BAYESIAN CLUSTERING OF MANY GARCH MODELS

L. Bauwens[1] , *J.V.K. Rombouts*[1]

November 11, 2003

Preliminary version

Abstract

We consider the estimation of a large number of GARCH models, of the order of several hundreds. To achieve parsimony, we classify the series in a small number of groups. 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. Illustrations are provided.

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

JEL Classification: C11, C32

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

The authors would like to thank Michel Mouchart for interesting discussions.

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 (2003) for a survey. MGARCH models define the conditional variance matrix as a function of the past data. However, the number of parameters to estimate in a multivariate GARCH model rises fastly with J, rendering them useless for modelling more than a handful of series. For example, the BEKK model of Engle and Kroner (1995), in its simplest form with one lag, would have 625,250 parameters in the SP500 example. By using a technique called 'variance targeting' by Engle and Mezrich (1996), this number is reduced to 500,000, which is still unmanageable.

Recently, dynamic conditional correlation (DCC) models were proposed by Engle (2002) and by Tse and Tsui (2002). These models generalize the constant conditional correlation model (CCC) of Bollerslev (1990). They require much less parameters than the BEKK model (1,502 in the SP500 example if one uses GARCH(1,1) models for the conditional variances and 'correlation targeting'). An essential feature of the DCC (and CCC) models is that one specifies separately the conditional variances and the conditional correlations. This feature enables a two-step consistent estimation procedure where one first estimates the parameters of the conditional variances, without taking account of the correlation parameters. This boils down to estimating univariate GARCH models separately if there are no lagged shocks or volatility spillovers. 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 this allows us firstly to present our new methodology in a simple case, and secondly because univariate GARCH models are the cornerstone of MGARCH models of the DCC type. Although the estimation of a large number of univariate GARCH models does not raise technical difficulties, it raises at least the issue of reporting a large number of estimates. Our main idea is that estimating a large number of univariate (or even low-dimensional multivariate) GARCH models can be circumvented 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 address this problem by treating the data as a draw of a finite mixture of distributions,

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

where $\eta_1 + \ldots + \eta_G = 1$ and y^j is the j-th time series of returns, possibly a vector (see Section 2 for details of notations). This implies a difficult likelihood to work with because it contains G^J terms:

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

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 modelling is the use of a normal mixture $\sum_{g=1}^{G} \eta_g N(\mu_g, \sigma_g^2)$. In this paper we have minor interest in the location or the scale as such and 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 $(g = 1, \ldots, G)$, 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 imagine 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 matter of particular importance in this respect is classification, *i.e.* the allocation of each series to one of the groups.

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 straightforward manner. This happens through the data augmentation technique, 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. On the one hand, financial decisions that are somehow based on an estimated parameter vector can be made more precise because of the accumulation of the prior and data information. On the other hand, these financial decisions are again made by financial specialists who duly appreciate that their knowledge is incorporated in the model. 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 organised as follows. In Section 2, we specify the model and the prior distribution. Since the corresponding posterior distribution is too complicated to conduct Bayesian inference analytically, we explain in Section 3 how we solve this problem by using the Gibbs sampler. Under regularity conditions, the draws from this Markov chain Monte Carlo (MCMC) sampler can be regarded as draws from the posterior distribution. In Section 4 we address the choice of G, 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

We have a collection of J vectors of N elements of (financial) time series of length T_j

$$\left\{y_{it}^{j}\right\} \quad i=1,\ldots,N;\ t=1,\ldots,T_j;\ j=1,\ldots,J. \tag{3}$$

We denote $y_t = (y_t^1, y_t^2, \ldots, y_t^J)$ as containing the J vectors y_t^j of dimension $N \times 1$ at time t, and y^j the $T_j \times N$ matrix containing all data on vector j. The aim is to group the J vectors into G groups. Members of the same group have common parameter vectors for the model in consideration. Before defining the model we define a group indicator.

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

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

Definition 2

$$y_t^j = [H_t^j(\theta_{S_j})]^{1/2}\epsilon_t^j \qquad j=1,\ldots,J;\ t=1,\ldots,T_j, \tag{4}$$

where $H_t^j(\theta_{S_j})$ is defined by some MGARCH specification, and where ϵ_t^j are i.i.d. with $E(\epsilon_t^j) = 0$, $V(\epsilon_t^j) = I_N$, and $\epsilon_t^i \perp\!\!\!\perp \epsilon_v^j \quad \forall i \neq j \quad \forall t,\ \forall v$.

Therefore, the model includes G parameter vectors $\theta_1, \theta_2, \ldots, \theta_G$ which we collect in the vector θ for notational convenience. When we know the values that the group indicators take, the inference on θ becomes straightforward. One can for example estimate θ by maximum likelihood given that ϵ_t^j has a specified distribution.

The choice of G is discussed in Section 4. How to decide which pair belongs to which group, given G, follows the same idea as in Frühwirth-Schnatter (2001): 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:

$$P(S_j = s_j) = \eta_{s_j} \qquad s_j \in \{1, \ldots, G\}. \tag{5}$$

The parameter $\eta = (\eta_1, \ldots, \eta_{G-1})$ 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 $S^J = (S_1, \ldots, S_J)$ and $\psi = (S^J, \zeta)$.

Assumption 1 *The prior density is factorized as*

$$\varphi(\psi) = \varphi(S^J|\eta)\,\varphi(\zeta) \tag{6}$$

$$= \varphi(S^J|\eta)\,\varphi(\theta)\,\varphi(\eta) \tag{7}$$

where

$$\varphi(s^J|\eta) = \prod_{j=1}^{J} P(S_j = s_j) = \prod_{j=1}^{J} \eta_{s_j} = \prod_{g=1}^{G} \eta_g^{x_g} \tag{8}$$

denoting $x_g = \sharp(s_j = g)$ *and*

$$\varphi(\theta) = \prod_{g=1}^{G} \varphi(\theta_g). \tag{9}$$

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 the same time we can write this in (8) as a product over G factors. This is explained in more detail in Appendix 1. *Thirdly*, the prior density on θ is also factorized into a product of densities on the θ_g's. That is, we assume a priori independence between the θ_g's.

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

Likelihood: *S*uppose that the N-variate vector 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 matrix equal to $H_t(\theta_g)$. I_t^j is the information set until $t-1$ containing (at least) $y_1^j, \ldots, y_{t-1}^j$ and initial conditions (assumed known). A likelihood is available for each N-variate vector y_t^j:

$$\prod_{j=1}^{J} \prod_{t=1}^{T_j} f(y_t^j|\theta_{S_j}, I_t^j) = \prod_{j=1}^{J} f(y^j|\theta_{S_j}). \tag{10}$$

It is possible to relax the normality assumption, for example to allow for a higher kurtosis of the data, and take another family of component distributions. This implies the inclusion of an extra vector ν_g containing other parameters of f. One can think as if θ_{S_j} contains ν_g such that we do not loose any generality.

A crucial fact is that we do not know which pair belongs to which group. This is why we consider the S_j's as latent parameters in the model. See Tanner and Wong (1987) for more details. Notice however that the two polar case 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 latter 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 summarize this section by writing the posterior density.

Posterior density: If we denote bt y all the available data then the posterior density is written as

$$\begin{aligned} \varphi(\psi|y) &\propto \varphi(\eta)\prod_{g=1}^{G}\varphi(\theta_g)\prod_{j=1}^{J}f(y^j|\theta_{S_j})\eta_{S_j} && (11) \\ &= \varphi(\eta)\prod_{g=1}^{G}\eta_g^{x_g}\varphi(\theta_g)\prod_{j=1}^{J}f(y^j|\theta_{S_j}). && (12) \end{aligned}$$

3 Gibbs sampling for the posterior density

To take advantage of the properties of (12), 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 seen to be mutually independent. Using (8) and (12) we can write

$$\begin{aligned}\varphi(S_1, \ldots, S_J|\zeta, y) &\propto \prod_{j=1}^{J} f(y^j|\theta_{S_j})\, \varphi(S_j|\eta) \\ &= \varphi(S_1|\zeta, y)\, \varphi(S_2|\zeta, y)\, \ldots\, \varphi(S_J|\zeta, y). \qquad (13)\end{aligned}$$

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

$$P(S_j = g|\zeta, y^j) \propto f(y^j|\theta_g)\, \eta_g, \;\; g = 1 \ldots G, \qquad (14)$$

so that

$$P(S_j = g|\zeta, y^j) = \frac{f(y^j|\theta_g)\, \eta_g}{\sum_{l=1}^{G} f(y^j|\theta_l)\, \eta_l}. \qquad (15)$$

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

Notice that we just defined the posterior probability distribution of S_j and how to sample from it. We have to repeat this for all the S_j, which means J times. Furthermore the probabilities are calculated conditional on ζ and therefore we have to calculate the probability distribution each time in the loop of the Gibbs sampler.

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

To sample η notice first that the relevant part of (12) is

$$\varphi(\eta|S^J, \theta, y) = \varphi(\eta|S^J) \propto \varphi(\eta) \prod_{g=1}^{G} \eta_g^{x_g}. \qquad (16)$$

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, $Di(a_{10}, \ldots, a_{G0})$ with parameter vector $a_0 = (a_{10}, \ldots, a_{G0})'$. As a consequence, $\varphi(\eta|S^J)$ is also a Dirichlet, $Di(a_1, \ldots, a_G)$ with $a_g = a_{g0} + x_g$, $g = 1, \ldots, G$. More details about this and sampling from a Dirichlet distribution can be found in Appendix 2.

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

Using the prior assumption (9) we can write

$$\varphi(\theta|S^J, \eta, y) = \varphi(\theta|S^J, y) = \varphi(\theta_1|\tilde{y}^1)\ \varphi(\theta_2|\tilde{y}^2)\ \ldots\ \varphi(\theta_G|\tilde{y}^G) \tag{17}$$

where

$$\varphi(\theta_g|\tilde{y}^g) \propto \varphi(\theta_g) \prod_{j \in J_g} f(y^j|\theta_{S_j}) \tag{18}$$

and $J_g = \{j \mid S_j = g\}$, and $\tilde{y}^g = \left\{y^j \,|j \in J_g\right\}$, i.e. the collection of data series that belong to group g. Therefore, to sample θ one can simulate the θ_g independently. The latter can be done using the griddy-Gibbs sampler, see for example Bauwens, Lubrano, and Richard (1999, chap. 3). Notice that if group g is empty, $\varphi(\theta_g|\tilde{y}^g) = \varphi(\theta_g)$. A simple approach is to take uniform priors on θ_g. Therefore the only user specified prior parameters in this model are the bounds of the uniform distributions and a_0 of the Dirichlet distribution. However, more informative prior densities can be easily incorporated and do not complicate the Gibbs sampling algorithm.

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 for every θ_g. Therefore, there is a need for G burn-in phases which has large computing time consequences. 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 the nature of $S^J = (S_1, \ldots, S_J)$, the discrete latent process, problems may arise by sampling from the unconstrained posterior distribution on ψ. More precisely, the complete data likelihood, see (10), and the prior on S^J, see (8), are invariant to a relabeling of the groups which means that we can take the labeling $\{1, 2, \ldots, G\}$ and do a permutation $\{\rho(1), \rho(2), \ldots, \rho(G)\}$ 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. Notice that for a commom, *i.e.* invariant to relabeling, parameter vector, inference can be done without any problem.

To solve the multimodality problem, identifiability constraints have to be imposed. Robert and Mengersen (1999) apply succesfully 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. We propose an easy identifiability constraint that uses the fact that we work with GARCH models. By selecting rather non-overlapping supports for the parameters, we are able circumvent the multimodality problem, see Section 5 for more details.

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, as we will see later in Section 5 that presents some illustrative examples, we can still be rather non-informative by taking proper uniform priors.

4 Choosing G

4.1 Inference or model choice

G, the number of component distributions in the mixture, 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 favours 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 choose the second approach. To implement it, we have to compute the marginal likelihood of the data for each G that we want to 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.

Definition 3 *The marginal likelihood is defined as the integral of the likelihood with respect to the prior density*

$$m(y) = \int f(y|\zeta)\varphi(\zeta)d\zeta. \tag{19}$$

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

$$m(y) = \frac{f(y|\zeta)\varphi(\zeta)}{\varphi(\zeta|y)}. \tag{20}$$

Notice that (20) is an identity that holds for every ζ. For a given value ζ^*, the estimate, in logarithms, is

$$\ln \hat{m}(y) = \ln f(y|\zeta^*) + \ln \varphi(\zeta^*) - \ln \hat{\varphi}(\zeta^*|y). \tag{21}$$

We have to evaluate the likelihood in (2) only once and 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

$$\varphi(\zeta^*|y) = \varphi(\eta^*|y)\ \varphi(\theta^*|y, \eta^*) \tag{22}$$

with

$$\begin{aligned} \varphi(\eta^*|y) &= \int \varphi(\eta^*|y, \theta, S^J)\ \varphi(\theta, S^J|y)\ \mathrm{d}\theta\ \mathrm{d}S^J & (23) \\ \varphi(\theta^*|y, \eta^*) &= \int \varphi(\theta|y, \eta^*, S^J)\ \varphi(S^J|y, \eta^*)\ \mathrm{d}S^J. & (24) \end{aligned}$$

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 (23) by

$$\hat{\varphi}(\eta^*|y) = \frac{1}{D}\sum_{d=1}^{D} \varphi(\eta^*|S^J_{(d)}) \tag{25}$$

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, see Appendix 2, we know the integrating constant and it is possible to estimate (23) by the Gibbs estimate in (25).

Applying directly the same technique, *i.e.* averaging of the Gibbs draws, to estimate (24) is impossible: 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 (24) is

$$\hat{\varphi}(\theta^*|y,\eta^*) = \frac{1}{D}\sum_{d=1}^{D}\varphi(\theta^*|y,S^J_{(d)}). \tag{26}$$

Remark that $\{S^J_{(d)}\}_{d=1,\ldots,D}$ in (26) are different from $\{S^J_{(d)}\}_{d=1,\ldots,D}$ in (25) 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 in this paper. As a consequence, if we want to use (26) we are back at 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

$$m(\tilde{y}^g) = \int f(\tilde{y}^g|\theta^g)\,\varphi(\theta_g)\,\mathrm{d}\theta_g \tag{27}$$

for each lower dimensional model. For the example of univariate GARCH models in Section 5, the marginal likelihood in (27) is the solution of a two-dimensional integral. This opens the door for other techniques, like deterministic integration or a Laplace approximation. These two alternative methods are explained in Appendix 3. The method we propose has a non-negligable 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 (26), that we can write as

$$\begin{aligned}
\hat{\varphi}(\theta^*|y,\eta^*) &= \frac{1}{D}\sum_{d=1}^{D}\prod_{g=1}^{G}\frac{f(\tilde{y}^g|\theta^*_g)\,\varphi(\theta^*_g)}{m(\tilde{y}^g)} && (28)\\
&= \frac{1}{D}\sum_{d=1}^{D}\exp\left[\sum_{g=1}^{G}\left[\ln\left(f(\tilde{y}^g|\theta^*_g)\,\varphi(\theta^*_g)\right)-\ln\left(m(\tilde{y}^g)\right)\right]\right], && (29)
\end{aligned}$$

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

$$\begin{aligned}\ln \hat{m}(y) \quad = \quad & \sum_{j=1}^{J} \ln \left(\sum_{g=1}^{G} \eta_g^* \; f(y^j|\theta_g^*) \right) + \sum_{g=1}^{G} \ln \left(\varphi(\theta_g^*) \right) + \ln \left(\varphi(\eta^*) \right) - \ln \left(\frac{1}{D} \sum_{d=1}^{D} \varphi(\eta^* | S^J_{(d)}) \right) \\ & - \ln \left(\frac{1}{D} \sum_{d=1}^{D} \exp \left[\sum_{g=1}^{G} \left[\ln \left(f(\tilde{y}^g(S^J_{(d)})|\theta_g^*) \; \varphi(\theta_g^*) \right) - \ln \left(m(\tilde{y}^g(S^J_{(d)})) \right) \right] \right] \right). \qquad (30)\end{aligned}$$

5 Simulated examples

In this section we illustrate how the Gibbs sampler performs by the use of examples that mimic realistic financial settings.

5.1 Three groups for one hundred series

5.1.1 DGP

We consider $J = 100$ time series of size $T_j = 1000$ drawn from a mixture with $G = 3$ components:

$$\tilde{f}(y^j) = \sum_{g=1}^{3} \eta_g f(y^j|\theta_g) \qquad (31)$$

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

$$\begin{aligned} y_t^j|\theta_g, I_t^j \quad & \sim \quad N(0, h_t^j) \qquad (32) \\ h_t^j \quad & = \quad (1 - \alpha_g - \beta_g)\tilde{\omega}^j + \alpha_g (y_{t-1}^j)^2 + \beta_g h_{t-1}^j. \qquad (33) \end{aligned}$$

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 pair j is then

$$\theta_{S_j} = (\alpha_{S_j}, \beta_{S_j})'. \qquad (34)$$

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 volatility process. The first process has a high persistence in the variance because $\alpha_1 + \beta_1 = 0.94$ is close to 1, the bound for a weakly stationary process. In stock markets, these could be stocks with large market caps. The second process is less persistent, with

$\alpha_2 + \beta_2 = 0.72$. The third gives relatively less weight to the lagged conditional variance and is slightly less persistent ($\alpha_3 + \beta_3 = 0.60$) than the second process.

The number of series belonging to each group is fixed at its expectation ($J\eta_g$). That is, the first 25 series belong to the first group, the next 50 belong to the second and the last 25 to the third group. This order is of course not important but we choose it in this way to simplify the comparison with the posterior classification. For the simulation we can choose G between the two polar cases of overall pooling ($G = 1$) and no pooling ($G = J$). This means for our example of 100 series that the number of parameters in θ may vary between 2 and 200. We take $G = 3$, the real number of components in the mixture which implies that θ contains 6 elements. Therefore the augmented parameter vector ψ contains $100 + 6 + 2 = 108$ parameters.

5.1.2 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 DGP. As we mentioned in Section 3.4 we have to select a proper prior distribution on θ. Given the assumption on the prior distribution on θ in (9) we only need to specify prior distributions on $\theta_g,\ g = 1, \dots, 3$ which are bivariate distributions in our example. We can still simplify this further by imposing prior independence of the elements in θ_g, *i.e.* taking the prior on θ_g as a product of the prior on α_g and the prior on β_g. In this example we take uniform distributions for the priors. This implies that we only have to select the support of the uniform distributions in order to have a proper prior. These intervals are given in Table 1. Because of the stationarity condition, $\alpha_g + \beta_g < 1$, it may happen during the Gibbs sampling that the joint support is not a rectangle, created by the respective bounds, anymore. This induces therefore a prior dependence between α_g and β_g, i.e. in this case the prior is uniform over a trapezium rather than a rectangle. Notice that other prior distributions on θ_g are possible also. One could think of beta distributions for example. For the Dirichlet distribution on η we choose $a_0 = (2, 2, 2)'$ to exclude empty groups a priori.

To compute the posterior results, we have drawn 20000 realisations of ψ and we used a burn-in period of 1000 draws. The computing time is about 40 hours on a powerful computer (2.6Ghz Intel Xeon processor). We first discuss the results on η. Figure 1 displays the posterior marginals that are rather symmetric. The Dirichlet prior on η implies that the prior means are all equal to 1/3. Therefore, the data play an important role in rectifying this prior information. That the elements of η are negatively related because of the restriction $\sum_{g=1}^{3} \eta_g = 1$ is exemplified in the correlation matrix in Table 1. The correlation between η_3 and η_1 and η_2 is high because η_3 is centered around 0.5, leaving minor freedom to the other two parameters.

We focus next on θ_g. Figure 2 shows the posterior marginals of α_g and β_g. While α_1 is slightly skewed to the right, the converse is true for β_1. The reason for this skewness is that the upper bounds for α_1 and β_1, see Table 1, are not respecting the stationarity condition $\alpha_1 + \beta_1 < 1$. One can easily distinguish three clusters in the way we expect them to appear. There is no overlapping for the α_g by choice of the prior intervals.

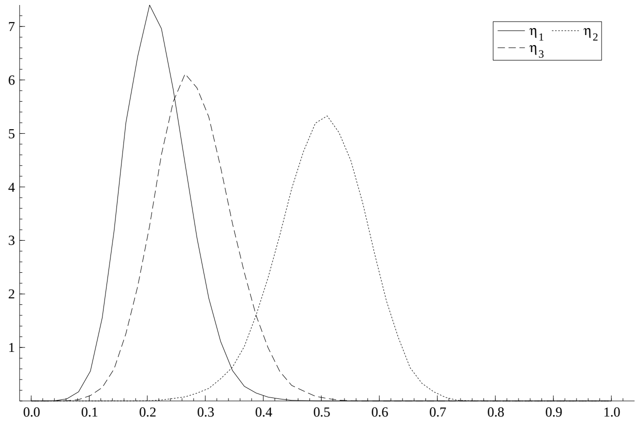


Figure 1: Posterior marginals of η_g $(G = 3)$

Figure 3 also reveals clearly that β_2 and β_3 are partially overlapping, to see this consider only the β_g axis (the horizontal axis). Nevertheless, as we already mentioned before, there is no identification problem because in the α_g direction no overlapping occurs. Table 1 provides some posterior summary statistics for θ. With respect to the posterior means we find values reasonably close to the values of the data generating process. The posterior standard deviations are rather small as we can also observe from Figure 2. The reason for the strong negative correlation within each θ_g is of the same nature as that for η, namely the parameter restriction $\alpha_g + \beta_g < 1$.

Until now we discussed the posterior results on ζ, *i.e.* the group probabilities and the parameters characterizing the component distributions. However, the fact that we use the data augmentation technique allows us to say something about the classification issue also. More pre-

Table 1: Posterior results on η and θ $(G = 3)$

		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
Correlation matrix		1	-0.4851	-0.2677
		-0.4851	1	-0.7127
		-0.2677	-0.7127	1
		$g = 1$	$g = 2$	$g = 3$
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

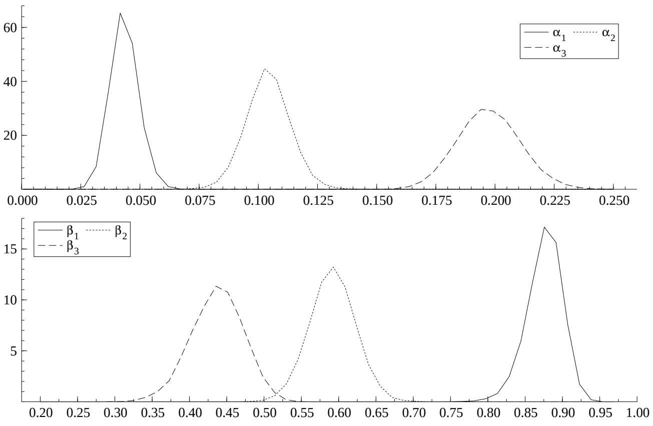


Figure 2: Posterior marginals of the elements of θ_g $(G = 3)$

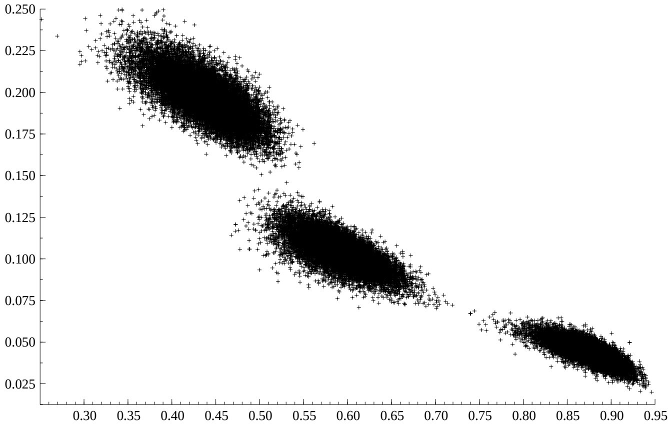


Figure 3: Scatterplot of the Gibbs draws of θ_g $(G = 3)$

Table 2: Hit Table $(G = 3)$

		Posterior group			
		1	2	3	
	1	19	6	0	25
Real group	2	0	45	5	50
	3	0	3	22	25
		19	54	27	100

The proportion of correct hits is 0.86.

cisely, we can use the posterior draws on S^J to identify, by some classification rule, the members of the three clusters. We propose a straightforward and simple classification rule: the data series belongs to the group to which it belongs most frequently a posteriori. For instance, of the 19000 draws of S^J in our example the last data series never belonged to the first group, 16 times to the second and hence 18984 times to the third group. As a consequence, the last data series is said to belong to group three which for this series is a correct classification. We applied this rule to all the 100 series and we summarize the classification results in Table 2 (we do not report the detailed results for the 100 S^J because of space limitations). We draw attention to two points. Firstly, when there is a misclassification this occurs only with the neighboring group. For instance the real third group was (wrongly) classified in the second group 3 times but it was never classified in the first group. Secondly, the total number of correct classifications, *i.e.* the sum of the diagonal elements of the 3 × 3 matrix in Table 2, amounts to 86 out of 100 which is a satisfactory result.

Table 3: 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	*-48028.65*	-48004.57	8	*-48032.20*
4	-48035.09	-48004.17	11	-48042.16
100	-48064.48	-47836.94	200	-48527.72

5.1.3 Results for incorrect numbers of groups

Next, we consider the case when the number of groups in the estimated model is wrong: we take four cases: one group, two groups, four groups, and one hundred groups. In the first case, all the series are considered as generated from the same GARCH(1,1) model, in the last case, they are considered to be all different, whereas they come from three different groups.

We report in Table 3 (second column) the value of the marginal likelihood for the different values of G. They were computed using formula (30), using the posterior mean as a high density point (using ML estimates, we obtain results that differ only in the decimals). Not surprisingly, the preferred model is the correct one. We can also use asymptotic model choice criteria, more easy to compute, to choose a preferred model. The Bayesian information criterion (BIC), see Schwarz (1978), 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 $\log(T)/2$ ($T = 1000$ in this example). Notice that the value of the maximized log-likelihood function increases with G since a model with given G embeds a model with smaller G.

In Tables 4, 5 and 6, we report the posterior results for different numbers of groups (comparable to Table 1, except that we do not report the correlation matrix of the group probabilities). The support of the prior uniform density were adjusted for each case. Obviously, for one group, we take as prior support for the GARCH parameters the union of the intervals for the case of three groups. For two groups we divide the prior support of one group for the α_g parameters in two intervals of equal length. In the case of four groups, we adjust the prior used for three groups by splitting the support of the parameters of the middle group in two pieces. The posterior results are not surprising. For one group, the posterior means of the GARCH parameters are roughly in the middle of the corresponding prior interval: the likelihood information forces a global homogeneity that has to be in the middle given the features of the DGP (50 series in the middle group, 25 in each of the other groups). For two groups, the series that belong to the middle group are forced to belong to one of the two outside groups: the posterior expected group probabilities are close to 0.5.. Hence, the posterior means are pulled toward the middle of the corresponding prior intervals. Notice how this increases the posterior correlation between α_1 and β_1 (-0.90) and to a lesser extent between α_2 and β_2 (-0.75), compared to the values for three groups. In the case of four groups, the middle group is split in two sub-groups, as is most clearly seen on the graph of the posterior densities of the GARCH parameters, see Figure 4. Notice how this artificially pulls toward zero the posterior correlations between α_2 and β_2 (-0.11), and between α_3 and β_3 (-0.04). The posterior results for the two outside groups are very much like in the case of three groups.

Finally, for 100 groups, we do not report the posterior results, but we show in Figure 5 the posterior densities of the 100 α and β GARCH parameters. We bet that for someone who does not know the DGP, it would not be clear that the DGP has three groups.

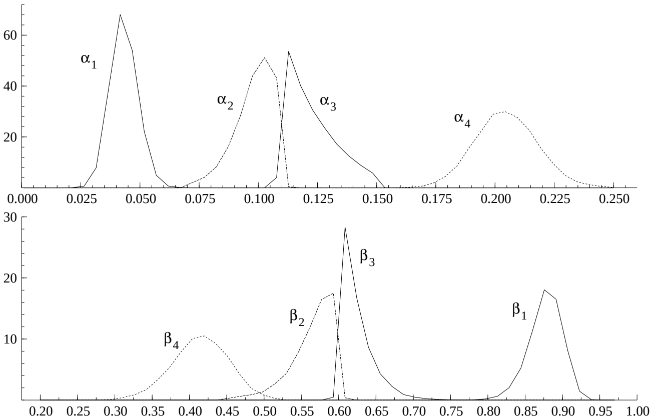


Figure 4: Posterior marginals of the elements of θ_g $(G = 4)$

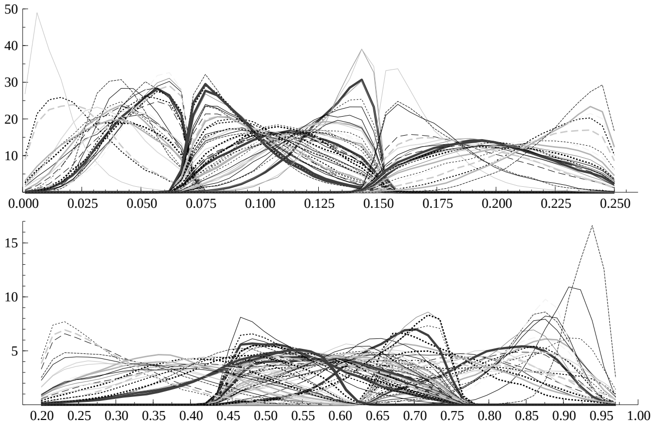


Figure 5: Posterior marginals of the elements of θ_g $(G = 100)$

Table 4: Posterior results on η and θ ($G = 1$)

Prior interval	α	0.001,0.25
	β	0.20,0.97
Mean	α	0.1137
	β	0.5979
Standard deviation	α	0.0069
	β	0.0264
Correlation α, β	-0.7263	

5.2 Many groups for one hundred series

In the next example we change the setting: the 100 data series are now drawn from a mixture with 25 components given by

$$\alpha_g = 0.06 + 0.01 \times (g - 1) \tag{35}$$

$$\beta_g = 0.88 - 0.02 \times (g - 1) \qquad g = 1, \ldots, 25. \tag{36}$$

The idea is to mimic a case where all the series are practically different, but not to a large extent. Hence it may be of interest, if only for practical reasons, to use a model with a small number of parameters, and we fix the number of groups to 3 for the inference. The choice of the prior is of particular importance in this setting because it determines which heterogeneous data series cluster together. The prior bounds on θ_g are given in Table 7. One can deduce from (35)-(36) that out of the 100 series, 32 series fall into the first group, 36 in the second and 32 in the third group. Therefore, this should be reflected in the posterior results on η, which is indeed the case as can be seen from the posterior results reported in Table 7.

We concentrate next on θ. Given that we did the inference as if there were only 3 components in the mixture, but in reality there are 25 of them, which posterior means should we expect? As we can see in (35) and (36) α_g and β_g are defined by using a fixed increment within a support. Given the prior bounds on θ_g, this implies that the posterior mean should not be too far away from the prior mean. Said differently, we expect that the posterior marginal densities are centrally located in the prior supports (see Figure 6). Posterior moments are reported in Table 7.

Table 5: Posterior results on η and θ ($G = 2$)

		η_1	η_2
Mean		0.5193	0.4807
Standard deviation		0.0918	0.0918
		$g = 1$	$g = 2$
Prior interval	α_g	0.01,0.125	0.125,0.22
	β_g	0.60,0.95	0.25,0.70
Mean	α_g	0.0659	0.1704
	β_g	0.7466	0.4750
Standard deviation	α_g	0.0086	0.0121
	β_g	0.0391	0.0285
Correlation α_g, β_g		-0.9014	-0.7589

Table 6: Posterior results on η and θ ($G = 4$)

		η_1	η_2	η_3	η_4
Mean		0.2097	0.3315	0.2181	0.2408
Standard deviation		0.0497	0.1166	0.1180	0.0602
		$g = 1$	$g = 2$	$g = 3$	$g = 4$
Prior interval	α_g	0.001,0.07	0.07,0.11	0.11,0.25	0.15,0.25
	β_g	0.65,0.97	0.45,0.60	0.60,0.75	0.20 ,0.60
Mean	α_g	0.0432	0.0982	0.1228	0.2047
	β_g	0.8772	0.5646	0.6240	0.4145
Standard deviation	α_g	0.0058	0.0081	0.0101	0.0131
	β_g	0.0223	0.0286	0.0220	0.0378
Correlation α_g, β_g		-0.7662	-0.1144	-0.0430	-0.7005

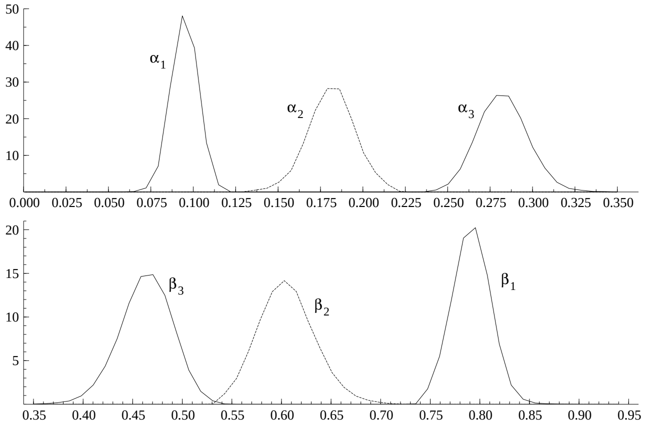


Figure 6: Posterior marginals of the elements of θ_g

Table 7: Posterior results on η and θ ($G = 3$)

		η_1	η_2	η_3
Mean		0.3423	0.3941	0.2637
Standard deviation		0.0675	0.0722	0.0690
		$g = 1$	$g = 2$	$g = 3$
Prior interval	α_g	0.001,0.13	0.13,0.22	0.22,0.35
	β_g	0.74,0.94	0.54,0.74	0.35,0.54
Mean	α_g	0.0945	0.1805	0.2824
	β_g	0.7922	0.6080	0.4629
Standard deviation	α_g	0.0079	0.0130	0.0144
	β_g	0.0189	0.0260	0.0260
Correlation α_g, β_g		-0.8708	-0.8392	-0.8211

6 Application to US stocks

We work with the returns on 131 stocks belonging to the biggest US companies. Each stock is observed from 29/09/99 to 30/07/03 implying 1000 observations each. Table 8 presents a summary of the descriptive statistics of all the series, which are given in Table 11. This table shows that there is a lot of 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. Hence, we expect also quite some heterogeneity in the estimates of GARCH(1,1) models for each series, which are also presented in Table 11. The overall reason for this data heterogeneity may be that individual stocks react differently to general news and specific company announcements.

Table 8: Descriptive statistics

	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, which can be found in Table 11.

Table 9: Marginal log-likelihood for application

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

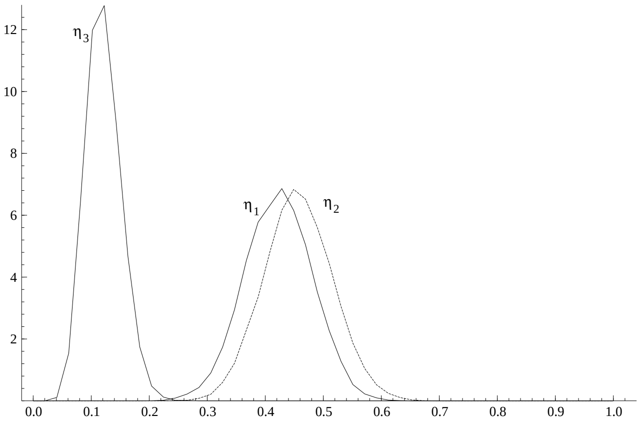


Figure 7: Posterior marginals of η_g $(G = 3)$

To select the number of groups, we allow a priori G to take the values, 1, 2, 3, and 132. Table 9 reports the corresponding values of the marginal log-likelihood. We come to the conclusion that the appropriate number of groups is three. We therefore report the results for three groups, based on 20000 draws from the MCMC sampler described in Section 3, out of which we dropped the first 1000. We do not report the values of the maximized log-likelihood because we were unable to obtain the convergence of the algorithm for ML estimation for $G = 2$ and $G = 3$.

The posterior means of η and θ can be found in Table 10. The posterior marginals of η are given in Figure 7, and those of the GARCH parameters α_g and β_g are in Figure 8.

The densities of η_1 and η_2 are quite similar and centered around 0.45. This forces the density of η_3 to be more concentrated on 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.

The prior intervals on α_g and β_g were chosen after some initial experiments to avoid too much zero mass in the densities (otherwise the numerical integrals in the griddy Gibbs sampler are wasting a lot of points). The posterior means of β_g are markedly different from each other. Compared to β_1 and β_2 the posterior standard deviation of β_3 is rather large. With respect to α_g we can see that α_2 and α_3 are close to each other. This does not imply that we should merge groups two and three. For example the persistence $\alpha_g + \beta_g$, 0.96 and 0.79 respectively, is clearly different between these groups. The high persistence for the first group, *i.e.* 0.99 , is forcing the correlation between α_1 and β_1 to be close to minus one. This is much less the case for group three.

Table 10: Posterior results on η and θ ($G = 3$)

		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

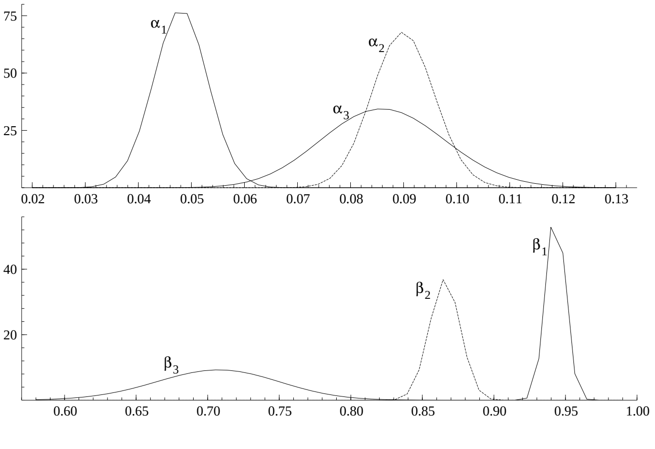


Figure 8: Posterior marginals of the elements of θ_g ($G = 3$)

We investigate the whole shape of the posterior marginal densities of α_g and β_g. The range of the density of α_3 covers that of the density of α_2 but the standard deviation of the former is almost twice as high. One might think that this causes identification problems. This is unlikely to be true because the posterior marginal densities of β_g are clearly separated. Notice the difference between α_3 and β_3, and the other groups. The densities for this group are still unimodal and we consequently do not find it necessary to split it up and to add a fourth group.

Finally, we can use the same simple classification rule as in Section 5 that a data series belongs to the group to which it belongs most frequently a posteriori. After applying this rule using the realisations of the group indicator S^J simulated by the algorithm, we find that 56 series belong to the first group, 60 series to the second and 15 series to the third group. In the last but one column of Table 8, one finds the posterior probability that each series belongs to its group, indicated in the last column. A large majority of the series, actually 93 (i.e. 71 percent), have a probability larger than 0.9 to belong to their group, while only 8 series (6 percent) have a probability less than 0.6 to belong to their group. According to this rule, 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.

The question may be asked 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 have not collected). 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.

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 0.99), 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 using 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 to use the griddy-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 (griddy-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. *Fourthly*, the clustering idea can also be used to identify clusters of pairs of series with similar correlation dynamics. *Fifthly*, 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.

Table 11: Descriptive statistics

j	stock	mean	std	min	max	skew	kurt	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_j$	g
1	3M Co	0,0393	1,82	-6,89	10,51	0,51	5,67	0.08	0.87	0.99	2
2	Abbott Labs	0,0046	2,21	-17,60	11,75	-0,39	8,85	0.04	0.94	1.00	1
3	Aflac	0,0423	2,29	-12,22	14,91	0,50	8,26	0.20	0.39	1.00	3
4	Alcoa	-0,0125	2,67	-11,66	12,22	0,10	4,62	0.04	0.94	0.97	1
5	Alltel	-0,0386	2,10	-12,54	11,78	-0,10	6,03	0.03	0.96	0.99	1
6	Altria Gp	0,0150	2,48	-14,90	15,06	-0,27	8,80	0.05	0.94	1.00	1
7	American Express	-0,0004	2,62	-14,61	10,44	-0,16	4,59	0.07	0.87	0.99	2
8	Anadarko Ptl	0,0320	2,53	-11,46	10,31	-0,04	4,43	0.04	0.96	1.00	1
9	Analog Devices	0,0302	4,67	-13,93	20,66	0,33	3,76	0.05	0.94	0.97	1
10	Anheuser-Busch	0,0398	1,85	-8,60	7,43	-0,29	5,46	0.09	0.89	0.83	1
11	AOL Time Warner	-0,1281	3,65	-18,79	14,88	-0,22	5,49	0.08	0.88	0.99	2
12	Applied Mats	-0,0037	4,36	-15,10	22,81	0,39	4,26	0.06	0.91	0.57	2
13	Atandt	-0,1057	3,12	-21,17	20,84	0,14	8,32	0.09	0.65	1.00	3
14	Avon Products	0,0878	2,29	-11,06	17,57	0,95	9,94	0.08	0.91	1.00	1
15	Baker Hughes	0,0080	2,98	-15,60	17,10	0,04	5,49	0.05	0.95	0.99	1
16	Bank of America	0,0420	2,24	-9,06	7,98	0,03	4,48	0.09	0.90	0.78	1
17	Bank of New York	-0,0099	2,68	-16,85	14,99	-0,05	7,09	0.09	0.82	0.93	2
18	Bank One	0,0132	2,25	-11,40	12,01	0,19	5,91	0.08	0.89	0.95	2
19	BB&T	0,0134	1,84	-8,17	10,31	0,17	6,12	0.11	0.87	0.55	2
20	Bellsouth	-0,0588	2,44	-19,98	10,90	-0,57	9,71	0.05	0.94	0.86	1
21	Boeing	-0,0260	2,47	-19,39	8,59	-0,58	7,43	0.08	0.87	1.00	2
22	Boston Scientific	0,0985	2,93	-32,74	15,11	-1,08	20,24	0.15	0.73	0.98	2
23	Bristol Myers	-0,0855	2,66	-25,38	13,67	-1,17	15,14	0.07	0.91	0.66	2
24	Cardinal Health	0,0593	2,34	-17,13	11,74	-0,46	8,49	0.12	0.78	0.99	2
25	Caterpillar	0,0196	2,27	-12,86	8,03	0,01	4,61	0.04	0.89	0.99	3
26	Cendant	0,0032	3,24	-20,97	31,38	0,85	14,84	0.09	0.85	0.99	2
27	Charles Schwab	-0,0811	4,09	-21,14	23,25	0,44	5,49	0.14	0.50	1.00	3
28	Chevron Texaco	-0,0206	1,63	-6,92	9,04	0,05	5,01	0.08	0.89	0.95	2
29	Cisco Systems	-0,0550	4,05	-14,07	21,82	0,32	5,70	0.07	0.91	0.83	1
30	Citigroup	0,0415	2,47	-17,11	11,90	-0,21	6,78	0.07	0.89	0.96	2
31	Clear Chl Comms	-0,0655	3,16	-18,03	13,70	-0,40	6,12	0.08	0.88	1.00	2
32	Coca Cola	-0,0074	1,99	-10,60	9,20	-0,02	5,94	0.03	0.96	1.00	1
33	Colgate Palmolive	0,0180	2,00	-17,33	18,50	0,08	17,26	0.11	0.88	0.80	1
34	Comcast	-0,0211	3,16	-15,21	13,58	0,15	5,04	0.05	0.95	1.00	1
35	Conagra	0,0008	1,91	-21,70	9,30	-1,47	21,52	0.03	0.96	1.00	1
36	Conocophillips	0,0070	1,75	-8,58	9,91	-0,13	5,17	0.06	0.91	0.62	1
37	CVS	-0,0333	2,83	-26,13	16,73	-0,79	14,56	0.08	0.73	1.00	3
38	Deere & Co	0,0237	2,36	-11,82	14,87	0,35	6,37	0.07	0.86	0.81	2
39	Dominion Res	0,0304	1,82	-13,68	8,38	-1,16	11,74	0.24	0.60	1.00	2
40	Dow Chemicals	-0,0050	2,56	-11,18	10,77	0,09	4,98	0.09	0.82	0.97	2
41	Duke Energy	-0,0446	2,45	-16,14	14,98	-0,24	8,55	0.14	0.73	0.99	1

j	stock	mean	std	min	max	skew	kurt	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_j$	g
42	Du Pont	-0,0321	2,17	-11,70	9,41	0,16	5,50	0.07	0.89	0.98	2
43	EMC Mass	-0,1215	4,75	-32,95	22,20	-0,40	7,22	0.04	0.95	0.97	1
44	Emerson Electric	-0,0186	2,12	-14,86	8,95	-0,17	5,92	0.10	0.65	1.00	3
45	Exelon	0,0456	1,95	-12,55	8,66	-0,25	6,66	0.16	0.75	1.00	2
46	Exxon Mobil	-0,0057	1,76	-8,84	10,48	0,19	6,12	0.07	0.91	0.65	2
47	Fannie Mae	0,0037	2,06	-7,12	9,11	0,32	4,92	0.06	0.93	1.00	1
48	Fifth Third Bancorp	0,0337	2,09	-8,53	10,62	0,26	4,49	0.12	0.85	0.81	2
49	First Data	0,0576	2,40	-8,53	14,21	0,28	5,62	0.12	0.83	1.00	2
50	Fleetboston Finl	-0,0123	2,62	-11,18	11,68	0,32	5,01	0.09	0.87	0.99	2
51	Ford Motor	-0,0937	2,78	-15,89	14,51	0,21	6,66	0.12	0.72	0.60	3
52	Forest Labs	0,1494	2,73	-26,90	14,19	-0,85	13,86	0.02	0.98	1.00	1
53	Fpl Group	0,0211	1,68	-9,03	8,74	-0,25	6,87	0.15	0.73	0.99	2
54	Freddie Mac	-0,0043	2,18	-17,50	11,14	0,00	8,99	0.05	0.90	0.96	2
55	Gannett	0,0112	1,66	-8,44	6,69	0,02	4,31	0.07	0.88	0.73	2
56	General Dynamics	0,0238	2,14	-13,21	8,73	-0,36	6,39	0.10	0.73	1.00	3
57	General Eelectric	-0,0328	2,34	-11,29	11,74	0,07	5,27	0.07	0.85	0.66	3
58	General Motors	-0,0529	2,43	-14,54	9,84	-0,16	5,49	0.06	0.90	0.83	2
59	Gillette	-0,0068	2,03	-9,02	14,97	0,35	7,72	0.04	0.95	0.99	1
60	Golden West Finl	0,0954	2,00	-11,02	12,02	0,08	6,41	0.07	0.92	0.95	1
61	Harley-Davidson	0,0615	2,39	-9,11	11,27	0,21	4,79	0.07	0.86	0.95	2
62	Heinz HJ	-0,0102	1,76	-8,60	13,51	0,48	9,44	0.16	0.77	0.98	2
63	Hewlett Packard	-0,0508	3,54	-20,70	19,01	0,04	6,53	0.02	0.96	0.55	1
64	Home Depot	-0,0356	2,99	-33,87	12,14	-1,47	20,53	0.13	0.77	0.94	2
65	Honeywell Intl	-0,0737	3,05	-19,57	25,38	-0,22	12,60	0.24	0.46	1.00	3
66	IBM	-0,0394	2,58	-16,89	12,26	-0,16	8,10	0.05	0.93	0.92	1
67	Illinois Tool Wks	-0,0058	2,04	-9,03	10,04	0,29	5,23	0.08	0.85	0.92	2
68	Intel	-0,0428	3,79	-24,88	18,34	-0,38	6,95	0.09	0.86	1.00	2
69	International Paper	-0,0212	2,35	-11,00	11,24	0,35	5,04	0.06	0.92	0.78	1
70	Johnson & Johnson	0,0011	2,07	-18,63	10,80	-0,89	13,12	0.10	0.83	1.00	2
71	Kellogg	-0,0082	2,11	-9,69	10,29	0,39	5,58	0.06	0.93	0.85	1
72	Keycorp	0,0055	2,08	-8,27	10,79	0,04	5,05	0.06	0.92	0.98	1
73	Kimberly-Clark	-0,0097	1,92	-11,55	10,08	-0,20	8,33	0.07	0.91	0.72	2
74	Kohls	0,0558	2,73	-10,59	10,51	0,19	4,29	0.08	0.89	0.87	2
75	Kroger	-0,0322	2,61	-28,25	9,46	-1,52	18,82	0.12	0.82	1.00	2
76	Linear Tech	0,0142	4,41	-14,69	16,35	0,30	3,43	0.07	0.92	0.66	1
77	Lowe's Cos	0,0697	2,82	-11,57	16,94	0,33	5,36	0.04	0.95	1.00	1
78	Marsh & Mclennan	0,0377	2,29	-13,50	12,88	0,25	6,63	0.10	0.86	1.00	2
79	Maxim Integ. Prod.	0,0126	4,52	-30,31	20,89	0,12	5,89	0.07	0.92	0.66	1
80	Mbna Corp	0,0386	2,98	-14,76	19,19	0,14	6,68	0.09	0.87	1.00	2
81	Mcgraw-Hill Co	0,0254	1,99	-11,93	13,40	0,32	7,52	0.07	0.91	0.70	2
82	Medtronic	0,0377	2,17	-9,05	10,60	-0,07	4,77	0.06	0.93	0.97	1
83	Mellon Finl	-0,0082	2,41	-10,55	10,15	0,10	4,51	0.09	0.87	0.99	2
84	Merck	-0,0148	2,03	-9,86	9,16	0,10	5,38	0.06	0.88	0.57	3

j	stock	mean	std	min	max	skew	kurt	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_j$	g
85	Merrilllynch	0,0468	2,88	-12,17	11,06	0,11	4,09	0.05	0.91	0.48	1
86	Microsoft	-0,0534	2,80	-16,97	17,86	-0,11	7,79	0.08	0.89	0.98	2
87	Motorola	-0,1185	3,95	-26,24	17,53	-0,60	8,59	0.08	0.88	0.98	2
88	National City	0,0223	1,99	-9,48	9,44	-0,03	5,22	0.08	0.91	0.97	1
89	Nextel Comms A	-0,0584	6,00	-33,44	28,60	0,04	6,82	0.05	0.95	1.00	1
90	Northrop Grumman	0,0434	2,17	-12,50	14,58	0,01	7,19	0.04	0.87	1.00	3
91	Omnicom	-0,0043	2,68	-21,94	12,13	-0,51	9,34	0.10	0.87	1.00	2
92	Oracle	0,0093	4,35	-23,63	19,31	0,09	5,24	0.05	0.95	0.99	1
93	Paychex	0,0351	3,08	-14,14	13,35	-0,01	4,54	0.05	0.93	0.90	1
94	Pepsico	0,0402	1,86	-10,73	13,86	0,33	8,59	0.05	0.94	0.94	1
95	Pfizer	-0,0022	2,17	-11,23	6,93	-0,19	4,60	0.08	0.90	0.76	2
96	Pnc Finl. Svs	-0,0040	2,22	-16,05	11,65	-0,30	7,69	0.10	0.86	1.00	2
97	Procter & Gamble	-0,0080	2,20	-37,66	9,09	-5,20	90,36	0.04	0.96	1.00	1
98	Progressive Corp	0,0854	2,43	-21,36	18,99	0,05	15,75	0.03	0.97	1.00	1
99	Qualcomm	-0,0229	4,63	-18,45	27,01	0,24	5,10	0.03	0.97	0.90	1
100	Raytheon New	-0,0455	3,45	-57,28	23,71	-4,78	85,19	0.13	0.63	1.00	3
101	Royal Dutch	-0,0303	1,84	-9,69	5,99	-0,48	5,17	0.10	0.85	1.00	2
102	Safeway	-0,0730	2,53	-19,06	12,66	-0,64	9,10	0.04	0.94	0.50	1
103	Sara Lee	-0,0192	1,89	-10,35	12,32	0,18	6,67	0.04	0.93	0.83	1
104	Sbc Communications	-0,0750	2,52	-13,54	8,85	-0,03	4,74	0.05	0.92	0.59	1
105	Schering-Plough	-0,0885	2,65	-15,82	11,14	-0,24	5,67	0.24	0.45	0.50	2
106	SLM	0,1058	1,96	-9,08	8,78	0,05	5,10	0.04	0.96	1.00	1
107	Southern	0,0615	1,72	-8,85	8,78	-0,02	5,66	0.16	0.79	0.92	2
108	Sprint	-0,1321	3,30	-24,42	18,82	-0,44	9,17	0.21	0.67	1.00	2
109	Statestreet	0,0353	2,57	-12,11	16,43	0,10	6,24	0.10	0.85	1.00	2
110	Stryker	0,1103	2,27	-19,26	18,41	0,05	13,13	0.04	0.96	1.00	1
111	Sun Microsystems	-0,1812	4,69	-31,09	26,02	-0,13	6,59	0.04	0.94	0.86	1
112	Suntrust Banks	-0,0063	1,92	-9,49	9,44	-0,04	5,70	0.07	0.92	0.86	1
113	Sysco	0,0548	1,92	-8,52	12,38	0,18	6,44	0.15	0.77	1.00	2
114	Target	0,0270	2,76	-11,19	12,23	0,07	4,77	0.04	0.95	1.00	1
115	Texas Instruments	-0,0811	4,20	-20,12	21,55	0,29	4,38	0.06	0.91	0.55	2
116	Tribune	-0,0019	2,09	-18,82	9,30	-0,71	11,34	0.13	0.83	1.00	2
117	Union Pacific	0,0233	1,92	-7,28	6,42	-0,06	3,94	0.05	0.93	0.92	1
118	United Health Gp	0,1286	2,25	-21,77	11,10	-1,10	13,35	0.11	0.86	0.53	2
119	United Technologies	0,0278	2,54	-33,20	9,38	-2,28	32,48	0.03	0.81	1.00	3
120	US Bancorp Del.	0,0024	2,53	-17,42	14,06	-0,15	7,51	0.11	0.85	1.00	2

j	stock	mean	std	min	max	skew	kurt	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_j$	g
121	Verizon Comms	-0,0625	2,32	-12,61	11,57	0,08	5,89	0.05	0.93	0.93	1
122	Viacomb	0,0018	2,97	-13,90	15,68	0,22	5,16	0.06	0.91	0.60	2
123	Wachovia	0,0208	2,23	-9,12	8,37	0,01	4,50	0.11	0.87	1.00	2
124	Walgreen	0,0198	2,15	-9,69	8,95	-0,14	4,69	0.03	0.96	0.94	1
125	Walmart	0,0203	2,37	-9,24	9,02	0,23	4,71	0.04	0.96	1.00	1
126	Walt Disney	-0,0189	2,74	-20,29	14,20	-0,42	8,65	0.04	0.84	1.00	3
127	Washington Mutual	0,0775	2,23	-11,68	11,54	0,06	5,44	0.15	0.78	1.00	2
128	Wasteman	0,0255	2,45	-14,23	23,32	0,65	13,50	0.04	0.95	1.00	1
129	Wellsfargo & Co	0,0286	1,90	-9,20	9,53	0,08	5,57	0.10	0.88	0.71	2
130	Weyerhaeuser	0,0003	2,32	-12,72	11,11	0,13	5,22	0.05	0.93	0.87	1
131	Wyeth	0,0097	2,73	-27,77	11,47	-1,34	18,70	0.06	0.93	1.00	1

Column 2 indicates the names of the firms in our sample.

Columns 3 to 8 report usual descriptive statistics of the returns (mean, standard deviation, minimum, maximum, skewness coefficient, kurtosis coefficient).

Columns $\hat{\alpha}$ and $\hat{\beta}$ report the QML estimates of the GARCH(1,1) parameters of $h_t = (1-\alpha-\beta)\tilde{\omega} + \alpha y_{t-1}^2 + \beta h_{t-1}$ for each series, when $\tilde{\omega}$ is fixed at the unconditional variance of the data.

Column $\hat{p}_j$ gives the probability that series j belongs to the group indicated in the last column for the model with three groups. The probability is estimated by the relative frequency of generated S_j (in the Gibbs sample) equal to the value indicated in the last column.

Appendix 1: The multinomial process

The multinomial process is basically a generalization of the binomial process. Let E_i $(i = 1, \ldots, G)$ be a partition of the sample space E. Consider an experiment whose outcomes must belong to one of the E_i's. This experiment is described by a vector $\theta = (\theta_1, \ldots, \theta_G)'$ where $\theta_i = P(\omega \in E_i) \geq 0$ and $\sum_{i=1}^{G} \theta_i = 1$.

Consider now n independent repetitions of the same experiment and let X_i be the number of outcomes that belong to E_i. Then the vector $X = (X_1, \ldots, X_G)'$ has a multinomial distribution with parameter (n, θ) and we write $X \sim M(n, \theta)$.

Characteristics

1. Probability distribution :

$$P(X = x|n, \theta) = \frac{n!}{\prod_{i=1}^{G} x_i!} \prod_{i=1}^{G} \theta_i^{x_i} \tag{37}$$

2. Sample space

$$\mathcal{S}_n = \left\{ x \in N^n | \sum_{i=1}^{G} x_i = n \right\} \tag{38}$$

3. Parameter space

$$\mathcal{S}_G = \left\{ \theta \in R^G \mid \theta_i \geq 0 \quad i = 1, \ldots, G \quad \text{and} \quad \sum_{i=1}^{G} \theta_i = 1 \right\} \tag{39}$$

4. First two moments

$$\begin{aligned} E(X_i|n, \theta) &= n\,\theta_i \\ V(X_i|n, \theta) &= n\,\theta_i\,(1 - \theta_i) \\ cov(X_i, X_j|n, \theta) &= -n\,\theta_i\,\theta_j \end{aligned} \tag{40}$$

To reconcile with the S_g's introduced in Section 2, see Table 12, where y_{gj} is an indicator variable taking the value 1 with probability η_g. It is easy to see that

$$\begin{aligned} P(S_j = g) &= P(y_{1j} = 0, y_{2j} = 0, \ldots, y_{gj} = 1, \ldots, y_{Gj} = 0) \\ &= \eta_1^{y_{1j}}\,\eta_2^{y_{2j}} \ldots \eta_g^{y_{gj}} \ldots \eta_G^{y_{Gj}} \\ &= \eta_g. \end{aligned} \tag{41}$$

	Group indicator				
group	S_1	S_1	$\dots$	S_J	
1	y_{11}	y_{12}	$\dots$	y_{1J}	$X_1 = \sum_{j=1}^J y_{1j} = x_1$
2	y_{21}	y_{22}	$\dots$	y_{2J}	$X_2 = \sum_{j=1}^J y_{2j} = x_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	y_{G1}	y_{G2}	$\dots$	y_{GJ}	$X_G = \sum_{j=1}^J y_{Gj} = x_G$
					$\sum_{g=1}^G x_g = J$

Table 12: Link between S_g and X_g

Appendix 2: The Dirichlet distribution

For fixed n the Dirichlet distribution is the conjugate prior of the parameters of the multinomial distribution $M(n, \theta)$.

Characteristics

1. Density function :

$$f_{Di}(\theta|\, a) = \frac{\Gamma(A)}{\prod_{i=1}^G \Gamma(a_i)} \prod_{i=1}^G \theta_i^{a_i-1} \; \mathbb{1}_{\mathcal{S}_G}(\theta) \tag{42}$$

   where $a = (a_1, \dots, a_G)$ is the parameter of the Dirichlet distribution such that $a_i > 0$ $(i = 1, \dots, G)$ and A is defined as $A = \sum_{i=1}^G a_i$. We write $\theta \sim Di(a)$.

   Note that the density function, given that θ is a parameter of the multinomial distribution, changes under the restriction $\sum_{i=1}^G \theta_i = 1$, for example like

$$\frac{\Gamma(A)}{\prod_{i=1}^G \Gamma(a_i)} \theta_1^{a_1-1} \; \theta_2^{a_2-1} \dots \theta_{G-1}^{a_{G-1}-1} \; (1 - \theta_1 - \theta_2 - \dots - \theta_{G-1})^{a_G-1} \; \mathbb{1}_{\mathcal{S}_G}(\theta_1, \dots, \theta_{G-1}) \tag{43}$$

2. Sample space

$$\mathcal{S}_G = \left\{ (\theta_1, \dots, \theta_{G-1}) \in R^{G-1} \mid \theta_i \geq 0 \quad i = 1, \dots, G-1 \quad \text{and} \quad \sum_{i=1}^{G-1} \theta_i \leq 1 \right\} \tag{44}$$

3. First two moments

$$\begin{aligned} E(\theta_i | a) &= \frac{a_i}{A} \\ V(\theta_i | a) &= \frac{a_i \; (A - a_i)}{A^2 (A+1)} \end{aligned}$$

$$cov(\theta_i, \theta_j | a) = -\frac{a_i a_j}{A^2(A+1)} \tag{45}$$

4. When $G = 2$ then the Dirichlet distribution is a beta distribution. Let us give the classic coin tossing example. The likelihood for J coin tosses (Bernouilli trials) with k heads is

   $$p(k|\theta, J) = C_J^k \theta^k (1-\theta)^{J-k} \tag{46}$$

   from which we compute directly that the maximum likehood estimator is $\hat{\theta} = k/J$. We put a prior distribution over the parameter θ and are interested in

   $$p(\theta|k) = \frac{p(k|\theta)p(\theta)}{p(k)} \tag{47}$$

   where

   $$p(\theta) = C(\alpha_1, \alpha_2)\theta^{\alpha_1 - 1}\,(1-\theta)^{\alpha_2 - 1} \tag{48}$$

   and

   $$C(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) + \Gamma(\alpha_2)}. \tag{49}$$

   We can reparametrize the α's as follows

   $$\begin{aligned} \alpha_1 &= p\,S + 1 \\ \alpha_2 &= (1-p)\,S + 1 \end{aligned} \tag{50}$$

   so that (47) becomes

   $$p(\theta|k) = \frac{C_J^k C(\alpha_1, \alpha_2)\theta^{k+p\,S}(1-\theta)^{(J-k)+(1-p)S}}{p(k)} \tag{51}$$

   or

   $$p(\theta|k) \propto \theta^{k+p\,S}(1-\theta)^{(J-k)+(1-p)S}. \tag{52}$$

   Deriving with respect to θ and solving yields the posterior mode $\theta^* = \frac{k+p\,S}{J+S}$.

   Notice that $p(\theta|k)$ is a beta distribution with parameters $k+p\,S+1$ and $(J-k)+(1-p)S+1$. Its mean is

   $$\frac{k+p\,S+1}{J+S+2}. \tag{53}$$

This is not equal to the posterior mode.

So, when $G = 2$, we have to sample from a beta distribution which is done by sampling independently $y_1 \sim G(k + p\,S + 1, 1)$, $y_2 \sim G((J - k) + (1 - p)S + 1, 1)$ where G means the gamma distribution and taking $\frac{y_1}{y_1 + y_2}$ as the beta variate.

Posterior distribution

Let

$$\begin{aligned} X|\theta &\sim M(n, \theta) \qquad (54) \\ \theta &\sim Di(a) \qquad (55) \end{aligned}$$

then

$$\theta|X = x \sim Di(a^*) \qquad (56)$$

$a^* = a + x$. This may indeed be seen using Bayes theorem:

$$\varphi(\theta|x) \propto \prod_{g=1}^{G} \theta_g^{a_g + x_g - 1} \qquad (57)$$

Sampling from a Dirichlet distribution

Suppose that $X_1, \ldots, X_G$ are independent random variables having each a gamma distribution $G(a_g, 1)$, $g = 1, \ldots, G$ and let

$$\begin{aligned} \theta_i &= \frac{X_i}{X_1 + \ldots + X_G} \quad i = 1, \ldots, G - 1 \\ \theta_G &= 1 - \theta_1 - \theta_2 - \ldots - \theta_{G-1}. \end{aligned}$$

Then $(\theta_1, \ldots, \theta_G) \sim Di(a_1, \ldots, a_G)$. Other results about the Dirichlet distribution can be found in Wilks (1962).

Appendix 3: Marginal likelihood

This appendix focuses on the calculation of the marginal likelihood $m(\tilde{y}^g)$.

Deterministic integration

We only discuss the Simpson rule but notice that many other deterministic integration methods may be used, see Bauwens, Lubrano, and Richard (1999, chap. 3). The interest lies in the integral

$$\int_{\theta_L}^{\theta_U} h(\theta)\,\mathrm{d}\theta. \tag{58}$$

With $2n$ intervals of equal length $d = \theta_j - \theta_{j-1} = 1/2n$ based on $2n+1$ points $\theta_0 (= \theta_L), \theta_1, \ldots, \theta_{2n} (= \theta_U)$ one can approximate (58) by

$$(d/3)\{h(\theta_0) + \sum_{i=1}^{2n-1} (3 + (-1)^{i+1}) h(\theta_i) + h(\theta_{2n})\}. \tag{59}$$

The Simpson rule can be generalized for higher dimensions. For instance, in two dimensions we write

$$\int_{\theta_L}^{\theta_U} \int_{\xi_L}^{\xi_U} h(\theta, \xi)\,\mathrm{d}\theta\,\mathrm{d}\xi = \int_{\theta_L}^{\theta_U} \mathrm{d}\theta \int_{\xi_L}^{\xi_U} h(\theta, \xi)\,\mathrm{d}\xi. \tag{60}$$

In words, we integrate the function with respect to ξ for all the possible values of θ, implying many $(2n+1)$ one-dimensional integrals. The integral of the resulting one-dimensional function of θ yields the answer.

Laplace approximation

Let us define $\exp(h(\theta_g)) = f(\tilde{y}^g|\theta^g)\ \varphi(\theta_g)$ and $\theta = \theta_g$ for notational convenience. The Laplace approximation is based on a second order Taylor expansion of $h(\theta)$ around $\hat{\theta} = \arg\max \ln f(\tilde{y}^g|\theta)$

$$h(\theta) \approx h(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})'\ \frac{\partial^2 h(\theta)}{\partial\theta\,\partial\theta'}|_{\theta=\hat{\theta}}\ (\theta - \hat{\theta}). \tag{61}$$

Therefore the marginal likelihood can be computed as

$$\int \exp h(\theta) \mathrm{d}\theta \approx \exp(h(\hat{\theta}))\ \int \exp\left(\frac{1}{2}(\theta - \hat{\theta})'\ \frac{\partial^2 h(\theta)}{\partial\theta\,\partial\theta'}|_{\theta=\hat{\theta}}\ (\theta - \hat{\theta})\right) \mathrm{d}\theta. \tag{62}$$

or

$$m(\tilde{y}^g) = f(\tilde{y}^g|\hat{\theta})\ \varphi(\hat{\theta})\ (2\pi)^{k/2}\ |\ \Sigma(\hat{\theta})\ |^{1/2} \tag{63}$$

where k is the dimension of θ and

$$\Sigma(\hat{\theta}) = \left[-\frac{\partial^2 \ln f(\tilde{y}^g|\theta)\ \varphi(\theta)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}} \right]^{-1}. \tag{64}$$

Examples

We now present some marginal likelihood computation examples to compare the deterministic integration with the Laplace approximation. We consider four cases.

1. Univariate normal with known variance.

$$Y|\theta \quad \sim \quad N(\theta, 1) \tag{65}$$

$$\theta \quad \sim \quad N(0, 1) \tag{66}$$

We draw an *i.i.d.* sample from (65) with $\theta = 0$ resulting in $\{y_1, y_2, \ldots, y_n\}$. The likelihood and the prior density are

$$f(y|\theta) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\theta)^2}{2}\right) \tag{67}$$

$$\varphi(\theta) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\theta^2}{2}\right) \tag{68}$$

We want to calculate the marginal likelihood $\int f(y|\theta)\varphi(\theta)d\theta$ which can be done analytically by noticing that

$$f(y|\theta)\varphi(\theta) = (2\pi)^{-\frac{n+1}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\theta)^2 + \theta^2}{2}\right) \tag{69}$$

$$= (2\pi)^{-\frac{n+1}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\bar{y})^2}{2}\right) \exp\left(-\frac{n\bar{y}^2}{2(n+1)}\right) \times \tag{70}$$

$$\exp\left(-\frac{1}{2\frac{1}{n+1}}\left(\theta - \frac{n}{n+1}\bar{y}\right)^2\right) \tag{71}$$

which implies that

$$\int f(y|\theta)\varphi(\theta)d\theta = c\int \exp\left(-\frac{1}{2\frac{1}{n+1}}\left(\theta - \frac{n}{n+1}\bar{y}\right)^2\right) d\theta \tag{72}$$

$$= c(2\pi)^{\frac{1}{2}}(n+1)^{-\frac{1}{2}} \tag{73}$$

where c contains everything in (70).

The Laplace approximation in (63) results in

$$\begin{aligned}\int f(y|\theta)\varphi(\theta)d\theta &= f(y|\hat{\theta})\ \varphi(\hat{\theta})\ (2\pi)^{\frac{1}{2}}(n+1)^{-\frac{1}{2}} \qquad (74)\\ &= (2\pi)^{-\frac{n}{2}}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\hat{\theta})^2}{2}\right)(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{\hat{\theta}^2}{2}\right)(2\pi)^{\frac{1}{2}}(n+1)^{-\frac{1}{2}}\end{aligned}$$

which is exactly the same as (73) because $\hat{\theta} = \frac{n}{n+1}\bar{y}$. This comes as no suprise since the Laplace method approximates a quadratic function by a Taylor expansion of order two.

2. Beta distribution

$$Y|\theta \sim \text{beta}(\theta, 3) \qquad (75)$$

$$\theta \sim \text{constant} \qquad (76)$$

The likelihood is

$$f(y|\theta) = \left(\frac{\Gamma(\theta+3)}{\Gamma(\theta)\Gamma(3)}\right)^n \prod_{i=1}^{n}\left(y_i^{\theta-1}(1-y_i)^2\right). \qquad (77)$$

3. Product of two univariate normal distributions

$$Y_i|\theta_i \sim N(\theta_i, 1) \quad i=1,2 \qquad (78)$$

$$\theta_i \sim N(0,1) \quad i=1,2 \qquad (79)$$

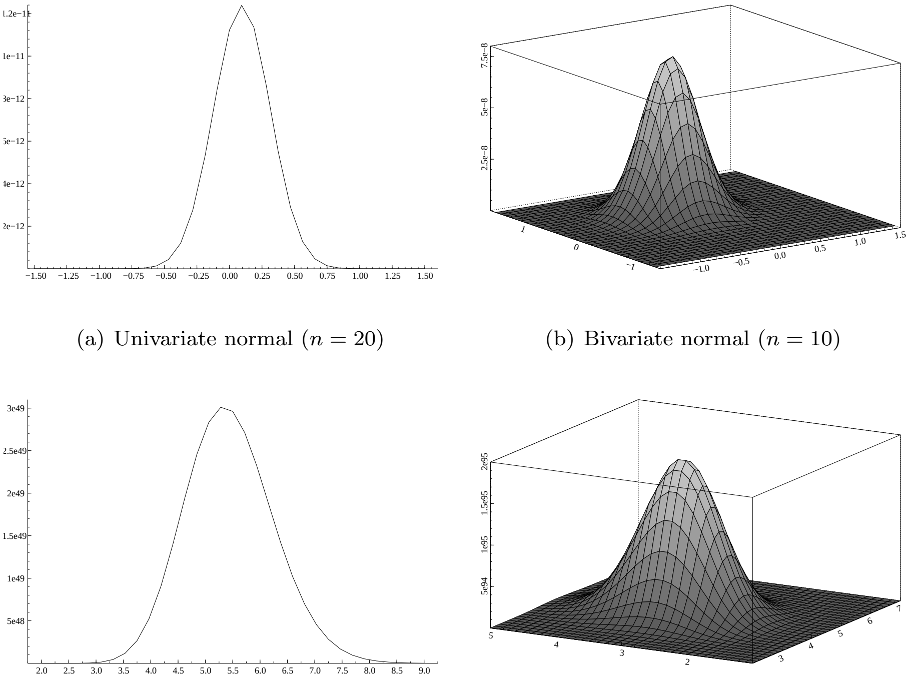
4. Product of two univariate beta distributions

$$Y_1|\theta_1 \sim \text{beta}(\theta_1, 3) \qquad (80)$$

$$Y_2|\theta_2 \sim \text{beta}(\theta_2, 3) \qquad (81)$$

$$\theta_i \sim \text{constant} \quad i=1,2 \qquad (82)$$

The posterior kernel for these examples are displayed in Figure 9. The marginal likelihoods by deterministic integration and the Laplace approximation are displayed in Table 13. Apparently, both techniques deliver almost the same results.



(a) Univariate normal ($n = 20$)

(b) Bivariate normal ($n = 10$)

(c) Univariate beta ($n = 20$)

(d) Bivariate beta ($n = 20$)

Figure 9: Posterior kernels

Table 13: Marginal likelihoods

	n	θ	Simpson (1)	Laplace (2)	(1)/(2)
Univariate normal	20	0	6.7755e-012	6.7755e-012	1
Bivariate normal	10	$(0\ \ 0)'$	4.4295e-008	4.42969e-008	0.99996
Univariate beta	20	3	6.0487e+049	6.04159e+049	1.0012
Bivariate beta	20	$(3\ \ 3)'$	3.8004e+095	3.79763e+095	1.0007

References

Bauwens, L., S. Laurent, and J. Rombouts (2003): "Multivariate GARCH Models: A Survey," CORE DP 2003/31.

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

Bollerslev, T. (1990): "Modeling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH model," *Review of Economics and Statistics*, 72, 498–505.

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.

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 Economics Statistics*, 20, 339–350.

Engle, R., and F. Kroner (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150.

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.

Kearney, C., and A. Patton (2000): "Multivariate GARCH Modelling of Exchange Rate Volatility Transmission in the European Monetary System," *Financial Review*, 41, 29–48.

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. Mengersen (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-regresive Conditional Heteroskedasticity model with time-varying correlations," *Journal of Business and Economic Statistics*, 20, 351–362.

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