CORE DISCUSSION PAPER
2006/12

MULTIVARIATE MIXED NORMAL CONDITIONAL HETEROSKEDASTICITY

L. Bauwens[1], C.M. Hafner[2] and J.V.K. Rombouts[3]

February 20, 2006

Abstract

We propose a new multivariate volatility model where the conditional distribution of a vector time series is given by a mixture of multivariate normal distributions. Each of these distributions is allowed to have a time-varying covariance matrix. The process can be globally covariance-stationary even though some components are not covariance-stationary. We derive some theoretical properties of the model such as the unconditional covariance matrix and autocorrelations of squared returns. The complexity of the model requires a powerful estimation algorithm. In a simulation study we compare estimation by maximum likelihood with the EM algorithm and Bayesian estimation with a Gibbs sampler. Finally, we apply the model to daily U.S. stock returns.

Keywords: Multivariate volatility, Finite mixture, EM algorithm, Bayesian inference.
JEL Classification: C11, C22, C52.

[1]Université catholique de Louvain, CORE and Department of Economics.
[2]Université catholique de Louvain, Institut de statistique and CORE.
[3]HEC Montréal, CIRANO, CIRPEE and CREF.

Bauwens's work was supported in part by the European Community's Human Potential Programme under contract HPRN-CT-2002-00232, MICFINMA and by a FSR grant from UCL. This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. Rombouts's work was supported by a HEC Montréal Fonds de démarrage and by the Centre for Research on e-Finance. The scientific responsibility is assumed by the authors.

Correspondence to Jeroen Rombouts at HEC Montréal, 3000 chemin de la Cte-Sainte-Catherine, H3T 2A7, Montreal, Canada. Email: jeroen.rombouts@hec.ca

1 Introduction

Several authors have argued in favour of adding flexibility to the family of GARCH models by using the idea of mixture models. For example, extending the model of Wong and Li (2000) and Wong and Li (2001), Haas, Mittnik, and Paolella (2004a) propose a mixed normal conditional heteroskedastic model where the conditional distribution of returns is a mixture of normal distributions, each of which has a regime specific conditional variance specified as a GARCH equation. In this way, they avoid the problem of path-dependence of the conditional variance of regime-switching GARCH models outlined by Gray (1996). Other related papers are those of Haas, Mittnik, and Paolella (2004b) and Alexander and Lazar (2004). All these articles deal with a univariate setting.

Multivariate mixture models have been frequently used in an iid context, but not, to the best of our knowledge, for time series models of conditional volatility, in particular multivariate GARCH models. In this paper, we try to fill this gap by extending the univariate model of Haas, Mittnik, and Paolella (2004a) to the multivariate case. Mixing two or more conditionally normal and heteroskedastic components can generate quite complex stochastic behavior, similar to the one often observed in financial time series. For example, it may be that a component is covariance stationary, another is not, but mixing them might again generate a covariance stationary process. It is possible that mixing many components, of which some are non-stationary, produces behavior similar to processes with long memory, but we have not investigated this issue further.

Note that our approach is different from the regime-switching model of Pelletier (2005), where the unobserved state variable follows a Markov chain and where within a regime correlations are constant.

The paper is organized as follows. In Section 2, we define the model and derive its properties. In Section 3, we present the estimation methods. In Section 4, we illustrate the estimation methods on simulated data, and in Section 5, we present an application using daily data for two stocks. Proofs are relegated in an Appendix.

2 The Model

Consider an N-dimensional vector time series $\{\varepsilon_t, t \in \mathbb{N}\}$. A flexible model for the distribution of ε_t conditional on the information set $\mathcal{F}_{t-1}$ is given by

$$\begin{aligned} \varepsilon_t|\mathcal{F}_{t-1} &\sim f(\lambda_1, \ldots, \lambda_k, \mu_1, \ldots, \mu_k, \Sigma_{1t}, \ldots, \Sigma_{kt}) && (1) \\ &= \sum_{j=1}^{k} \lambda_j f(\varepsilon_t|\mu_j, \Sigma_{jt}) && (2) \end{aligned}$$

where $\lambda_j > 0, j = 1, \ldots, k$, $\sum_{j=1}^{k} \lambda_j = 1$ and $f(\varepsilon_t|\mu_j, \Sigma_{jt})$ is a multivariate density with mean vector μ_j and variance-covariance matrix Σ_{jt}. Note that λ_j is the probability of being in state j, characterized by the density $f(\varepsilon_t|\mu_j, \Sigma_{jt})$, and λ_j is constant over time. Similarly, the means of each state density, $\mu_1, \ldots, \mu_k$, are assumed constant over time. If ε_t is an error term, one would like to impose a restriction

on the μ_j such that the conditional mean of ε_t is zero. For example, one such condition is

$$\mu_k = -\sum_{j=1}^{k-1}(\lambda_j/\lambda_k)\mu_j. \tag{3}$$

The first two conditional moments of ε_t are then be given by

$$\mathrm{E}[\varepsilon_t \mid \mathcal{F}_{t-1}) = 0 \tag{4}$$

$$\mathrm{Var}[\varepsilon_t \mid \mathcal{F}_{t-1}] = \sum_{j=1}^{k}\lambda_j\Sigma_{jt} \tag{5}$$

The process ε_t is conditionally heteroskedastic as every Σ_{jt} is allowed to depend on the information set. We model this dependence using multivariate GARCH (MGARCH) specifications. In particular, we assume that Σ_{jt} is a function of ε_{t-1} and of $\Sigma_{j,t-1}$, which can be called a 'diagonality' restriction since the conditional variance of state j depends only on its own past. In principle, any MGARCH model (VEC, BEKK, DCC,..., see Bauwens, Laurent, and Rombouts (2006)) can be used, but we focus here on the VEC model. Each matrix Σ_{jt} is a VEC model, such that

$$h_{jt} = \mathrm{vech}(\Sigma_{jt}) \tag{6}$$

has the dynamic structure

$$h_{jt} = \omega_j + A_j\eta_{t-1} + B_j h_{j,t-1} \tag{7}$$

where ω_j is a vector of $N^* = N \times (N+1)/2$ parameters, A_j and B_j are square matrices of order N^*, and

$$\eta_t = \mathrm{vech}(\varepsilon_t\varepsilon_t'). \tag{8}$$

In words, we have k VEC models with common shocks that are a function of ε_t. We can write the model compactly as

$$h_t = \omega + A\eta_{t-1} + Bh_{t-1}, \tag{9}$$

where

$$h_t = \begin{pmatrix} h_{1t} \\ h_{2t} \\ \vdots \\ h_{kt} \end{pmatrix}_{kN^*\times 1} \quad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_k \end{pmatrix}_{kN^*\times 1} \quad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_k \end{pmatrix}_{kN^*\times N^*} \quad B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_k \end{pmatrix}_{kN^*\times kN^*}. \tag{10}$$

We refer to the process defined by equations (1)-(10) as the MN-MGARCH(VEC), for mixed normal MGARCH (in VEC version), model.

For later reference, we provide the uncentered conditional second moment of ε_t,

$$\mathrm{E}[\eta_t \mid \mathcal{F}_{t-1}] = \Lambda' h_t + c \tag{11}$$

where $c = \sum_{i=1}^{k}\lambda_i \mathrm{vech}(\mu_i\mu_i')$, and

$$\Lambda = \begin{pmatrix} \lambda_1 I_{N^*} \\ \vdots \\ \lambda_k I_{N^*} \end{pmatrix}_{kN^*\times N^*} \tag{12}$$

Theorem 1 *The process $\{\varepsilon_t\}$ defined by (1)-(10) is covariance stationary if and only if the eigenvalues of the matrix*

$$C = A\Lambda' + B \tag{13}$$

are smaller than one in modulus. In that case,

$$h = E[h_t] = (I_{kN^*} - C)^{-1}(\omega + Ac) \tag{14}$$

and the unconditional covariance matrix is given by

$$E[\eta_t] = \Lambda'(I_{kN^*} - C)^{-1}(\omega + Ac) + c \tag{15}$$

The crucial matrix to check is therefore C, written explicitly

$$C = \begin{pmatrix} \lambda_1 A_1 + B_1 & \lambda_2 A_1 & \cdots & \lambda_k A_1 \\ \lambda_1 A_2 & \lambda_2 A_2 + B_2 & \cdots & \lambda_k A_2 \\ \vdots & & \ddots & \\ \lambda_1 A_k & \cdots & & \lambda_k A_k + B_k \end{pmatrix}_{kN^* \times kN^*}$$

We can get results on the fourth moment structure of the model by assuming that the densities of the individual states are spherical. For simplicity we assume that they are Gaussian with mean zero. Some more notation is necessary. Denote by $\tilde{\Lambda}$ the $N^{*2} \times kN^{*2}$ matrix

$$\tilde{\Lambda} = \left(\lambda_1 I_{N^{*2}}, \cdots, \lambda_k I_{N^{*2}}\right).$$

Furthermore, let P_{kq} be the $kq^2 \times (kq)^2$ permutation matrix such that for any $kq \times kq$ matrix A, $P_{kq}\text{vec}A = (\text{vec}(A_1)', \ldots, \text{vec}(A_k)')'$, where A_j is the j-th $q \times q$ matrix on the block-diagonal of A.

Theorem 2 *For the process defined by (1)-(10), assume that $f(\varepsilon_t \mid \mu_j, \Sigma_{jt}) = N(0, \Sigma_{jt})$. Then a necessary and sufficient condition for finite fourth moments of ε_t is that the eigenvalues of the matrix*

$$Z = (A \otimes A)G_N\tilde{\Lambda}P_{kN^*} + B \otimes B + B \otimes A\Lambda' + A\Lambda' \otimes B \tag{16}$$

have modulus smaller than one, where

$$G_N = 2\{(D_N^+ \otimes D_N^+)(I_N \otimes C_{NN} \otimes I_N)(D_N \otimes D_N) + I_{N^{*2}}\},$$

C_{NN} is the commutation matrix, D_N the duplication matrix and D_N^+ its generalized inverse. In that case, the unconditional fourth moments of ε_t are given by

$$vec(\Sigma_\eta) = G_N\tilde{\Lambda}P_{kN^*}(I_{N^{*2}} - Z)^{-1}\gamma,$$

where

$$\gamma = vec(\omega\omega' + \omega h'\Lambda A' + \omega h'B + A\Lambda' h\omega' + B'h\omega')$$

and h is given by (14). Moreover, the autocovariance function of η_t, $\Gamma(\tau) = E[\eta_t\eta_{t-\tau}] - E[\eta_t]E[\eta_t]'$ is given by

$$\Gamma(\tau) = \Lambda' C^{\tau-1}\left\{A\Sigma_\eta + B\Sigma_h\Lambda - C(I_{kN^*} - C)^{-1}\omega\omega'(I_{kN^*} - C)^{-1}\Lambda\right\},$$

where $\Sigma_h = E[h_t h_t']$.

3 Estimation

We describe how we perform estimation by the maximum likelihood (ML) method (section 3.1), by the expectation-maximization (EM) algorithm (section 3.2) of Dempster, Laird, and Rubin (1977), and by Bayesian inference (section 3.3). We assume that T observation vectors y_t, for $t = 1$ to T, are available for estimation. The link between y_t and ε_t in (1) is given by $\varepsilon_t = y_t - \mathrm{E}(y_t|\mathcal{F}_{t-1})$. We suppose for ease of presentation that the conditional mean is either known or estimated consistently in a first step, so that the residuals ε_t are available for estimation of the parameters of the MN-MGARCH(VEC) model in the second step. We denote by ε the vector of observations $(\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T')'$. We do not write explicitly the observations before $t = 1$, which are used as initial conditions where they should appear. The complete parameter vector, called Ψ, regroups the parameters λ_j, μ_j, and θ_j for $j = 1, \ldots, k$, where θ_j' is the row vector containing all the parameters of ω_j, A_j and B_j, see equation (7). Thus, $\Psi = (\mu', \theta', \lambda')'$, where $\lambda' = (\lambda_1, \lambda_2, \ldots, \lambda_k)$, $\mu' = (\mu_1', \mu_2', \ldots, \mu_k')$, and $\theta' = (\theta_1', \theta_2', \ldots, \theta_k')$.

3.1 ML estimation

The log-likelihood of ε for the MN-MGARCH(VEC) model is given by

$$\mathcal{L}(\Psi; \varepsilon) = \sum_{t=1}^{T} \log \left(\sum_{j=1}^{k} \lambda_j \phi(\varepsilon_t | \mu_j, \theta_j) \right), \tag{17}$$

where $\phi(\cdot|\mu_j, \theta_j)$ denotes a multivariate normal density with mean μ_j and variance-covariance matrix denoted by h_{jt}, see equation (6), h_{jt} being a function of θ_j.

Numerical methods are needed to obtain $\hat{\Psi} = \arg\max \mathcal{L}(\Psi; \varepsilon)$. To avoid the problem of label-switching, we impose the identifying restrictions

$$\lambda_1 > \lambda_2 > \ldots > \lambda_k. \tag{18}$$

Because of these restrictions, we use the FSQP algorithm of Lawrence and Tits (2001) which allows optimisation subject to constraints.

3.2 EM algorithm

In the EM framework, the observed data vector ε is considered as incomplete since we do not know from which component of the mixture each observation is generated. This information is given by the latent variable $z_t = (z_{t1}, z_{t2}, \ldots, z_{tK})'$ where z_{tk} is a dichotomous variable taking the value 1 if ε_t comes from the k-th mixture component, and 0 otherwise. The complete data log-likelihood is given by

$$\mathcal{L}_c(\Psi; \varepsilon) = \sum_{t=1}^{T} \sum_{j=1}^{k} z_{tj} \left[\log \lambda_j + \log \phi_j(\varepsilon_t | \mu_j, \theta_j) \right]. \tag{19}$$

This simplifies the expression of the log-likelihood in (17) because we do not take the logarithm over the entire sum but a sum of logarithms. Because z_t is not observed, we proceed in two steps.

E-step: Suppose that Ψ is known and equal to $\Psi^{(i)}$. We compute the expectation of the unobserved value z_{tj} given all the observations y. This is given by

$$\mathrm{E}(z_{tj}|\varepsilon, \Psi^{(i)}) = \tau_{tj}(\varepsilon; \Psi^{(i)}) = \frac{\lambda_j^{(i)} \phi(\varepsilon_t|\mu_j^{(i)}, \theta_j^{(i)})}{\sum_{j=1}^{k} \lambda_j^{(i)} \phi(\varepsilon_t|\mu_j^{(i)}, \theta_j^{(i)})}. \tag{20}$$

Next, we substitute the latter for z_{tk} in (19). This yields the observed complete data log-likelihood:

$$Q(\Psi, \Psi^{(i)}; \varepsilon) = \sum_{t=1}^{T} \sum_{j=1}^{k} \tau_{tj}(\varepsilon; \Psi^{(i)}) \left[\log \lambda_j + \log \phi(\varepsilon_t|\mu_j, \theta_j)\right]. \tag{21}$$

M-step: We maximize numerically $Q(\Psi, \Psi^{(i)}; \varepsilon)$ with respect to Ψ to get updated estimates of the parameters, denoted by $\Psi^{(i+1)}$. Notice that we have to impose the constraints (18) and (3), so that the maximization has to be done numerically with respect to all the parameters, including the probabilities.

The E-step and M-step are alternated repeatedly until convergence, see McLachlan and Peel (2000) for a detailed description of the application of the EM algorithm to mixture models.

3.3 Bayesian estimation

We introduce for each observation a state variable $S_t \in \{1, 2, \ldots, k\}$ that takes the value j if the observation ε_t belongs to component j. Notice that S_t conveys the same information as z_t in the EM algorithm. The vector S contains the state variables for the T observations. The model specification assumes that the state variables are independent given the group probabilities, and the probability that S_t is equal to j is equal to λ_j. Thus, the joint density of the states given the parameters is

$$\varphi(S|\lambda) = \prod_{t=1}^{T} \varphi(S_t|\lambda) = \prod_{t=1}^{T} \lambda_{S_t}. \tag{22}$$

Given S and Ψ, the joint density of ε is

$$f(\varepsilon|\Psi, S) = \prod_{t=1}^{T} \phi(\varepsilon_t|\mu_{S_t}, \theta_{S_t}). \tag{23}$$

This would be the likelihood function to use if the states were observed. Since they are not, we treat S as a parameter of the model. This technique is called data augmentation, see Tanner and Wong (1987) for more details. Although the augmented model contains more parameters, inference is feasible by making use of Markov chain Monte Carlo (MCMC) methods. In this paper we implement a Gibbs sampling algorithm that allows to sample from the posterior distribution of S and Ψ by sampling from the full conditional posterior densities of subsets of parameters, which are called the blocks of the Gibbs sampler. The joint posterior distribution is given by

$$\varphi(S, \mu, \theta, \lambda|\varepsilon) \propto \varphi(\mu)\, \varphi(\theta)\, \varphi(\lambda) \prod_{t=1}^{T} \lambda_{S_t} \phi(\varepsilon_t|\mu_{S_t}, \theta_{S_t}), \tag{24}$$

where $\varphi(\mu)$, $\varphi(\theta)$, $\varphi(\lambda)$ are the corresponding prior densities. Thus we assume prior independence between λ, μ and θ. We define these prior densities below while we explain the different blocks of the Gibbs sampler.

3.3.1 Sampling S from $\varphi(S|\mu,\theta,\lambda,\varepsilon)$

Given μ,θ,λ and ε, the posterior density of S is proportional to $\varphi(S|\lambda)f(\varepsilon|\Psi,S)$. It turns out that the S_t's are mutually independent, so that we can write the relevant conditional posterior density as

$$\varphi(S|\mu,\theta,\lambda,\varepsilon)=\prod_{t=1}^{T}\varphi(S_t|\mu,\theta,\lambda,\varepsilon), \tag{25}$$

where $\varphi(S_t|\mu,\theta,\lambda,\varepsilon)$ is a discrete distribution explicitly defined as

$$P(S_t=j|\mu,\theta,\lambda,\varepsilon)=\frac{\lambda_j\phi(\varepsilon_t|\mu_j,\theta_j)}{\sum_{i=1}^{k}\lambda_j\phi(\varepsilon_t|\mu_j,\theta_j)},\ (j=1,\ldots,k). \tag{26}$$

To sample S_t we draw a random number from a uniform distribution on $(0,1)$ and decide which group j to take according to (26).

3.3.2 Sampling λ from $\varphi(\lambda|S^T,\mu,\theta,\varepsilon)$

The full conditional posterior density of λ is given by

$$\varphi(\lambda|S,\varepsilon)=\varphi(\lambda|S)\propto\varphi(\lambda)\prod_{j=1}^{k}\lambda_j^{x_j} \tag{27}$$

where x_j is the number of times that $S_t=j$. The prior $\varphi(\lambda)$ is chosen to be a Dirichlet distribution, $Di(a_{10},a_{20}\cdots a_{k0})$ with parameter vector $a_0=(a_{10},a_{20}\cdots a_{k0})$. As a consequence, $\varphi(\lambda|S,\varepsilon)$ is also a Dirichlet distribution, $Di(a_1,a_2\cdots a_k)$ with $a_j=a_{j0}+x_j$, $j=1,2,\ldots,k$. Notice that it does not depend on μ and θ. To sample a $Di(a_1,a_2\cdots a_k)$ distribution, we sample k independent gamma random variables, $X_j\sim G(a_j,1)$, and transform them to (see Wilks (1962))

$$\begin{aligned}\lambda_j &= \frac{X_j}{X_1+\ldots+X_k} \quad j=1,\ldots,k-1\\ \lambda_k &= 1-\lambda_1-\lambda_2-\ldots-\lambda_{k-1}.\end{aligned}$$

3.3.3 Sampling μ from $\varphi(\mu|S,\lambda,\theta,\varepsilon)$

We sample $\tilde{\mu}'=(\mu_1',\mu_2',\ldots,\mu_{k-1}')$ and recover μ_k by use of (3) since λ is known. The likelihood contribution to the full conditional posterior density of $\tilde{\mu}$, given in (23), can be shown (see the Appendix) to be proportional to a multivariate normal density with variance-covariance matrix A^{-1} and mean $A^{-1}b$ defined below.

Theorem 3 *$f(\varepsilon|\Psi,S)\propto\exp\big[-\frac{1}{2}(\tilde{\mu}-A^{-1}b)'A(\tilde{\mu}-A^{-1}b)\big]$, where $p=(k-1)N$,*

$$A=diag\left(\sum_{t\in\{S_t=1\}}\Sigma_{1t}^{-1},\ldots,\sum_{t\in\{S_t=k-1\}}\Sigma_{k-1,t}^{-1}\right)+\frac{\tilde{\lambda}\tilde{\lambda}'}{\lambda_k^2}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}, \tag{28}$$

denoting $\tilde{\lambda}=(\lambda_1,\ldots,\lambda_{k-1})$, and

$$b=\begin{bmatrix}\sum_{t\in\{S_t=1\}}\Sigma_{1t}^{-1}\varepsilon_t-\frac{\lambda_1}{\lambda_k}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\varepsilon_t\\ \vdots\\ \sum_{t\in\{S_t=k-1\}}\Sigma_{k-1,t}^{-1}\varepsilon_t-\frac{\lambda_{k-1}}{\lambda_k}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\varepsilon_t\end{bmatrix}. \tag{29}$$

The variance-covariance matrix A^{-1} is not block diagonal because of the restriction (3). From the proposition, we deduce that if $\varphi(\tilde{\mu})$ is either a normal density or is non-informative (i.e. proportional to a constant), then $\varphi(\mu|S,\lambda,\theta,\varepsilon)$, the full conditional posterior of μ, is also a normal density.

3.3.4 Sampling θ from $\varphi(\theta|S,\mu,\lambda,\varepsilon)$

By assuming prior independence between the θ_j's, i.e. $\varphi(\theta) = \prod_{j=1}^{k} \varphi(\theta_j)$, it follows that

$$\varphi(\theta|S,\mu,\lambda,\varepsilon) = \varphi(\theta|S,\mu,\varepsilon) = \varphi(\theta_1|\mu_1,\widetilde{\varepsilon}^1)\varphi(\theta_2|\mu_2,\widetilde{\varepsilon}^2)\cdots\varphi(\theta_k|\mu_k,\widetilde{\varepsilon}^k) \tag{30}$$

where $\widetilde{\varepsilon}^j = \{\varepsilon_t|S_t = j\}$ and

$$\varphi(\theta_j|\mu_j,\widetilde{\varepsilon}^j) \propto \varphi(\theta_j) \prod_{t\in\{S_t=j\}} \phi(\varepsilon_t|\mu_j,\theta_j). \tag{31}$$

Since we condition on the state variables, we can simulate each block θ_j separately. We do this with the griddy-Gibbs sampler, see Bauwens, Lubrano, and Richard (1999) for details. Note that lower and upper bounds for each parameter must be selected. The choice of these bounds needs to be fine tuned in order to cover the range of the parameter over which the posterior is relevant. The prior for each individual parameter can be uniform between these bounds.

4 Illustration with simulated data

We illustrate the estimation methods on two bivariate two component data generating processes for which we simulate one dataset each. The first one has one stable component with high probability and one unstable component. The parameters are given by

DGP1

$$\lambda_1 = 0.8, \quad \mu_1 = \begin{pmatrix} 0.1 \\ 0.05 \end{pmatrix}, \quad \omega_1 = \begin{pmatrix} 0.001 \\ 0.005 \\ 0.02 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0.05 & 0.0 & 0.0 \\ 0.0 & 0.04 & 0.0 \\ 0.0 & 0.0 & 0.06 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0.92 & 0.0 & 0.0 \\ 0.0 & 0.9 & 0.0 \\ 0.0 & 0.0 & 0.85 \end{pmatrix}$$

$$\lambda_2 = 0.2, \quad \mu_2 = \begin{pmatrix} -0.4 \\ -0.2 \end{pmatrix}, \quad \omega_2 = \begin{pmatrix} 0.015 \\ 0.01 \\ 0.05 \end{pmatrix} \quad A_2 = \begin{pmatrix} 0.25 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.3 \end{pmatrix} \quad B_2 = \begin{pmatrix} 0.85 & 0.0 & 0.0 \\ 0.0 & 0.75 & 0.0 \\ 0.0 & 0.0 & 0.8 \end{pmatrix}.$$

The largest eigenvalue of the matrix C in (13) is 0.96162 which is smaller than 1 so the overall process is stationary, even if for example $A_{2,11} + B_{2,11}$ is larger than 1. The implied unconditional standard deviations for the first and second series are respectively 0.648 and 0.662 and the unconditional correlation is 0.305.

The second DGP has the same first component as DGP1 but the second component is now less persistent than the first one. This is done by lowering the values in A_2 and B_2. The parameters are given by

DGP2

$$\lambda_1 = 0.8, \quad \mu_1 = \begin{pmatrix} 0.1 \\ 0.05 \end{pmatrix}, \quad \omega_1 = \begin{pmatrix} 0.001 \\ 0.005 \\ 0.02 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0.05 & 0.0 & 0.0 \\ 0.0 & 0.04 & 0.0 \\ 0.0 & 0.0 & 0.06 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0.92 & 0.0 & 0.0 \\ 0.0 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.85 \end{pmatrix}$$

$$\lambda_2 = 0.2, \quad \mu_2 = \begin{pmatrix} -0.4 \\ -0.2 \end{pmatrix}, \quad \omega_2 = \begin{pmatrix} 0.015 \\ 0.01 \\ 0.05 \end{pmatrix} \quad A_2 = \begin{pmatrix} 0.15 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.2 \end{pmatrix} \quad B_2 = \begin{pmatrix} 0.45 & 0.0 & 0.0 \\ 0.0 & 0.35 & 0.0 \\ 0.0 & 0.0 & 0.5 \end{pmatrix}$$

The largest eigenvalue of the matrix C in (13) is given by 0.96021 which is smaller than 1 so the overall process is stationary which is not surprising here since both components are stable. The implied unconditional standard deviations for the first and second series are respectively 0.353 and 0.477 and the unconditional correlation is 0.316.

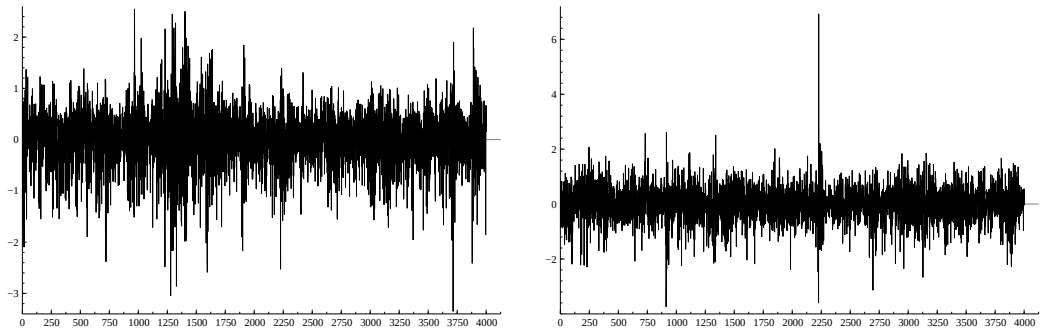
We simulate $T = 4000$ observations for DGP1 and DGP2. The sample paths, marginal kernel density estimates and sample autocorrelation functions of the data simulated using DGP 1 are given in Figure 1. A bivariate kernel density estimate is given in Figure 2. This estimate is based on a Gaussian product kernel with a scalar bandwidth computed using the rule of thumb. From the graphs we see that the sample autocorrelations for the squared data persist less longer for the second series as we expect given the DGP1 parameter values, and that there is a more negative skewness in the first series than in the second. This is indeed confirmed by the summary statistics given in Table 1. The estimated kurtosis coefficient is higher for the second series, though this is likely due to the high maximum in that series. Note that the empirical second moments match the theoretical second moments reasonably well, for example the estimated and theoretical correlation are respectively given by 0.314 and 0.305.

Table 1: DGP 1 summary statistics

	$T = 4000$	
	first series	**second series**
Mean	0.0462	0.0416
Standard Deviation	0.5695	0.63826
Maximum	2.5481	6.922
Minimum	−3.3493	−3.7382
Skewness	−0.45257	−0.03566
Kurtosis	5.298	7.8619

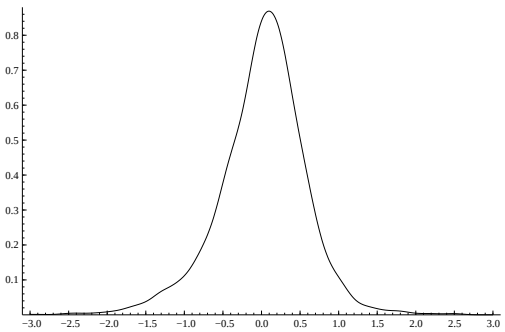
Descriptive statistics of the data simulated using DGP 1. The estimated correlation coefficient is 0.31433.

We estimate the parameters of DGP1 using maximum likelihood (ML), the EM algorithm and by Bayesian inference (Bayes), see Section 3 for details. The results are given in Table 2. The ML estimates

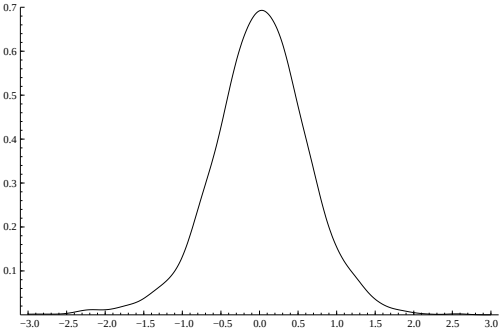


(a) sample path of first series

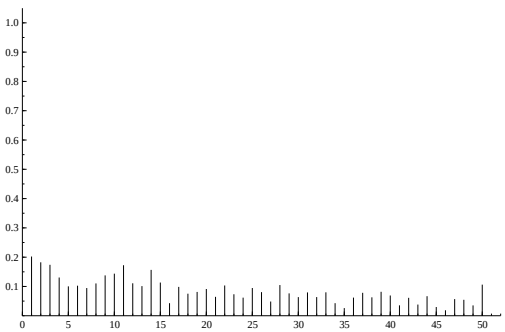
(b) sample path of second series



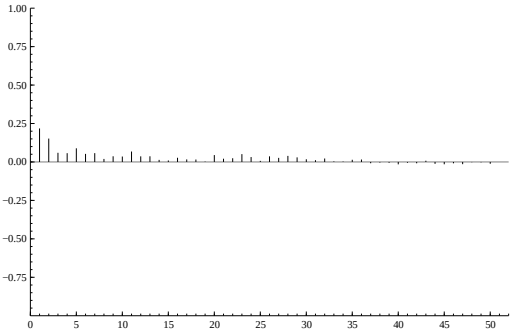
(c) kernel density estimate of first series



(d) kernel density estimate of second series



(e) sample ACF of first squared series



(f) sample ACF of second squared series

Figure 1: Sample paths, kernel density estimates and sample autocorrelation functions of the data simulated using DGP 1 (4000 observations).

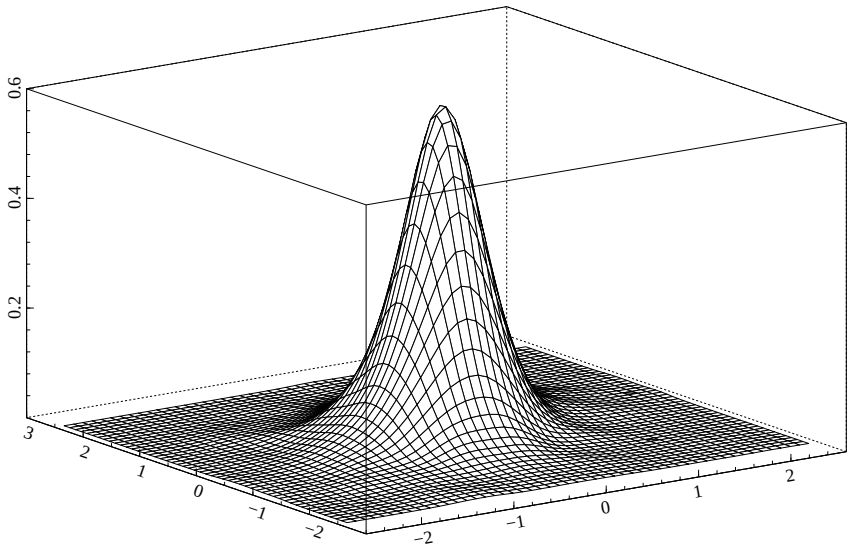


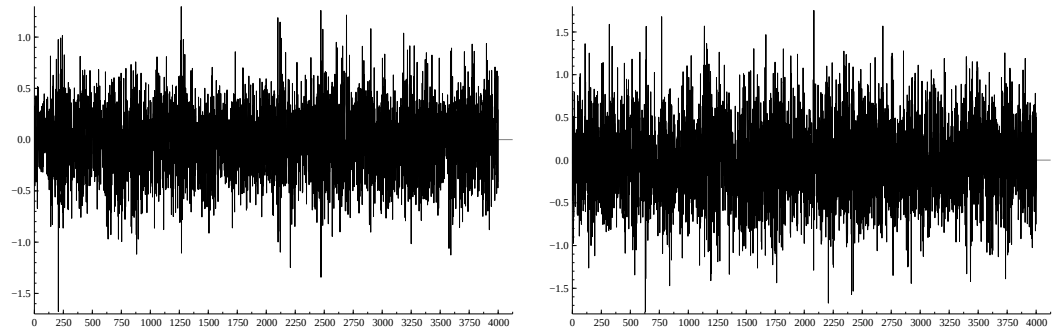
Figure 2: kernel density estimate of the data simulated using DGP 1 (4000 observations)

are reasonably close to the true parameter values given the standard errors which are computed via the evaluation of the Hessian at the estimates. Note that the standard errors of the parameters in the second component are drastically higher compared to those of the first component. The likelihood curvature is indeed much smaller for the second component because only $800 = 0.2 \times 4000$ observations are expected to come from that component. The EM estimates are almost identical to the ML estimates. The EM standard errors are computed in the same way as for the ML estimates so they also hardly differ. The Bayes' results are based on 2400 draws of which 400 were discarded to warm up the sampler. Though these results are only indicative in the sense that the marginal posterior standard deviations are too different from the ML standard errors. This is due to too tightly chosen supports, not displayed here, for the parameters drawn using the griddy Gibbs sampler. Therefore, the bounds should be adapted to fully cover the parameter supports. Nevertheless, the posterior standard deviations for the parameters λ_1 and μ_1 which are sampled with an uninformative prior are reasonably close to their ML standard errors.

We now turn to DGP2. The sample paths, marginal kernel density estimates and sample autocorrelation functions of the simulated data are given in Figure 3. A bivariate kernel density estimate is given in Figure 4. Descriptive statistics are given in Table 3. The lower autocorrelations in the squared data compared to DGP1 are not surprising given the now much less persistent second component in the mixture. The standard deviations are also smaller compared to DGP1 because we keep the same values in DGP2 for ω_1 and ω_2. Estimation results for DGP2 are given in Table 4. The ML parameter estimates are again reasonably close to the DGP values. Regarding the EM estimates we find that the parameters of the first component, that is $\hat{\omega}_1, \hat{A}_1$ and $\hat{B}_1$ are very close to the ML estimates. The other EM parameter estimates, that is $\hat{\lambda}, \hat{\mu}_1, \hat{\omega}_2, \hat{A}_2$ and $\hat{B}_2$ are slightly closer to the true parameter values than the ML estimates for this simulated dataset.

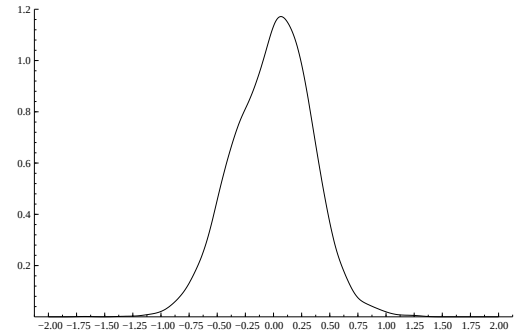
Table 2: DGP1 Two components results

	DGP1	ML		EM		Bayes	
		estimate	std error	estimate	std error	mean	std dev.
λ_1	0.8	0.81174	0.020487	0.81167	0.02050	0.81011	0.02110
$\mu_{1,11}$	0.1	0.10528	0.00803	0.10529	0.00803	0.10496	0.00868
$\mu_{1,21}$	0.05	0.05824	0.00923	0.05823	0.00924	0.05840	0.00870
$\omega_{1,11}$	0.001	0.00078	0.00053	0.00078	0.00053	0.00465	0.00017
$\omega_{1,22}$	0.005	0.00379	0.00097	0.00379	0.00097	0.00427	0.00012
$\omega_{1,33}$	0.02	0.01499	0.00378	0.01498	0.00378	0.01449	0.00244
$A_{1,11}$	0.05	0.04380	0.00470	0.04380	0.00470	0.05330	0.00523
$A_{1,22}$	0.04	0.03262	0.00516	0.03262	0.00516	0.03092	0.00095
$A_{1,33}$	0.06	0.05240	0.00878	0.05239	0.00878	0.05031	0.00508
$B_{1,11}$	0.9	0.92969	0.00752	0.92969	0.00753	0.89169	0.00795
$B_{1,22}$	0.9	0.91584	0.01478	0.91585	0.01478	0.90754	0.00206
$B_{1,33}$	0.85	0.87628	0.02097	0.87630	0.02097	0.87855	0.01327
$\omega_{2,11}$	0.015	0.01983	0.01023	0.01984	0.01023	0.01755	0.00228
$\omega_{2,22}$	0.01	-0.00283	0.00968	-0.00278	0.00971	0.00150	0.00056
$\omega_{2,33}$	0.05	0.09691	0.03842	0.09673	0.03836	0.09754	0.00704
$A_{2,11}$	0.25	0.23442	0.05209	0.23441	0.05208	0.25606	0.04056
$A_{2,22}$	0.2	0.19382	0.05696	0.19393	0.05692	0.19337	0.02325
$A_{2,33}$	0.3	0.30244	0.07403	0.30198	0.07389	0.30289	0.02642
$B_{2,11}$	0.85	0.82361	0.03878	0.82360	0.03877	0.82661	0.01794
$B_{2,22}$	0.75	0.70363	0.09606	0.70273	0.09659	0.69606	0.02533
$B_{2,33}$	0.8	0.72776	0.06823	0.72815	0.06815	0.72917	0.01470

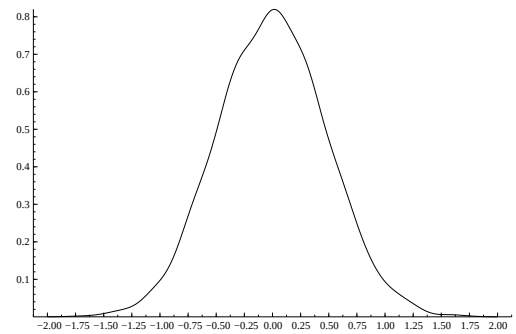


(a) sample path of first series

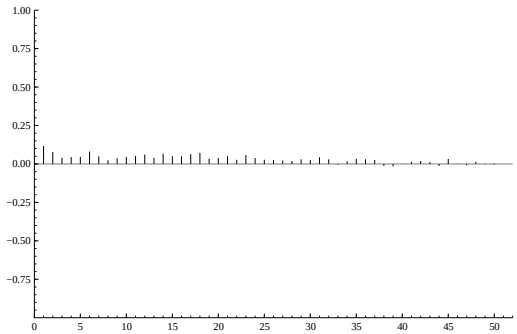
(b) sample path of second series



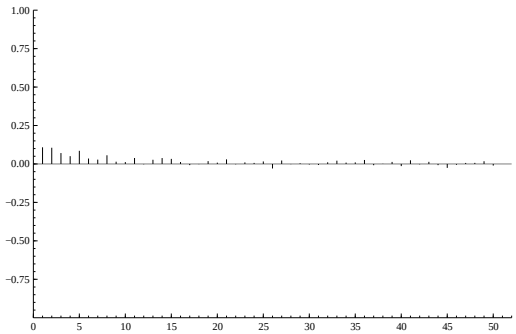
(c) kernel density estimate of first series



(d) kernel density estimate of second series



(e) sample ACF of first squared series



(f) sample ACF of second squared series

Figure 3: Sample paths, kernel density estimates and sample autocorrelation functions of the data simulated using DGP 2 (4000 observations).

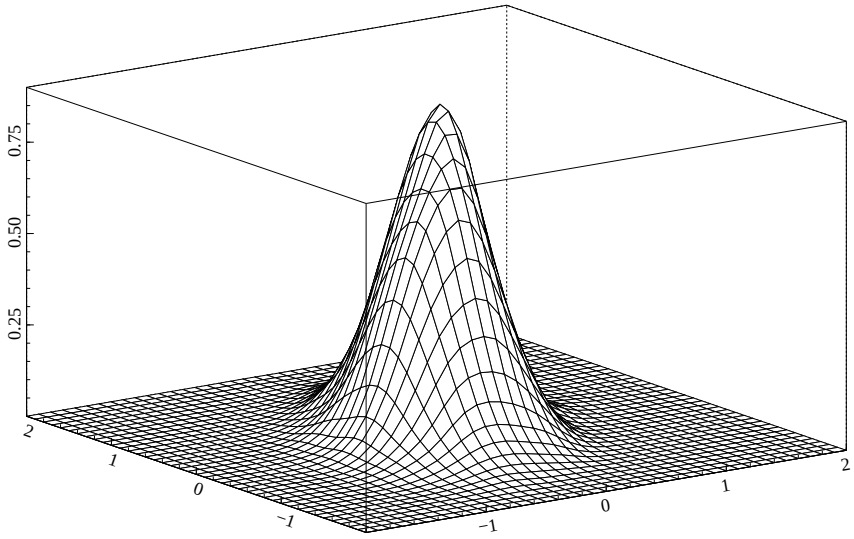


Figure 4: kernel density estimate of the data simulated using DGP 1 (4000 observations)

Table 3: DGP 2 summary statistics

	$T = 4000$	
	first series	**second series**
Mean	−0.01015	−0.00514
Standard Deviation	0.34385	0.48067
Maximum	1.2971	1.7524
Minimum	−1.6771	−1.7772
Skewness	−0.12201	0.01965
Kurtosis	3.2058	3.0245

Descriptive statistics of the data simulated using DGP 2. The estimated correlation coefficient is 0.32746.

Table 4: DGP2 Two components results

	DGP2	ML		EM	
		estimate	std error	estimate	std error
λ_1	0.8	0.81942	0.02863	0.81080	0.03019
$\mu_{1,11}$	0.1	0.08860	0.01207	0.09211	0.01247
$\mu_{1,21}$	0.05	0.05404	0.01172	0.05394	0.01203
$\omega_{1,11}$	0.001	0.00113	0.00057	0.00106	0.00056
$\omega_{1,22}$	0.005	0.00159	0.00082	0.00159	0.00082
$\omega_{1,33}$	0.02	0.02760	0.00688	0.02743	0.00691
$A_{1,11}$	0.05	0.05721	0.00685	0.05700	0.00682
$A_{1,22}$	0.04	0.03147	0.00877	0.03134	0.00876
$A_{1,33}$	0.06	0.07890	0.01427	0.07813	0.01434
$B_{1,11}$	0.9	0.91040	0.01030	0.91052	0.01029
$B_{1,22}$	0.8	0.90690	0.03026	0.90759	0.03000
$B_{1,33}$	0.85	0.79543	0.03871	0.79753	0.03880
$\omega_{2,11}$	0.015	0.01171	0.00555	0.01191	0.00560
$\omega_{2,22}$	0.01	0.00398	0.00462	0.00526	0.00494
$\omega_{2,33}$	0.05	0.04923	0.01948	0.05009	0.01931
$A_{2,11}$	0.15	0.17075	0.04100	0.16976	0.04060
$A_{2,22}$	0.1	0.09478	0.03682	0.09636	0.03637
$A_{2,33}$	0.2	0.24119	0.05945	0.23755	0.05842
$B_{2,11}$	0.45	0.43228	0.14142	0.44277	0.13984
$B_{2,22}$	0.35	0.34712	0.19099	0.35254	0.18539
$B_{2,33}$	0.5	0.44934	0.12669	0.45841	0.12265

To be sure that the results are correct, we generate some extra sample paths for both DGP1 and DGP2 of the same sample size and then we estimate the model parameters again, the results of which are not reported here. It follows that the conclusions are the same as described above in this section.

5 Application

We model daily return data from the Bank of America and Boeing stocks using a sample from 01/01/1980 to 30/07/2003 implying 6152 observations downloaded from Datastream. Daily returns are measured by log-differences of closing prices. The sample paths, marginal kernel density estimates and sample autocorrelation functions of the data are given in Figure 5. A bivariate kernel density estimate is given in Figure 6. Both companies share similar summary statistics which are given in Table 5. Some important events between 1980 and 2003 give rise to several extreme values for both companies. These values are not discarded from the sample. We start by fitting univariate one and two component models to learn

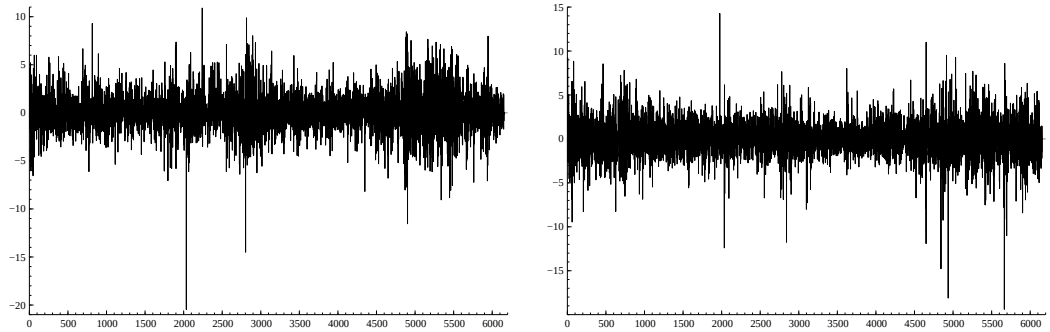
Table 5: Bank of America - Boeing summary statistics

	01/01/1980−30/07/2003 $T = 6152$	
	Bank of America	**Boeing**
Mean	0.05184	0.03044
Standard Deviation	1.8922	1.9732
Maximum	10.903	14.278
Minimum	−20.458	−19.389
Skewness	−0.17633	−0.28139
Kurtosis	8.0566	9.1044

Descriptive statistics for the Bank of America - Boeing data. The estimated correlation coefficient is 0.25448.

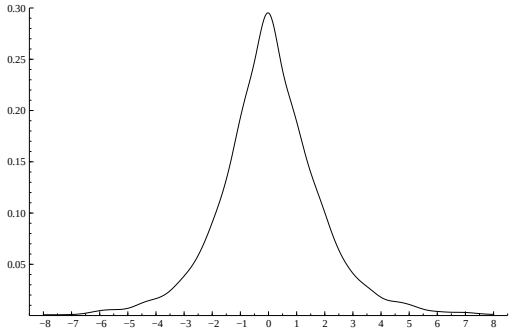
more about the individual time series dynamics of both companies and also to get an idea of good starting values for the multivariate mixture model. The ML estimates for the univariate models are displayed in Table 6. We can also apply Bayesian inference or the EM algorithm but the results are very similar to the ML estimates and are not reported. The one component model, or the usual GARCH model, estimates for both Bank of America and Boeing imply stationary but highly persistent processes. The two component mixture model parameter estimates reveal indeed that for both companies the second component is not stable with probabilities belonging to that component respectively given by 0.165 and 0.079.

The estimation results for the bivariate one and two component models are given in Table 7. The largest eigenvalue of the estimated matrix C in (13) is given by 0.98435 which implies a stationary process. The implied estimated unconditional standard deviations for Bank of America and Boeing are

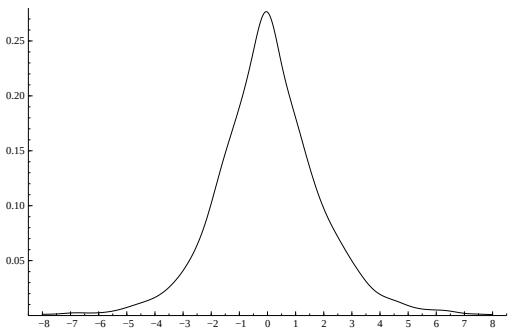


(a) sample path of Bank of America

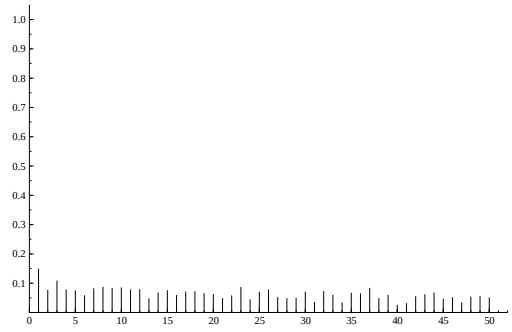
(b) sample path of Boeing



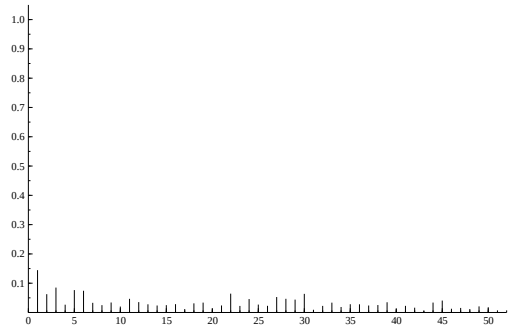
(c) kernel density estimate of Bank of America



(d) kernel density estimate of Boeing



(e) sample ACF of Bank of America squared series



(f) sample ACF of Boeing squared series

Figure 5: Sample paths, kernel density estimates and sample autocorrelation functions for the Bank of America - Boeing data

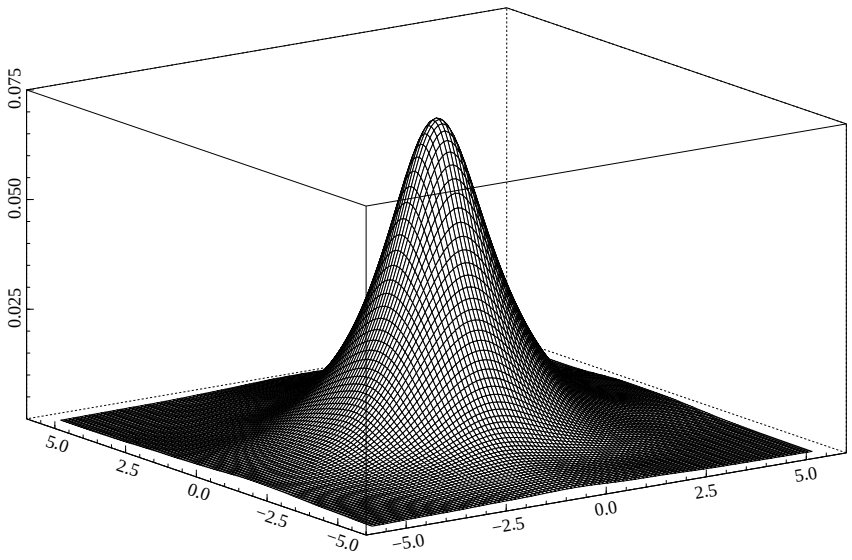


Figure 6: kernel density estimate for the Bank of America - Boeing data

Table 6: Univariate estimation results

	Bank of America				Boeing			
	estimate	std error	estimate	std error	estimate	std error	estimate	std error
λ_1	-	-	0.83546	0.02812	-	-	0.92084	0.01597
μ_1	-	-	0.02161	0.01960	-	-	-0.00607	0.01678
ω_1	0.13539	0.02952	0.03144	0.01298	0.05496	0.00172	0.02882	0.00879
$A_{1,11}$	0.08175	0.01214	0.04652	0.00925	0.04009	0.00927	0.03010	0.00476
$B_{1,11}$	0.88053	0.01901	0.91757	0.01721	0.94618	0.01640	0.94770	0.00859
ω_2	-	-	3.1304	0.94430	-	-	2.6525	1.2180
$A_{2,11}$	-	-	0.69763	0.18304	-	-	0.51102	0.20203
$A_{2,11}$	-	-	0.41636	0.12488	-	-	0.73011	0.09004

Results for the one component (first two columns for each company) and two component (last two columns for each company) univariate mixture GARCH(1,1) models. All the parameters are estimated by ML.

respectively given by 1.8653 and 1.9716 and the unconditional correlation is 0.209, which is close to the summary statistics reported in Table 5. Comparing the univariate one component estimates with their equivalents in the bivariate one component model, or the usual diagonal VEC model, we see that they differ only marginally as expected. Generally speaking, this is also true for the bivariate two component model but to a lesser extent so for the second component which is now stable. The large difference in the loglikelihood function values evaluated at their ML estimates between the one and the two component models allows to reject easily a likelihood ratio test in favor of the more general model.

6 Conclusion

The multivariate mixture model we have proposed in this paper can be extended in several ways. One can use other multivariate GARCH models for the components than the VEC formulation. We refer to the survey of Bauwens, Laurent, and Rombouts (2006) for other multivariate GARCH models. One advantage of the VEC specification is the ease with which moments can be derived. One could also think of using non-normal distributions, but this may not be worth the effort since a mixture of normal distributions allows for a lot of flexibility. The most important challenge at this stage is to improve upon the estimation algorithms (especially the Bayesian one) and to test them with time series of higher dimension. Another topic for future research is to evaluate the models on statistical and economic criteria, in comparison with one-component models.

Appendix

Proof of Theorem 1: Let $u_t = \eta_t - \Lambda' h_t - c$ and note that $\mathrm{E}[u_t \mid I_{t-1}] = 0$. Write (9) as

$$\begin{aligned} h_t &= \omega + A(\Lambda' h_{t-1} + c + u_{t-1}) + B h_{t-1} \\ &= (\omega + Ac) + (A\Lambda' + B) h_{t-1} + A u_{t-1} \end{aligned}$$

Denoting the lag operator by L and $C = A\Lambda' + B$, this can be written as

$$(I_{kN^*} - CL) h_t = (\omega + Ac) + A u_{t-1}. \tag{32}$$

The linear operator $(I_{kN^*} - CL)$ is invertible if and only if all eigenvalues of C have modulus smaller than one. In that case we can write $h_t = (I_{kN^*} - C)^{-1}(\omega + Ac) + (I_{kN^*} - CL)^{-1} A u_{t-1}$, which is a VMA($\infty$) representation of $\{h_t\}$ from which we directly deduce $h = \mathrm{E}[h_t] = (I_{kN^*} - C)^{-1}(\omega + Ac)$. Premultiplying both sides of (32) by the adjoint, $(I_{kN^*} - CL)^*$, we obtain

$$\det(I_{kN^*} - CL) h_t = (I_{kN^*} - C)^*(\omega + Ac) + (I_{kN^*} - CL)^* A u_{t-1}$$

Premultiplying by Λ' and using $\Lambda' h_t = \eta_t - u_t - c$ gives

$$\det(I_{kN^*} - CL)(\eta_t - u_t - c) = \Lambda'(I_{kN^*} - C)^*(\omega + Ac) + \Lambda'(I_{kN^*} - CL)^* A u_{t-1}$$

Table 7: Bank of America - Boeing results

	ML				EM	
	estimate	std error	estimate	std error	estimate	std error
λ_1	-	-	0.85592	0.01889	0.84833	0.01954
$\mu_{1,11}$	-	-	0.03675	0.01771	0.03720	0.01801
$\mu_{1,21}$	-	-	-0.01075	0.01907	-0.01290	0.01937
$\omega_{1,11}$	0.14562	0.02857	0.03041	0.00996	0.02957	0.00986
$\omega_{1,22}$	0.02617	0.00753	0.00372	0.00159	0.00365	0.00158
$\omega_{1,33}$	0.07300	0.01677	0.02810	0.00880	0.02768	0.00879
$A_{1,11}$	0.08353	0.01124	0.04588	0.00754	0.04536	0.00746
$A_{1,22}$	0.02587	0.00506	0.01146	0.00270	0.01137	0.00266
$A_{1,33}$	0.04122	0.00535	0.02748	0.00439	0.02730	0.00441
$B_{1,11}$	0.87564	0.01786	0.92578	0.01243	0.92609	0.01244
$B_{1,22}$	0.93974	0.01312	0.97453	0.0052841	0.97451	0.00528
$B_{1,33}$	0.94011	0.00891	0.94771	0.0088205	0.94761	0.00894
$\omega_{2,11}$	-	-	3.5169	1.0715	3.3482	0.95338
$\omega_{2,22}$	-	-	0.41860	0.29750	0.41166	0.27223
$\omega_{2,33}$	-	-	2.2500	0.83084	2.1491	0.77734
$A_{2,11}$	-	-	0.60068	0.15838	0.58961	0.15030
$A_{2,22}$	-	-	0.13174	0.07384	0.12842	0.06905
$A_{2,33}$	-	-	0.25788	0.08024	0.25432	0.07685
$B_{2,11}$	-	-	0.36883	0.14227	0.38091	0.12996
$B_{2,22}$	-	-	0.78676	0.12018	0.78481	0.11659
$B_{2,33}$	-	-	0.72074	0.07796	0.72330	0.07600

Results for the one component (first two columns) and two component (last four columns) bivariate mixture model. The value of the loglikelihood function evaluated at the ML estimates of the one and two component models are respectively given by -24663.714 and -24177.475.

The process is stable if and only if all roots of the characteristic equation $\det(I_{kN^*} - Cz) = 0$ lie outside the unit circle or, equivalently, all eigenvalues of C have modulus smaller than one. Finally, dividing both sides by $\det(I_{kN^*} - CL)$ and rearranging yields

$$\eta_t = \Lambda'(I_{kN^*} - C)^{-1}(\omega + Ac) + c + \Lambda'(I_{kN^*} - CL)^{-1}Au_{t-1} + u_t$$

This is the VMA(∞) representation of $\{\eta_t\}$ and we deduce directly the unconditional variance of $\{\varepsilon_t\}$, i.e.

$$\text{vech}(\text{Var}(\varepsilon_t)) = \text{E}[\eta_t] = \Lambda'(I_{kN^*} - C)^{-1}(\omega + Ac) + c$$

Proof of Theorem 2: First, $\text{vec}(\text{E}[\eta_t\eta_t' \mid \mathcal{F}_{t-1}]) = G_N \sum_{j=1}^{k} \lambda_j \text{vec}(h_{jt}h_{jt}')$ by application of Theorem 1 of Hafner (2003). Taking the expectation operator on both sides yields

$$\text{vec}(\Sigma_\eta) = G_N\tilde{\Lambda}P_{kN^*}\text{vec}(\Sigma_h) \tag{33}$$

where $\Sigma_\eta = \text{E}[\eta_t\eta_t']$ and $\Sigma_h = \text{E}[h_th_t']$. Substituting the model for h_t in Σ_h, one obtains

$$\begin{aligned}
\text{vec}(\Sigma_h) &= \text{vec}(\omega\omega' + \omega h'\Lambda A' + \omega h'B + A\Lambda' h\omega' + B'h\omega') \\
&+ (A \otimes A)\text{vec}(\text{E}[\eta_{t-1}\eta_{t-1}']) + (B \otimes B)\text{vec}(\text{E}[h_{t-1}h_{t-1}']) \\
&+ (B \otimes A)\text{vec}(\text{E}[\eta_{t-1}h_{t-1}']) + (A \otimes B)\text{vec}(\text{E}[h_{t-1}\eta_{t-1}]) \\
&= \gamma + (A \otimes A)\text{vec}(\Sigma_\eta) + (B \otimes B)\text{vec}(\Sigma_h) \\
&+ (B \otimes A)\text{vec}(\text{E}[\text{E}(\eta_{t-1}h_{t-1}' \mid \mathcal{F}_{t-1})]) + (A \otimes B)\text{vec}(\text{E}[\text{E}(h_{t-1}\eta_{t-1} \mid \mathcal{F}_{t-1})]) \\
&= \gamma + (A \otimes A)G_N\tilde{\Lambda}P_{kN^*}\text{vec}(\Sigma_h) + (B \otimes B)\text{vec}(\Sigma_h) \\
&+ (B \otimes A)\text{vec}(\text{E}[\Lambda' h_{t-1}h_{t-1}']) + (A \otimes B)\text{vec}(\text{E}[h_{t-1}h_{t-1}'\Lambda]) \\
&= \gamma + (A \otimes A)G_N\tilde{\Lambda}P_{kN^*}\text{vec}(\Sigma_h) + (B \otimes B)\text{vec}(\Sigma_h) + (B \otimes A\Lambda')\text{vec}(\Sigma_h) + (A\Lambda' \otimes B)\text{vec}(\Sigma_h) \\
&= \gamma + Z\text{vec}(\Sigma_h).
\end{aligned}$$

Rearranging gives the result provided that $I_{N^{*2}} - Z$ is invertible, which is the case if and only if all eigenvalues of Z have modulus smaller than one. Finally, application of (33) yields the desired result for Σ_η.

For the second part of the theorem, note that

$$\begin{aligned}
\text{E}[h_t \mid \mathcal{F}_{t-\tau}] &= (I_{kN^*} + C + \cdots + C^{\tau-1})\omega + C^{\tau-1}(A\eta_{t-\tau} + Bh_{t-\tau}) \\
&= (I_{kN^*} - C^\tau)(I_{kN^*} - C)^{-1}\omega + C^{\tau-1}(A\eta_{t-\tau} + Bh_{t-\tau})
\end{aligned}$$

Now,

$$\begin{aligned}
\text{E}[\eta_t\eta_{t-\tau}] &= \text{E}[\text{E}(\eta_t \mid \mathcal{F}_{t-1})\eta_{t-\tau}'] \\
&= \text{E}[\Lambda' h_t\eta_{t-\tau}'] \\
&= \text{E}[\Lambda'\text{E}(h_t \mid \mathcal{F}_{t-\tau})\eta_{t-\tau}'] \\
&= \text{E}[\Lambda'\{(I_{kN^*} - C^\tau)(I_{kN^*} - C)^{-1}\omega + C^{\tau-1}(A\eta_{t-\tau} + Bh_{t-\tau})\}\eta_{t-\tau}']
\end{aligned}$$

$$= \quad \Lambda'(I_{kN^*} - C^\tau)(I_{kN^*} - C)^{-1}\omega\omega'(I_{kN^*} - C')^{-1}\Lambda + \Lambda' C^{\tau-1}(A\Sigma_\eta + B\Sigma_h\Lambda)$$

Subtracting $\mathrm{E}[\eta_t]\mathrm{E}[\eta_t]' = \Lambda'\{(I_{kN^*} - C)^{-1}\omega\omega'(I_{kN^*} - C')^{-1}\Lambda$, the result for $\Gamma(\tau)$ is obtained.

Proof of Theorem 3: We start from (23), where we substitute $-\sum_{j=1}^{k-1}(\lambda_j/\lambda_k)\mu_j$ for μ_k and we neglect all the factors that do not depend on $\tilde{\mu}$. Given the state variables, we know to which group each observation ε_t belongs and we denote by $\{S_t = j\}$ the set of indices of the observations belonging to group j. Thus, taking the logarithm of (23) and multiplying it by -2, we get

$$\begin{aligned}
&-2\sum_{t=1}^{T}\log\phi(\varepsilon_t|\mu_{S_t},\theta_{S_t}) - C \;=\; -2\sum_{j=1}^{k}\sum_{t\in\{S_t=j\}}\log\phi(\varepsilon_t|\mu_j,\theta_j) - C \\
&= \sum_{j=1}^{k-1}\sum_{t\in\{S_t=j\}}(\varepsilon_t-\mu_j)'\Sigma_{jt}^{-1}(\varepsilon_t-\mu_j) \;+\; \sum_{t\in\{S_t=k\}}\Big(\varepsilon_t+(\sum_{j=1}^{k-1}\frac{\lambda_j}{\lambda_j}\mu_j)\Big)'\Sigma_{jt}^{-1}\Big(\varepsilon_t+(\sum_{j=1}^{k-1}\frac{\lambda_j}{\lambda_j}\mu_j)\Big) \\
&= \sum_{j=1}^{k-1}\sum_{t\in\{S_t=j\}}\Big[C_j+\mu_j'\Big(\sum_{t\in\{S_t=j\}}\Sigma_{jt}^{-1}\Big)\mu_j - 2\mu_j'\Big(\sum_{t\in\{S_t=j\}}\Sigma_{jt}^{-1}\varepsilon_t\Big)\Big] \\
&+ \sum_{t\in\{S_t=k\}}\Big[C_k+\Big(\sum_{j=1}^{k-1}\frac{\lambda_j}{\lambda_k}\mu_j\Big)'\Big(\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\Big)\Big(\sum_{j=1}^{k-1}\frac{\lambda_j}{\lambda_k}\mu_j\Big)+2\Big(\sum_{j=1}^{k-1}\frac{\lambda_j}{\lambda_k}\mu_j\Big)'\Big(\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\varepsilon_t\Big)\Big] \\
&= \sum_{j=1}^{k-1}\Big[\mu_j'\Big(\sum_{t\in\{S_t=j\}}\Sigma_{jt}^{-1}+\frac{\lambda_j^2}{\lambda_k^2}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\Big)\mu_j\Big] \;+\; \sum_{j=1}^{k-1}\sum_{i\neq j}\mu_j'\Big[\Big(\frac{\lambda_j\lambda_i}{\lambda_k^2}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\Big)\mu_i\Big] \\
&-2\sum_{j=1}^{k-1}\Big[\mu_j'\Big(\sum_{t\in\{S_t=j\}}\Sigma_{jt}^{-1}\varepsilon_t-\frac{\lambda_j}{\lambda_k}\sum_{t\in\{S_t=k\}}\Sigma_{kt}^{-1}\varepsilon_t\Big)\Big]+\sum_{j=1}^{k}C_j \\
&= \tilde{\mu}'A\tilde{\mu}-2\tilde{\mu}'b+\sum_{j=1}^{k}C_j = (\tilde{\mu}-A^{-1}b)'A(\tilde{\mu}-A^{-1}b)+\sum_{j=1}^{k}C_j - b'A^{-1}b
\end{aligned}$$

where C and the C_j's are constants that do not depend on $\tilde{\mu}$, while A and b are defined in (28) and (29). Therefore, by taking the exponential of minus one half of the the last expression, and neglecting the two irrelevant constant terms, we get

$$\exp -\frac{1}{2}(\tilde{\mu}-A^{-1}b)'A(\tilde{\mu}-A^{-1}b), \tag{34}$$

which is the kernel of a $N_p(A^{-1}b, A^{-1})$ density for $\tilde{\mu}$.

References

Alexander, C., and E. Lazar (2004): "Normal Mixture GARCH(1,1)," forthcoming in *Journal of Applied Econometrics*.

Bauwens, L., S. Laurent, and J. Rombouts (2006): "Multivariate GARCH Models: A Survey," *Journal of Applied Econometrics*, 21, forthcoming.

Bauwens, L., M. Lubrano, and J. Richard (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, Oxford.

Dempster, A., N. Laird, and D. Rubin (1977): "Maximimum Likelihood From Incomplete Data via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

Gray, S. (1996): "Modeling the conditional distribution of interest rates as a regime-switching process," *Journal of Financial Economics*, 42, 27–62.

Haas, M., S. Mittnik, and M. Paolella (2004a): "Mixed Normal Conditional Heteroskedasticity," *Journal of Financial Econometrics*, 2, 211–250.

——— (2004b): "A New Approach to Markov-Switching GARCH Models," *Journal of Financial Econometrics*, 2, 493–530.

Hafner, C. (2003): "Fourth Moment Structure of Multivariate GARCH processes," *Journal of Financial Econometrics*, 1, 26–54.

Lawrence, C., and A. Tits (2001): "A Computationally Efficient Feasible Sequential Quadratic Programming Algorithm," *SIAM Journal of Optimization*, 11, 1092–1118.

McLachlan, G., and D. Peel (2000): *Finite Mixture Models*. Wiley Interscience, New York.

Pelletier, D. (2005): "Regime Switching for Dynamic Correlations," forthcoming in the Journal of Econometrics.

Tanner, M., and W. Wong (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Wilks, S. (1962): *Mathematical Statistics*. Wiley, New York.

Wong, C., and W. Li (2000): "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society, Series B*, 62, 95–115.

——— (2001): "On a Mixture Autoregressive Conditional Heteroscedastic Model," *Journal of the American Statistical Association*, 96, 982–995s.