Multimodality and the GARCH Likelihood

Jurgen A. Doornik*
Nuffield College, University of Oxford

Marius Ooms
Econometric Institute, Erasmus University Rotterdam

26th January 2000

Abstract

We investigate several aspects of GARCH(p, q) models which are relevant for empirical applications. In particular, we note that the inclusion of a dummy variable as regressor can lead to multimodality in the GARCH likelihood. This makes standard inference on the estimated coefficient impossible. Next, we investigate the implementation of different restrictions on the GARCH parameter space. We present a small refinement to the Nelson and Cao (1992) conditions for a GARCH($2, q$) model, and show how these can be implemented by parameter transformations. We argue that these conditions are also too restrictive, and consider restrictions which are formulated in terms of the unconditional variance. These are easier to work with and understand. Finally, we show that multimodality is a real concern for models of the £/$ exchange rate, and should be taken account of, especially when $p \geq 2$.

1 Introduction

The ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models have found widespread application since their introduction. Despite this, the literature has paid relatively little attention to the implementation aspects of GARCH models, and largely ignored the possibility of multimodality in the likelihood.

In this note we illustrate how multimodality in the likelihood of GARCH-type models is induced when correcting for an additive outlier in the mean equation through a dummy variable. We also show in §2.2 that adding the corresponding dummy to the variance equation can exacerbate the problem. Section 3 discusses restrictions on the GARCH parameter space. The next section then investigates whether multimodality is of practical relevance, and if it depends on the adopted restrictions. We conclude that multimodality is a potential problem in empirical applications, and recommend the adoption of a limited search using random starting values whenever estimating a GARCH model.

*Correspondence to: jurgen.doornik@nuffield.ox.ac.uk

The regression model with normal-GARCH(p,q) errors is defined as:

$$\begin{aligned} y_t &= x_t'\zeta + \varepsilon_t, \\ \varepsilon_t &= \xi_t h_t^{1/2}, \quad \xi_t|\mathcal{F}_{t-1} \sim N(0,1), \\ h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-1}^2 + \sum_{i=1}^{p} \beta_i h_{t-1}, \quad t = 1, \ldots T. \end{aligned} \tag{1}$$

$\mathcal{F}_t$ is the filtration up to time t. The ARCH(q) model corresponds to GARCH$(0,q)$. Recent surveys include Bollerslev, Engle, and Nelson (1994), Shephard (1996), and Gourieroux (1997).

The equation for h_t can be written in ARMA form using $u_t = \varepsilon_t^2 - h_t = (\xi_t^2 - 1)h_t$:

$$\varepsilon_t^2 = \alpha_0 + \sum_{i=1}^{m} (\alpha_i + \beta_i)\varepsilon_{t-i}^2 - \sum_{i=1}^{p} \beta_i u_{t-i} + u_t, \tag{2}$$

where $m = \max(p,q)$ and $\beta_i = 0$ for $i > p$, $\alpha_i = 0$ for $i > q$; note that $\mathsf{E}[u_t|\mathcal{F}_{t-1}] = 0$.

We also consider the EGARCH(p,q) specification (see Nelson, 1991):

$$\log h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \left(\vartheta_1 \xi_{t-1} + \vartheta_2 |\xi_{t-1}|\right) + \sum_{i=1}^{p} \beta_i \log h_{t-i}, \tag{3}$$

with $\alpha_1 = 1$.

In the remainder, dummies are always variables assuming value one for a single observation, and zero otherwise.

2 Multimodality caused by dummy variables

In a normal regression model, the effect of a dummy variable is to set the residual for that observation to zero. At first sight, it may be assumed that the same effect is achieved in the regression model with ARCH or GARCH errors.

To illustrate that this is not the case, we use the Dow–Jones index (Dow Jones Industrial Average: close at midweek from 1980 to September 1994);[1] the first differences of the logs are given in Figure 1a. We start by estimating an ARCH(1) model, where the mean equation consists of a constant and a dummy variable for 21-Oct-1987 (value one for the Wednesday after the crash, zero otherwise):

$$\begin{aligned} y_t &= c + \gamma d_{\text{crash}} + \varepsilon_t, \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2. \end{aligned}$$

Let $\hat{c}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\gamma}$ be the maximum likelihood estimates. Figure 1b plots the log-likelihood values as a function of γ, with the remaining coefficients kept fixed at $\hat{c}, \hat{\alpha}_0, \hat{\alpha}_1$. We were initially very surprised to see the pronounced bimodal shape of the likelihood. Adding an ARCH term to a regression model with a dummy variable clearly changes the role of the dummy.[2]

[1] We used 770 observations from 2-Jan-1980 to 28-Sep-1994; the first observation is lost when growth rates are used. The figures are for Wednesday, or Tuesday if the stock market was closed on Wednesday. The Dow Jones data are available from www.economagic.com.

[2] The role of the intercept is also changed: the residuals will not have mean zero despite the presence of a constant term.

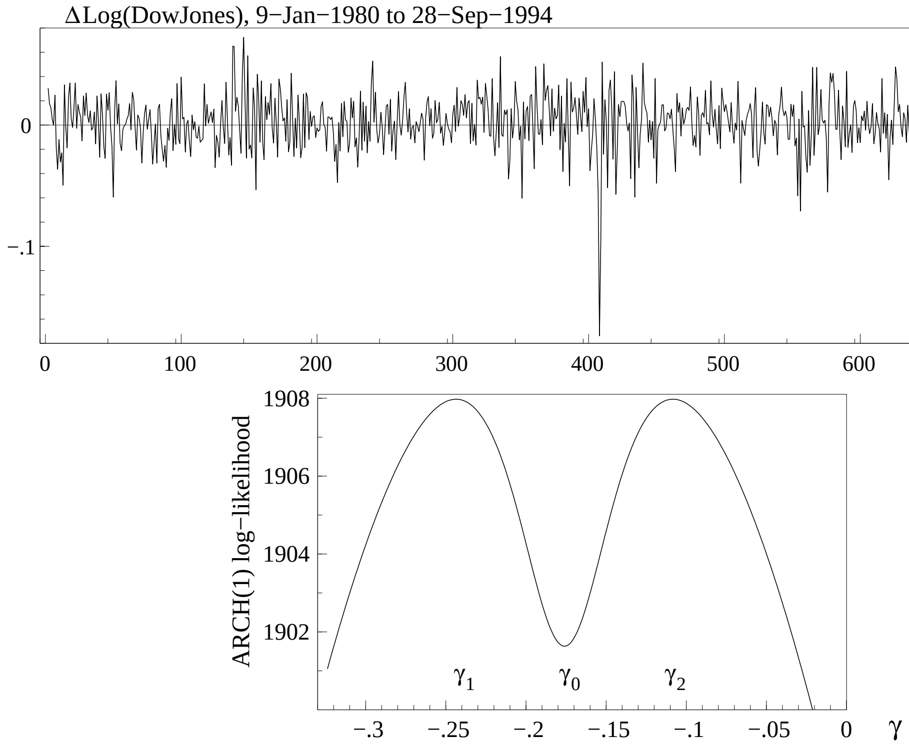

Figure 1: Log-returns on Dow–Jones index (top), with likelihood grid for the dummy parameter corresponding to the 1987 crash (bottom).

2.1 ARCH(1) with a dummy variable in the mean

The following theorem explains the effect of the dummy variable for the ARCH(1) model.

Theorem 1 *Consider the ARCH(1) regression model with mean specified as* $y_t = x_t'\zeta + d_t\gamma + \varepsilon_t$*. The additional regressor is a dummy* d_t*, where* $d_t = 1$ *when* $t = s, 1 < s \leq T$*, and* $d_t = 0$ *otherwise. Define*

$$G_s = \frac{1}{2}h_s\left[\left(1 + \frac{4\varepsilon_{s+1}^2}{\alpha_1 h_s}\right)^{1/2} - 1\right] - \frac{\alpha_0}{\alpha_1}.$$

(a) When $G_s \leq 0$ *the log-likelihood* $\ell(\theta)$ *has a* unique solution *for* γ*:*

$$\hat{\gamma}_0 = y_s - x_s'\hat{\zeta},$$

with $\hat{\varepsilon}_s = 0$*.*

(b) When $G_s > 0$*,* $\ell(\theta)$ *has* two maxima*, which are only different in the value of* γ*:*

$$\hat{\gamma}_{1,s} = y_s - x_s'\hat{\zeta} - G_s^{1/2},$$
$$\hat{\gamma}_{2,s} = y_s - x_s'\hat{\zeta} + G_s^{1/2}.$$

Both modes have identical likelihood values and second derivatives, and have otherwise the same parameter values. In this case $\hat{\gamma}_{0,s} = y_s - x_s'\hat{\zeta}$ *corresponds to a local minimum.*

The derivation of G_s and the properties of the likelihood are given in Appendix 1.

Theorem 1 indicates that the dummy does not always lead to multimodality. If G_s is negative or zero, $\hat{\gamma} = y_s - x_s'\hat{\zeta}$, and the dummy plays a similar role as in the regression model without ARCH errors.

However, when G_s is positive, there are two identical maxima. The value of G_s depends on the parameter values, and, because $h_s = \alpha_0 + \alpha_1\varepsilon_{s-1}^2$, on the residuals immediately after and before the time of the impulse. Theorem 1 shows that the likelihood derivatives are identical at both maxima. As a consequence, both estimates of γ have the same estimated standard error, which results in different t-values. The estimation procedure may pick either maximum, but deciding significance by looking at the t-value is problematic.[3] There are also two residuals: $\hat{\varepsilon}_{1,s} = G_s^{1/2}$ and $\hat{\varepsilon}_{2,s} = -G_s^{1/2}$. Diagnostic tests based on the residuals (or standardized residuals: there is one value for h_s), will have different outcomes, unless only the squared values are used.

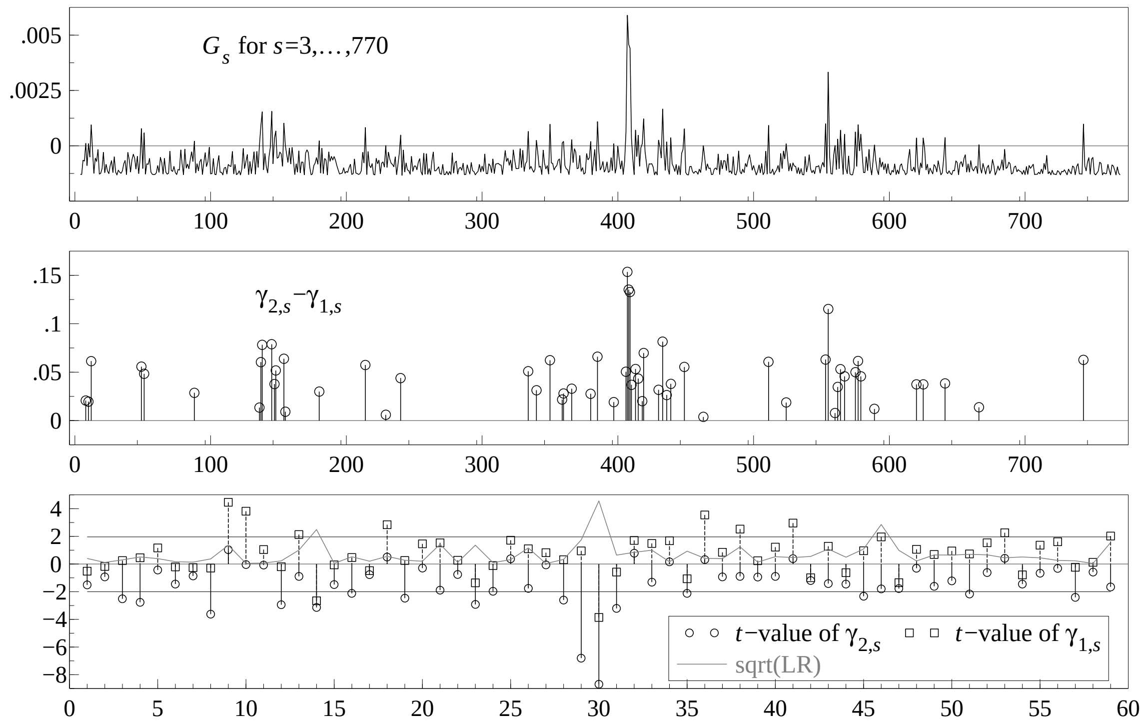

Figure 2: ARCH(1) model for growth rates of Dow-Jones with moving dummy variable: G_s (top), $\hat{\gamma}_{2,s} - \hat{\gamma}_{1,s}$ (middle), and t-values and square root of likelihood-ratio test (bottom; only for observations with multimodality).

To assess the empirical relevance of Theorem 1, we run a singly dummy through the data, re-estimating the ARCH(1) model every time (the mean is specified as $c + \gamma d_t,\ d_t = 1$ for $t = s,\ s =$

[3]Note that, when the constant is omitted, and only a dummy is included as regressor, standard econometric software may find the local minimum instead of one of the maxima: the OLS-based starting value for the dummy parameter would have a zero derivative, so that the estimate of its coefficient may not move in subsequent iterations.

$3, \ldots, 770$). Figure 2a plots the value of G_s for the ARCH(1) model, with positive values indicating multiple maxima. In this case, there are 59 cases with $\hat{G}_s > 0$, and correspondingly with two solutions for γ; the second graph displays the difference $\hat{\gamma}_{2,s} - \hat{\gamma}_{1,s} = G_s^{1/2}$ for the cases with multimodality. The bottom graph shows the t-values when $\hat{G}_s > 0$. In this graph, the observations with $G_s \leq 0$ are omitted. Using a critical value of two, there are several cases with one t-statistic insignificant, and the other significant. The graph also shows the square root of the likelihood-ratio test, which has one degree of freedom. Now only three of the displayed observations are significant.

2.2 ARCH(1) with a dummy variable in the conditional variance

It may be considered that adding the corresponding dummy to the variance equation would provide a solution:

$$\begin{aligned} y_t &= x_t'\zeta + \gamma d_t + \varepsilon_t, \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \tau d_t, \end{aligned}$$

where as before $d_t = 1$ when $t = s$, and zero otherwise. In the ARCH(1) case, an analytical solution immediately follows from (7) in Appendix 1:

$$\hat{\tau} = \varepsilon_s^2 - \alpha_0 - \alpha_1 \varepsilon_{s-1}^2.$$

When γ is such that $\varepsilon_s = 0$, then $h_s = 0$ and the log-likelihood is minus infinity. Consequently, the maximum likelihood estimate does not exist when $G_s \leq 0$. When $G_s > 0$, bimodality is still present, as Table 2.2 shows. In this case, the outer-product of gradients can not be used to estimate the variance-covariance matrix, because singularity is induced by $\sum_t (\partial \ell_t(\theta)/\partial \gamma)^2 = 0$. The last line in the table has another striking illustration of the fact that two different t-values are obtained. In this case the t-value and likelihood-ratio test also conflict when the dummy only enters in the variance equation: the t-value has a p-value of 49%, while that of the LR test is around $10^{-5}\%$.

Table 1: ARCH(1) model for log-returns on Dow Jones with a constant in the mean. Dummy variable for the 1987 crash entering in all possible ways.

Dummy in mean equation						in variance		log-likelihood
$\hat{\gamma}_{1,s}$	$t_{\hat{\gamma}_{1,s}}$	$t^{BW}_{\hat{\gamma}_{1,s}}$	$\hat{\gamma}_{2,s}$	$t^{BW}_{\hat{\gamma}_{2,s}}$	$t_{\hat{\gamma}_{2,s}}$	$\hat{\tau}$	$t_{\hat{\tau}}$	$\ell(\hat{\theta})$
–	–		–	–		–	–	1897.5743
–	–		–	–		0.0302	0.69	1910.9354
−0.244	−13.7	−3.1	−0.108	−6.1	−1.4	–	–	1907.9715
−0.350	−3.50		−0.0017	−0.017		0.0296	0.53	1910.9355

The t-values are based on the Hessian matrix; t^{BW} of Bollerslev and Wooldridge (1992) type.

2.3 Dummy variable in GARCH and EGARCH models

Figure 3 shows that a dummy variable in the mean can also cause bimodality for both GARCH and EGARCH models. As before, we plot the likelihood grid as a function of γ, with the other parameters kept fix at the value found at the maximum.

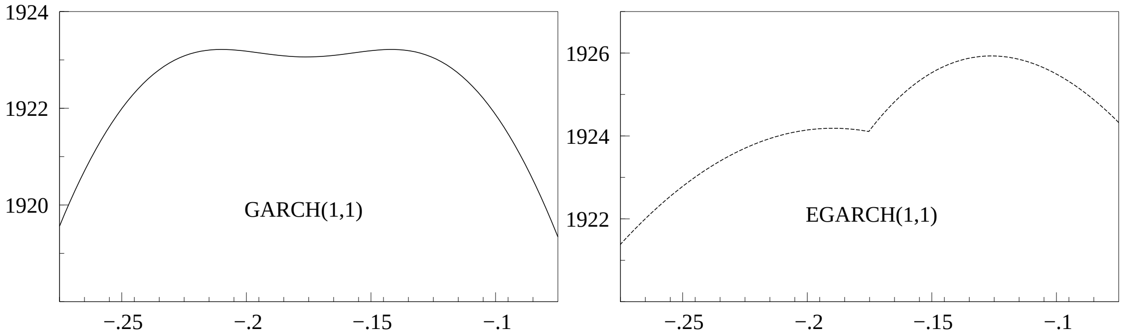

Figure 3: Likelihood grid for the dummy parameter corresponding to the 1987 crash, GARCH(1,1) (left) and EGARCH(1,1) (right).

For EGARCH, the two maxima are at different likelihood values, owing to the asymmetry term (when $\vartheta_1 = 0$ in (3) both modes are at the same likelihood value). Because of the absolute value, the local minimum is at a point where the likelihood is non-differentiable. In this case it matters which of the two maxima is found.

We also estimated a GARCH(1,1) model with the same dummy variable both in the mean and in the variance equation. When trying all possibly dummies, we did not encounter bimodality, but found that over 50% of cases did not converge.

Figure 4 plots $\hat{\gamma}_{2,s} - \hat{\gamma}_{1,s}$ for the GARCH(1,1) and EGARCH(1,1) models. Now there are about 30 cases with two modes in the likelihood.

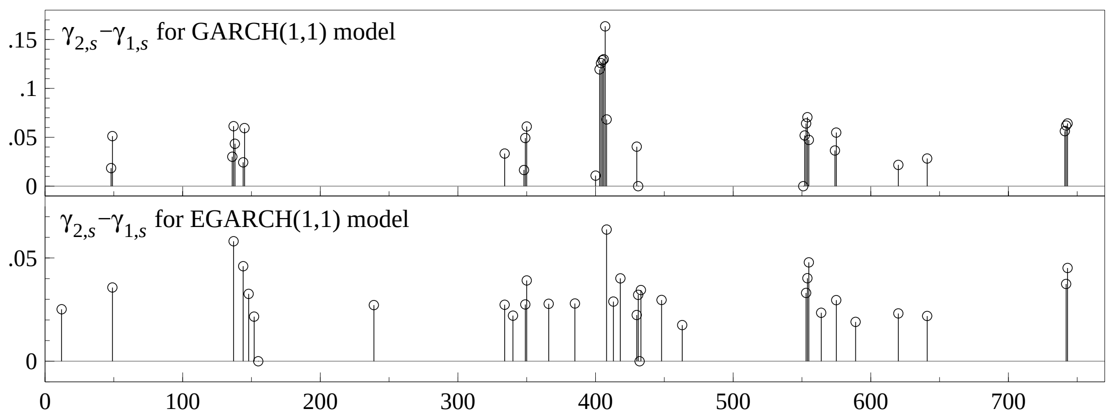

Figure 4: Estimates of $\hat{\gamma}_{2,s} - \hat{\gamma}_{1,s}$ for the GARCH(1,1) model (top) and the EGARCH(1,1) model (bottom).

3 Parameter restrictions

In order to investigate the incidence of multimodality, it is important to know what restrictions are imposed on the parameter space. In practice, the GARCH model is often estimated without restrictions, but Bollerslev (1986) formulated the model with $\alpha_0 > 0$, and the remaining parameters nonnegative.

Nelson and Cao (1992) argued that imposing all coefficients to be nonnegative is overly restrictive, and that negative estimates occur in practice (they list several examples). Subsequently, He and Teräsvirta (1999) have shown that such negative coefficients allow for richer shapes of the autocorrelation function. Nelson and Cao (1992) gave sufficient conditions such that the conditional variance is always nonnegative for the GARCH$(1, q)$, and GARCH$(2, q)$ case.[4]

Define the lag polynomials $\beta(L) = 1 - \sum_{i=1}^{p} \beta_i L^i$, and $\alpha(L) = \sum_{i=1}^{q} \alpha_i L^i$. The restrictions are imposed in the ARCH(∞) form:

$$h_t = \beta(L)^{-1} \left(\alpha_0 + \alpha(L)\varepsilon^2\right) = \alpha_0^* + \sum_{i=1}^{\infty} \delta_i \varepsilon_{t-i}^2, \tag{4}$$

such that $\alpha_0^* = \alpha_0/\beta(1) > 0$ and $\delta_i \geq 0 \ \forall i$. This requires that the roots of $\beta(z) = 0$ lie outside the unit circle. Furthermore, $\beta(z)$ and $\alpha(z)$ are assumed to have no common roots.

In Appendix 3 we refine the conditions for the GARCH$(p = 2, q)$ case by removing redundant conditions. Table 2 summarizes the restrictions for low-order GARCH models. The conditions on the roots when $p = 2$ can also be expressed as $\beta_2 + \beta_1 < 1$, $\beta_1^2 + 4\beta_2 \geq 0$. The restriction for GARCH(2,2) which is unnecessary is $\beta_1(\alpha_2 + \beta_1\alpha_1) + \alpha_1 \geq 0$; also $\alpha_0^* > 0$ reduces to $\alpha_0 > 0$.[5] In addition, Appendix 3 shows how the restrictions can be imposed by parameter transformations for $p \leq 2$, which allows implementation in the form of unconstrained optimization.

Table 2: Nelson & Cao conditions for some GARCH models

GARCH(1,1)	$\alpha_0 > 0,\ \alpha_1 \geq 0$	$0 \leq \rho_1 < 1.$	
GARCH(1,2)	$\alpha_0 > 0,\ \alpha_1 \geq 0$	$0 \leq \rho_1 < 1$	$\alpha_2 + \rho_1\alpha_1 \geq 0$
GARCH(2,1)	$\alpha_0 > 0,\ \alpha_1 \geq 0$	$0 \leq \lvert\rho_2\rvert \leq \rho_1 < 1$, ρ_1, ρ_2 real	
GARCH(2,2)	$\alpha_0 > 0,\ \alpha_1 \geq 0$	$0 \leq \lvert\rho_2\rvert \leq \rho_1 < 1$, ρ_1, ρ_2 real	$\alpha_2 + (\rho_1 + \rho_2)\alpha_1 \geq 0$
			$\alpha_2 + \rho_1\alpha_1 > 0$

Notes: $p = 1$: $\beta(L) = (1 - \rho_1 L),\ \beta_1 = \rho_1$;
$p = 2$: $\beta(L) = (1 - \rho_1 L)(1 - \rho_2 L),\ \beta_1 = \rho_1 + \rho_2,\ \beta_2 = -\rho_1\rho_2$.
$\alpha(L)$ and $\beta(L)$ have no common roots; ρ_1 is largest absolute (inverse) root.

It could be argued that even the Nelson and Cao (1992) conditions are too restrictive.[6] For example, the restrictions imply $h_t \geq \alpha_0^*$. Also, when the initial δ_i are positive and dominate the coefficients

[4] Instead of nonnegative h_t, we use positive; when h_t is zero, the log-likelihood is minus infinity.

[5] This slightly simplifies the derivations in the Appendix of Engle and Lee (1999), where, in a component GARCH(1,1) model, the component (which itself follows a GARCH(2,2) process) is shown to be positive.

[6] This point was also made by Drost and Nijman (1993).

at higher lags, the probability of obtaining a negative conditional variance becomes essentially zero. This is coupled with the fact that the constraints are very complex for higher order models. Therefore we suggest another set of constraints which relax the positivity restrictions, but are easier to implement and interpret. Defining $m = \max(p, q)$, $\beta_i = 0$ for $i > p$, $\alpha_i = 0$ for $i > q$:

$$\begin{aligned} \alpha_0 &> 0, \\ \alpha_i + \beta_i &\geq 0, \qquad \text{for } i = 1, \ldots, m. \\ 0 < \textstyle\sum_{i=1}^{m} \alpha_i + \beta_i &< 1. \end{aligned} \tag{5}$$

In terms of (2), these restrictions imply that the unconditional variance exists, and is always positive. Note that estimation automatically ensures that in-sample values of h_t are positive, otherwise the log-likelihood would be minus infinity or undefined. The restrictions (5) could be combined with imposing invertibility of $\beta(L)$.

4 Searching for multiple modes

Section 2 showed how a dummy variable can induce multimodality. It may be that, when the mean only consists of a constant term, multimodality is not likely to occur. We have not found much discussion of this issue in the literature.We consider the following parameter restrictions:

(UNR) Unrestricted: $\alpha_0 > 0$;

(N&C) Positive conditional variance: conditions (DO1)–(DO4) as explained in Appendix 3.

(UV) Positive and finite unconditional variance: restrictions (5) as explained in Appendix 4.

(POS) All coefficients positive: $\alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0$;

The choice of restrictions will affect the outcome: restricting the parameter space may reduce the number of modes, but could also introduce additional solutions on the boundary of the parameter space.

To look for multimodality, we estimate a GARCH model, giving parameter estimates $\hat{\theta}$ (say). We then re-estimate with $\hat{\theta} + \epsilon$ as starting values, with ϵ drawn from the standard normal distribution. In case restrictions are imposed, the transformed parameters are randomized, which keeps the new starting values within the constraints. We sample starting value until 250 GARCH models have been successfully estimated. If any local solutions are found, the models are then re-estimated to look at specific properties. For example, the second derivative at the solution must be negative definite for a local maximum.

Initially, we look at a GARCH(2,2) model for a short sample of 500 observations, from 7-Jun-1973 to 9-Jun-1975, of the British pound to US dollar daily exchange rate.[7] Next, we use a sample of 2915 observations (7-Jun-1973 to 28-Jan-1985), which is similar to some of the estimations in Nelson and Cao (1992).

[7]The data source is: Federal Reserve Statistical Release H.10, available on the web from www.frbchi.org/econinfo /finance/for-exchange/welcome.html

Table 3 shows the solutions which were found to the GARCH(2,2) likelihood maximization problem at sample size 500, estimating the model 250 times with random starting values. The first column lists the obtained log-likelihood value. The next four then indicate under which set of restrictions that particular solution was found. We see, for example, that in unrestricted estimation we found -202.85 in 85% of the 250 successful estimations. In a small number of cases, a higher likelihood was obtained. The final three columns give an indication of the properties of the solution; e.g. $\pi_1 < 0$ rules this solution out from UV, while $\rho_1 = -1.001$ violates N&C. In the case of all positive parameters (POS), 28% of the solutions are not listed in the table. These converged to a likelihood which was far removed from the optimal solution.

Table 3: Likelihood values at located maxima for GARCH$(2,2)$ models for growth rates of £/$ daily exchange rates at sample size 500. Based on 250 model estimates from random starting values.

log-like-	Parameter restrictions				Properties of solution		
lihood	UNR	N&C	UV	POS	$0 < \pi_i < 1$	$S < 1$	$0 < \rho_1 < 1$
-201.33	6%				$\hat{\pi}_1 < 0$		$\hat{\rho}_1 = -1.001$
-202.85	**82%**				$\hat{\pi}_2 < 0$		
-202.88		**99%**	**99%**		$\hat{\pi}_2 = 0$		
-203.06			1%		$\hat{\pi}_2 = 0$	$\hat{S} = 1$	
-203.96				**72%**			
-205.11	11%						$\hat{\rho}_1 = -0.9$

$\pi_i = \alpha_i + \beta_i$, $S = \sum_i \alpha_i + \beta_i$, ρ_1 is largest absolute root of $\beta(z^{-1})$

Table 5 illustrates that the multimodality does not disappear at larger sample size. For GARCH(1,1), GARCH(2,1), and GARCH(1,2) we found no multimodality. However, for higher order models, we did find multiple solutions. In the unrestricted case in particular, the random search delivered coinsiderably higher likelihoods. Testing down the lag length is problematic: it can easily happen that a sequence of nested hypotheses is not nested in terms of likelihood values. This would be an obvious sign of trouble. For the other cases, the solutions are very close in terms of the log-likelihood.

Each parameterization selects a different model according to the AIC criterion, see Table 4.

Table 4: GARCH model selected by AIC, for GARCH($p \leq 3, q \leq 3$).

	$T = 2915^{*}$	$T = 2915^{**}$
unrestricted	$(3,2)$	$(2,3)$
Positive conditional variance	$(3,2)$	$(3,2)$
Positive and finite unconditional variance	$(2,2)$	$(2,2)$
All coefficients positive	$(2,1)$	$(2,1)$

* is outcome using most commonly found solution.

**is outcome using best solution.

Table 5: Likelihood values at located maxima for growth rates of £/$ daily exchange rates at sample size 2915. Based on 250 model estimates from random starting values.

log-like-lihood	Parameter restrictions				Properties of solution		
	UNR	N&C	UV	POS	$0 < \pi_i < 1$	$S < 1$	$0 < \rho_1 < 1$
GARCH(1,1)							
-2147.17	**100%**	**100%**	**100%**	**86%**			
GARCH(1,2)							
-2142.32	**100%**	**100%**			$\hat{\alpha}_2 < 0$		
-2147.17			**100%**	**70%**			
GARCH(2,1)							
-2143.86	**100%**	**100%**	**100%**	**84%**			
GARCH(2,2)							
-2113.10	7%				$\hat{\pi}_1 < 0$		$\hat{\rho}_1 = -1.0014$
-2134.78	0.5%					$\hat{S} = 1.014$	$\hat{\rho}_1 = -0.998$
-2139.03			6%		$\hat{\pi}_1 = 0$	$\hat{S} = 1$	$\hat{\rho}_1 = -0.998$
-2142.56	**92%**	**100%**	**90%**				
-2143.86				**71%**	$\hat{\alpha}_2 = 0$		
-2144.94			4%		$\hat{\pi}_2 = 0$	$\hat{S} = 1$	$\hat{\rho}_1 = -0.998$
GARCH(2,3)							
-2095.92	16%				$\hat{\pi}_3 < 0$	$\hat{S} = 1.01$	$\hat{\rho}_1 = -1.001$
-2112.66	2%				$\hat{\pi}_1 < 0, \hat{\beta}_3 < 0$	$S = 1.016$	$\hat{\rho}_1 = -0.999$
-2139.03			6%		$\hat{\pi}_1 = 0$	$S = 1$	$\hat{\rho}_1 = -0.998$
-2141.31	**82%**	**100%**			$\hat{\pi}_3 < 0$		
-2142.56			**94%**		$\hat{\alpha}_3 = 0$		
-2143.86				**78%**	$\hat{\alpha}_2 = \hat{\alpha}_3 = 0$		
GARCH(3,2)							
-2099.76	0.5%				$\hat{\pi}_1 > 1, \hat{\beta}_3 < 0$		
-2102.73	1.5%				$\hat{\pi}_1 < 0, \hat{\beta}_3 < 0$	$\hat{S} = 1.004$	$\hat{\rho}_1 = -1$
-2133.77	**66%**	**75%**			$\hat{\pi}_2 < 0, \hat{\pi}_1 > 1$		
-2141.05	32%				$\hat{\pi}_2 < 0$		
-2142.30				**99%**	$\hat{\alpha}_2 = \hat{\beta}_2 = 0$		
-2142.33		25%	**97%**				
-2144.65			2.5%		$\pi_2 = 0$		

$\pi_i = \alpha_i + \beta_i$, $S = \sum_i \alpha_i + \beta_i$, ρ_1 is largest absolute root of $\beta(z^{-1})$.

5 Planned extensions

Recently, Gan and Jiang (1999) re-interpreted White (1982)'s information matrix test as a test for a global maximum. Although there are a couple of potential problems (the GARCH model may be used in a QML setting, as well as the notoriously bad size properties of the information-matrix test), we intend to investigate its usefulness for GARCH models.

In addition, we are looking at autoregressive conditional-duration models (Engle and Russell, 1998), which have a close similarity to GARCH models.

6 Conclusion

We found that inclusion of a dummy variable in the mean equation of a GARCH regression model could lead to multimodality in the likelihood. We believe that this curiosity, while of empirical relevance, has not yet been noted in the literature.

This finding has important consequences for empirical modelling. Firstly, a t-test on the coefficient of a dummy variable cannot be used in GARCH regression models. When there are two maxima, at $\hat{\gamma}_{1,s}$ and $\hat{\gamma}_{2,s}$, they will both have the same estimated standard errors, and hence potentially very different t-values. Consequently, it is possible that one is significant, and the other insignificant. Secondly, all model statistics which involve the value of the dummy are affected. Next, we noted that with only dummies as regressors, standard software may find a local minimum of the likelihood. Finally, asymptotic likelihood theory is affected by this violation of the regularity conditions.

We considered several types of restrictions on the GARCH parameters. In particular, we presented a small refinement to the Nelson & Cao constraints, and showed how these can be made operational within an unconstrained maximization setting. We also suggested

We have shown that multimodality of the GARCH likelihood is of practical relevance. It is likely that many applied results have been published without the authors being aware of the possibility of multiple modes. Our results indicate that, especially when going beyond the GARCH(1,1) model, a search for local maxima is important. We have also investigated the role of different restrictions of the parameter space. Unrestricted estimation is especially likely to show multimodality (for example with a unit root in the β lag-polynomial, or with the sum of the coefficients greater than one). However, no set of restrictions is clearly better.

Acknowledgements

We wish to thank Neil Shephard and David Hendry for helpful discussions and suggestions. Financial support from the UK Economic and Social Research Council (grant R000237500) is gratefully acknowledged by JAD.

Appendix 1 Proof of Theorem 1

The log-likelihood of (1) is given by:

$$\ell(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = c - \frac{1}{2} \sum_{t=1}^{T} \left(\log(h_t) + \frac{\varepsilon_t^2}{h_t} \right). \tag{6}$$

Assuming that the start-up of the recursive process does not depend on the parameters:

$$\frac{\partial \ell_t(\theta)}{\partial \theta} = -\frac{\varepsilon_t}{h_t} \frac{\partial \varepsilon_t}{\partial \theta} - \frac{1}{2} \frac{1}{h_t^2} \left(h_t - \varepsilon_t^2 \right) \frac{\partial h_t}{\partial \theta}. \tag{7}$$

In this case $\varepsilon_t = y_t - x_t' \zeta - d_t \gamma$ and $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$, so:

$$\frac{\partial \varepsilon_t}{\partial \gamma} = -d_t, \; \frac{\partial h_t}{\partial \gamma} = -2\alpha_1 \varepsilon_{t-1} d_{t-1},$$

Since $d_t = 0$ for $t \neq s$ and $d_s = 1$, the score with respect to γ is:

$$\frac{\partial \ell(\theta)}{\partial \gamma} = \frac{\varepsilon_s}{h_s} + \frac{1}{h_{s+1}^2} \left(h_{s+1} - \varepsilon_{s+1}^2 \right) \alpha_1 \varepsilon_s. \tag{8}$$

Finding the zeros of this expression gives $\varepsilon_s = 0$, with the remaining zeros found from:

$$h_{s+1}^2 + h_s \alpha_1 \left(h_{s+1} - \varepsilon_{s+1}^2 \right) = 0. \tag{9}$$

Solving the quadratic in h_{s+1} gives:

$$\tilde{h}_{s+1} = \frac{1}{2} h_s \alpha_1 \left[-1 \pm \left(1 + \frac{4\varepsilon_{s+1}^2}{\alpha_1 h_s} \right)^{1/2} \right].$$

Since $\tilde{h}_{s+1}$ must be positive, the negative term can be dropped. In terms of ε_s the additional solutions to $\partial \ell(\theta)/\partial \gamma = 0$ can be written as

$$\tilde{\varepsilon}_s^2 = \frac{1}{2} h_s \left[\left(1 + \frac{4\varepsilon_{s+1}^2}{\alpha_1 h_s} \right)^{1/2} - 1 \right] - \frac{\alpha_0}{\alpha_1} = G_s.$$

If the G_s does not have a positive value, then the only solution is $\varepsilon_s = 0$, with $\hat{\gamma} = y_s - x_s' \zeta$. Otherwise the additional two solutions are $\tilde{\gamma} = y_s - x_s' \zeta \pm G_s^{1/2}$. In that case, the likelihood and its derivatives are identical for both values.

The second derivative of the log-likelihood with respect to γ is:

$$\frac{\partial^2 \ell(\theta)}{(\partial \gamma)^2} = \frac{4\alpha_1^2 \varepsilon_s^2}{h_{s+1}^2} \left(\frac{1}{2} - \frac{\varepsilon_{s+1}^2}{h_{s+1}} \right) - \frac{1}{h_s} - \frac{\alpha_1}{h_{s+1}^2} \left(h_{s+1} - \varepsilon_{s+1}^2 \right). \tag{10}$$

Consider $\varepsilon_s = 0$. In that case the second derivative is block diagonal with respect to γ: all terms in the derivative of (8) w.r.t. the ARCH parameters involve ε_s. The first term in (10) drops out; the remaining term is equal to (9) divided by $-h_s h_{s+1}$. If $G_s \leq 0$, (9) has no feasible solution, and is always positive. This makes the Hessian element negative, required for a maximum. If $G_s > 0$ and $\varepsilon_s = 0$, h_{s+1} is at its minimum, where (9) takes on negative values. This creates a positive diagonal element in the Hessian, violating the conditions for a maximum. □

Appendix 2 Implementing the GARCH likelihood

Implementation of the GARCH likelihood involves several decisions, often only summarily discussed in the literature:

1. How to select initial values for the variance recursion;
   Evaluation of the likelihood requires presample values for ε_t^2 and h_t. Bollerslev (1986) suggested to use the mean of the squared residuals:
   $$\varepsilon_i^2 = h_i = T^{-1} \sum_{t=1}^{T} \varepsilon_t^2, \text{ for } i \leq 0. \tag{11}$$
   An alternative is to use the recursion (2): since u_t has mean zero, it can be started up from $\varepsilon_1^2 \dots \varepsilon_m^2$. In that case, the likelihood is evaluated from $t = m+1$ onwards, conditional on the m presample values; the first term then is:
   $$h_{m+1} = \alpha_0 + \sum_{i=1}^{m} (\alpha_i + \beta_i) \varepsilon_{m+1-i}^2.$$
   Other methods include: adding the missing $\varepsilon_1^2 \dots \varepsilon_m^2$ as parameters which are to be estimated, using the unconditional variance provided it exists, backwards forecasting of the missing values; all these make the likelihood derivatives more complex.
2. Which restrictions to impose;
   Bollerslev (1986) proposed the GARCH model with $\alpha_0 > 0$, $\alpha_i \geq 0$, and $\beta_i \geq 0$. This ensures that $h_t > 0$, and can easily be implemented. Let $\phi_0, \dots, \phi_{q+p}$ be the parameters used in estimation, then $\alpha_0, \alpha_1, \dots, \beta_p = e^{\phi_0}, \dots, e^{\phi_{q+p}}$ will ensure that all coefficients are positive. The Jacobian matrix of this transformation is dg$(\alpha_0, \alpha_1, \dots, \beta_p)$. More general formulations are discussed in §3, and below.
3. Which maximization technique to use;
   We prefer BFGS (see e.g. Fletcher, 1987 or Gill, Murray, and Wright, 1981). This avoids the need for second derivatives, while being one of the most robust methods available. This is supplemented by a line search when, at an iteration step, the likelihood does not increase. BFGS was not considered by Fiorentini, Calzolari, and Panattoni (1996), but we found 100% convergence when replicating their Table 1 with 1000 replications (requiring about 17 iterations on average, whether starting from the DGP values, or from a starting value routine).
4. How to compute starting values fot the parameters;
   We apply the method of Galbraith and Zinde-Walsh (1997) applied to the squared data (after removing regressors in the mean). If necessary, the resulting parameter values are reduced to enforce the unconditional variance to exist.
5. Whether to use numerical or analytical derivatives;
   Numerical derivatives are more convenient, but less accurate than analytical derivatives (see Fiorentini, Calzolari, and Panattoni, 1996). The latter are to be preferred, but convenience

often dictates the use of the former. In simple GARCH models, we found numerical derivatives sufficiently effective, with model estimation taking the same amount of time, and convergence achieved as frequently. All estimates in this paper use analytical derivatives, except when the Hessian matrix is required for the variance-covariance matrix, and for EGARCH-type models.

6. Which estimate of the variance-covariance matrix to use.
   A comparison of various estimators is given in Fiorentini, Calzolari, and Panattoni, 1996.

Appendix 3 Positive conditional variance

Nelson and Cao (1992) (hereafter NC) formulated conditions so that the coefficients in (4) are always positive. The conditions, expressed in terms of the lag polynomials $\beta(L)$ and $\alpha(L)$, require that the roots of $\beta(z) = \prod_{i=1}^{p}(1-\rho_i z) = 0$ lie outside the unit circle. Furthermore, $\beta(z)$ and $\alpha(z)$ are assumed to have no common roots. The δ_i in (4) can be derived recursively for $i = 1, 2, \dots$:

$$\begin{aligned} &\delta_i = 0, && i < 1, \\ &\delta_i = \textstyle\sum_{j=1}^{p} \beta_j \delta_{i-j} + \alpha_i, && i \le q, \\ &\delta_i = \textstyle\sum_{j=1}^{p} \beta_j \delta_{i-j}, && i > q. \end{aligned} \tag{12}$$

So $\delta_1 = \alpha_1$.

GARCH($\le 2, q$) case

The necessary and sufficient conditions for $\delta_i \ge 0 \ \forall i$ for the GARCH($2, q$) case are:

$$\begin{aligned} \alpha_0 > 0; & \qquad \text{(DO1)} \\ 0 < \rho_1 < 1, \quad \rho_1 \text{ is real}; & \qquad \text{(DO2.1)} \\ |\rho_2| \le \rho_1, \quad \rho_2 \text{ is real}, & \qquad \text{(DO2.2)} \\ \delta_i \ge 0,\ i = 1, \dots, q; & \qquad \text{(DO3)} \\ \textstyle\sum_{j=1}^{q} \rho_1^{q-j} \alpha_j > 0. & \qquad \text{(DO4)} \end{aligned}$$

NC Theorem 2 gives these conditions as:

$$\begin{aligned} \alpha_0^* > 0; & \qquad \text{(NC1)} \\ 0 < \rho_1, \quad \rho_1, \rho_2 \text{ are real}; & \qquad \text{(NC2)} \\ \delta_i \ge 0,\ i = 1, \dots, q; & \qquad \text{(NC3.1)} \\ \delta_{q+1} \ge 0; & \qquad \text{(NC3.2)} \\ \textstyle\sum_{j=1}^{q} \rho_1^{1-j} \alpha_j > 0. & \qquad \text{(NC4)} \end{aligned}$$

Where it is assumed that $|\rho_2| \le |\rho_1|$ without loss of generality. In the next theorem we show that these two sets of conditions are identical.

Theorem 2 *Conditions (NC1)–(NC3.2) and (DO1)–(DO3) are equivalent when* $|\rho_2| \le |\rho_1| < 1$.

Proof (DO2.1) and (DO2.2) combine (NC2) with the assumption that $\beta(L)$ is invertible, and ρ_1 is the largest root in absolute value. Next, (DO2.x) imply that $\beta(1) = 1 - \rho_1 - \rho_2 + \rho_1\rho_2 > 0$, reducing (NC1) to (DO1).

To see that (NC3.2) is redundant when ρ_2 is negative use

$$\delta_{q+1} = \beta_1\delta_q + \beta_2\delta_{q-1} = (\rho_1 + \rho_2)\delta_q - \rho_1\rho_2\delta_{q-1},$$

and $\delta_{q+1} \geq 0$ follows from (NC3.1) and $0 < -\rho_2 \leq \rho_1$.
If the roots are real and distinct (NC equation A.9):

$$\delta_i = (\rho_1 - \rho_2)^{-1} \sum_{j=1}^{\min(i,q)} \left(\rho_1^{1+i-j} - \rho_2^{1+i-j}\right)\alpha_j, \quad i = 1, \ldots .$$

Writing $a_i = \sum_{j=1}^{\min(i,q)} \rho_1^{1-j}\alpha_j$ and $b_i = \sum_{j=1}^{\min(i,q)} \rho_2^{1-j}\alpha_j$:

$$\delta_i^* = \delta_i\,(\rho_1 - \rho_2) = \rho_1^i a_i - \rho_2^i b_i.$$

Then $\delta_q^* \geq 0$ and $\rho_2 > 0$ implies $\rho_2\rho_1^q a_q \geq \rho_2^{q+1} b_q$. Combining this with (NC4), which is $a_q > 0$:

$$\delta_{q+1}^* = \rho_1^{q+1}a_q - \rho_2^{q+1}b_q \geq \rho_1^{q+1}a_q - \rho_2\rho_1^q a_q = \rho_1^q a_q\,(\rho_1 - \rho_2) \geq 0.$$

When the roots are equal, $\rho_1 = \rho_2 = \rho > 0$ (NC equation A.6):

$$\delta_i = \sum_{j=1}^{\min(i,q)} (1 + i - j)\rho^{1+i-j}\alpha_j, \quad i = 1, \ldots .$$

So

$$\rho^{-1}\delta_{q+1} = \sum_{j=1}^{q} \rho^{1+q-j}(1 + q - j)\alpha_j + \sum_{j=1}^{q} \rho^{1+q-j}\alpha_j = \delta_q + \rho^{-q}a_q,$$

which is positive by (NC4) and (NC3.1). □

(DO1)–(DO4) has one restriction more than the number of parameters. However, ρ_1^{q-1}(NC4) = (DO4) is not always binding. For example, when $q = 1$, it is automatically satisfied. In the GARCH(2,2) case:

$$\begin{array}{ll} \rho_1\alpha_1 + \alpha_2 > 0, & \text{(NC4)}, \\ (\rho_1 + \rho_2)\alpha_1 + \alpha_2 > 0, & \text{from } \delta_q \text{ in (12).} \end{array}$$

When ρ_2 is negative (making β_2 positive), the first restriction is not binding.

The set of restrictions can implemented by transformation when (DO4) and $\delta_q \geq 0$ are combined in one restriction, obviating the need for constrained estimation. The conditions

$$\begin{array}{l} \sum_{j=1}^{p} \beta_j\delta_{q-j} + \alpha_q > 0, \\ \sum_{j=1}^{q-1} \rho_1^{q-j}\alpha_j + \alpha_q > 0, \end{array}$$

are both satisfied when α_q is sufficiently large. Therefore, we estimate the product as a parameter $\exp(\phi_q)$ which is always positive, and take α_q as the largest root.

To restrict any coefficient between $-\rho$ and ρ we can use:[8]

$$x = \rho \frac{1 - e^{\phi}}{1 + e^{\phi}},\ -\rho < x < \rho \quad \Leftrightarrow \quad \phi = \log\left(\frac{1 - x/\rho}{1 + x/\rho}\right),\ -\infty < \phi < \infty.$$

See Marriott and Smith (1992) for the application of such Fisher-type transformations to impose stationarity in ARMA models.

The restrictions can be implemented as follows. Let $\phi_0, \phi_1 \ldots, \phi_q, \varphi_1, \varphi_2$ be the unrestricted parameters. Then:

$$\begin{array}{ll}
(a) & \alpha_0 = \exp(\phi_0), \\
(b) & \rho_1 = \dfrac{\exp(\varphi_1)}{1 + \exp(\varphi_1)},\ \rho_2 = \rho_1 \dfrac{1 - \exp(\varphi_2)}{1 + \exp(\varphi_2)}, \\
(c) & \beta_1 = \rho_1 + \rho_2,\ \beta_2 = -\rho_1\rho_2, \\
(d) & \alpha_i = \delta_i - \sum_{j=1}^{p} \beta_j \delta_{i-j} \text{ using } \delta_i = \exp(\phi_i) \text{ for } 1 \le i \le q-1,\ \delta_i = 0 \text{ for } i < 1, \\
(e) & \alpha_q = -\frac{1}{2}(x + y) + \frac{1}{2}\left[(x - y)^2 + 4\exp(\phi_q)\right]^{1/2}, \quad x = \sum_{j=1}^{p} \beta_j \delta_{q-j},\ y = \sum_{j=1}^{q-1} \rho_1^{q-j} \alpha_j.
\end{array}$$

This transformation imposes the necessary and sufficient conditions for GARCH($\le 2, q$) models.

As NC point out, starting the recursion with the sample mean (11) will ensure positive conditional variance. This is not necessarily the case when conditioning on initial values.

Appendix 4 Positive and finite unconditional variance

Estimation under restrictions (5) is achieved by transforming the GARCH parameters. Write $\pi_i = \alpha_i + \beta_i$, and s_i for the partial sums: $s_i = \sum_{j=1}^{i} \pi_j$. The restrictions imply that $0 < s_1 \le s_2 \cdots \le s_m < 1$, $m = \max(p, q)$. This can be implemented by introducing $0 < \theta_i < 1$:

$$\sum_{i=1}^{k} \pi_i = \prod_{i=1}^{m+1-k} \theta_i.$$

For example, for $m = 3$:

$$\begin{array}{lcr}
\pi_1 & = & \theta_1\theta_2\theta_3, \\
\pi_1 + \pi_2 & = & \theta_1\theta_2, \\
\pi_1 + \pi_2 + \pi_3 & = & \theta_1.
\end{array}$$

An unrestricted parameter ϕ is mapped to $(0, 1)$ using $\theta_i = [1 + \exp(-\phi)]^{-1}$.

If the unconstrained version is $\theta_u = \alpha_0, \pi_1, \ldots, \pi_m, \beta_1, \ldots, \beta_n$, $n = \min(p, q)$, and the transformed parameterization $\phi = \log \alpha_0, \phi_1, \ldots, \phi_m, \beta_1, \ldots, \beta_n$, using $\phi_i = \log[\theta_1/(1 - \theta_1)]$, then the

[8]Numerically, it is better to use $\frac{1 - e^{\phi}}{1 + e^{\phi}}$ when $\phi \le 0$, and $\frac{e^{-\phi} - 1}{e^{-\phi} + 1}$ otherwise. This prevents overflow when evaluating the exponential.

Jacobian matrix can be used to move backwards and forwards. For example, when $m = 3$:

$$\frac{\partial \theta}{\partial \pi'} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (\pi_1 + \pi_2 + \pi_3)^2 & 0 \\ 0 & 0 & (\pi_1 + \pi_2)^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 1 \\ \pi_3 & \pi_3 & -1 \\ \pi_2 & -1 & 0 \end{pmatrix},$$

and $\partial \phi_i / \partial \theta_i = [\phi_i(1 - \phi_i)]^{-1}$.

This allows the use of standard derivatives, as given in Fiorentini, Calzolari, and Panattoni (1996) for example. This representation also makes it easy to impose $S = 1$, which estimates the IGARCH(p, q) model.

References

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *J. Econometrics 51*, 307–327.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994). ARCH models. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 49. Amsterdam: North-Holland.

Bollerslev, T. and J. M. Wooldridge (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews 11*, 143–172. NL.

Drost, F. C. and T. E. Nijman (1993). Temporal aggregation of GARCH processes. *Econometrica 61*, 909–927.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflations. *Econometrica 50*, 987–1007.

Engle, R. F. and G. G. J. Lee (1999). A permanent and transitory component model of stock return volatility. In *Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger*, Chapter 20, pp. 475–497. Oxford: Oxford University Press.

Engle, R. F. and J. R. Russell (1998). Forecasting transaction rates: the autoregressive conditional duration model. *Econometrica 66*, 1127–1162.

Fiorentini, G., G. Calzolari, and L. Panattoni (1996). Analytic derivatives and the computation of GARCH estimates. *Journal of Applied Econometrics 11*, 399–417. [CO.NL.HS.

Fletcher, R. (1987). *Practical Methods of Optimization*, (2nd ed.). New York: John Wiley & Sons.

Galbraith, J. W. and V. Zinde-Walsh (1997). On some simple, autoregression-based estimation and identification techniques for ARMA models. *Biometrika 84*, 685–696.

Gan, L. and J. Jiang (1999). A test for global maximum. *Journal of the American Statistical Association 94*, 847–854.

Gill, P. E., W. Murray, and M. H. Wright (1981). *Practical Optimization*. New York: Academic Press.

Gourieroux, C. (1997). *ARCH Models and Financial Applications*. New York: Springer Verlag.

He, C. and T. Teräsvirta (1999). Properties of the autocorrelation function of squared observations for second-order GARCH processes under two sets of parameter constraints. *Journal of Time Series Analysis 20*, 23–30.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Business and Economic Statist. 12*, 371–417.

Marriott, J. M. and A. F. M. Smith (1992). Reparameterization aspects of numerical Bayesian methodology for autoregressivce moving-average models. *J. Time Series Analysis 13*, 327–43.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset pricing: a new approach. *Econometrica 59*, 347–370.

Nelson, D. B. and C. Q. Cao (1992). Inequality constraints in the univariate GARCH model. *Journal of Business and Economic Statistics 10*, 229–235.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Nielson (Eds.), *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman & Hall.

White, H. (1982). Instrumental variables regression with independent observations. *Econometrica 50*, 483–499.