

BAYESIAN CLUSTERING OF MANY GARCH MODELS

*L. Bauwens*¹, *J.V.K. Rombouts*²

28 December 2004, revised May 30, 2006

Resubmitted to *Econometric Reviews* Special Issue “Bayesian Dynamic Econometrics”

Abstract

We consider the estimation of a large number of GARCH models, of the order of several hundreds. Our interest lies in the identification of common structures in the volatility dynamics of the univariate time series. To do so, we classify the series in an unknown number of clusters. Within a cluster, the series share the same model and the same parameters. Each cluster contains therefore similar series. We do not know a priori which series belongs to which cluster. The model is a finite mixture of distributions, where the component weights are unknown parameters and each component distribution has its own conditional mean and variance. Inference is done by the Bayesian approach, using data augmentation techniques. Simulations and an illustration using data on US stocks are provided.

Keywords: Bayesian inference, Clustering, GARCH, Gibbs sampling, Mixtures.

JEL Classification: C11, C32

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

²HEC Montréal and CIRANO.

The authors would like to thank Michel Mouchart for interesting discussions, two anonymous referees and the editor in charge (Gary Koop) for useful comments and suggestions. Work supported in part by the European Community’s Human Potential Programme under contract HPRN-CT-2002-00232, MICFINMA.

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

An important question still standing in modelling the volatility of asset returns is how to deal with a large number of series, for example all the stocks of the SP500 (let us say that in general we consider J of them). It is well known that financial return series are dynamically interrelated and that this has to be taken into account for example in the construction of optimal portfolios. Multivariate GARCH models (MGARCH) are potentially useful in this respect; see Bauwens, Laurent, and Rombouts (2006) for a survey. MGARCH models define the conditional variance matrix as a function of the past returns. However, the number of parameters they require rises fastly with J . Dynamic conditional correlation (DCC) models were proposed by Engle (2002) and by Tse and Tsui (2002). They require much less parameters than the conventional MGARCH models. An essential feature of the DCC models is that one specifies separately the conditional variances and the conditional correlations. This feature enables a two-step estimation procedure where one first estimates the parameters of the conditional variances, without taking account of the correlation parameters. In the second step, one estimates the parameters of the conditional correlations given the parameters estimated in the first step.

In this paper, we focus on the estimation of a large number of univariate GARCH models, because firstly this allows us to present the methodology we use in a simple case, and secondly because univariate GARCH models are specified as the first stage of DCC models. Our main interest lies in the identification of common structures in the volatility dynamics of the univariate time series. This can be done by postulating the existence of a finite number of groups, say G of them, such that the members or the data series of each group have the same parameter vector determining the conditional variance specification. The overall problem to be solved is twofold: the inference on the number of groups, and given this number the inference on the parameters of the different groups.

We adopt the modeling framework of Frühwirth-Schnatter and Kaufmann (2005) by treating the data as draws of a finite mixture of distributions,

$$\tilde{f}(y^j) = \sum_{g=1}^G \eta_g f(y^j | \theta_g), \quad (1)$$

where y^j is the j -th time series of returns, $\eta_1 + \dots + \eta_G = 1$, and θ_g are group specific parameters.

This implies a difficult likelihood to work with because it contains G^J terms:

$$L(\eta, \theta | y) \propto \prod_{j=1}^J \left(\sum_{g=1}^G \eta_g f(y^j | \theta_g) \right). \quad (2)$$

A mixture problem involves making inferences about the group probabilities and the component distributions given only a sample from the mixture. The closer the component distributions are to each other, the more difficult this is because of problems related to identifiability and computational instability. For more details on finite mixtures, see the contributions of Diebolt and Robert (1994) and Richardson and Green (1997). See also Chib and Hamilton (2000) for an application to treatment models and Frühwirth-Schnatter (2001) for an application to US quarterly real gross national product data.

Popular in applied finite mixture modeling is the use of a normal mixture $\sum_{g=1}^G \eta_g \mathcal{N}(\mu_g, \sigma_g^2)$. In this paper we focus on the differentiation between the component distributions via different conditional heteroskedasticity structures by the use of GARCH models. We illustrate that in this complicated dynamic structure the use of finite mixtures is very promising. For the sake of exposition we use a normal mixture but extensions, for example the use of the Student t -distribution, are not difficult to cope with.

One can think of finite mixtures in two ways, see for example Richardson and Green (1997). Firstly, we can postulate a heterogenous population of G components of sizes proportional to η_g ($g = 1, \dots, G$), from which the data is drawn. Secondly, we can consider (1) as a parsimonious representation of a non-standard density. Take again the example of the SP500. Even if we believe that the 500 stocks are all different (*i.e.* no two stocks are driven by the same volatility process), it may be convenient to consider that there are for example three groups of stocks, those with low persistence in the variance, those with high persistence, and stocks with in-between persistence. An additional issue of particular interest in this respect is classification, *i.e.* the allocation of each series to one of the groups.

The central problem of this paper could also be approached using the more general class of product partition models but this is left for further research. We refer to Hartigan (1990), Crowley (1997), MacEachern and Muller (1998), and Quintana and Iglesias (2003) for more details.

The paradigm of inference in this paper is Bayesian for several reasons. A *first* reason is that in the approach of finite mixtures, Bayesian inference allows to treat classification in a straight-

forward manner. This happens through the data augmentation technique of Tanner and Wong (1987), whereby group indicators are created and treated as parameters that facilitate the numerical integration of the posterior density. Simulated values of these parameters provide posterior densities of these indicators. A *second* reason is that mixture models are inherently difficult to estimate, due to identification difficulties. The Bayesian approach helps to identify the model by inputting adequate prior information. A *third* reason is the need to infer the number of groups. This is conveniently done by computing posterior probabilities on the range of values deemed a priori plausible. For each number, the marginal likelihood must be computed, after which posterior probabilities are easily obtained. A *fourth* reason is inherent to Bayesian inference: information coming from financial specialists can be incorporated into statistical models through the prior distribution. For example a financial specialist may have information on the persistence in volatility for a stock that he trades all the time.

The paper is organized as follows. In Section 2, we specify the model and the prior distribution. We explain in Section 3 how to compute posterior results by using the Gibbs sampler with data augmentation. In Section 4 we address the choice of the number of groups. In Section 5, we show simulated examples to illustrate the feasibility and the reliability of the procedure and in Section 6 we apply it to a set of 131 return series of the SP500 index. New paths to explore and conclusions are mentioned in Section 7.

2 Model and prior specification

Let y_t^j denote the financial return of asset j (with $j = 1, \dots, J$) for period t ($t = 1, \dots, T_j$), and y^j the $T_j \times 1$ vector containing all the observed returns for asset j . The aim is to group the J time series y^j into a small number (G) of groups. The series in the same group have identical parameters for the model in consideration. Before defining the model we define a group indicator.

Definition 1 *The group indicator S_j takes the value $s_j = g$ when the j -th series ($j = 1, \dots, J$) belongs to group g , where $g \in \{1, \dots, G\}$.*

The model is then defined as a GARCH model with conditional variance $h_t^j(\theta_g)$ if y^j belongs to group g .

Definition 2 *The model is defined by*

$$y_t^j = [h_t^j(\theta_{S_j})]^{1/2} \epsilon_t^j \quad j = 1, \dots, J; t = 1, \dots, T_j, \quad (3)$$

where $h_t^j(\theta_{S_j})$ is defined by some GARCH specification with parameter vector θ_{S_j} , ϵ_t^j for $t = 1, \dots, T_j$ and a given j are independent normal random variables with zero mean and unit variance, and $\epsilon_t^i \perp \epsilon_t^j$ for all $i \neq j$ and for all t and v .

Therefore, the model consists of G univariate GARCH equations, and includes the parameters $\theta_1, \theta_2, \dots, \theta_G$ that we collect in the vector θ . Obviously, the normal distribution for the error terms can be replaced by another more appropriate distribution (and including any additional parameter in θ). If we knew the values of the group indicators, the inference on θ would be straightforward. One could estimate θ by maximum likelihood given an assumption about the distribution of ϵ_t^j .

The choice of G is discussed in Section 4. How to specify which series belongs to which group, given G , follows the same idea as in Frühwirth-Schnatter (2001) and Frühwirth-Schnatter and Kaufmann (2005): if we assume that a priori nothing can be said about group membership, then the prior probability that the j -th series belongs to group g is assumed to be equal to the proportion of vectors in group g :

$$P(S_j = s_j) = \eta_{s_j} \quad s_j \in \{1, \dots, G\}. \quad (4)$$

The parameter $\eta = (\eta_1, \dots, \eta_{G-1})$ has to be estimated and η_G is determined as $\eta_G = 1 - \sum_{l=1}^{G-1} \eta_l$.

For notational purposes we define $\zeta = (\theta, \eta)$. Furthermore, because the S_j 's are not observed they will have to be estimated also. We define $S^J = (S_1, \dots, S_J)$ and $\psi = (S^J, \zeta)$.

Prior density: *It is factorized as*

$$\varphi(\psi) = \varphi(s^J | \eta) \varphi(\theta) \varphi(\eta) \quad (5)$$

where

$$\varphi(s^J | \eta) = \prod_{j=1}^J P(S_j = s_j) = \prod_{j=1}^J \eta_{s_j} = \prod_{g=1}^G \eta_g^{x_g} \quad (6)$$

denoting $x_g = \#(s_j = g)$, and

$$\varphi(\theta) = \prod_{g=1}^G \varphi(\theta_g). \quad (7)$$

We draw attention to several important issues. *Firstly*, the prior density on ζ is factorized into the products of the priors on θ and η , which means that the group probabilities do not affect the parameter vectors of the G groups and vice versa. *Secondly*, when the group probabilities (η) are known then the prior density on S^J can be factorized into the product of the prior densities on each S_j . Since each of the J vectors can only belong to one group at a time, we can write this product in (6) as a product over G factors. *Thirdly*, the prior density on θ is also factorized into a product of densities on the different θ_g 's. That is, we assume a priori independence between the parameters of the G GARCH components.

The fact that the high dimensional prior density on ψ is factorized by assuming several independence properties simplifies the evaluation of the posterior density. The precise choice of each density is described in Section 3. Next we define the likelihood function.

Likelihood: Suppose that y_t^j belongs to group g . Then its likelihood contribution is given by $f(y_t^j | \theta_g, I_t^j)$, which is a normal density with zero conditional mean and conditional variance equal to $h_t^j(\theta_g)$. I_t^j is the information set until $t - 1$ containing y_1^j, \dots, y_{t-1}^j and initial conditions (assumed known). The likelihood of all y_t^j is then

$$\prod_{j=1}^J \prod_{t=1}^{T_j} f(y_t^j | \theta_{S_j}, I_t^j) = \prod_{j=1}^J f(y^j | \theta_{S_j}). \quad (8)$$

Notice however that the two polar cases of overall pooling ($G = 1$) and no pooling ($G = J$) make the S_j 's redundant. In the former case there is only one model parameter vector that is the same for every data vector y^j while in the case of no pooling the model parameter is data vector specific which implies that the likelihood is just the product of the J individual likelihoods. We end this section by writing the posterior density.

Posterior density: If we denote by y all the available data then the posterior density is

$$\varphi(\psi | y) \propto \varphi(\eta) \prod_{g=1}^G \varphi(\theta_g) \prod_{j=1}^J f(y^j | \theta_{s_j}) \eta_{S_j} \quad (9)$$

$$= \varphi(\eta) \prod_{g=1}^G \eta_g^{x_g} \varphi(\theta_g) \prod_{j=1}^J f(y^j | \theta_{s_j}). \quad (10)$$

3 Gibbs sampling for the posterior density

To take advantage of the properties of (10), it is convenient to split ψ into three blocks and to use the following Gibbs sampling mechanism:

1. Sample S^J from $\varphi(s^J|\theta, \eta, y)$.
2. Sample η from $\varphi(\eta|S^J, \theta, y)$.
3. Sample θ from $\varphi(\theta|S^J, \eta, y)$.

We iterate over these blocks until convergence to the stationary distribution. See Diebolt and Robert (1994) for details on the convergence of MCMC samplers. We discuss the three blocks in detail in the next subsections.

3.1 Sampling S^J from $\varphi(s^J|\zeta, y)$

Given ζ and y the S_j 's are mutually independent. Using (6) and (10) we can write

$$\begin{aligned} \varphi(s_1, \dots, s_J|\zeta, y) &\propto \prod_{j=1}^J f(y^j|\theta_{s_j}) \varphi(s_j|\eta) \\ &= \varphi(s_1|\zeta, y) \varphi(s_2|\zeta, y) \dots \varphi(s_J|\zeta, y). \end{aligned} \quad (11)$$

The random vector $\{S_j\}_{j=1}^J$ is equivalent to a multinomial process, so we have to sample from a discrete distribution where the G probabilities are based on

$$P(S_j = g|\zeta, y^j) \propto f(y^j|\theta_g) \eta_g, \quad g = 1 \dots G, \quad (12)$$

so that

$$P(S_j = g|\zeta, y^j) = \frac{f(y^j|\theta_g) \eta_g}{\sum_{l=1}^G f(y^j|\theta_l) \eta_l}. \quad (13)$$

3.2 Sampling η from $\varphi(\eta|S^J, \theta, y)$

To sample η notice first that the relevant part of (10) is

$$\varphi(\eta|S^J, \theta, y) = \varphi(\eta|S^J) \propto \varphi(\eta) \prod_{g=1}^G \eta_g^{x_g}. \quad (14)$$

Indeed, knowing y and which vectors belong to each of the G groups implies that the likelihood is constant with respect to η . The prior on η is chosen to be a Dirichlet distribution, $Di(a_{10}, \dots, a_{G0})$ with parameter vector $a_0 = (a_{10}, \dots, a_{G0})'$, such that $a_{i0} > 0$ ($i = 1, \dots, G$) and

$$\begin{aligned} E(\eta_i|a_0) &= \frac{a_{i0}}{A_0} \\ V(\eta_i|a_0) &= \frac{a_{i0} (A_0 - a_{i0})}{A_0^2 (A_0 + 1)} \end{aligned}$$

$$\text{cov}(\eta_i, \eta_j | a_0) = -\frac{a_{i0}a_{j0}}{A_0^2(A_0 + 1)} \quad (15)$$

where $A_0 = \sum_{i=1}^G a_{i0}$. As a consequence, $\varphi(\eta | S^J)$ is also a Dirichlet, $Di(a_1, \dots, a_G)$ with $a_g = a_{g0} + x_g$, $g = 1, \dots, G$.

3.3 Sampling θ from $\varphi(\theta | S^J, \eta, y)$

Using the prior assumption (7) we can write

$$\varphi(\theta | S^J, \eta, y) = \varphi(\theta | S^J, y) = \varphi(\theta_1 | \tilde{y}^1) \varphi(\theta_2 | \tilde{y}^2) \dots \varphi(\theta_G | \tilde{y}^G) \quad (16)$$

where

$$\varphi(\theta_g | \tilde{y}^g) \propto \varphi(\theta_g) \prod_{j \in J_g} f(y^j | \theta_{S_j}) \quad (17)$$

with $J_g = \{j | S_j = g\}$ and $\tilde{y}^g = \{y^j | j \in J_g\}$, i.e. the collection of data series that belong to group g . Therefore, to sample θ one can simulate the θ_g independently. Notice that if group g is empty, $\varphi(\theta_g | \tilde{y}^g) = \varphi(\theta_g)$. A simple approach is to take proper uniform priors on θ_g . Therefore, the only user specified prior parameters in this model are the finite bounds of the uniform distributions and the parameter a_0 of the Dirichlet distribution. However, more informative (than uniform) prior densities can be easily incorporated and do not complicate the Gibbs sampling algorithm.

To sample θ_g from its conditional posterior (17), one can consider to use a Metropolis algorithm or the griddy-Gibbs sampler described for example in Bauwens, Lubrano, and Richard (1999, chap. 3). We have opted for the latter because it is less labor intensive than the former (although it is usually more computer intensive). The Metropolis algorithm samples from a proposal density which is not necessarily easy to design without a big dose of fine tuning. In particular, since the conditional posterior targeted by the proposal depends on the group indicators, its parameters should in principle be adjusted at each iteration of the Gibbs sampler (for improved numerical efficiency). Moreover, the elements of θ are the parameters of a GARCH(1,1) equation, like $h_t = \omega + \alpha \epsilon_{t-1}^2 + \beta h_{t-1}$, implying that we want to restrict them to be in the stationarity area and to be positive. In practice, the area of location of the likelihood is in the corner of a triangle, since α is close to 0 and β is close to 1, while $\alpha + \beta$ has to be smaller than 1. Hence a normal (or Student) density does not seem a convenient choice. Either it has to have very small variances, something not desirable, or it has too big variances, leading to many rejections. These rejections are needed to

avoid explosive GARCH processes. They are not the usual rejections of the Metropolis algorithm to decide whether a draw from the proposal is accepted as a draw from the conditional posterior.

Conversely, the griddy-Gibbs sampler directly uses the exact shape of the conditional posterior of each element of θ_g given the other parameters, albeit at the cost of numerical integrations. It avoids the search for proposal densities that are good approximations of the targeted conditional posterior ones. Nevertheless, it also requires some fine tuning, but directly on the integration (or prior) interval of each parameter. Our experience of this type of tuning is that it is easy to do. After running the overall Gibbs sampler for the model, one has just to look at the marginal posterior densities and check whether some of them are truncated due to an inadequate choice of some integration intervals. If this is the case, one can then redefine the intervals and run the algorithm again. We discuss further this issue in the next sub-section since it is related to the way we avoid the labeling problem inherent to mixture models.

Notice that for the griddy-Gibbs sampler, like every MCMC sampler, a burn-in phase is necessary in order to sample from the stationary distribution. More precisely, for every draw of ψ and thus of θ we apply the griddy-Gibbs sampler to every θ_g . Therefore, there is in principle a need for G burn-in phases, which has the consequence of making the computing time quite large. After some experiments of sampling from different settings we came to the conclusion that a burn-in phase for the overall Gibbs sampler suffices. This makes sense because the next draw of ψ is conditional on the last one implying that every time we draw θ we do not use some fixed starting value. Hence, the fact that the griddy-Gibbs sampler for every θ_g is a sub-chain of the overall Gibbs sampler in our model helps to reduce the overall computing time.

3.4 Multimodality and identification issues

Inherent to mixture models is an identification problem due to the arbitrary labeling of the mixture components. More precisely, the data likelihood (8) and the prior (6) on S^J are invariant to a relabeling of the groups which means that we can do a permutation of the groups without changing the value of the function. If the prior $\varphi(\zeta)$ is also invariant to relabeling then the posterior $\varphi(\psi|y)$ has this property also. As a result, the posterior may have $G!$ different modes. Because S^J , θ and η depend on this labeling we may expect that the sampling results are difficult to use for the calculation of posterior moments.

To solve the multimodality problem, identifiability constraints have to be imposed. Robert and Mengersen (1999) apply reparameterisations and multistep algorithms to G -component normal mixtures. Frühwirth-Schnatter (2001) explores first the unconstrained posterior distribution using the random permutation sampler. The aim of this sampler is to explore all the possible modes of the posterior distribution. Based on the resulting draws, she is able to graphically find identification restrictions on some parameters. One can then run a permutation sampler taking into account these restrictions.

To circumvent the labeling problem, we propose to use an identifying constraint on the parameters of the GARCH equations. This works through the prior density of θ , which must be proper at least for some parameters. The basic idea is to separate sufficiently the prior densities of the θ_g parameters of the different groups. Given the assumption of prior independence in (7), we need to specify independent distributions on $\theta_g = (\omega_g, \alpha_g, \beta_g)$, $g = 1, \dots, G$. We can simplify this further by imposing prior independence between ω_g , α_g and β_g . We choose proper distributions by selecting a finite support for each parameter. Convenient choices are (proper) uniform densities, or more generally beta densities. Therefore we may opt for a separation of the supports of the α_g parameters (as in the simulation example of Section 5) or of the β_g parameters as in the application to real data in Section 6. There is no guarantee that this type of simple separation is feasible, although in our applications to real data, we were able to find a separation based on the information conveyed by the likelihood function. If necessary, one could consider to separate the pairs of parameters (α_g, β_g) rather than just the α_g or the β_g .

Another identification problem is due to the possibility of empty groups. In Section 3.3 on the sampling of θ we mentioned that if group g is empty then $\varphi(\theta_g|\tilde{y}^g) = \varphi(\theta_g)$. Therefore an improper prior is not allowed for θ_g . However, this issue is solved by the need to define finite intervals for each GARCH parameter, in order to use the gridy-Gibbs sampler. Thus, the prior density of θ_g is always defined on a bounded region, and thus can easily be proper.

4 Choosing G

4.1 Estimation or model choice

The number of component distributions in the mixture (G) is of particular importance. There are two modelling approaches to take care of G . First, one can treat G as an extra 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 G in such a way that it favors a small number of components. A second approach is to treat the choice of G as a problem of model selection. By so-doing one separates the issue of the choice of G from estimation with G fixed (Section 3 deals with estimation with G fixed). For example, one can take $G = 2$ and $G = 3$ and do the estimation separately for the two models. Then Bayesian model comparison techniques can be applied, for instance by the calculation of the Bayes factor, see Cowles and Carlin (1996) and Chib (1995) for more details. We adopt the second approach. To implement it, we have to compute the marginal likelihood of the data for each G we consider. Once this is done, posterior probabilities for every value of G are easily computed, and one may opt for the value of G with the highest probability. In this framework, the model parameter is $\zeta = (\theta, \eta)$, not ψ which includes also S^J because of the data augmentation.

The marginal likelihood is related to the prior, posterior, and data density by

$$m(y) = \frac{f(y|\zeta)\varphi(\zeta)}{\varphi(\zeta|y)}. \quad (18)$$

Since (18) holds for any ζ in the admissible parameter space, Chib (1995) argues that one can pick a value ζ^* , and estimate the marginal likelihood in logarithm as

$$\ln \hat{m}(y) = \ln f(y|\zeta^*) + \ln \varphi(\zeta^*) - \ln \hat{\varphi}(\zeta^*|y). \quad (19)$$

Chib (1995) advises to take ζ^* as a point of high posterior mass, e.g. the posterior mean or mode. We have to evaluate the likelihood in (2) only once. The evaluation of the prior is straightforward. How to estimate the posterior at ζ^* is explained below.

4.2 Calculation of $\hat{\varphi}(\zeta^*|y)$

We start by the fact that the posterior density can be expressed as

$$\varphi(\zeta^*|y) = \varphi(\eta^*|y) \varphi(\theta^*|y, \eta^*) \quad (20)$$

with

$$\varphi(\eta^*|y) = \int \varphi(\eta^*|y, \theta, S^J) \varphi(\theta, S^J|y) d\theta dS^J \quad (21)$$

$$\varphi(\theta^*|y, \eta^*) = \int \varphi(\theta|y, \eta^*, S^J) \varphi(S^J|y, \eta^*) dS^J. \quad (22)$$

This can be further simplified because $\varphi(\eta^*|y, \theta, S^J) = \varphi(\eta^*|S^J)$ and $\varphi(\theta|y, \eta^*, S^J) = \varphi(\theta|y, S^J)$, see Section 3. One can estimate (21) by

$$\hat{\varphi}(\eta^*|y) = \frac{1}{D} \sum_{d=1}^D \varphi(\eta^*|S_{(d)}^J) \quad (23)$$

where D denotes the number of Gibbs draws. Therefore, we have to evaluate D times a Dirichlet density with parameter $a_g^{(d)}$ in the vector η^* . Because there is a closed form expression of the Dirichlet density, we know the integrating constant and it is possible to estimate (21) by the Gibbs estimate in (23).

Applying directly the same technique, *i.e.* averaging of the Gibbs draws, to estimate (22) is not so easy: we do not have Gibbs draws from $\varphi(S^J|y, \eta^*)$, we only have Gibbs draws from $\varphi(S^J|y, \eta)$ for different values of η . A solution is to apply a new Gibbs sampling to $\varphi(S^J|y, \eta^*, \theta)$ and $\varphi(\theta|y, S^J)$ so that the estimate for (22) is

$$\hat{\varphi}(\theta^*|y, \eta^*) = \frac{1}{D} \sum_{d=1}^D \varphi(\theta^*|y, S_{(d)}^J). \quad (24)$$

Notice that $\{S_{(d)}^J\}_{d=1, \dots, D}$ in (24) is different from $\{S_{(d)}^J\}_{d=1, \dots, D}$ in (23) because the former draws are sampled from a distribution with η fixed to η^* . In Chib (1995) it is necessary that all the conditional densities used in the Gibbs sampler have closed form expressions. In our model, there is no closed form expression for the density $\varphi(\theta|y, S^J)$, which is the reason why we use the griddy-Gibbs sampler. As a consequence, if we want to use (24) we are back to the initial problem of the calculation of the integrating constant of $\varphi(\theta|y, S^J)$ for each draw. However, this problem can be solved more easily than before by noticing that $\varphi(\theta|y, S^J) = \prod_{g=1}^G \varphi(\theta_g|\tilde{y}^g)$. This decomposition implies that we have to calculate the marginal likelihood

$$m(\tilde{y}^g) = \int f(\tilde{y}^g|\theta^g) \varphi(\theta_g) d\theta_g \quad (25)$$

for each lower dimensional model. For the univariate GARCH models in Sections 5 and 6, the marginal likelihood in (25) is the solution of a two-dimensional integral. This opens the door for other techniques, like deterministic integration (which we use) or a Laplace approximation. The method we propose has a non-negligible computational cost: for every draw from $\varphi(\theta|y, S^J)$ we have to calculate the G marginal likelihoods in order to have a correct estimate in (24), which we can write as

$$\hat{\varphi}(\theta^*|y, \eta^*) = \frac{1}{D} \sum_{d=1}^D \prod_{g=1}^G \frac{f(\tilde{y}^g|\theta_g^*) \varphi(\theta_g^*)}{m(\tilde{y}^g)} \quad (26)$$

$$= \frac{1}{D} \sum_{d=1}^D \exp \left[\sum_{g=1}^G [\ln (f(\tilde{y}^g|\theta_g^*) \varphi(\theta_g^*)) - \ln (m(\tilde{y}^g))] \right], \quad (27)$$

where actually \tilde{y}^g depends on $S_{(d)}^J$. Collecting all terms, the estimated marginal likelihood in (19) is given by

$$\begin{aligned} \ln \hat{m}(y) &= \sum_{j=1}^J \ln \left(\sum_{g=1}^G \eta_g^* f(y^j|\theta_g^*) \right) + \sum_{g=1}^G \ln (\varphi(\theta_g^*)) + \ln (\varphi(\eta^*)) - \ln \left(\frac{1}{D} \sum_{d=1}^D \varphi(\eta^*|S_{(d)}^J) \right) \\ &\quad - \ln \left(\frac{1}{D} \sum_{d=1}^D \exp \left[\sum_{g=1}^G [\ln (f(\tilde{y}^g(S_{(d)}^J)|\theta_g^*) \varphi(\theta_g^*)) - \ln (m(\tilde{y}^g(S_{(d)}^J)))] \right] \right). \end{aligned} \quad (28)$$

5 Simulated examples

In this section we illustrate how the Gibbs sampler performs using simulated data. We consider $J = 100$ time series of size $T_j = 1000$ drawn from a mixture with $G = 3$ components:

$$\tilde{f}(y^j) = \sum_{g=1}^3 \eta_g f(y^j|\theta_g) \quad (29)$$

with $\eta_1 = 0.25$ and $\eta_2 = 0.5$. Remember that $f(y^j|\theta_g) = \prod_{t=1}^{T_j} f(y_t^j|\theta_g, I_t^j)$, and we take

$$y_t^j|\theta_g, I_t^j \sim N(0, h_t^j) \quad (30)$$

$$h_t^j = (1 - \alpha_g - \beta_g)\tilde{\omega}^j + \alpha_g(y_{t-1}^j)^2 + \beta_g h_{t-1}^j. \quad (31)$$

For the simulation of the data we fix $\tilde{\omega}^j = 1$ which implies that the unconditional variance for every generated data series is equal to one. However, the constant $\tilde{\omega}^j$ in the conditional variance is not subject to inference, rather it is fixed at the empirical variance of the data. This technique of forcing the estimated unconditional variance to be equal to the empirical variance is called variance targeting (see Engle and Mezrich, 1996). The parameter vector for series j is then

$$\theta_{S_j} = (\alpha_{S_j}, \beta_{S_j})'. \quad (32)$$

The chosen true values for the α 's and β 's are given in Table 1. We clearly cover three different situations with respect to the persistence of the conditional variance (high, intermediate, and low persistence). The number of series belonging to each group is fixed at its expectation ($J\eta_g$). The first 25 series belong to the first group, the next 50 belong to the second and the last 25 to the third group.

5.1 Results for a correct number of groups

We discuss first the case when the model is correctly specified, in particular when the number of groups is equal to three, like in the data generation process (DGP). For the Dirichlet distribution on η we choose $a_0 = (2, 2, 2)'$, implying prior means equal to 1/3 for each group probability.

As explained in Section 3.4, we select well separated independent proper prior distributions on the different θ_g parameters to avoid the labeling problem. Given that the $\tilde{\omega}_g$ are fixed by variance targeting, we use a product of uniform densities on α_g and β_g when this is compatible with the stationarity condition ($\alpha_g + \beta_g < 1$). We take proper uniform distributions by selecting a finite support for each parameter. These intervals are given in Table 1. However, to impose the stationarity condition, we truncate the rectangular support of the prior on (α_g, β_g) implied by the independent uniform densities. In this case the prior is uniform over a trapezium rather than a rectangle. This induces therefore a prior dependence between α_g and β_g . Notice that we have chosen to separate clearly the supports of the α_g parameters, while there is some overlap of the intervals for the β_g parameters. Clearly, given the chosen DGP, we could have opted for a separation of the β_g parameters and obtain the same results, since the posterior densities of the β_g parameters are clearly separated as well (see Figure 1).

We report in Table 1 posterior moments computed from a Gibbs sample of 20000 realizations (the first 1000 being used as burn-in sample). We see that the sample information moves the equal prior means of η in the direction of the true values and that the posterior means of all parameters are reasonably close to the values of the DGP.

We can easily use the posterior draws of S^J to classify the series into the three clusters. An obvious classification rule is to attribute a series to the group to which it belongs most frequently a posteriori. For instance, the last data series never belongs to the first group, belongs 16 times to the second group and 18984 times to the third group. Hence it is attributed to group three

Table 1: Posterior results on η and θ for simulated DGP

		η_1	η_2	η_3
True value		0.25	0.50	0.25
Mean		0.2166	0.4981	0.2853
Standard deviation		0.0555	0.0763	0.0692
		$g = 1$	$g = 2$	$g = 3$
True value	α_g	0.04	0.12	0.20
	β_g	0.90	0.60	0.40
Prior interval	α_g	0.001,0.07	0.07,0.15	0.15,0.25
	β_g	0.65,0.97	0.45,0.75	0.20,0.60
Mean	α_g	0.0435	0.1041	0.1975
	β_g	0.8758	0.5917	0.4369
Standard deviation	α_g	0.0060	0.0092	0.0132
	β_g	0.0238	0.0306	0.0350
Correlation α_g, β_g		-0.7849	-0.71409	-0.7184

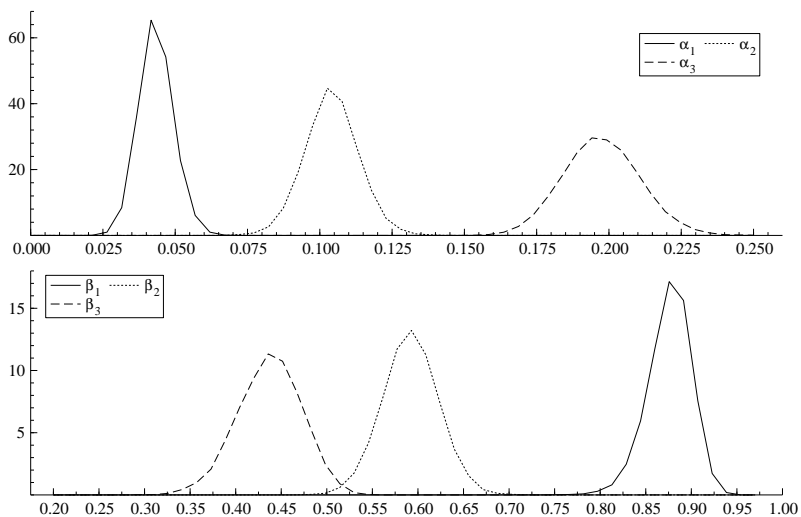


Figure 1: Posterior densities of the elements of θ_g for simulated DGP

Table 2: Model choice criteria for simulated DGP

G	Marginal log-lik.	Maximized log-lik.	# par.	BIC
1	-48085.20	-48078.49	2	-48085.40
2	-48035.39	-48019.68	5	-48036.95
3	<i>-48028.65</i>	-48004.57	8	<i>-48032.20</i>
4	-48035.09	-48004.17	11	-48042.16
100	-48064.48	-47836.94	200	-48527.72

(which is correct). For the 100 series, the number of correct classifications amounts to 86, which is a good score.

5.2 Results for incorrect numbers of groups

We report in Table 2 (second column) the values of the marginal log-likelihood for different values of G . They are computed using formula (28), using the posterior mean as a high density point (using the ML estimates, we obtain results that differ only in the decimals). Not surprisingly, the preferred model is the correct one (3 groups). The Bayesian information criterion (BIC), see Schwarz (1978), also selects the correct model, see the last column of the table. The BIC is equal to the maximized log-likelihood value less a penalty term equal to the number of parameters times $\ln(T)/2$ ($T = 1000$ in this example). Posterior results for 1, 2, 4, and 100 groups are available in Bauwens and Rombouts (2003), the discussion paper version of this article, available on line.

In Figure 2, we show, for 100 groups, the posterior densities of the α and β GARCH parameters. We bet that someone who does not know the DGP would not guess that it has three groups.

6 Application to US stocks

We apply the clustering idea to 131 stock return series belonging to the biggest US companies. The data source is Datastream and the stock list can be found in Bauwens and Rombouts (2003). Each return is observed from 29/09/99 to 30/07/03 corresponding to 1000 observations. Table 3 presents a summary of the descriptive statistics of the series. It shows that there is a lot of

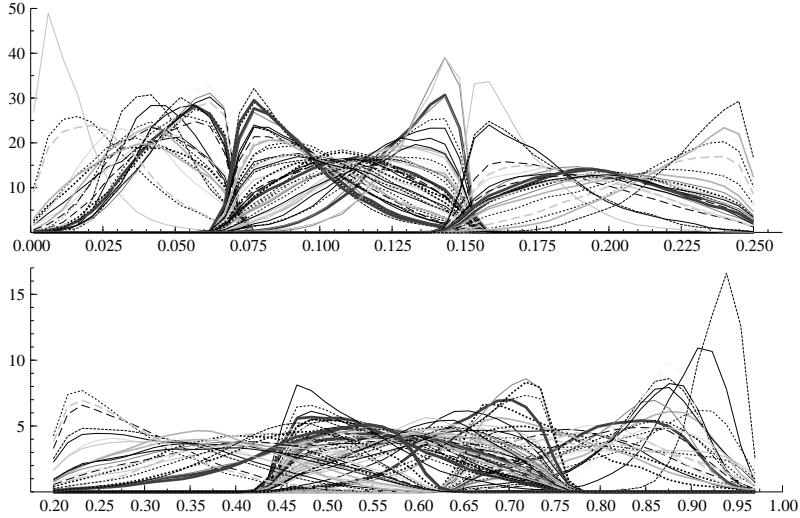


Figure 2: Posterior densities of the elements of θ_g ($G = 100$)

variation in the different empirical characteristics of the stocks. For example, the mean kurtosis for all the series is 8.83 but it ranges from 3.43 to 90.4 with a standard deviation of 10.7. We also found a lot of heterogeneity in the estimates of GARCH(1,1) models for each series, see Bauwens and Rombouts (2003) for details. The most likely reason for this heterogeneity is that individual stocks react differently to general news and specific company announcements.

In this paper we are only interested in the conditional variances of the series. Taking into account the conditional correlations in a second step, in the spirit of dynamic conditional correlation (DCC) models (see Engle, 2002, and Tse and Tsui, 2002) would be necessary to build a complete multivariate GARCH model. We leave this for future research.

To select the number of groups, we allow a priori G to take the values, 1, 2, 3, 4, and 131. Table 4 reports the corresponding values of the marginal log-likelihood. We come to the clear conclusion that the appropriate number of groups is three. We therefore report the results for three groups, based on 20000 draws from the Gibbs sampler described in Section 3, out of which we dropped the first 1000.

The posterior means of η and θ can be found in Table 5. The posterior densities of η are given in Figure 3, and those of the GARCH parameters α_g and β_g are in Figure 4. The prior on η is a Dirichlet with parameters equal to 2, as in the example of Section 5. The posterior densities of

Table 3: Summary of descriptive statistics for 131 series

	mean	st. dev.	minimum	maximum
mean	-0.0007	0.05	-0.18	0.15
std	2.56	0.78	1.63	6.00
min	-15.75	7.64	-57.3	-6.89
max	13.38	4.93	5.99	31.4
skew	-0.17	0.77	-5.20	0.96
kurt	8.83	10.70	3.43	90.4

Each line of this table reports the mean, standard deviation (st. dev.), minimum, and maximum of the descriptive statistics (mean, std, min, max, skew, kurt) of the 131 series, see Bauwens and Rombouts (2003) for details.

Table 4: Marginal log-likelihood for 131 series

G	Marginal log-lik.	# par.
1	-179457.40	2
2	-179230.35	5
3	-179129.93	8
4	-180559.68	11
131	-179357.60	262

η_1 and η_2 are quite similar and centered around 0.45. Thus the density of η_3 is located around 0.12. The negative correlation between η_1 and η_2 is relatively high while the correlations between η_1 and η_3 , and η_2 and η_3 are less pronounced.

After some trials, we could define three almost non-intersecting intervals for the β_g parameters (see Table 5). Given this, the prior intervals for the α_g parameters were adjusted to avoid too much zero mass in the tails of the densities. This avoids wasting points in computing the numerical integrals for the griddy-Gibbs steps. The prior densities are all uniform on finite intervals.

The posterior marginal densities of the β_g parameters are clearly separated, despite the large standard deviation of β_3 compared to β_1 and β_2 . We can also see that the posterior densities of

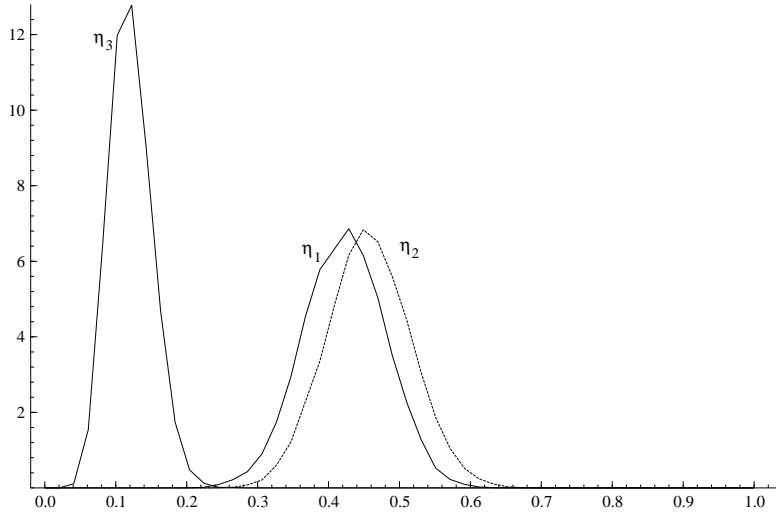


Figure 3: Posterior marginals of η_g for 131 series

α_2 and α_3 are centered in the same area, with a much larger standard deviation for α_3 . This does not imply that we should merge groups two and three. Indeed, the persistence $\alpha_g + \beta_g$ is clearly different between these groups (0.96 for group 2, 0.79 for group 3). Notice also how the high persistence for the first group (0.98) is forcing the correlation between α_1 and β_1 to be close to minus one.

Finally, applying the classification rule that a data series belongs to the group to which it belongs most frequently a posteriori (see Section 5), we find that 56 series belong to the first group, 60 series to the second and 15 series to the third group. This low number for the third group helps to understand the large posterior standard deviations of α_3 and β_3 , compared to the other groups. Since the posterior densities for group three are clearly unimodal, there is no need to split it up and to add a fourth group. This is in agreement with the marginal likelihood values.

The posterior probability that a series belongs to its group can be estimated by the mean of the corresponding group indicator, provided by the Gibbs output. For a large majority of the series, actually 93 (i.e. 71 percent), this posterior probability is larger than 0.9, while it is less than 0.6 for only 8 series (6 percent). Thus, the allocation of the series to the groups is rather clear, but it should be kept in mind that the model does not imply a sure classification, since each series has a non-zero probability to belong to each group.

Table 5: Posterior results on η and θ for 131 series

		η_1	η_2	η_3
Mean		0.4248	0.4513	0.1239
Standard deviation		0.0594	0.0598	0.0312
Correlation matrix		1	-0.8632	-0.2502
		-0.8632	1	-0.2726
		-0.2502	-0.2726	1
		$g = 1$	$g = 2$	$g = 3$
Prior interval	α_g	0.02,0.07	0.07,0.12	0.055,0.13
	β_g	0.90,0.99	0.82,0.92	0.58,0.80
Mean	α_g	0.0474	0.0908	0.0862
	β_g	0.9438	0.8653	0.7095
Standard deviation	α_g	0.0037	0.0044	0.0083
	β_g	0.0047	0.0081	0.0304
Correlation α_g, β_g		-0.9674	-0.8733	-0.6635

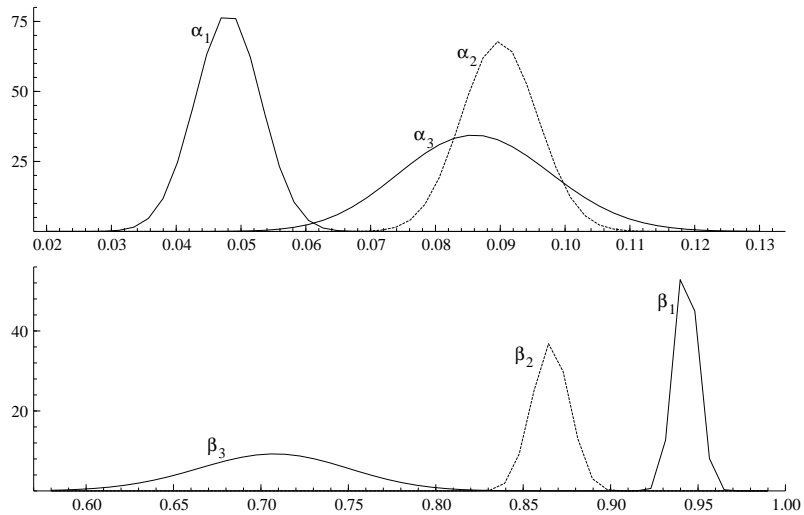


Figure 4: Posterior densities of the elements of θ_g for 131 series

The question may be raised if there is an economic or financial interpretation of the groups (e.g. in terms of sectors). Searching for an interpretation of this kind would require to analyze the classification in relation to observable characteristics of the firms (which we do not have). We do not believe that this would be a fruitful exercise, since the model is not designed for this purpose. A possible extension of our model would be to parameterize the group probabilities as functions of observable variables, but this is beyond the scope of this paper. Recent papers which make use of time-varying probabilities for mixture models are Frühwirth-Schnatter and Kaufmann (2005), Geweke and Keane (2005), and Bauwens, Preminger, and Rombouts (2006).

The interpretation of the groups, according to the classification rule we have proposed, can only be done in terms of the GARCH parameters. Group 1 corresponds to highly persistent conditional variances ($\alpha_1 + \beta_1$ estimated at 0.98), and group 3 to less persistent processes ($\alpha_3 + \beta_3$ estimated at 0.79). In terms of persistence, group 2 is closer to group 1 than to group 3, with $\alpha_2 + \beta_2$ estimated at 0.96. The difference between groups 1 and 2 lies in the relative importance of the impact of the lagged shock (0.05 in group 1, 0.09 in group 2) and of the autoregressive parameter of the conditional variance (0.94 in group 1, 0.87 in group 2).

7 Conclusion

We have addressed the problem of estimating a large number of GARCH models. The approach consists in pooling similar series in a cluster and using a small number of clusters. The model specifies the distribution of each series as a mixture of a small number of GARCH models. We have illustrated that inference is feasible using the Bayesian approach by data augmentation and the Gibbs sampler. The Gibbs sampler has two levels: at the first level, we have three blocks (corresponding to group indicators, group probabilities, and parameters of the GARCH components), and at the second level, for the GARCH parameters, we have used the grid-Gibbs sampler within each group. We have illustrated with simulated and real data that the approach is feasible and delivers sensible results.

Several extensions and applications are on our agenda. *Firstly*, more flexible specifications of the component distributions could be considered. We use normal densities for ease of illustration. Student t and skew-t densities could, and probably, should be used. Even a non-parametric

specification can be considered. *Secondly*, the same method can be used to cluster a large number of small multivariate GARCH models. One application of this approach would be to adapt the study of Kearney and Patton (2000). The practical limit is the length of computations given that the numerical burden of the second level Gibbs sampler (grid-Gibbs) is proportional to the number of parameters of each GARCH component. As an alternative approach, one can try and replace the second level Gibbs sampler by a Metropolis step. *Thirdly*, in principle, our algorithm can be used to split a single long (univariate or multivariate) series in different groups corresponding to different regimes: the latent variables would indicate to which regime each observation belongs. See Bauwens, Preminger, and Rombouts (2006) for a related regime-switching GARCH model. *Fourthly*, our medium term more ambitious objective is to construct and estimate a multivariate GARCH model for a large number of series. One idea is to find the members of the clusters by the approach of this paper. Given the clusters, we can then specify correlation (or covariance) models within each cluster. The last task would be to correlate the clusters by a higher level link.

References

- BAUWENS, L., S. LAURENT, AND J. ROMBOUTS (2006): “Multivariate GARCH Models: A Survey,” *Journal of Applied Econometrics*, 21, 79–109.
- BAUWENS, L., M. LUBRANO, AND J. RICHARD (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, Oxford.
- BAUWENS, L., A. PREMINGER, AND J. ROMBOUTS (2006): “Regime-switching GARCH models,” Revision of CORE Discussion Paper 2006-11.
- BAUWENS, L., AND J. ROMBOUTS (2003): “Bayesian Clustering of Similar GARCH Models,” CORE DP 2003/87.
- CHIB, S. (1995): “Marginal Likelihood From the Gibbs Output,” *Journal of the American Statistical Association*, 90, 1313–1321.
- CHIB, S., AND B. HAMILTON (2000): “Bayesian Analysis of Cross-Section and Clustered Data Treatment Models,” *Journal of Econometrics*, 97, 25–50.
- COWLES, M., AND B. CARLIN (1996): “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” *Journal of the American Statistical Association*, 91, 883–904.
- CROWLEY, E. (1997): “Product Partition Models for Normal Means,” *Journal of the American Statistical Association*, 92, 192–198.
- DIEBOLT, J., AND C. ROBERT (1994): “Estimation of Finite Mixture Distributions through Bayesian Sampling,” *Journal of the Royal Statistical Society, Series B*, 56, 363–375.
- ENGLE, R. (2002): “Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models,” *Journal of Business and Economic Statistics*, 20, 339–350.
- ENGLE, R., AND J. MEZRICH (1996): “GARCH for Groups,” *RISK*, 9, 36–40.
- FRÜHWIRTH-SCHNATTER, S. (2001): “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models,” *Journal of the American Statistical Association*, 96, 194–209.

- FRÜHWIRTH-SCHNATTER, S., AND S. KAUFMANN (2005): “Model-based clustering of multiple time-series,” Working Paper, Johannes Kepler Universität Linz.
- GEWEKE, J., AND M. KEANE (2005): “Smoothly mixing regressions,” *Forthcoming in Journal of Econometrics*.
- HARTIGAN, J. (1990): “Partition Models,” *Communications in Statistics, Part A*, 19, 2745–2756.
- KEARNEY, C., AND A. PATTON (2000): “Multivariate GARCH Modelling of Exchange Rate Volatility Transmission in the European Monetary System,” *Financial Review*, 41, 29–48.
- MACEachern, S., AND P. MULLER (1998): “Estimating Mixture of Dirichlet Process Models,” *Journal of Computational and Graphical Statistics*, 7, 223–238.
- QUINTANA, A., AND P. IGLESIAS (2003): “Bayesian Clustering and Product Partition Models,” *Journal of the Royal Statistical Society, Series B*, 65, 557–574.
- 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.
- ROBERT, C., AND K. MENGENSEN (1999): “Reparametrisation Issues in Mixture Modelling and their bearing on MCMC algorithms,” *Computational Statistics and Data Analysis*, 29, 325–343.
- SCHWARZ, G. (1978): “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.
- TANNER, M., AND W. WONG (1987): “The calculation of posterior distributions by data augmentation,” *Journal of the American Statistical Association*, 82, 528–540.
- TSE, Y., AND A. TSUI (2002): “A multivariate Generalized Auto-regressive Conditional Heteroskedasticity model with time-varying correlations,” *Journal of Business and Economic Statistics*, 20, 351–362.