mathbf{A}_0) - \Psi_n(\mathbf{A}) = \Psi(\mathbf{A}_0) - \Psi(\mathbf{A}) + O_P(n^{-1/2}) \le -aD(\mathbf{A}_0, \mathbf{A}) + O_P(n^{-1/2}). \tag{6.8}$$

Now by substituting $\mathbf{A}$ by $\widehat{\mathbf{A}}$, the left hand side of (6.8) must be non-negative by the definition of $\widehat{\mathbf{A}}$. The right hand side of (6.8) would be negative unless

$$D(\mathbf{A}_0, \widehat{\mathbf{A}}) = O_P(n^{-1/2}).$$

This completes the proof.

References

Alexander, C. (2001). Orthogonal GARCH. In *Mastering Risk*. Financial Times-Prentice Hall: London; **2**, 21-38.

Arcones, M.A. and Yu, B. (1994). Central limit theorems for empirical processes and U-processes of stationary mixing sequences. *J. Theor. Probab.*, **7**, 47–71.

Back, A. and Weigend, A.S. (1997). A first application on independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, **8**,473-484.

Bauwens, L., Laurent, S. and Rombouts, J.V.K. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, **21**, 79-109.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. *Review of Economics and Statistics*, **72**, 498-505.

Bollerslev, T.R., Engle, R. and Wooldridge, J. (1998). A capital asset pricing model with time varying covariances. *Journal of Political Economy*, **96**, 116-131.

Cheung, Y.W. and Ng, R.P. (1996). A causality in variance test and its application to financial market prices. *Journal of Econometrics*, **72**, 33-48.

Chow, Y.S. and Teicher, H. (1997). *Probability Theory* (3rd edition). Springer, New York.

Engle, R. (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 generalized ARCH. *Econometric Theory* , **11**, 122-150.

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.

Engle, R.F. and Sheppard, K. (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. *A preprint*.

Fan, J., Fan, Y. and Lv, J. (2007). Aggregation of nonparametric estimators for volatility matrix. *Journal of Financial Economics*, **5**, 321-357.

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

Giraitis, L., Kokoszka, P., and Leipus, R. (2000). Stationary ARCH models: Dependence structure and central limit theorem. *Econometric Theory*, **16**, 3–22.

Granger, C.W.J., Robins, R.P. and Engle, R.F. (1984). Wholesale and retail prices: bivariate time series modeling with forecastable error variances. In *Model Reliability* (D. Belsley and E. Kuh, eds.). MIT Press, Cambridge, MA.

Hafner, C.M. and Herwartz, H. (2000). Testing for linear autoregressive dynamics under heteroscedasticity. *The Econometrics Journal*, **3**, 177-197.

Hafner, C.M. and Herwartz, H. (2006). A Lagrange multiplier test for causality in variance. *Economics Letters*, **93**, 137-141.

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

Harvey, A., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, **61**, 247-264.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.

Jerez, M., Casals, J. and Sotoca, S. (2001). The likelihood of multivariate GARCH models is ill-conditioned. *A preprint*.

Kiviluoto, K. and Oja, E. (1998). Independent component analysis for parallel financial time series. In *Proc. Int. Conf. on Neural Information Processing (ICONIP'98)*, vol.2, pp.895-989, Tokyo.

Mălăroiu, S., Kiviluoto, K. and Oja, E. (2000). Time series prediction with independent component analysis. *A preprint*.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, **21**, 255-285.

McLeod, A. I. and Li, W.K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, **4**, 269-273.

Pelletier, D. (2006). Regime switching for dynamic correlations. *Journal of Econometrics*, **131**, 445-473.

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

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

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

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields* (D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielsen, eds.). Chapman and Hall, London. pp. 1-67.

Straumann, D. (2005). *Estimation in Conditional Heteroscedastic Time Series Models*. Springer, Heidelberg.

Tiao, G.C. and Box, G.E.P. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association*, **76**, 802-816.

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

Tse, Y. K. and Tsui, A.K.C. (1999). A note on diagnosing multivariate conditional heteroscedasticity models. *Journal of Time Series Analysis*, **20**, 679-691.

Vilenkin, N. (1968). Special functions and the theory of group representation, translations of mathematical monographs. *American Math. Soc.*, Providence, Rhode Island, 22.

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

van der Weide, R. (2002). GO-GARCH: a multivariate generalized orthogonal GARCH model. *Journal of Applied Econometrics*, **17**, 549-564.

Wang, M. and Yao, Q. (2005). Modelling multivariate volatilities: an ad hoc approach. To appear in "*Contemporary Multivariate Analysis and Experimental Designs*" J. Fan, G. Li & R. Li (edit.) World Scientific, Singapore.

Wu, J.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *The Annals of Statistics*, **14**, 1261-1350.

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

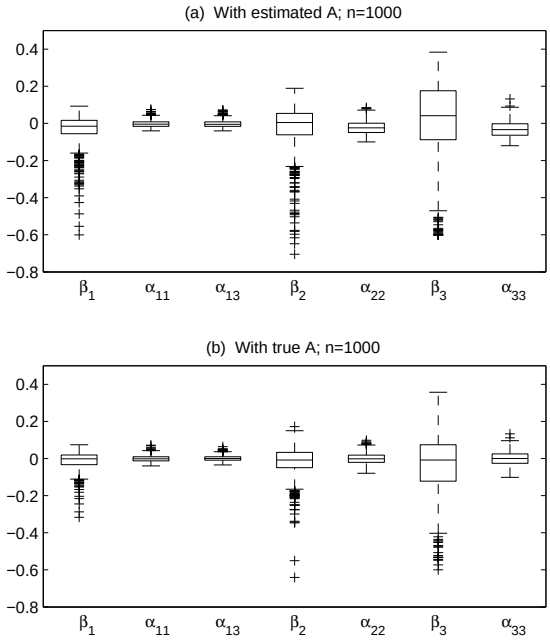

Figure 1: *Boxplots of the estimation errors for CUC-GARCH(1,1) model (3.1) with* $\mathbf{A} = \widehat{\mathbf{A}}$ *estimated (upper panel) and the true* $\mathbf{A}$ *(lower panel). The sample size is* $n = 1000$.

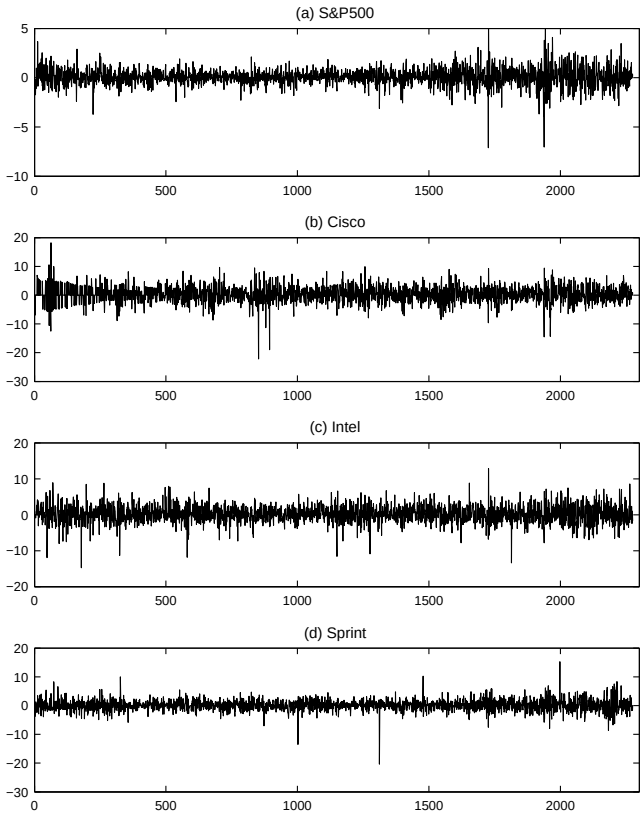

Figure 2: *Plots of daily log returns of (a) S&P 500 index, (b) Cisco Systems , (c) Intel Corporation and (d) Sprint in 2 January 1991 – 31 December 1999.*

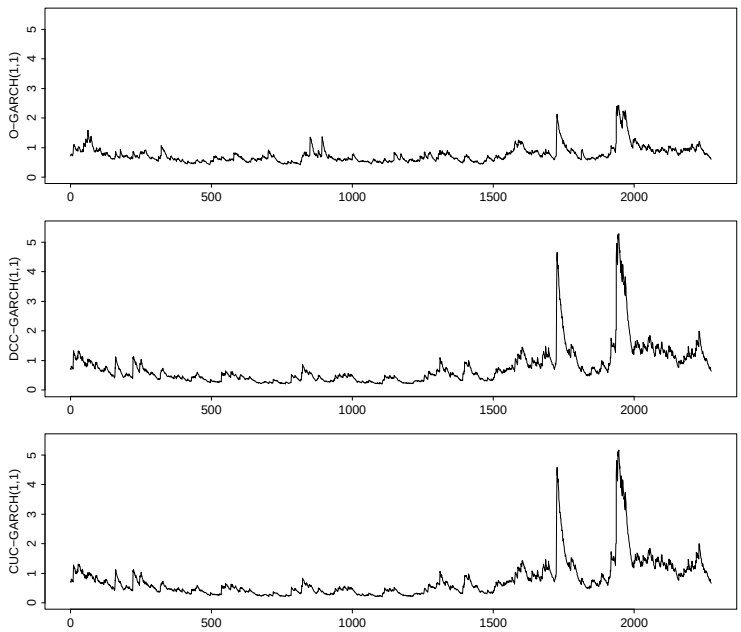

Figure 3: *The estimated volatility processes for the S&P500 return.*

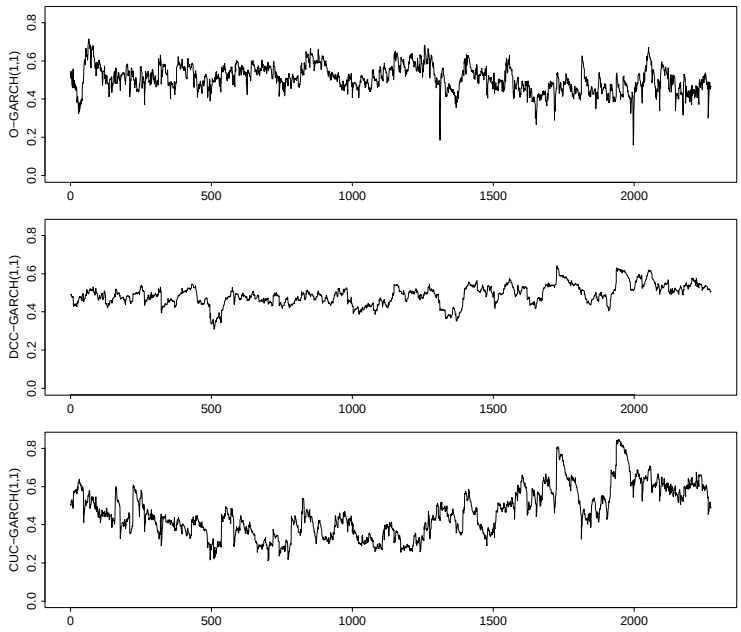

Figure 4: *The estimated conditional correlation processes between the S&P500 return and the Intel return.*