CORE DISCUSSION PAPER

2005/85

BAYESIAN INFERENCE FOR THE MIXED CONDITIONAL HETEROSKEDASTICITY MODEL

L. Bauwens[1] and J.V.K. Rombouts[1]

December 1, 2005

Abstract

We estimate by Bayesian inference the mixed conditional heteroskedasticity model of (Haas, Mittnik, and Paolella 2004a). We construct a Gibbs sampler algorithm to compute posterior and predictive densities. The number of mixture components is selected by the marginal likelihood criterion. We apply the model to the SP500 daily returns.

Keywords: Finite mixture, ML estimation, Bayesian inference, Value at Risk.

JEL Classification: C11, C15, C32

[1]CORE and Department of Economics, Université Catholique de Louvain.

[2]HEC Montréal, CIRANO, CIRPEE and CREF.

We thank Viorel Maxim for excellent research assistance and Arie Preminger for his comments. 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. Rombouts's work was supported by a HEC Montréal Fonds de démarrage and by the Centre for Research on e-Finance.

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. The scientific responsibility is assumed by the authors.

1 Introduction

Finite mixture models, see e.g. (McLachlan and Peel 2000), are more and more used in statistics and econometrics. Their main advantage lies in the flexibility they provide in model specification, compared to the use of a more simple distribution. On the other hand, these models are more difficult to estimate than corresponding models without a mixture, but their estimation becomes more and more feasible as computational power increases. However, computational power is not sufficient, one needs also good algorithms. Maximum likelihood estimation of mixture models is not at all as easy as for non-mixture models, and not very reliable in some cases. The EM algorithm was initially developed in this perspective, see (Dempster, Laird, and Rubin 1977). Bayesian estimation is also very efficient for mixture models, see (Marin, Mengersen, and Robert 2005).

Conditionally heteroskedastic models are very widespread for modelling time-series of financial returns. The most used class of model is the GARCH family, see e.g. (Bollerslev, Engle, and Nelson 1994) for a survey. A lot of research has has been devoted to refine the dynamic specification of the conditional variance equation, for which the benchmark is the linear GARCH specification of (Bollerslev 1986). The conditional distribution of the model error term is chosen by most researchers among the Gaussian, Student-t, and too a smaller extent skewed versions of these and the GED distribution, see (Nelson 1991). Empirical models typically include around five parameters to fit time-series of a few thousands observations. This may be considered as a powerful way to represent the data. Simultaneously such parsimonious models may be too restrictive: one should be able to fit the data better by using a more flexible model, like a mixture model. Mixture GARCH models have been recently developed, see (Haas, Mittnik, and Paolella 2004a), who build on the results of (Wong and Li 2000) and (Wong and Li 2001), and (Haas, Mittnik, and Paolella 2004b) and (Alexander and Lazar 2004). All these authors use ML estimation, while (Bauwens, Bos, van Oest, and van Dijk 2004) propose a particular two-component mixture GARCH model and estimate it by Bayesian inference.

Bayesian inference for the mixed normal GARCH model of (Haas, Mittnik, and Paolella 2004a) is the subject of this paper. The model is defined in Section 2. In Section 3 we explain how this model can be estimated in the Bayesian framework. We design a Gibbs sampler, and discuss how to obtain predictive densities and how to choose the number of components

of the mixture. In Section 4, we illustrate all this on simulated data, and in Section 5, we apply the approach to returns of the SP500 index.

2 Mixed conditional heteroskedasticity

(Haas, Mittnik, and Paolella 2004a) define a mixture model on a demeaned series $y_t = Y_t - E(Y_t|\mathcal{F}_t)$ where $\mathcal{F}_t$ is the information set up to time t and the conditional mean does not depend on the components of the mixture. They call this model (diagonal) MN-GARCH (MN for mixed normal). The conditional cdf of y_t is the K-component mixture

$$F(y_t|\mathcal{F}_t) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{y_t - \mu_k}{\sqrt{h_{k,t}}}\right) \tag{1}$$

where

$$h_{k,t} = \omega_k + \alpha_k y_{t-1}^2 + \beta_k h_{k,t-1} \tag{2}$$

and $\Phi(\cdot)$ is the standard Gaussian cdf. Note that the parameter π_k is positive for all k and $\sum_{k=1}^{K} \pi_k = 1$, which is imposed by setting $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. The other Greek letters denote the other parameters. The zero mean assumption on y_t is ensured by the restriction

$$\mu_K = -\sum_{k=1}^{K-1} \frac{\pi_k \mu_k}{\pi_K}. \tag{3}$$

(Haas, Mittnik, and Paolella 2004a) also consider a more general model where the $h_{k,t}$'s are GARCH(p_k, q_k) and more importantly may depend on other $h_{j,t}$'s, $k \neq j$ (contrary to the diagonal specification defined above). The weak stationarity condition for a (diagonal) MN-GARCH model is

$$\left[\sum_{k=1}^{K} \frac{\pi_k}{\widetilde{\beta}_k}\Big(1 - \alpha_k - \beta_k\Big)\right] \prod_{k=1}^{K} \widetilde{\beta}_k > 0. \tag{4}$$

where $\widetilde{\beta}_k = 1 - \beta_k$. Its unconditional variance is then given by

$$E(y_t^2) = \frac{c + \sum_{k=1}^{K} \pi_k \omega_k / \widetilde{\beta}_k}{\sum_{k=1}^{K} \pi_k (1 - \alpha_k - \beta_k)/\widetilde{\beta}_k} \tag{5}$$

where $c = \sum_{k=1}^{K} \pi_k \mu_k^2$. One can check that the process may be stationary even if some components are not stationary provided that these components have sufficiently low corresponding component weights.

3 Bayesian inference

The likelihood of the MN-GARCH model for T observations is given by

$$\mathcal{L}(\Psi \mid y) = \prod_{t=1}^{T} \sum_{k=1}^{K} \pi_k \phi(y_t|\mu_k, \theta_k) \tag{6}$$

where Ψ is the vector regrouping the parameters π_k, μ_k, θ_k for $k = 1, \ldots, K$, $y = (y_1, y_2, \ldots, y_T)$, $\phi(\cdot|\mu_k, \theta_k)$ denotes a normal density with mean μ_k and variance $h_{k,t}$ that depends on $\theta_k = (\omega_k, \alpha_k, \beta_k)$. A direct evaluation of the likelihood function is difficult because it consists of a product of sums. To alleviate this evaluation, we introduce for each observation a state variable $S_t \in \{1, 2, \ldots, K\}$ that takes the value k if the observation y_t belongs to component k. The vector S^T contains the state variables for the T observations. We assume that the state variables are independent given the group probabilities, and the probability that S_t is equal to k is equal to π_k:

$$\varphi(S^T|\pi) = \prod_{t=1}^{T} \varphi(S_t|\pi) = \prod_{t=1}^{T} \pi_{S_t}, \tag{7}$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$. Given S^T and y the likelihood function is

$$\mathcal{L}(\Psi \mid S^T, y) = \prod_{t=1}^{T} \pi_{S_t} \phi(y_t|\mu_{S_t}, \theta_{S_t}), \tag{8}$$

which is easier to evaluate than (6). Since S^T is not observed we treat it 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 becomes easier 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 by sampling from its conditional posterior densities, which are called blocks. The blocks of the Gibbs sampler, and the prior densities, are explained in the next subsections, using the parameter vectors π, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, and $\mu = (\mu_1, \mu_2, \ldots, \mu_K)$. The joint posterior distribution is given by

$$\varphi(S^T, \mu, \theta, \pi|y) \propto \varphi(\mu)\, \varphi(\theta)\, \varphi(\pi) \prod_{t=1}^{T} \pi_{S_t} \phi(y_t|\mu_{S_t}, \theta_{S_t}), \tag{9}$$

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

3.1 Sampling S^T from $\varphi(S^T|\mu,\theta,\pi,y)$

Given μ,θ,π and y, the posterior density of S^T is proportional to $\mathcal{L}(\Psi \mid S^T,y)$. It turns out that the S_t's are mutually independent, so that we can write the relevant conditional posterior density as

$$\varphi(S^T|\mu,\theta,\pi,y) = \varphi(S_1|\mu,\theta,\pi,y)\cdots\varphi(S_T|\mu,\theta,\pi,y). \tag{10}$$

As the sequence $\{S_t\}_{t=1}^T$ is equivalent to a multinomial process, we simply have to sample from a discrete distribution where the K probabilities are given by

$$P(S_t = k|\theta,\mu,\pi,y) = \frac{\pi_k\phi(y_t|\mu_k,\theta_k)}{\sum_{k=1}^K \pi_k\phi(y_t|\mu_k,\theta_k)},\ (k=1,\ldots,K). \tag{11}$$

To sample S_t we draw one observation from a uniform distribution on $(0,1)$ and decide which group k to take according to (11).

3.2 Sampling π from $\varphi(\pi|S^T,\mu,\theta,y)$

The full conditional posterior density of π is given by

$$\varphi(\pi|S^T,y) = \varphi(\pi|S^T) \propto \varphi(\pi)\prod_{k=1}^K \pi_k^{x_k} \tag{12}$$

where x_k is the number of times that $S_t = k$. The prior $\varphi(\pi)$ 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, $\varphi(\pi|S^T,y)$ is also a Dirichlet distribution, $Di(a_1,a_2\cdots a_K)$ with $a_k = a_{k0} + x_k$, $k = 1,2,\ldots,K$. Notice that it does nor depend on μ and θ. The Dirichlet density function is given by

$$f_{Di}(\pi|\ a_1,a_2\cdots a_K) = \frac{\Gamma(A)}{\prod_{k=1}^K\Gamma(a_k)}\ \prod_{k=1}^K \pi_k^{a_{k-1}}\ \mathbb{1}_{\mathcal{S}_K}(\pi) \tag{13}$$

where $a_k > 0$ $(k=1,\ldots,K)$, $A = \sum_{i=1}^K a_i$ and $\mathcal{S}_K = \{\pi_k, k=1,\ldots,K|\pi_k > 0\,\forall k, \sum_{k=1}^K \pi_k = 1\}$. The first two moments are given by $E(\pi_i|a) = \frac{a_i}{A}$, $V(\pi_i|a) = \frac{a_i\,(A-a_i)}{A^2(A+1)}$ and $cov(\pi_i,\pi_j|a) = -\frac{a_i a_j}{A^2(A+1)}$ respectively.

We sample a Dirichlet distribution by sampling K independent gamma random variables, $X_k \sim G(a_k,1)$, and transforming them to

$$\begin{aligned}\pi_i &= \frac{X_i}{X_1+\ldots+X_K} \quad i=1,\ldots,K-1\\ \pi_K &= 1-\pi_1-\pi_2-\ldots-\pi_{K-1}.\end{aligned}$$

It follows that $(\pi_1, \ldots, \pi_K) \sim Di(a_1, \ldots, a_K)$. Other properties of the Dirichlet distribution can be found in (Wilks 1962).

3.3 Sampling μ from $\varphi(\mu|S^T, \pi, \theta, y)$

Since the mean of the mixture is equal to zero, see (3), the μ_k's cannot be drawn independently. We illustrate this for $K = 3$. Minus two times the log-kernels for the first two components are given by

$$\sum_{t \in S_t = k} \left(\frac{y_t - \mu_k}{\sqrt{h_{k,t}}} \right)^2 = c_k + \mu_k^2 \sum_{t \in S_t = k} \frac{1}{h_{k,t}} - 2\mu_k \sum_{t \in S_t = k} \frac{y_t}{h_{k,t}} \quad (k = 1, 2), \tag{14}$$

where c_k is a constant that does not depend on μ_k. The third mixture component contributes in the following way:

$$\begin{aligned} \sum_{t \in S_{k=3}} \left(\frac{y_t + \frac{\pi_1}{\pi_3}\mu_1 + \frac{\pi_2}{\pi_3}\mu_2}{\sqrt{h_{3t}}} \right)^2 = c_3 + \mu_1^2 \left(\frac{\pi_1}{\pi_3} \right)^2 \sum_{t \in S_t = 3} \frac{1}{h_{3t}} + \\ \mu_2^2 \left(\frac{\pi_2}{\pi_3} \right)^2 \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} + 2\mu_1 \frac{\pi_1}{\pi_3} \sum_{t \in S_t = 3} \frac{y_t}{h_{3,t}} + \\ 2\mu_2 \frac{\pi_2}{\pi_3} \sum_{t \in S_t = 3} \frac{y_t}{h_{3,t}} + 2 \frac{\pi_1 \pi_2 \mu_1 \, u_2}{\pi_3^2} \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} \end{aligned} \tag{15}$$

The sum of (14) and (15) can be written compactly as

$$(\mu - \bar{\mu})' A (\mu - \bar{\mu}) + c, \tag{16}$$

where c is a constant not depending on μ, by defining the matrix A as

$$\begin{bmatrix} \sum_{t \in S_t = 3} \frac{1}{h_{1,t}} + \left(\frac{\pi_1}{\pi_3} \right)^2 \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} & \frac{\pi_1 \pi_2}{\pi_3^2} \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} \\ \frac{\pi_1 \pi_2}{\pi_3^2} \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} & \sum_{t \in S_t = 3} \frac{1}{h_{2,t}} + \left(\frac{\pi_2}{\pi_3} \right)^2 \sum_{t \in S_t = 3} \frac{1}{h_{3,t}} \end{bmatrix}, \tag{17}$$

and the vector $\bar{\mu}$ as $-A^{-1}b$, where

$$b = \begin{bmatrix} \frac{\pi_1}{\pi_3} \sum_{t \in S_{k=3}} \frac{y_t}{h_{3t}} - \sum_{t \in S_{k=1}} \frac{y_t}{h_{1t}} \\ \\ \frac{\pi_2}{\pi_3} \sum_{t \in S_{k=3}} \frac{y_t}{h_{3t}} - \sum_{t \in S_{k=2}} \frac{y_t}{h_{2t}} \end{bmatrix}. \tag{18}$$

Minus one half times the first term of (16) is the log-kernel of a bivariate normal density with mean $\bar{\mu}$ and covariance matrix A^{-1}.

In general, for K components, in this block of the Gibbs sampler, the $K-1$ first parameters μ_k are drawn from a multivariate normal density with mean $-A^{-1}b$ and covariance matrix A^{-1}, where

$$A = \text{diag}\left(\sum_{t\in S_t=1}\frac{1}{h_{1,t}},\ldots,\sum_{t\in S_t=K-1}\frac{1}{h_{K-1,t}}\right)+\frac{\tilde{\pi}\tilde{\pi}'}{\pi_K^2}\sum_{t\in S_t=K}\frac{1}{h_{K,t}}, \tag{19}$$

denoting $\tilde{\pi}=(\pi_1,\ldots,\pi_{K-1})$, and

$$b=\begin{bmatrix}\frac{\pi_1}{\pi_K}\sum_{t\in S_t=K}\frac{y_t}{h_{K,t}}-\sum_{t\in S_t=1}\frac{y_t}{h_{1,t}}\\ \vdots \\ \frac{\pi_{K-1}}{\pi_K}\sum_{t\in S_t=K}\frac{y_t}{h_{K,t}}-\sum_{t\in S_t=K-1}\frac{y_t}{h_{K-1,t}}\end{bmatrix}. \tag{20}$$

3.4 Sampling θ from $\varphi(\theta|S^k,\mu,\pi,y)$

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

$$\varphi(\theta|S^T,\pi,y)=\varphi(\theta|S^T,y)=\varphi(\theta_1|\widetilde{y}^1)\varphi(\theta_2|\widetilde{y}^2)\cdots\varphi(\theta_K|\widetilde{y}^K) \tag{21}$$

where $\widetilde{y}^k=\{y_t|S_t=k\}$ and

$$\varphi(\theta_k|\widetilde{y}^k)\propto\varphi(\theta_k)\prod_{t\in S_t=k}\phi(y_t|\mu_k,\theta_k). \tag{22}$$

Since we condition on the state variables, we can simulate each block θ_k separately. We do this with the griddy-Gibbs sampler. The algorithm works as follows at iteration $n+1$ (for lighter notations, we drop the index k):

1. Using (22), compute $\kappa(\omega|\alpha^n,\beta^n,\tilde{y})$, the kernel of the conditional posterior density of ω given the values of α and β sampled at iteration n, over a grid $(\omega_1,\omega_2\cdots,\omega_G)$, to obtain the vector $G_\kappa=(\kappa_1,\kappa_2,\cdots,\kappa_G)$.

2. By a deterministic integration rule using M points, compute $G_f=(0,f_2,\ldots,f_G)$ where

$$f_i=\int_{\omega_1}^{\omega_i}\kappa(\omega|\alpha^n,\beta^n,\tilde{y})\,d\omega,\quad i=2,\ldots,G. \tag{23}$$

3. Generate $u\sim U(0,f_G)$ and invert $f(\omega|\alpha^{(n)},\beta^{(n)},\tilde{y})$ by numerical interpolation to get a draw $\omega^{(n+1)}\sim\varphi(\omega|\alpha^{(n)},\beta^{(n)},\tilde{y})$.

4. Repeat steps 1-3 for $\varphi(\alpha|\omega^{(n+1)},\beta^n,\tilde{y})$ and $\varphi(\beta|\omega^{(n+1)},\alpha^{(n+1)},\tilde{y})$.

Note that intervals of values for ω, α and β must be defined. The choice of these bounds (such as ω_1 and ω_G) needs sometimes to be fine tuned in order to cover the range of the parameter over which the posterior is relevant. For the deterministic integration we used thirty-three points, which proved to be enough according to several experiments. For further details and remarks on the griddy-Gibbs sampler we refer to (Bauwens, Lubrano, and Richard 1999).

3.5 Predictive densities

Predictive densities are essential for financial applications such as portfolio optimization and risk management. Unlike prediction in the classical framework, predictive densities take into account parameter uncertainty by construction. The predictive density of y_{T+1} is given by

$$f(y_{T+1} \mid y) = \int f(y_{T+1} \mid \Psi)\, \varphi(\Psi \mid y)\, \mathrm{d}\Psi \tag{24}$$

where $f(y_{T+1} \mid \Psi) = \sum_{k=1}^{K} \pi_k \phi(y_t|\mu_k, \theta_k)$ as implied by (1). An analytical solution to (24) is not available but it can be approximated by

$$\frac{1}{N} \sum_{j=1}^{N} \left(\sum_{k=1}^{K} \pi_k^{(j)} \phi \left(y_{T+1} | \mu_k^{(j)}, \theta_k^{(j)} \right) \right) \tag{25}$$

where the superscript (j) indexes the draws generated with the Gibbs sampler and N is the number of draws. Therefore, simultaneously with the Gibbs sampler, we repeat N times the following two-step algorithm

step 1: simulate $\Psi^{(j)} \sim \varphi(\Psi \mid y)$. This is done by the Gibbs sampler.

step 2: simulate $y_{T+1}^{(j)} \sim f(y_{T+1} \mid \Psi^{(j)})$. Go to step 1.

Extending the idea used for y_{T+1}, the predictive density for y_{T+s} may be written as

$$\begin{aligned} f(y_{T+s} \mid y) \;=\; & \int \Big[\int \int \cdots \int f(y_{T+s} \mid y_{T+s-1}, \ldots, y_{T+1}, y, \Psi) \times \\ & f(y_{T+s-1} \mid y_{T+s-2}, \ldots, y_{T+1}, y, \Psi) \times \\ & \ldots \times \\ & f(y_{T+1} \mid y, \Psi) \mathrm{d}y_{T+s-1} \mathrm{d}y_{T+s-2} \mathrm{d}y_{T+1} \Big]\, \varphi(\Psi \mid y)\, \mathrm{d}\Psi \end{aligned} \tag{26}$$

for which draws can be obtained by extending the above algorithm to a (s+1)-step algorithm. The draw of y_{T+1} serves as conditioning information to draw y_{T+2}, both realisations serve

to draw y_{T+3}, etc. All these draws are easily generated from the finite mixture of normal densities. A non-Bayesian procedure typically proceeds by conditioning on a point estimate of Ψ, which ignores the estimation uncertainty.

3.6 Marginal likelihood

The marginal likelihood of y, also called predictive density, is useful for selecting the number of components K in the mixture. For example, Bayes factors are ratios of marginal likelihoods, see (Kass and Raftery 1995) for a detailed explanation. The marginal likelihood is defined as the integral of the likelihood with respect to the prior density

$$m(y) = \int \mathcal{L}(\Psi \mid y)\varphi(\Psi)d\Psi. \tag{27}$$

Since this is the normalizing constant in Bayes' theorem we can also write

$$m(y) = \frac{\mathcal{L}(\Psi \mid y)\varphi(\Psi)}{\varphi(\Psi|y)}. \tag{28}$$

Notice that (28) is an identity that holds for every Ψ. Deterministic numerical integration of (27) is computationally too demanding for the finite mixture model of this paper. Instead, we calculate the marginal likelihood by the Laplace approximation, see (Tierney and Kadane 1986). To explain this, let us define $\exp(h(\Psi)) = \mathcal{L}(\Psi \mid y)\varphi(\Psi)$. The Laplace approximation is based on a second order Taylor expansion of $h(\Psi)$ around the posterior mode $\hat{\Psi} = \arg\max \ln \mathcal{L}(\Psi \mid y)$, so that the first order term in the expansion vanishes:

$$h(\Psi) \approx h(\hat{\Psi}) + \frac{1}{2}(\Psi - \hat{\Psi})' \left.\frac{\partial^2 h(\Psi)}{\partial\Psi\ \partial\Psi'}\right|_{\Psi=\hat{\Psi}} (\Psi - \hat{\Psi}). \tag{29}$$

Therefore the marginal likelihood can be computed as

$$\int \exp h(\Psi)\mathrm{d}\Psi \approx \exp(h(\hat{\Psi})) \int \exp\left(\frac{1}{2}(\Psi - \hat{\Psi})' \left.\frac{\partial^2 h(\Psi)}{\partial\Psi\partial\Psi'}\right|_{\Psi=\hat{\Psi}} (\Psi - \hat{\Psi})\right) \mathrm{d}\Psi \tag{30}$$

or

$$m(y) = \mathcal{L}(\hat{\Psi} \mid y)\ \varphi(\hat{\Psi})\ (2\pi)^{k/2} \mid \Sigma(\hat{\Psi})\mid^{1/2}, \tag{31}$$

where k is the dimension of Ψ and

$$\Sigma(\hat{\Psi}) = \left[-\left.\frac{\partial^2 \ln \mathcal{L}(\Psi \mid y)\ \varphi(\Psi)}{\partial\Psi\partial\Psi'}\right|_{\Psi=\hat{\Psi}}\right]^{-1}. \tag{32}$$

We choose the model with the highest marginal likelihood value.

Another possibility to choose the number of components is to treat K as an additional parameter in the model as is done in (Richardson and Green 1997) who make use of the reversible jump MCMC methods. In this way, the prior information on the number of components can be taken explicitly into account by specifying for example a Poisson distribution on K in such a way that it favours a small number of components.

4 Illustration on simulated data

The purpose of this section is to validate, using simulated data, the Gibbs sampler described in the preceding section and to compare Bayesian results with maximum likelihood estimates. We simulate one dataset from the following two component model:

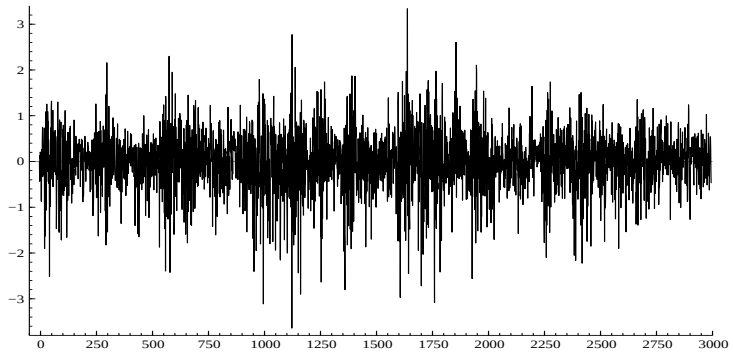
$$F(y_t|\mathcal{F}_t) = 0.8\Phi\left(\frac{y_t - 0.08}{\sqrt{h_{1,t}}}\right) + 0.2\Phi\left(\frac{y_t - 0.32}{\sqrt{h_{2,t}}}\right) \tag{33}$$

where

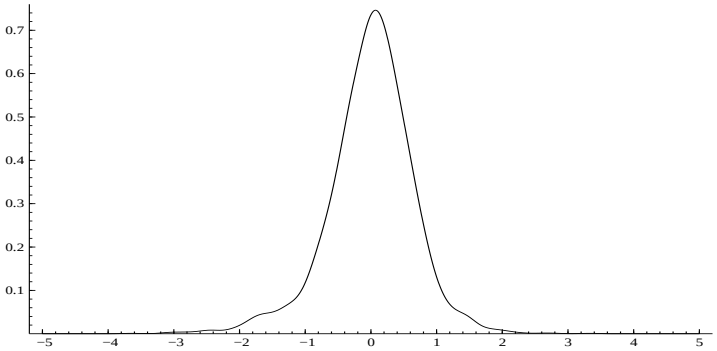
$$\begin{aligned} h_{1,t} &= 0.003 + 0.03y_{t-1}^2 + 0.94h_{1,t-1} \\ h_{2,t} &= 0.03 + 0.25y_{t-1}^2 + 0.85h_{2,t-1}. \end{aligned} \tag{34}$$

The sample size is fixed at 3000 and the conditional mean to zero. Although the second GARCH component is explosive, the model is weakly stationary because the expression given in equation (4) is equal to 0.0024. The parameters are chosen to be close to the estimates obtained for the same model using a comparable amount of real data in the empirical illustration described in Section 5. Table 1 provides descriptive statistics for the simulated data. The parameter values for this process clearly generate unconditional negative skewness and excess kurtosis, in addition to high persistence in the conditional variance process. This is also visible in Figure 1, which shows the sample path, the estimated kernel density, and the correlogram of the squared data.

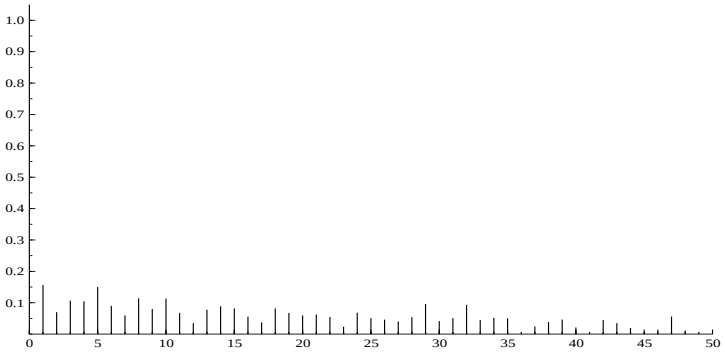
In Table 2, we report the parameter estimates for the two component model by maximum likelihood (ML) and by Bayesian inference, using the simulated data. The ML estimator is obtained by maximizing the natural logarithm of (6) taking into account the restrictions on the component probabilities. The standard errors are obtained from the Hessian matrix evaluated at the ML estimates. The Bayesian results are the posterior means and standard



(a) Simulated data



(b) Kernel density of the data



(c) Correlogram of squared data

Figure 1: Simulated data for the Gaussian two component mixture GARCH(1,1) model defined in (33)-(34).

Table 1: Descriptive statistics - simulated data

Observations	3000
Mean	-0.0048
Standard Deviation	0.65
Maximum	3.34
Minimum	-3.64
Skewness coefficient	-0.53
Kurtosis coefficient	5.39

Statistics for the simulated data of the two component model in (33)-(34).

deviations computed using 6500 draws of which the first 500 ones are discarded to warm up the Gibbs sampler. The parameters a_{0k} of the Dirichlet prior for π are all equal to 1, implying that the prior density of π_1 is uniform on $(0, 1)$. The prior densities for the other parameters are all independent and uniform on finite ranges, chosen to be wide enough not to truncate the posterior density but narrow enough not to waste computational time.

We see from Table 2 that the parameters estimates for both estimation methods are close to each other and of the same order of magnitude as the true values. Generally speaking, we also notice that the bias and the variance of the Bayes estimates are somewhat smaller, although some care has to be taken since the table contains only results for one simulated data set. We did a more detailed analysis of these estimators by running a Monte Carlo study, the results of which are reported in (Bauwens and Rombouts 2006), and it turns out that the smaller bias and variance for the Bayes estimator indeed are confirmed.

In Table 3, we report the marginal likelihood values, see Section 3.6, for the one, two and three component model. As expected, the marginal likelihood is maximized for the two component model since this is the true data generating process. To compare with the marginal likelihood, we also compute the Bayesian information criterion (BIC), defined as $-2\mathcal{L}(\hat{\Psi} \mid y) + k \log(T)$, using the maximum likelihood estimator $\hat{\Psi}$. Again, the two component model is preferred because it minimizes the BIC.

This illustration on simulated data shows that the Gibbs sampler for the mixture GARCH

Table 2: Estimation results - simulated data

	DGP	MLE		Bayes	
		estimate	std error	mean	std dev
π_1	0.8	0.79824	0.040065	0.76908	0.046114
μ_1	0.08	0.088729	0.011766	0.084313	0.0087951
ω_1	0.003	0.0042042	0.0017198	0.0043303	0.0015511
α_1	0.03	0.054952	0.0092736	0.054893	0.0081617
β_1	0.94	0.89565	0.016395	0.89236	0.015149
ω_2	0.03	0.019888	0.010659	0.024543	0.0094535
α_2	0.25	0.18725	0.056536	0.19938	0.0494
β_2	0.85	0.89226	0.027749	0.88094	0.025564

Results for two component mixture GARCH(1,1) model in (33)-(34).

Table 3: Model choice criteria - simulated data

K	Marginal log-lik.	Maximized log-lik.	# par.	BIC
1	-2772.59	-2761.2	3	5546.4
2	-2675.68	-2653.2	8	5370.5
3	-2680.74	-2651.0	13	5406.1

K is the number of components of the Gaussian mixture GARCH(1,1) model.

model performs well. In the next section we apply the model to a real dataset.

5 Application to S&P500 data

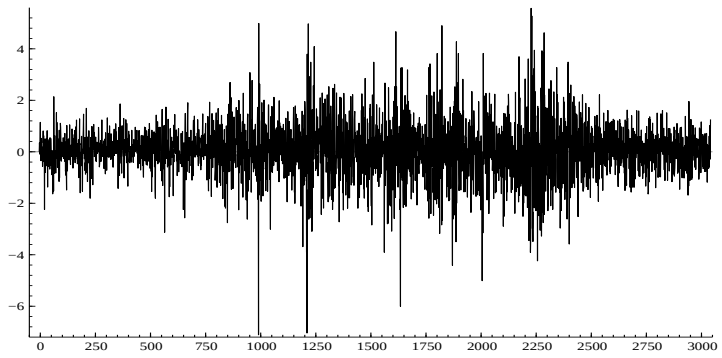
We fit the two component mixture model to daily S&P500 percentage return data from 01/03/1994 to 09/06/2005 (3047 observations). Descriptive statistics are given in Table 4. Figure 2 displays the sample path, estimated kernel density for the data and the correlogram for the squared data. It is clear from this that excess kurtosis and volatility clustering are present in the data. We analyzed whether a dynamic specification for the conditional mean is necessary and we found evidence for an autoregressive model of order three. The data are filtered for these effects in the rest of the empirical application.

Table 4: Descriptive statistics - S&P 500 returns

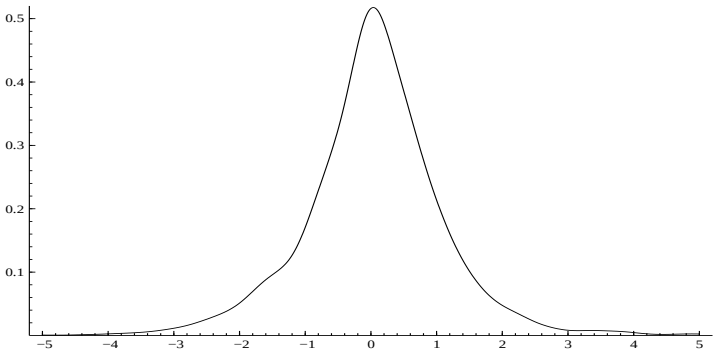
Observations	3047
Mean	0.0389
Standard Deviation	1.07
Maximum	5.58
Minimum	-7.11
Skewness	-0.11
Kurtosis	6.74

Statistics for S&P500 percentage daily returns from 01/03/1994 to 09/06/2005.

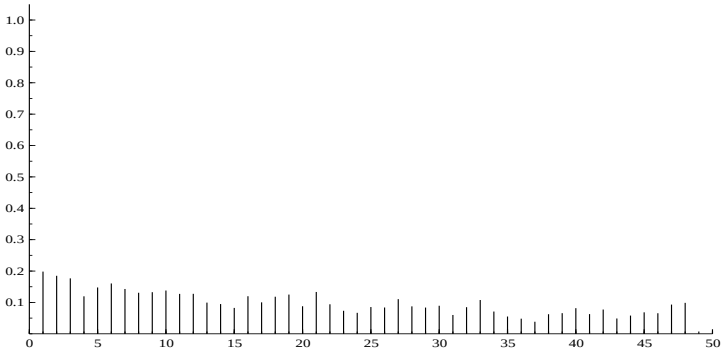
The ML estimates and the Bayes' first two marginal posterior moments are given in Table 5. The parameters a_{k0} of the Dirichlet prior for π are all equal to 1 like in the simulation example. The prior densities for the other parameters are all independent and uniform on finite ranges given by $0.0001 < \omega_1 < 0.0097$, $0.0005 < \alpha_1 < 0.08$, $0.89 < \beta_1 < 0.99$, $0.001 < \omega_2 < 0.13$, $0.0001 < \alpha_2 < 0.73$, $0.73 < \beta_2 < 0.99$. These values are the bounds used in the griddy-Gibbs sampler part of the algorithm described in Section 3.4. The posterior marginal distributions for all the parameters are given in Figure 3. The x-axes for the GARCH parameters are the prior intervals reported above. Note that the posterior marginals for ω_1



(a) S&P 500 returns



(b) Kernel density of the data



(c) Correlogram of squared data

Figure 2: S&P 500 graphs

and ω_2 are somewhat truncated at zero given that they are restricted to be positive.

From Table 5, we conclude that the parameter estimates are close to each other but that the posterior standard deviations (std dev.) are smaller than the ML standard errors (std error). The latter are computed from the Hessian matrix evaluated at the ML estimates. The estimated probability is about 0.8 for the first component which is driven by a persistent $\alpha_1 + \beta_1 = 0.98$ GARCH process. The second component of the mixture has a conditional variance process where $\alpha_2 + \beta_2 = 1.14$ with a probability of about 0.2.

Table 5: Estimation results - S&P 500

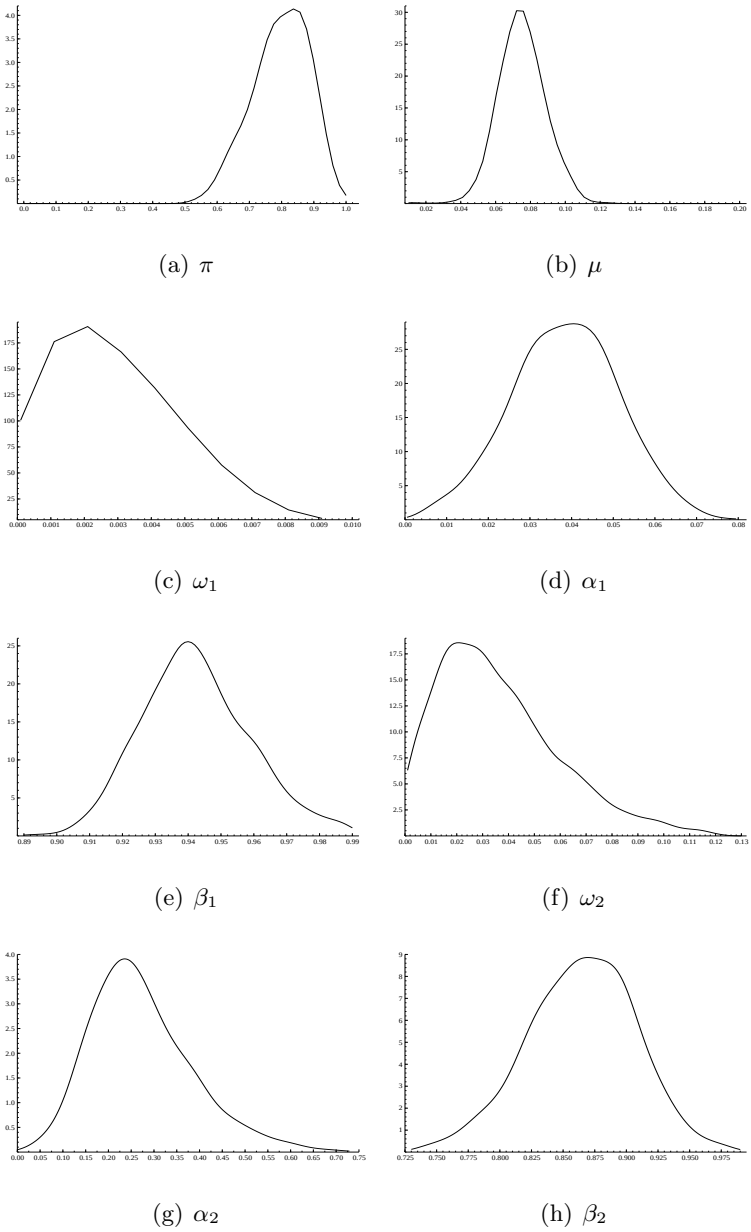
	MLE		Bayes	
	estimate	std error	mean	std dev.
π_1	0.83496	0.13179	0.79347	0.085364
μ_1	0.074463	0.023198	0.074918	0.013654
ω_1	0.0025423	0.0024439	0.0028809	0.0019193
α_1	0.036845	0.016711	0.038411	0.012836
β_1	0.94662	0.017437	0.94241	0.016584
ω_2	0.030760	0.029664	0.03589	0.023328
α_2	0.27255	0.14932	0.273	0.11191
β_2	0.87141	0.047557	0.86448	0.042872

Results for two component Gaussian mixture GARCH(1,1) model.

Figure 4 displays convergence plots for all the parameters. The convergence statistics for a parameter ρ are computed as follows:

$$CS_t = \frac{\left(\frac{1}{t}\sum_{n=1}^{t}\rho_n\right) - \mu_\rho}{\sigma_\rho}, \tag{35}$$

where μ_ρ and σ_ρ are the empirical mean and standard deviation, respectively, of the N draws $\rho_1, \rho_2, \ldots, \rho_N$. If the sampler converges, the graph of CS_t against t should converge smoothly



(a) π (b) μ

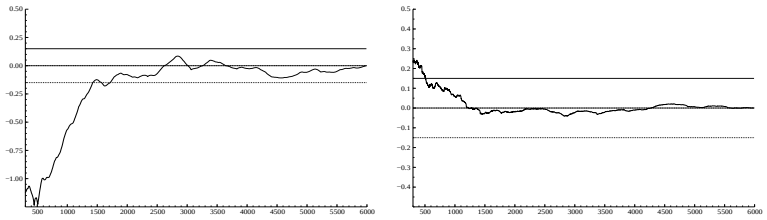
(c) ω_1 (d) α_1

(e) β_1 (f) ω_2

(g) α_2 (h) β_2

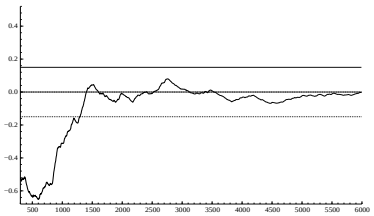
Figure 3: Posterior densities (kernel estimates from Gibbs output) for two component Gaussian mixture GARCH(1,1) model.

to zero. One can see from Figure 4 that convergence is indeed achieved for all the parameters.

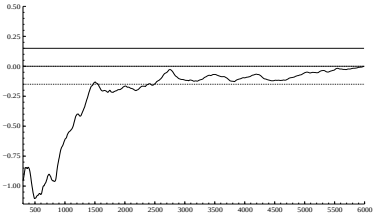


(a) π

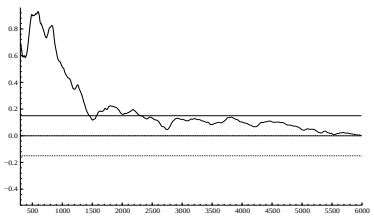
(b) μ



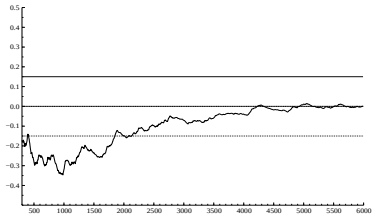
(c) ω_1



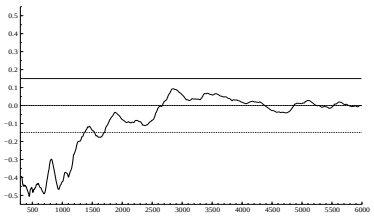
(d) α_1



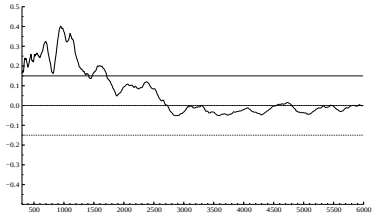
(e) β_1



(f) ω_2



(g) α_2



(h) β_2

Figure 4: Convergence plots of Gibbs estimates of posterior means

As a comparison, we also estimate the one-component mixture model, i.e. the conventional GARCH(1,1) model. The maximum likelihood estimates and the Bayes' first two marginal posterior moments are given in Table 6. The process looks like highly persistent, given that $\alpha_1 + \beta_1$ is estimated as 0.996. This may be interpreted as a compromise between the less

persistent and explosive components of the mixture model. We obtain a similar result when we estimate the GARCH(1,1) model with the data simulated from the two component mixture of Section 4. Thus the observation that a quasi-integrated GARCH model ($\hat{\alpha}_1 + \hat{\beta}_1 \approx 1$) is obtained in many empirical results can be explained by a lack of flexibility of this model.

Table 6: Estimation results (one component) - S&P 500

	MLE		Bayes	
	estimate	std dev.	mean	std error
ω_1	0.0054295	0.0019993	0.0057050	0.001763
α_1	0.062177	0.0082521	0.063294	0.0079359
β_1	0.93494	0.0085043	0.93373	0.008198

Resulst for Gaussian GARCH(1,1) model.

in Table 7, we report the marginal likelihood and the BIC values for the one and two component models. The results indicate a strong preference for the two component model.

Table 7: Model choice criteria - S&P500 data

K	Marginal log-lik.	Maximized log-lik.	# par.	BIC
1	-4139.13	-4127.1	3	8278.2
2	-4090.87	-4071.0	8	8206.1

K is the number of components of the Gaussian mixture GARCH(1,1) model.

As for any time series model, prediction is essential. As we explained in Section 3.5 Bayesian inference allows to obtain predictive densities that by construction incorporate parameter uncertainty. Furthermore, they can be easily computed together with the Gibbs sampler for the model parameters. We calculate predictive densities out of sample for a horizon up to five days, that for September 7, 2005 until September 11, 2005. Kernel density estimates for the predictive densities are given in Figure 5. The dotted line represents the two component model, the solid line represents the one component model. Eyeballing Figure

Table 8: Features of predictive densities

	h	One component	Two components
Mean	1	0.0035362	0.010062
	2	0.012670	-0.0033736
	3	-0.010694	-0.0012910
	4	0.0034478	0.0028309
	5	-0.0067062	0.017732
Std Dev.	1	0.57949	0.58562
	2	0.5845	0.59397
	3	0.57378	0.57801
	4	0.5902	0.57529
	5	0.58926	0.57141
VaR	1	-1.3428	-1.6054
	2	-1.3621	-1.6473
	3	-1.2906	-1.5556
	4	-1.3461	-1.5505
	5	-1.3648	-1.5812

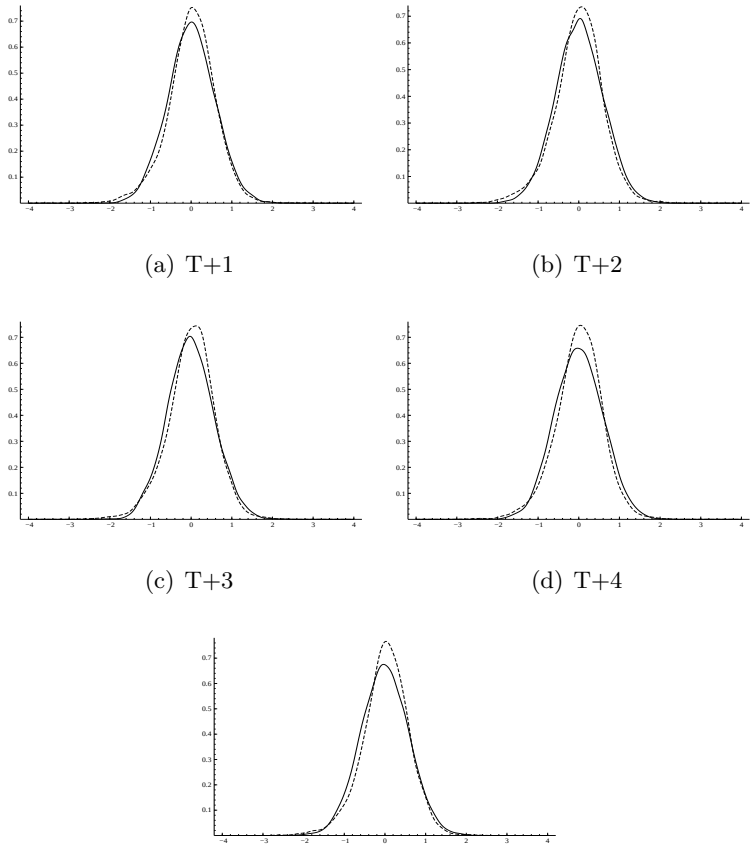
h is the post-sample prediction horizon. VaR is the 5 percent value-at-risk quantile.

5, we see that the left tail of the predictive densities are fatter for the two component model compared to the simple GARCH model.

In Table 8 we give the mean, standard deviation and value-at-risk at 5 percent (VaR) for the five days. Because of the fatter left tail in the two component model, the VaR is smaller than for the one component model.

6 Conclusion

We have shown how a certain type of mixture GARCH model can be estimated by Bayesian inference. ML estimation is typically not easy because of the complexity of the likelihood



(a) T+1 (b) T+2

(c) T+3 (d) T+4

(e) T+5

Figure 5: Kernel density estimates of predictive densities from September 7, 2005 to September 11, 2005. The dotted line represents the two component model, the solid line represents the one component model.

function. In Bayesian estimation, this is taken care of by enlarging the parameter space with state variables, so that a Gibbs sampling algorithm is easy to implement. Despite a higher computing time, the Bayesian solution is more reliable since estimation does not fail, while this may happen in MLE. Moreover, as we show in Section 3, the Gibbs algorithm can be extended to include the computation of predictive densities, which takes care of estimation uncertainty. Prediction in the ML approach is typically done by conditioning on the ML estimate and therefore ignores estimation uncertainty.

Bayesian estimation of other types of mixture GARCH models, including multivariate models, can probably be handled in a similar way as in this paper. Such extensions are on our research agenbda.

References

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

Bauwens, L., C. Bos, R. van Oest, and H. van Dijk (2004): "Adaptive radial-based direction sampling: a class of flexible and robust Monte Carlo integration methods," *Journal of Econometrics*, 123/2, 201–225.

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

Bauwens, L., and J. Rombouts (2006): "A comparison of estimators for the mixed conditional heteroskedasticity model," forthcoming CORE DP.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Bollerslev, T., R. Engle, and D. Nelson (1994): "ARCH Models," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden, chap. 4, pp. 2959–3038. North Holland Press, Amsterdam.

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

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.

Kass, R., and A. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Marin, J., K. Mengersen, and C. Robert (2005): *Bayesian Modelling and Inference on Mixtures of Distributions,* Handbook of Statistics 25. D. Dey and C.R. Rao (eds), Elsevier-Sciences.

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

Nelson, D. (1991): "Conditional Heteroskedasticity in Asset Returns: a New Approach," *Econometrica*, 59, 349–370.

Richardson, S., and P. Green (1997): "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society, Series B*, 59, 731–792.

Tanner, M., and W. Wong (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Tierney, L., and J. Kadane (1986): "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

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.