Bayesian inference on GARCH models using the Gibbs sampler

LUC BAUWENS¹, MICHEL LUBRANO²

¹CORE, Université Catholique de Louvain
E-mail: bauwens@core.ucl.ac.be

²GREQM, EHESS-CNRS
E-mail: milu@ehess.cnrs-mrs.fr

Received: April 1997, in final form January 1998

Summary This paper explains how the Gibbs sampler can be used to perform Bayesian inference on GARCH models. Although the Gibbs sampler is usually based on the analytical knowledge of the full conditional posterior densities, such knowledge is not available in regression models with GARCH errors. We show that the Gibbs sampler can be combined with a unidimensional deterministic integration rule applied to each coordinate of the posterior density. The full conditional densities are evaluated and inverted numerically to obtain random draws of the joint posterior. The method is shown to be feasible and competitive compared with importance sampling and the Metropolis–Hastings algorithm. It is applied to estimate an asymmetric Student–GARCH model for the return on a stock exchange index, and to compute predictive option prices on the index. We prove, moreover, that a flat prior on the degrees of freedom parameter leads to an improper posterior density.

Keywords: Bayesian inference, GARCH, Gibbs sampler, Monte Carlo, Option pricing.

1. INTRODUCTION

Bayesian inference on ARCH models has been implemented using importance sampling—see Geweke (1989) or Kleibergen and van Dijk (1993)—and more recently using a Metropolis algorithm—see Geweke (1994). A drawback of these kinds of methods is that they require a careful choice of a good approximation to the posterior density. Tuning is often necessary, so the method cannot be ‘automatic’. Moreover, as they are purely based on a sampling approach, accurate graphs of the marginal densities require a large number of draws.

The Gibbs sampler—see Geweke (1995) for a survey—uses analytical properties of the posterior density as much as possible. Gibbs sampling is simple when the vector of parameters, say \( \theta \), can be split into \( \theta_1, \theta_2 \) and the analytical expressions of both conditional posterior densities are known. Sampling sequentially in each conditional posterior density produces a Markov chain which converges to a draw of the joint posterior density. In some cases the conditional posterior density of \( \theta_1 | \theta_2 \) is known, but that of \( \theta_2 | \theta_1 \) is not. So direct sampling in \( \psi(\theta_1 | \theta_2, y) \) is easy, but sampling in \( \psi(\theta_2 | \theta_1, y) \) is more complicated. The traditional solution is then to use a rejection technique for sampling in the latter density, for example the Metropolis–Hastings (MH) algorithm.
After the introduction of ARCH models by Engle (1982), many extensions and applications of this class of models were proposed in the literature. Bollerslev (1986) put forward the generalized ARCH (GARCH), which combines parsimony in the parameters and flexibility in the lag structure of the conditional variance. Engle and Bollerslev (1986) used a Student distribution instead of a normal to account better for the excess kurtosis found in financial data. Several non-linear generalizations of the conditional variance equation were proposed, in particular to allow for an asymmetric response of the conditional variance to a past shock. Surveys on ARCH model formulation and properties are provided by Bera and Higgins (1993), and Bollerslev et al. (1994). Bollerslev et al. (1992) survey the application of ARCH models in finance.

GARCH models correspond to a particular form of heteroskedasticity. A multiplicative heteroskedasticity model can be written as

\[
\begin{align*}
  y_t &= x_t' \beta + u_t \\
  u_t &\sim N(0, \sigma^2_t) \\
  \sigma^2_t &= h(z_t, \gamma).
\end{align*}
\]  

(1)

A Gibbs sampler can be combined with a rejection step in this model. As soon as \( z_t \) and \( \gamma \) are given, the posterior density of \( \beta \) is a Student density (assuming a flat prior). However, the posterior density of \( \gamma \) given \( \beta \) is not of a known form: a Metropolis algorithm for instance has to be used for sampling from this posterior. A GARCH model is complicated by the fact that \( z_t = (y_{t-1} - x'_{t-1} \beta)^2 \). So there is no separability between the regression parameters and the heteroskedastic parameters. The previous algorithm cannot be used.

The aim of this paper is to use the Griddy–Gibbs sampler of Ritter and Tanner (1992) in order to conduct Bayesian inference on GARCH models. In Section 3, we consider the issue of integrability of the posterior density of a simple GARCH model with Student errors, since it turns out that the degrees-of-freedom parameter of the Student density has a crucial role in this respect. In Section 2, we explain the Griddy–Gibbs algorithm and we illustrate its feasibility in GARCH models with Student errors. In Section 4, we compare the Griddy–Gibbs sampler with importance sampling and the MH algorithm. In Section 5, we estimate an asymmetric GARCH model for a series of returns on a stock market index and we compute predictive option prices. The last section concludes.

2. GARCH WITH STUDENT ERRORS: PRIOR DENSITIES AND POSTERIOR INTEGRABILITY

We wish to conduct Bayesian inference on a regression model with GARCH errors. To ease exposition, we consider a parsimonious model which is quite representative of GARCH models used in finance. It is a GARCH(1, 1) model with Student errors (called Student–GARCH in the following). As reported in Bollerslev et al. (1992), the Student density, with its fat tails, is more able than the normal to account for the excess kurtosis present in financial data. The model is

\[
\begin{align*}
  y_t &= \epsilon_t \sqrt{h_t}, \quad t = 1, 2, \ldots, T \\
  \epsilon_t | I_{t-1} &\sim \text{Student}(0, 1, \nu) \\
  h_t &= \omega + \alpha \epsilon_{t-1}^2 + \beta h_{t-1}.
\end{align*}
\]  

(2)

© Royal Economic Society 1998
The $\epsilon_t$ sequence is assumed independent. The distribution of $y_t$, given the past information $I_{t-1}$, is Student with mean zero, and variance $h_t\nu/(\nu - 2)$ (assuming $\nu > 2$). The initial variance $h_0$ is treated as a known constant. The parameters of the variance equation are restricted by

$$\omega \geq 0, \alpha > 0, \beta \geq 0.$$  \hfill (3)

to ensure the positivity of $h_t$. Other restrictions may be considered: for the $y_t$ process to be covariance stationary, one must impose that

$$0 \leq \frac{\nu}{\nu - 2} \alpha + \beta < 1 \quad (\text{assuming } \nu > 2).$$  \hfill (4)

As underlined by Nelson (1990), while not being covariance stationary, the $y_t$ process can still be strongly stationary since the condition for strong stationarity, which is $E\log(\alpha \epsilon_t^2 + \beta) < 0$, is less stringent than the condition for weak stationarity, which ensures the existence of the unconditional variance. A necessary condition for strong (and weak) stationarity is $\beta < 1$.

Let $\theta$ denote the parameter vector $(\omega, \alpha, \beta, \nu)$. For a sample of $T$ observations, the posterior density is

$$\varphi(\theta|y) \propto \varphi(\theta) l(\theta|y)$$  \hfill (5)

where the likelihood function is given by

$$l(\theta|y) \propto \prod_{t=1}^{T} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{y_t^2}{\nu h_t}\right]^{-\frac{\nu+1}{2}}.$$  \hfill (6)

The prior density $\varphi(\theta)$ should respect at least the positivity restrictions on the parameters, and the condition $\beta < 1$. One important issue of Bayesian inference is the integrability of the posterior density. If an integrable (or proper) prior is used, and the likelihood function is not pathological (e.g. infinite for some parameter values), the posterior is proper. However, non-integrability of the posterior may arise with an improper prior.

If we consider the behaviour of the likelihood function (5) with respect to the parameters $\omega$, $\alpha$, and $\beta$, we see that no pathology arises if every $h_t$ is strictly positive, as the Student density is finite and positive under this condition. A flat (uniform) prior for each of these parameters can be used on whatever region of the parameter space is considered adequate (depending on what other restrictions one wishes to impose). For the degrees-of-freedom parameter $\nu$, the conclusion is different.

**Theorem 1.** With a flat prior on $\nu$ on $(0, \infty)$, the posterior density (5) is not integrable.

**Proof.** Since the Student$(0, h_t, \nu)$ density tends to the $N(0, h_t)$ density when $\nu$ tends to infinity, $l(\theta|y)$ is $O(1)$ in $\nu$: the likelihood function (6) tends to

$$\prod_{t=1}^{T} (2\pi h_t)^{-\frac{1}{2}} \exp -\frac{y_t^2}{2h_t}$$  \hfill (7)

which is finite and strictly positive.

This problem is not due to the GARCH model but to the Student density: in (1) if we replace the $N(0, \sigma_t^2)$ distribution on $u_t$ by the Student$(0, \sigma_t^2, \nu)$, the same problem arises.\footnote{This explains why Geweke had to use an exponential prior density on $\nu$, and states that ‘uninformative prior distributions for $\nu$ can be more troublesome’ (Geweke 1993, p. S27).} Kleibergen

© Royal Economic Society 1998
and van Dijk (1993) report posterior results under a flat prior on $\nu$ for their analysis of the Student–GARCH model. This does not contradict our theoretical result: since they used an importance sampling algorithm for integrating the posterior, with a finite Monte Carlo sample size, they have implicitly truncated the region of integration (there is no arbitrarily large drawn value of $\nu$ from the importance function).

The conclusion is that sufficient prior information is needed on $\nu$ to force the posterior to tend to zero quickly enough at the tail, in order to be integrable. The prior should at least be $O(\nu^{1+d})$, for small positive $d$, at the right tail; for example $1/\nu^2$, an improper prior which is obtained by being flat on $1/\nu$. Unfortunately, this prior creates a problem at the left tail, since $l(\theta|y)/\nu^2 \to \infty$ when $\nu \to 0$. Therefore, one should in principle truncate this prior to the interval $(m, \infty)$ where $m$ is a small positive constant (0.001 for example).

A proper prior that behaves like the previous improper one at the right tail but is finite at $\nu = 0$ is

$$\psi(\nu) \propto (1 + \nu^2)^{-1} \quad \text{if } \nu > 0$$

It is the half-right side of a Cauchy centred at 0, and thus is proper. A third possible prior is a flat prior on $\nu$ on a finite range $(0, M)$. A fourth possible prior is an exponential density, as in Geweke (1993), with parameter chosen subjectively to fix the prior mean (and variance) of $\nu$. Other classes could be considered.

### 3. GRIDDY–GIBBS SAMPLER FOR GARCH MODELS

The Gibbs sampler applied to $\psi(\theta|y)$ produces, by simulation, a Markov chain $\{\theta^n\}_{n=1}^N$ of draws with equilibrium distribution $\psi(\theta|y)$. The draws can be used to estimate consistently characteristics of interest of the posterior (such as moments and marginal densities). See for example Casella and George (1992) for an intuitive explanation and Tierney (1994) for a detailed treatment.

Let us suppose that we can partition $\theta$ into $\theta_1$ and $\theta_2$ so that the posterior densities $\phi(\theta_1|\theta_2, y)$ and $\phi(\theta_2|\theta_1, y)$ have known analytical forms (e.g. Student densities) from which it is easy to draw random numbers. This ability to sample from conditional distributions is called ‘conjugacy’ by Carlin and Gelfand (1991) and is central to the Gibbs sampler. The next draw $\theta^n$ of the joint posterior density is generated using the previous draw $\theta^{n-1}$ by the following cycle:

$$\begin{align*}
\theta_1^n &\sim \phi(\theta_1|\theta_2^{n-1}, y) \\
\theta_2^n &\sim \phi(\theta_2|\theta_1^n, y)
\end{align*}$$

($x^n \sim \phi(x)$ means that $x^n$ is a draw from the distribution $\phi(x)$). The procedure is started using any value $\theta_2^0$ in the support of the posterior distribution. The influence of the starting conditions vanishes after a certain number of draws, say $ns$. If the algorithm converges, after $ns$ draws it produces random draws of the posterior density of $\theta$; the sample is, however, not independent because of the Markov chain structure.

© Royal Economic Society 1998
3.1. Griddy–Gibbs sampler

The previous algorithm cannot be applied to the GARCH model even if the error term is (conditionally) normal. To demonstrate this, let us use the simple ARCH model

\[
\begin{align*}
  y_t &= \mu + \epsilon_t \sqrt{h_t} \\
  \epsilon_t | I_{t-1} &\sim N(0, 1) \\
  h_t &= 1 + \alpha (y_{t-1} - \mu)^2.
\end{align*}
\]

There are two parameters, \( \mu \) for the regression function, and \( \alpha \) for the skedastic function. Let us assume a flat prior. The conditional posterior density \( \phi(\mu | \alpha, y) \) has a kernel given by

\[
\kappa(\mu | \alpha, y) = \prod_t [h_t(\alpha, \mu)]^{-1/2} \exp \left\{ -\frac{1}{2} \frac{(y_t - \mu)^2}{h_t(\alpha, \mu)} \right\}
\]

when \( \alpha \) is a given fixed value. The conditional posterior density \( \phi(\alpha | \mu, y) \) has exactly the same expression, but for given \( \mu \). If \( h_t \) were fixed, (11) would be a normal density. As \( h_t \) is a function of both \( \alpha \) and \( \mu \), the conditional posterior density of \( \mu \) contains \( h_t \) which is also a function of \( \mu \). Consequently, it cannot be a normal or any other well known density from which random numbers could be easily generated. There is no property of conjugacy.

However, the kernel of \( \phi(\mu | \alpha, y) \) (respectively \( \phi(\alpha | \mu, y) \)), conditionally on a previous draw of the conditioning parameter, can be evaluated over a grid of points. One can then compute the corresponding distribution function using a deterministic integration rule—see for example Davis and Rabinowitz (1975). Afterwards, one can generate a draw of \( \mu \) (respectively \( \alpha \)) by inversion of the distribution at a random value sampled uniformly in \([0, 1]\). This technique, called the Griddy–Gibbs sampler (hereafter GGS), was applied by Ritter and Tanner (1992) to bivariate posterior densities.

It is easy to incorporate conditioning, which is a variance reduction technique, in the algorithm. Conditioning means that to estimate \( E(\alpha | y) \), one uses \( \sum_{n=1}^N E(\alpha | \mu^n, y) / N \) instead of \( \sum_{n=1}^N \alpha^n / N \). Although both estimators are root-\(N\) consistent (as in all Monte Carlo techniques), the former usually has a smaller variance than the latter.\(^2\)

Conditioning is very useful in order to get accurate graphs of the marginal densities for a relatively small number of draws, as one computes \( \phi(\alpha_\ast | y) = \int \phi(\alpha_\ast | \mu, y) \phi(\mu | y) d\mu \), (where \( \alpha_\ast \) denotes one point of the grid of values of \( \alpha \)) by

\[
\phi(\alpha_\ast | y) \approx \sum_{n=1}^N \phi(\alpha_\ast | \mu^n, y) / N.
\]

In our experiments, precise graphs were obtained usually with \( N = 1000 \), whereas ten times as many draws were required to get good-looking graphs using smoothed histograms of the draws.

The Appendix provides a detailed description of the Griddy–Gibbs algorithm with conditioning.

\(^2\)When the draws are independent, this results from the property that the variance of the first estimator—\( \text{Var}(\alpha | y) / N \)—is larger than the variance of the second estimator—\( \text{Var}(E(\alpha | \mu, y)) / N \). When the draws are dependent, the variances of both estimators have to account for the autocorrelation structure of the draws (for the first estimator) or of the conditional expectations (for the second estimator). To quote Tierney (1994): ‘Whether this form of conditioning leads to a reduction in variance depends on the correlation structure.’ If the autocorrelation structure of the conditional expectations is close to that of the draws, conditioning reduces the variance.
Table 1. ML and posterior results with Griddy–Gibbs sampler for Student–Garch model (2) (simulated data). Sample of 150 simulated observations. True parameter values are given in the first column. Flat prior on $\omega$, $\alpha$, $\beta$, and half Cauchy prior (8) on $\nu$. (1) Maximum likelihood estimate with asymptotic standard errors. (2) Posterior means [standard deviations] computed as sample moments of draws. (3) Results computed by averaging conditional moments. Same draws as in (1). (4) CPU time on a Pentium Pro 200 (with GAUSS, version 3.2.30), including the 1000 discarded initial draws (except for ML).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ML (1)</th>
<th>GGS (2)</th>
<th>GGS (3)</th>
<th>GGS (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$ (0.1)</td>
<td>0.19 [0.10]</td>
<td>0.24 [0.09]</td>
<td>0.24 [0.09]</td>
<td>0.23 [0.09]</td>
</tr>
<tr>
<td>$\alpha$ (0.4)</td>
<td>0.24 [0.12]</td>
<td>0.30 [0.13]</td>
<td>0.30 [0.13]</td>
<td>0.30 [0.13]</td>
</tr>
<tr>
<td>$\beta$ (0.4)</td>
<td>0.38 [0.25]</td>
<td>0.28 [0.16]</td>
<td>0.28 [0.16]</td>
<td>0.29 [0.17]</td>
</tr>
<tr>
<td>$N$</td>
<td>1000</td>
<td>1000</td>
<td>10000</td>
<td></td>
</tr>
<tr>
<td>Time (4)</td>
<td>0.10 s</td>
<td>155 s</td>
<td>155 s</td>
<td>14.3 min</td>
</tr>
</tbody>
</table>

3.2. A pilot study for a simple Student-GARCH model

The applications of the Griddy–Gibbs sampler in Ritter and Tanner (1992) are interesting but mainly illustrative. If the posterior density is bivariate, usual deterministic integration rules can be used and are much more efficient. The method is really useful if it can deal with higher dimensional posterior densities (the model analysed in Section 5 has seven parameters).

To monitor in detail the performance of the algorithm, we tried it on an artificial sample of 150 observations generated from the simple Student–GARCH model (2) with $\omega = 0.1$, $\alpha = \beta = 0.4$, and $\nu = 5$. We dropped the first 1000 draws and computed the results with 1000 and also with 10 000 draws. In Table 1, we report maximum likelihood estimates (with asymptotic standard errors), and posterior means (with standard deviations). Figure 1 shows the marginal posterior densities. The ranges of the graphs are the intervals of integration of each parameter. This means that the prior is flat over the product of these intervals, except that for $\nu$, the half Cauchy prior (8) has been used on the interval $(0, 18)$.

In Figure 1, one can see the densities of the parameters $\omega$, $\alpha$, and $\beta$ after 1000 and 10 000 draws; for $\nu$, we do not report the density after 1000 draws, because it cannot be distinguished from the density after 10 000 draws (shown as the solid line on the figure). The other density of $\nu$ on the figure (see dotted line) is the posterior obtained when the prior is flat on the interval $(0, 30)$: the corresponding posterior mean is equal to 9.37, and the standard deviation to 5.09. The difference with respect to the results obtained using the half Cauchy prior is striking, and is a consequence of the relatively slow decrease of the right tail of the likelihood function when $\nu$ tends to infinity (see Section 2). In this respect, one should keep in mind that a sample size of 150 is rather small for a GARCH model.

3The posterior results for the other parameters are not much influenced by the change of the prior on $\nu$ (see Table 4, column 2, for moments). Indeed, the posterior correlations between $\nu$ and the other parameters are small (less than 0.20).
Bayesian inference on GARCH models using the Gibbs sampler

Figure 1. Posterior densities for Student–GARCH model (2) (simulated data).

This also explains the big differences between the estimation results and the true values (see Table 1).

Posterior moments of the parameters differ from the ML results, because the posterior densities are skewed to the right. On the graphs of the marginal posterior densities in Figure 1, one can check that the posterior modes are close to the ML estimates, except for $\beta$. In the case of $\beta$, the skewness is caused mainly by the positivity restriction, with two effects: the ML estimate is larger than the posterior mean, and the ML standard error is larger than the posterior standard deviation (because the ML evaluation of the standard error does not take the restriction into account).

It appears that the Griddy–Gibbs algorithm is successful in dealing with the skewness of the posterior. This is due to the fact that integration is done on a grid so that every direction can be explored in detail. The main cost of the method is of course the evaluation of the posterior density kernel: with 33 point$^4$ grids, 4 parameters, and 1000 draws, the algorithm requires

$^4$We also computed the results using grids of 17 and 65 points. Results with 33 point grids seem accurate enough.
132 000 functional evaluations. So it can be greedy in computational time. In Section 4, it is compared with two other methods: importance sampling and the MH algorithm.

3.3. Convergence diagnostics

In our pilot study, the algorithm seems to have converged since multiplying by ten the number of draws hardly changes the results: compare the last two columns of Table 1, and the graphs of the marginal densities after 1000 and 10 000 draws in Figure 1. Of course, the relative stability of the posterior results as the number of draws increases could be due to the fact that the Gibbs sampler is trapped for a long time in a small part of the parameter space. One can argue that if this were the case, univariate and bivariate histograms of the draws would reveal the trapping.\footnote{Marginal densities obtained by conditioning—see (12)—would not reveal the trapping: if the sampler is trapped in some region, it will average almost constant conditional densities, and thus yield a precise graph.} Figure 2 displays the scatter plot of the draws of $\alpha$ and $\beta$ after 100, 500, 1000 and 10 000 draws (always after the first 1000 discarded draws). This plot, and plots for the other parameter combinations (not reported), suggest clearly that no trapping has occurred.

Several convergence criteria have been proposed for Monte Carlo Markov chains (MCMCs). Gelman and Rubin (1992) introduce a diagnostic based on multiple chains (with very dispersed starting values), in order to check if the posterior results are not too sensitive to the starting

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scatter_plots.png}
\caption{Scatter plots of draws of $\alpha$ and $\beta$ for Student–Garch model (2) (simulated data).}
\end{figure}
Bayesian inference on GARCH models using the Gibbs sampler

values. This criterion is not very attractive as it requires to run many times the Griddy–Gibbs sampler, which is numerically demanding. Convergence criteria based on the examination of a single long run are thus preferable. Zellner and Min (1995) put forward three simple criteria which are useful only for the case of two blocks described in (9). Geweke (1992) provides a statistic that compares the estimate of a posterior mean from the first \( N_1 \) draws to the estimate from the last \( N_2 \) draws of the chain. The statistic is normally distributed if \( N \) is large and the chain has converged. A simple tool proposed by Yu and Mykland (1994) is the visual inspection of CUMSUM statistics. Robert (1995) reports that this criterion works relatively well. Suppose we have \( N \) draws of a MCMC. A standardized version of the statistic proposed by Yu and Mykland is,

\[
CS_t = \frac{\left(1 - \sum_{n=1}^{t} \theta^n - \mu_\theta\right)}{\sigma_\theta}, \text{ for } t = 50, 100, 150, \ldots, N
\]  

(13)

where \( \mu_\theta \) and \( \sigma_\theta \) are the empirical mean and standard deviation of the \( N \) draws. If the MCMC converges, the graph of \( CS_t \) against \( t \) should converge smoothly to zero. In contrast

© Royal Economic Society 1998
Table 2. Sample sizes for convergence of the posterior mean estimates for Student–Garch model (2) (simulated data). Option 1 means that the posterior expectations are estimated by the average of the draws. Option 2 means that they are estimated by the average of the conditional expectations. See Definition 1 in the text.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ϵ = 0.05</th>
<th>ϵ = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Option 1</td>
<td>Option 2</td>
</tr>
<tr>
<td>ω</td>
<td>5200</td>
<td>4750</td>
</tr>
<tr>
<td>α</td>
<td>1100</td>
<td>1050</td>
</tr>
<tr>
<td>β</td>
<td>5200</td>
<td>5200</td>
</tr>
<tr>
<td>ν</td>
<td>350</td>
<td>1</td>
</tr>
</tbody>
</table>

long and regular excursions away from zero are the indication of an absence of convergence. Standardizing the statistic is convenient if one wants to compare the convergence of different chains or parameters. A value of 0.05 for a CUMSUM (after t draws) means that the estimate of the posterior expectation diverges from the final estimate (after N draws) by 5% in units of the final estimate of the posterior standard deviation; so a divergence of even 0.25 would not indicate an imprecise estimate. We propose the following convergence criterion:

Definition 1. If CS_r remains in a band of ±ϵ for all t larger than K(ϵ), then the sampler has converged after K(ϵ) draws for the estimation of the posterior mean with a relative error of 100 × ϵ%.

The value of K(ϵ) depends on the parameter considered, so that the maximum of these values is the minimum sample size to be used. The relative error (ϵ) should be fixed at a low value, like 0.05 or 0.10.

In Figure 3, the CUMSUM evolution of the Monte Carlo estimates of the posterior mean of each parameter are displayed. Each figure contains the CUMSUM plots for the two types of estimates: one based on the sample mean of the draws (option 1), the other on the sample mean of the conditional means (option 2). Upon examining Figure 3a, for example, we can see that for option 1 (dotted line), K is equal to 5200, while for option 2 (solid line), it is equal to 4750 (for an error of 5%). Table 2 displays the results for the four parameters and for two error levels (0.05 and 0.10).

As the Griddy–Gibbs algorithm is relatively greedy in computational time, the question is of course to determine the minimal number of draws necessary to achieve a reasonable precision. With 1000 draws (after 1000 initial draws), we see that we have an error of 10% at most (in the sense of Definition 1).

Two other conclusions can be drawn from the examination of the CUMSUM figures:

- The convergence occurs more quickly for ν and α than for the other parameters. Actually, the convergence is slowest for the parameter that has the highest posterior correlations with other parameters, namely β.6 On Figure 1c, one can see that the posterior density of β after 1000 draws is still a bit imprecise compared to what it is after 10 000 draws.

6In the posterior correlation matrix, all correlation coefficients are lower than 0.25, except for ω and β (−0.67), and for α and β (−0.50). When parameters are highly correlated, the draws tend to cluster, so that it takes a lot of
• The convergence is slightly faster when conditioning (option 2) is used than when sample means of the draws are used (option 1)—see Table 2. The main interest of conditioning is that it yields much more precise graphs of the posterior densities, at no additional cost.

4. COMPARISONS WITH OTHER METHODS

4.1. Importance sampling

The importance sampling (IS) algorithm uses an approximation to the posterior density to generate random draws. This approximation is called the importance function. Bayesian computations require that we evaluate

$$
E[g(\theta)] = \frac{\int g(\theta) \psi(\theta | y) d\theta}{\int \psi(\theta | y) d\theta}
$$

(14)

where $\psi(\theta | y)$ is a kernel of the posterior density and $g(\cdot)$ is an integrable function. The integrals appearing in (14) can be expressed as expectations with respect to the importance function, say $I(\theta)$. For instance,

$$
\int \psi(\theta | y) d\theta = \int \frac{\psi(\theta | y)}{I(\theta)} I(\theta) d\theta = E[I(\theta) \cdot \frac{\psi(\theta | y)}{I(\theta)}]
$$

(15)

which can be approximated by the sample mean

$$
E[I(\theta) \cdot \frac{\psi(\theta | y)}{I(\theta)}] \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\psi(\theta_i | y)}{I(\theta_i)}
$$

(16)

The estimation converges provided the weight function $\psi(\theta | y)/I(\theta)$ is bounded. The smaller the variation coefficient of this ratio, the better the precision. The crux of the method is to find a good importance function. Most of the time the choice is based on the Student density, which is relatively easy to simulate. This density is rather flexible as the degrees-of-freedom parameter governs the thickness of its tails, but it is a symmetric density (no skewness). If $b$ is the location parameter, $M$ the scale matrix and $\lambda$ the degrees of freedom, a usual way to calibrate this density is to fix $b = \hat{\theta}$, the posterior mode, and $M = \{c^2 H^{-1}(\lambda - 2)/\lambda\}^{-1}$, where $H$ is the Hessian of the log posterior evaluated at its mode and $c$ is a tuning constant (e.g. 1.2 to inflate the standard deviations). The choice of $\lambda$ is experimental, unless one has specific information on the tails of the posterior, for example through conditions on existence of moments. A sometimes useful practice is to get a first evaluation of $E(\theta | y)$ and $\text{Var}(\theta | y)$ with a moderate number of draws ($N_1 = 1000$ to $5000$), to use this evaluation to recalibrate the importance function and then to launch a second run with a greater number of draws ($N_2 = 5000$ to $50000$).

The precision which can be obtained by the IS method can be measured as follows. An integral is evaluated by a sample mean which we call $m$. For $N$ large enough, by application draws to explore the posterior, and convergence is slower. This is revealed also by the autocorrelograms of the draws: the first autocorrelations are larger, and the autocorrelations go more slowly to zero, for $\beta$ and $\omega$ than for $\alpha$ and $\nu$. The autocorrelograms are not shown to save space.
of a central limit theorem, the estimator of the integral is distributed as a normal density with
mean equal to \( m \) and variance \( s^2/N \) (provided the sample is iid), where \( s^2 \) is the variance
of the function to estimate. With 95\% level of confidence, the relative error of the integral is
lower than
\[
\frac{1.96 \times 2 \times s}{\sqrt{N} \times m}.
\]
(17)

For the integral (15), the relative error is obtained by replacing \( s/m \) by the variation coefficient
of the weight function. It is also possible to compute the relative error obtained for the
expectation of each component of \( \theta \). This expectation is obtained as the ratio of two integrals
given in (14). The variance of this ratio can be approximated using formula (27.7.3) in Cramer
(1946) as suggested by Kloek and van Dijk (1978). As explained by Geweke (1992), this
method is also applicable to MCMC methods, but as the draws are not independent, the
variance of the estimator is given by \( S(0)/N \), where \( S(0) \) is the spectral density (evaluated at
zero) of the function whose integral is \( m \). The value \( S(0) \) can be estimated from the Gibbs
output by a spectral estimator—see Bauwens and Giot (1998) for an example of application in
the case of the Gibbs sampler with conditioning.

We launched an experiment on the same simulated sample as in Section 3.2, with \( N_1 = 10000, N_2 = 0 \) (no second run), \( \lambda = 5 \) and \( c = 1.2 \). A Student importance function cannot be
totally adequate for the GARCH model because it has to be truncated (at least for the positivity
restrictions). Since we used the same prior as in the pilot experiment with the Griddy–Gibbs
sampler, the parameter values that did not fall in the following intervals were rejected:
\[
0 < \omega < 1, \quad 0 < \alpha < 1, \quad 0 < \beta < 1, \quad 0 < \nu < 18 \quad (18)
\]
(with the half Cauchy prior on \( \nu \)). We got 17\% of rejections. The overall quality of the
importance function is not too bad, although it is not very good, since the variation coefficient
of the weight function is equal to 1.56 (a variation coefficient below 1 is a good benchmark).
With a confidence level of 95\%, this corresponds to a 6.1\% relative error for the integral in the
denominator of (14), and to about 3.5\% error for the estimates of the posterior means. Table 3
compares the posterior means and standard deviations obtained with GGS to those obtained
with IS. The results are very close. Computational time is moderate when one is interested
only in the posterior moments: 10 000 draws were enough in this case. However, this number
of draws is not at all sufficient to get smooth graphs of the marginal densities.

4.2. Metropolis–Hastings algorithm

The MH algorithm was proposed by Metropolis et al. (1953) and extended by Hastings
(1970). It is a Markov chain algorithm which, when convergence is achieved, draws from
the exact posterior density. It is very much related to the importance sampling method as it
uses an approximation to the posterior to draw random numbers. However, there is a rejection
mechanism (instead of a weighting) to decide if a draw belongs or not to the posterior density.
Given a previous draw \( \theta^{n-1} \), a new draw \( \theta^n \) is accepted with probability \( p \) given by
\[
p = \text{Min} \left\{ \frac{\phi(\theta^n|y)/I(\theta^n)}{\phi(\theta^{n-1}|y)/I(\theta^{n-1})}, 1 \right\},
\]
(19)
otherwise the previous draw is kept.
Table 3. Posterior results by different methods for Student–Garch model (2) (simulated data)–Half-Cauchy prior on \( \nu \). Same model, prior, and data as for Table 1. Results with the Griddy–Gibbs are reproduced from the penultimate column of Table 1. (1) CPU time on a Pentium Pro 200 (with GAUSS 3.2.30), including discarded or rejected draws.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Griddy–Gibbs</th>
<th>Importance sampling</th>
<th>MH</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>0.24 [0.09]</td>
<td>0.24 [0.09]</td>
<td>0.22 [0.08]</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.30 [0.13]</td>
<td>0.30 [0.12]</td>
<td>0.27 [0.11]</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.28 [0.16]</td>
<td>0.29 [0.16]</td>
<td>0.33 [0.17]</td>
</tr>
<tr>
<td>( N )</td>
<td>1000</td>
<td>10 000</td>
<td>10 000</td>
</tr>
<tr>
<td>Time (1)</td>
<td>155 s</td>
<td>10 s</td>
<td>10 s</td>
</tr>
</tbody>
</table>

We launched a Metropolis chain of 10 000 accepted draws with the same Student density as in IS. To respect the constraints on the parameters, 2070 draws had to be discarded (compared to 2104 in IS). In the Metropolis chain, 3144 draws were ‘rejected’, which means that in 31\% of the draws, the previous draw was kept. Such a rate of rejection does not seem too bad. Posterior results obtained by this method are also in Table 3. They are fairly close to the results of the two other methods, but the posterior means of \( \omega \), \( \alpha \), and \( \beta \) seem a bit off. Looking at the histograms of the draws (not reported), we found that the MH algorithm did not explore enough the tails of the distribution. Moreover, increasing the number of draws to 50 000 did not change the results.

4.3. Partial conclusions

In our pilot experiment, the three methods provided quite close results. We also compared the three methods with the same model, data, and prior, except that we used a flat prior on \( \nu \) on the interval \((0, 30)\). In this case, we did not fully succeed to get accurate results with the MH algorithm (see Table 4). For the Student importance function used in the MH algorithm, we had to inflate the Hessian approximation of the standard deviation of \( \nu \) by a factor of 4 (instead of 1.2 for the other parameters) to get fairly good results on the posterior moments of \( \nu \). However, this seems to be at the expense of the results for the other parameters, as can be seen in Table 4. Again, increasing the number of draws (to 150 000) for MH hardly changed the results. Our experience is thus not very favourable to the use of the MH algorithm in Student–GARCH models with a flat prior.\(^7\) One could object that a better importance function should be found, but this does not seem easy. We do not think it is very worthwhile to look for it, as the other algorithms are available.

The old fashioned importance sampling algorithm provides as accurate posterior moments as the Griddy–Gibbs sampler, but less precise graphs of the marginal densities. Although GGS is more costly in computer time, it seems easier to implement, since in our experiments,\(^7\)Geweke (1994) used the MH algorithm with success on a GARCH model with normal errors, which does not have the difficulty caused by the degrees-of-freedom parameter (the integrability problem explained in Section 1).
Table 4. Posterior results by different methods for Student–Garch model (2) (simulated data) – Flat prior on $\nu$. Same model, data and prior as for Table 3, except that the prior on $\nu$ is flat on $(0,30)$. (1) CPU time on a Pentium Pro 200 (with GAUSS 3.2.30), including discarded or rejected draws.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Griddy–Gibbs</th>
<th>Importance sampling</th>
<th>MH</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega$</td>
<td>0.26 [0.10]</td>
<td>0.26 [0.10]</td>
<td>0.21 [0.08]</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.30 [0.13]</td>
<td>0.31 [0.13]</td>
<td>0.26 [0.10]</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.31 [0.17]</td>
<td>0.31 [0.17]</td>
<td>0.37 [0.17]</td>
</tr>
<tr>
<td>$\nu$</td>
<td>9.70 [5.34]</td>
<td>9.60 [5.02]</td>
<td>9.90 [5.48]</td>
</tr>
<tr>
<td>$N$</td>
<td>1000</td>
<td>5000 + 50 000</td>
<td>50 000</td>
</tr>
<tr>
<td>time (1)</td>
<td>152 s</td>
<td>55 s</td>
<td>55 s</td>
</tr>
</tbody>
</table>

the calibration of a Student importance function was sometimes not obvious. If we take into account the labour time to be devoted to the fine tuning of an importance function, the Griddy–Gibbs sampler is a method worth considering.

5. AN ASYMMETRIC GARCH MODEL FOR THE BRUSSELS STOCK MARKET INDEX

To model an observed series, such as the returns of a stock index, the simple GARCH model (2) cannot capture some well known stylized facts. First, the model implies that the series has a constant zero mean. As documented by Campbell et al. (1997, Ch. 2), positive autocorrelation of order one is quite usual for a stock index and could be induced by non-synchronous trading of the index components; see Campbell et al. (1997, Ch. 3). Second, the model implies that a shock influences the next conditional variance only by its squared magnitude, not by its sign: the news impact curve, i.e. the graph of $h_t$ on the past shock—which is $\gamma_{t-1}$ in the simple model (2)—is symmetrical (around 0). This is at odds with the so-called ‘leverage effect’ noted by Black (1976), which refers to the negative correlation between changes in stock prices and volatility—see for example the survey by Bollerslev et al. (1992) for a discussion and further references. Several specifications have been proposed to introduce asymmetry, such as the exponential GARCH model of Nelson (1991), the ‘GJR’ model of Glosten et al. (1993), and the threshold ARCH model of Zakoian (1994).

The estimated model allows for a non-zero first-order autocorrelation of the returns, and a GJR-type asymmetric Student–GARCH equation:

\[
\begin{align*}
\gamma_t &= \mu + \rho \gamma_{t-1} + u_t \\
u_t &= \epsilon_t \sqrt{h_t} \\
\epsilon_t | I_{t-1} &\sim \text{Student}(0,1,\nu) \\
\hat{h}_t &= \omega + \alpha^+ u_{t-1}^2 + \alpha^- u_{t-1}^2 + \beta \hat{h}_{t-1} \\
u_t^+ &= u_t^+ 1_{[u_t > 0]} , \quad u_t^- = u_t^+ 1_{[u_t < 0]} .
\end{align*}
\]

We have used weekly data on the spot market index of shares of Belgian firms at the Brussels stock exchange, for the period 03/01/86–26/01/96. The data consist of closing quotes.
Bayesian inference on GARCH models using the Gibbs sampler

on Friday, providing 508 observations. The dependent variable is the index return, measured as the first difference of the logarithm of the index at time $t$. A time-series plot of the data is provided in Figure 4. The extreme returns just before observation 100 correspond to the October 1987 crash. The mean return is equal to 0.002246 and the standard deviation to 0.01909. The index exhibits an overall upward trend over the sample period.

5.1. Posterior results

All the results were computed with the Griddy–Gibbs sampler using 1000 draws after 1000 initial draws (dropped). A flat prior on finite intervals was used for each parameter. For $\nu$, we also used the half-Cauchy prior (8), and a dogmatic prior which says that $\nu$ is large (thus replacing the Student distribution by the normal in the specification of the GARCH model). Table 5 and Figure 5 provide the posterior results (and the prior intervals). They call for the following comments:

1. The results for the three specifications do not differ a lot. In the Student case, the prior on $\nu$ has a lot of influence on the posterior results for $\nu$, but not for the other parameters, which we find reassuring. When a normal density is used ($\nu = \infty$), the estimates (posterior means) of $\alpha^+, \alpha^-$ and $\beta$ are slightly higher. This adjustment accommodates for the excess kurtosis in the data, since in the Student case the degrees of freedom take up this effect.

2. Negative shocks have a stronger impact than positive shocks in all specifications. When a symmetric model is estimated (unreported results), the posterior mean of $\alpha$ is approximately equal to the average of the posterior means of $\alpha^+$ and $\alpha^-$ in the corresponding asymmetric model. The difference $\alpha^- - \alpha^+$ has a posterior mean equal to 0.13 with a standard deviation equal to 0.09. The posterior density of the $\alpha$ parameters and of their difference is reported in Figures 5a and c, respectively.

Figure 4. Brussels spot market index return (3/1/86–26/1/96).
Table 5. ML and posterior results for asymmetric GARCH model (20) on Brussels stock index. Entries in the first column of results (ML) are maximum likelihood estimates and asymptotic standard errors. In the last three columns, they are posterior means, standard deviations (between square brackets, under the means), and numerical standard errors (between parentheses, beside the means), except in the last line where they are posterior probabilities. Posterior moments were computed by conditioning (i.e. marginalizing conditional moments). WSC means weak stationarity condition (see text). Under the parameter names are the prior intervals (integration ranges). The value of $M$ is 30 for the half-Cauchy prior, and 50 for the flat prior. CPU time is on a Pentium Pro 200 (with GAUSS, version 3.2.30), for 2000 draws including the 1000 discarded initial draws (except for ML).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Student none (ML)</th>
<th>Student Half-Cauchy</th>
<th>Student Flat</th>
<th>Normal none</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.090</td>
<td>0.088</td>
<td>0.086</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>(-0.06, 0.24)</td>
<td>[0.035]</td>
<td>[0.036]</td>
<td>[0.036]</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.20</td>
<td>0.20 (1.7e-4)</td>
<td>0.20</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>(0.0, 0.40)</td>
<td>[0.046]</td>
<td>[0.047]</td>
<td>[0.047]</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.054</td>
<td>0.071 (8.8e-4)</td>
<td>0.073</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>(0.01, 0.20)</td>
<td>[0.020]</td>
<td>[0.026]</td>
<td>[0.027]</td>
</tr>
<tr>
<td>$\alpha^+$</td>
<td>0.071</td>
<td>0.099 (9.8e-4)</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>(1e-5, 0.30)</td>
<td>[0.038]</td>
<td>[0.044]</td>
<td>[0.045]</td>
</tr>
<tr>
<td>$\alpha^-$</td>
<td>0.19</td>
<td>0.23 (1.5e-3)</td>
<td>0.24</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>(0.01, 0.50)</td>
<td>[0.055]</td>
<td>[0.063]</td>
<td>[0.064]</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.75</td>
<td>0.70 (2.6e-3)</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(0.35, 0.95)</td>
<td>[0.058]</td>
<td>[0.069]</td>
<td>[0.064]</td>
</tr>
<tr>
<td>$\nu$</td>
<td>9.63</td>
<td>10.3 (4.9e-2)</td>
<td>18.2</td>
<td>$\infty$</td>
</tr>
<tr>
<td></td>
<td>(0.01, $M$)</td>
<td>[3.97]</td>
<td>[4.62]</td>
<td>[10.9]</td>
</tr>
<tr>
<td>WSC</td>
<td>0.082</td>
<td>0.089</td>
<td>0.089</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>na</td>
<td>[0.049]</td>
<td>[0.045]</td>
<td>[0.039]</td>
</tr>
<tr>
<td>P(WSC &gt; 0)</td>
<td>na</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
</tbody>
</table>

3. Figure 5b illustrates the effect of the prior specification on the posterior density of $\nu$. With the flat prior, the density is decreasing very slowly at the right tail, illustrating once again the problem of integrability outlined in Section 2. The long right tail draws the posterior mean to 18, with a large standard deviation. The posterior mode is of course much closer to what it is when the half-Cauchy prior is used. In the latter case, the posterior mean is much closer to the posterior mode. Although the data indicate a likely value of about 10 for $\nu$, they do not favour at all very low degrees of freedom, as there is very little probability on the interval $\nu < 3$. One can interpret the results as weakly favourable to a Student specification, compared with a normal. We have also estimated the model for individual stocks and found that the degrees-of-freedom parameter was close to 5 in many cases.
4. The weak stationarity condition (WSC) for the Student model requires that $\text{WSC} = 1 - (\alpha^+ + \alpha^-) \nu / [2(\nu - 2)] - \beta$ be positive – see (4). Its posterior probability is quite close to 1 (see also the density on Figure 5d). There is clear evidence in favour of weak stationarity.

5. The returns exhibit positive first-order autocorrelation of the order of 0.20 (the posterior mean of $\rho$). When a second lag of $y_t$ is introduced in the regression equation, the posterior mean of its coefficient is found to be very close to 0, with a high standard deviation.

As regards the convergence of the Griddy–Gibbs sampler for the results reported in Table 5, the examination of the CUMSUM plots (not reported) based on 1000 draws (after 1000 initial deleted draws) reveals that the number of draws after which convergence has been obtained for a relative error of 5% is about 600 at most (see Definition 1). It seems therefore sufficient to use 1000 draws. We also report in Table 5 ‘numerical’ standard errors, i.e. standard errors

© Royal Economic Society 1998
of the Gibbs estimates of the posterior means, computed by spectral analysis techniques—see Geweke (1992). The spectral density at zero has been computed using a Daniell window; see Bauwens and Giot (1998) for details. The corresponding estimated relative errors obtained by applying formula (17) are smaller than than 1% for \( \mu \) and \( \rho \), between 1 and 2% for \( \beta \) and \( \nu \), 2.5% for \( \alpha^- \), 3.8% for \( \alpha^+ \) and 4.8% for \( \omega \). Such levels are quite acceptable.

5.2. Predictive option prices

The cash price \( (C_{T+s}^-) \) of a European call option \( s \) units of time before maturity is evaluated by (see Ghysels et al. 1995)

\[
C_{T+s}^- = e^{-rs} E(S_{T+s} - K)^+
\]

where \( (S_{T+s} - K)^+ = \max(S_{T+s} - K, 0) \), \( S_{T+s} \) is the price of the underlying security at maturity (assumed without dividend), \( K \) is the exercise price (‘strike’), and \( r \) is the interest rate (assumed known, though for practical applications, it could be modelled). We choose to normalize \( C_{T+s}^- \) as

\[
C_{T+s} = \frac{C_{T+s}^-}{Ke^{-rs}} = E\left(\frac{S_{T+s}}{K} - 1\right)^+
\]

and to compute the expectation using the predictive distribution of \( S_{T+s} \) (short of knowing the ‘risk-neutral’ probability distribution that should be used). The predictive option price \( C_{T+s} \) is

\[
E\left\{\left(\frac{S_{T+s}}{K} - 1\right)^+ \bigg| y\right\} = \int (S_{T+s}/K - 1)^+ f(S_{T+s}|y) dS_{T+s}
\]

\[
= \int (S_{T+s}/K - 1)^+ \left\{ \int f(S_{T+s}|\theta, y) \psi(\theta|y) d\theta \right\} dS_{T+s}
\]

\[
= \int \left\{ \int (S_{T+s}/K - 1)^+ f(S_{T+s}|\theta, y) dS_{T+s} \right\} \psi(\theta|y) d\theta.
\]

If we can simulate \( R \) realizations of the distribution of \( S_{T+s}|\theta, y \), the inner integral of the last line of the above formula can be approximated by its empirical counterpart \( R^{-1} \sum_{r=1}^{R} (S_{T+s}/K - 1)^+ \) (given \( \theta \)). Then having \( N \) realizations of the posterior distribution of \( \theta \) (e.g. a Gibbs sample), we compute the simple mean of the \( N \) estimates of the inner integral to estimate the outside integral. To simulate the distribution of \( S_{T+s}|\theta, y \), we use the econometric model where \( \Delta \ln S_t = y_t \). The future price \( S_{T+s} \) is related to the last observed price \( S_T \) by

\[
S_{T+s} = S_T \exp \sum_{i=1}^{h} y_{T+i}
\]

so that we have to simulate the future returns \( y_{T+i} \). If \( s = 1 \), we just have to simulate \( y_{T+1} \) (the first post-sample return). If the econometric model is \( y_t \sim \text{Student}(0, h_t, v) \) with \( h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1} \), then given \( \theta \) and \( y \) (the observed sample), we know \( y_t \),

\( ^8 \)Numerical standard errors are not given for the results in the last two columns of Table 5, but they are very close to those reported in the second column.
In the same way as in the last line of (23), we have
\[ \theta \sim \text{Gamma}(a, b) \]
Since \( \sigma \) is unknown, it must be predicted to apply (25). So we can apply the Black–Scholes formula with the volatility predicted using the GARCH model on the return. This is known as the ‘plug-in’ method – see Noh et al. (1994), who find that it may be a profitable trading rule, although the Black–Scholes formula is only applicable in the context of deterministic volatility and other assumptions not respected by the GARCH model. When the time to maturity, \( s \), is larger than one, \( \sigma^2 \) is the average of the \( s \) future variances \( \nu \tau+i \):
\[ \sigma^2 = \frac{1}{s} \sum_{i=1}^{s} \nu \tau+i. \] (26)

With a GARCH–Student model like (20), \( \nu \tau+i = \nu \tau+i/(\nu-2) \), provided \( \nu \) is larger than two. Therefore, \( \sigma^2 \) is a function of the GARCH parameters and we can compute the predictive mean of \( BS_{T+s} \) for any given \( s \) and \( S_{T+s}/Ke^{-r s} \), as soon as we have a Gibbs (or any other) sample of the posterior distribution of the GARCH parameters. Actually, the computation of the predictive mean of \( BS_{T+s} \) is done by simulation of the predictive density of \( BS_{T+s} \).

In the same way as in the last line of (23), we have
\[ BS_{T+s} = \mathbb{E}(BS_{T+s} | y) = \mathbb{E}_{\theta | y}(\mathbb{E}(BS_{T+s} | \theta, y)). \] (27)

As explained above, \( BS_{T+s} \) is a function of the future conditional variances \( h^* = (h_{\tau+1}, \ldots, h_{T+s}) \), and therefore a function of the future standardized innovations \( \epsilon^* = (\epsilon_{\tau+1}, \ldots, \epsilon_{T+s}) \):
\[ BS_{T+s} = BS_{T+s}(\sigma) = BS_{T+s}[\sigma(h^*)] = BS_{T+s}[\sigma(h^*(\epsilon^*))]. \] (28)

Since \( \epsilon_i = u_i/\sqrt{\nu_i} \) has the known distribution \( \epsilon_i | \theta \sim \text{Student}(0, 1, \nu) \) (given the past), we can easily simulate \( \epsilon^* \) and compute \( BS_{T+s}(\epsilon^*) \).

We can now detail an algorithm for computing the predictive means \( C_{T+s} \) and \( BS_{T+s} \). For each draw of \( \theta \) provided by the GGS, we perform the following steps.

- **step 1**: compute \( u_T \) and \( h_T \) for the given \( \theta \) (period \( T \) is the last one of the observed sample).
- **step 2**: compute \( h_{\tau+1} = \omega + \alpha u_T^2 + \beta h_T \).
With $N$ minimized (see Hull and White 1987), the Black–Scholes formula is linear in volatility, and the bias due to stochastic volatility is the exact evaluation for at the money options is not surprising since for this type of options to a large value, like 50. Finding that the Black and Scholes evaluation hardly differs from and 3.3.4), because the latter specification does not make sense for long-run predictions. For the Brussels spot index, we report in Table 6, the predictive option prices of ‘at the money’ options for different maturities (one week: $s = 0.716$ and 13 weeks or 3 months). For the direct evaluation formula (23), we have set the strike at the last observed value of the stock index, in which case (25) simplifies into $BS_{T+s} = 2N(\sigma \sqrt{s/2}) - 1$, which is almost linear in $\sigma$. For the direct evaluation formula (23), we have set the strike at the last observed value of the stock index, in which case $ST + s = \exp \sum i=1 s/r + 1/K - 1$. For predicting future returns, we have assumed that $y_t$ has zero mean, instead of $\mu + \rho y_{t-1}$ (see steps 3.2 and 3.3.4), because the latter specification does not make sense for long-run predictions. The results of the two evaluations are almost identical whether we fix $R$ equal to 1 or to a large value, like 50. Finding that the Black and Scholes evaluation hardly differs from the exact evaluation for at the money options is not surprising since for this type of options the Black–Scholes formula is linear in volatility, and the bias due to stochastic volatility is minimized (see Hull and White 1987).

### Table 6. Predictive call option prices on Brussels index. Entries are in percent. $C_{T+s}$ is defined in (22), $BS_{T+s}$ in (27). $T$ is the last date of the observed data, $s$ is the horizon (e.g. $s = 13$ corresponds to 13 weeks or 3 months).

<table>
<thead>
<tr>
<th>Maturity (s)</th>
<th>1 week</th>
<th>3 months</th>
<th>6 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{T+s}$</td>
<td>0.716</td>
<td>2.624</td>
<td>3.762</td>
</tr>
<tr>
<td>$BS_{T+s}$</td>
<td>0.749</td>
<td>2.648</td>
<td>3.917</td>
</tr>
</tbody>
</table>

- **step 3.1:** $r = 1$; do while $r \leq R$.
- **step 3.2:** simulate $\epsilon_{T+i,r} \sim \text{Student}(0, 1, v)$, and compute $y_{T+i,r} = h_{T+i,r}^{-1} \epsilon_{T+i,r}$.
- **step 3.3.1:** $i = 2$; do while $i \leq s$.
- **step 3.3.2:** compute $h_{T+i,r} = \omega + h_{T+i-1,r}(\alpha \epsilon_{T+i-1,r}^r + \beta)$ (for $h_{T+i,r}$ always use $h_{T+1}$ from step 2).
- **step 3.3.3:** simulate $\epsilon_{T+i,r} \sim \text{Student}(0, 1, v)$.
- **step 3.3.4:** compute $y_{T+i,r} = h_{T+i,r}^{-1} \epsilon_{T+i,r}$.
- **step 3.3.5:** increment $i$ by 1 and go to step 3.3.2 unless $i > s$.
- **step 3.4:** for Black–Scholes evaluation: compute $\sigma_r^2 = s^{-1} \sum i=1 s/r \exp \sum i=1 s/r + 1/K - 1$. For the direct evaluation formula (23), we have set $ST + s = \exp \sum i=1 s/r + 1/K - 1$. For predicting future returns, we have assumed that $y_t$ has zero mean, instead of $\mu + \rho y_{t-1}$ (see steps 3.2 and 3.3.4), because the latter specification does not make sense for long-run predictions. The results of the two evaluations are almost identical whether we fix $R$ equal to 1 or to a large value, like 50. Finding that the Black and Scholes evaluation hardly differs from the exact evaluation for at the money options is not surprising since for this type of options the Black–Scholes formula is linear in volatility, and the bias due to stochastic volatility is minimized (see Hull and White 1987).

### 6. CONCLUSION

Drèze (1977) stated that for the resolution of a statistical problem in a Bayesian framework, ‘it has become acceptable to quit upon obtaining analytically a marginal posterior density for one or two parameters of interest, and declaring victoriously that it could be analysed by numerical
integration techniques’. However, as mentioned in a footnote, usual quadrature rules would require about sixteen years of computing time when the dimension of the integral is equal to nine.\(^9\) The paradox of our approach is that the integration method we use for high dimensions is based on quadrature rules, but that it does not suffer from the usual curse of dimensionality which affects these methods. The Griddy–Gibbs sampler evaluates a \(k\)-dimensional integral by splitting it into a series of \(k\) unidimensional integrals and finding an accurate result by averaging those sets of \(k\) univariate integrals. On powerful microcomputers, the route is now opened for analysing complex statistical models requiring numerical integration in a dimension not exceeding say, fifteen.

The method has been applied to estimate GARCH models for the Brussels stock market. Asymmetric effects of a shock on the next variance are found, but as these effects do not seem to be persistent, it makes sense to compute predictive option prices.

There are many econometric models which lend themselves to using the Griddy–Gibbs sampler. In many cases, it is not necessary to apply it to the complete parameter vector as in GARCH models. As an example, consider the heteroskedastic regression model (1): we can apply the Griddy–Gibbs sampler to the posterior density of \(\gamma\), and marginalize the posterior conditional density and moments of \(\beta\). Other examples are ARMA models, and simultaneous equation models.

A. APPENDIX

The Griddy–Gibbs sampler works as follows for \(N\) draws of the posterior (11).

- **step 1**: starting point of the chain at a value \(a^0\) for \(a\).
- **step 2**: loop starting at \(n = 1\).
- **step 3**: compute \(\kappa(\mu|a^{n-1}, y)\) over the grid \((\mu_1, \mu_2, \ldots, \mu_G)\) to obtain the vector \(G_\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_G)\).
- **step 4**: by a deterministic integration rule using \(G\) points, compute the values \(G_\Phi = (0, \Phi_2, \ldots, \Phi_G)\) where
  \[
  \Phi_i = \int_{\mu_1}^{\mu_i} \kappa(\mu|a^{n-1}, y) \, d\mu, \quad i = 2, \ldots, G. \]
  Compute (and cumulate for the marginal) the normalized pdf values \(G_\psi = G_\kappa / G_\Phi\) of \(\psi(\mu|a^{n-1}, y)\). Compute \(E(\mu|a^{n-1}, y)\) and \(\text{Var}(\mu|a^{n-1}, y)\) by the same type of integration rule as above, and store them in a table.
- **step 5**: generate \(u \sim U[0, \Phi_G]\) and invert \(\Phi(\mu|a^{n-1}, y)\) by numerical interpolation to get a draw of \(\mu|a^{n-1}, y\), indexed \(a^n\). Store this draw in a table.
- **step 6**: repeat steps 3–5 for \(a^n|\mu^n, y\).
- **step 7**: increment \(n\) by 1 and go to step 3 unless \(n > N\).
- **step 8**: compute the posterior moments of \(\mu\) and \(a\) from the tables where conditional moments are stored (by averaging). Likewise, plot the marginal densities (cumulated \(G_\psi / N\)). With the table containing \([\mu^n, a^n]_{n=1}^N\), one can compute posterior moments (as sample averages) and draw a histogram of any function of the parameters.

\(^9\)He assumed that it takes one millisecond to compute the posterior, and that a grid of twenty points is used for each coordinate (leading to \(20^9\) function evaluations). Even if nowadays microcomputers are 1000 times faster than computers we used twenty years ago, the computing time would still be six days.

\(\copyright\) Royal Economic Society 1998
Generalizing the Gibbs sampler to $m$ parameters implies that we consider a more complex Markov chain. Let $\theta$ denote the vector of $m$ parameters and let $\theta_{-k}$ denote the vector $\theta$ without its $k$th element $\theta_k$. A draw $\theta^n$ of the posterior density is generated using the previous draw $\theta^{n-1}$ by the following cycle:

\[
\begin{align*}
\theta_1^n & \sim \phi(\theta_1^n | \theta_{-1}^{n-1}, y) \\
\theta_2^n & \sim \phi(\theta_2^n | \theta_1^n, \theta_{-2}^{n-1}, \ldots, \theta_m^{n-1}, y) \\
& \vdots \\
\theta_k^n & \sim \phi(\theta_k^n | \theta_1^n, \ldots, \theta_{k-1}^{n-1}, \theta_{k+1}^{n-1}, \ldots, \theta_m^{n-1}, y) \\
& \vdots \\
\theta_m^n & \sim \phi(\theta_m^n | \theta_{-m}^{n-1}, y).
\end{align*}
\]

The procedure is started using any value $\theta_0^{-1}$ in the support of the posterior distribution. Steps 3 to 6 can be easily generalized to incorporate this more complex cycle.

- The choice of the grid of points has to be made carefully and constitutes the main difficulty in applying the method. One must have an idea of the bounds of integration. These may come from prior restrictions on the parameters, e.g. $0 < \alpha < 1$. However, even in such cases, one may consider to restrict the integration to the subset of the parameter space where the value of the posterior density is big enough to contribute to the integrals. Otherwise, too many points of the grids will be practically useless, at the expense of precision in the inversion of the cdf. It is worth exploring the shape of the conditional densities by varying the conditioning values enough, before fixing the integration ranges.

- It is quite evident that a variable grid could be more efficient since more points should be used in the area where the posterior density varies a lot. A variable grid can be used, but it may be difficult to devise a way to adapt it at each iteration of the loop of draws (since the cdf varies because of the renewal of the conditioning values). An automatic way to adapt the grid is to use an iterative method of integration (but see the next remark).

- Our implementation of the algorithm requires that the grid of points of each parameter remains constant throughout the $N$ draws. Otherwise, we would have to interpolate the value of the conditional pdf over a fixed grid (using the values of the pdf over the variable grid of the current cycle) in order to apply (12). This is not recommended if the functional evaluation of the posterior density is expensive (as in GARCH models). For this reason, iterative integration methods (like the iterative Simpson’s method) are less attractive.

- We use the trapezoidal rule of integration (see Davis and Rabinowitz 1975, p. 40) over fixed grids of (usually 33) points, and linear interpolation between adjacent points to invert the cdf. The cost of the method being proportional to the number of points in the grids, a smaller number of points could be compensated by a more sophisticated interpolation method to invert the cdf.
Bayesian inference on GARCH models using the Gibbs sampler

ACKNOWLEDGEMENTS

Support of the European Commission Human Capital and Mobility Program through the network 'Econometric inference using simulation methods' is gratefully acknowledged. While remaining responsible for any error, the authors wish to thank two anonymous referees, Ana Justel and seminar participants at the University of Toronto, Erasmus University Rotterdam, and Humboldt-University-Berlin, for useful remarks and suggestions. This paper 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 author.

REFERENCES


