Multimodality in the GARCH Regression Model

Jurgen A. Doornik*
Nuffield College, University of Oxford

Marius Ooms
Dept of Econometrics, Free University Amsterdam

October 9, 2003

Abstract

Several aspects of GARCH(p, q) models that are relevant for empirical applications are investigated. In particular, it is noted that the inclusion of dummy variables as regressors can lead to multimodality in the GARCH likelihood. This invalidates standard inference on the estimated coefficients. Next, the implementation of different restrictions on the GARCH parameter space is considered. A refinement to the Nelson and Cao (1992) conditions for a GARCH($2, q$) model is presented, and it is shown how these can then be implemented by parameter transformations. It is argued that these conditions may also be too restrictive, and a simpler alternative is introduced which is formulated in terms of the unconditional variance. Finally, examples show that multimodality is a real concern for models of the £/$ exchange rate, especially when $p \geq 2$.

Keywords: Dummy variable, EGARCH, GARCH, Multimodality.

JEL code: C22, C51.

*Correspondence to: Jurgen Doornik, Nuffield College, Oxford OX2 0LD, UK. Tel: +44-1865-278610. Email: jurgen.doornik@nuffield.ox.ac.uk

1 Introduction

The ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models have found widespread application since their introduction. Indeed, there have been so many publications involving GARCH models, that we expect that most users consider their estimation to be a routine operation. This paper should undermine that belief somewhat. Particular issues of practical relevance are multimodality of the likelihood, of which we shall give several examples, and adoption of restrictions on the parameter space — issues to which the literature has paid relatively little attention, despite the popularity of GARCH models.

We write the regression model with normal-GARCH(p, q) errors as:

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

where $\mathcal{F}_t$ 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).

At first sight, it would appear that variables entering the mean equation of a GARCH regression model do not seriously affect the properties of the model, and standard results for explanatory variables in linear dynamic regression models would apply. In this paper, however, 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. The correction for an additive outlier corresponds to treating one observation as missing. Surprisingly, this multimodality does not always happen. We provide analytical and empirical results in §2. Multimodality is more likely to occur when volatility, according to estimated GARCH parameters, is persistent and when dummies are added before or within volatile periods, i.e. precisely in those periods where they are considered most relevant. We show that the multimodality problem may remain when adding dummies that are nonzero for more than one period. Replacing a GARCH by an EGARCH specification does not remove the problem either. We do show in §2.5 that adding the corresponding dummy one period lagged in the variance equation can solve the problem of multimodality. Doornik and Ooms (2002) use this to implement a procedure for outlier detection in GARCH models.

Section 3 further investigates whether multimodality is of practical relevance, even without dummy variables in the mean equation. This requires us to be more specific about the model, in particular about possible restrictions on the GARCH parameter space. We present a refinement to the Nelson and Cao (1992) conditions, which relax the original Bollerslev (1986) positivity conditions, and show

how these can be implemented by parameter transformations. The major benefit of this is that we can estimate the model using standard unconstrained maximization, and that the original analytical derivatives can be used (see Fiorentini, Calzolari, and Panattoni, 1996), in combination with the Jacobian of the transformation. Because the Nelson and Cao (1992) constraints are very complex for higher order models, we suggest another set of constraints. These relax the positivity restrictions in a different way, and are easier to implement and interpret. We compare the impact of the different parameter sets in GARCH$(2,2)$ models. Using four choices of the parameter space, we then search for multimodality in samples from simulated GARCH$(2,2)$ processes and in an empirical application using daily British Pound/US Dollar exchange rates. We conclude that multimodality is a potential problem in applications, and recommend the adoption of a limited search using random starting values whenever estimating a higher-order GARCH model.

2 Multimodality caused by dummy variables

In a normal linear regression model, the effect of introducing a single dummy variable is to set the residual, $\widehat{\varepsilon}_s$, for that observation, y_s, to zero. The same effect is obtained by replacing y_s by $\widehat{y}_s$, its conditional expectation given all other observations: $\widehat{y}_s = E(y_s|y_1, \ldots, y_{s-1}, y_{s+1}, \ldots, y_T)$, and leaving all other values unchanged. Effectively, the observation is treated as missing and replaced by the ML estimate. Essentially, the same effect of introducing a dummy applies when ε_t follows a linear Gaussian time-series process, see Gómez, Maravall, and Peña (1999) for a systematic overview of this topic for ARMA processes. At first sight, it is not unreasonable to think that this also applies to a regression model with ARCH or GARCH errors. The next example, however, shows that this is not the case.

As an illustration, we use the Dow–Jones index (Dow Jones Industrial Average: close at midweek from Janary 1980 to September 1994, 770 observations in total); the figures are for Wednesday (or Tuesday if the stock market was closed on Wednesday; the data are from www.djindexes.com). The returns, $\log Y_t - \log Y_{t-1}$, are given in the top panel of Figure 1. The large negative return of -17.4% corresponds to the Black Monday crash of 19 October 1987.

We start by estimating an ARCH(1) model, where the mean equation consists of a constant and a dummy variable (or impulse intervention) for the 1987 crash:

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

where d_{crash} takes value one for the Wednesday after the crash, zero otherwise. Let $\widehat{c}, \widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\gamma}$ be

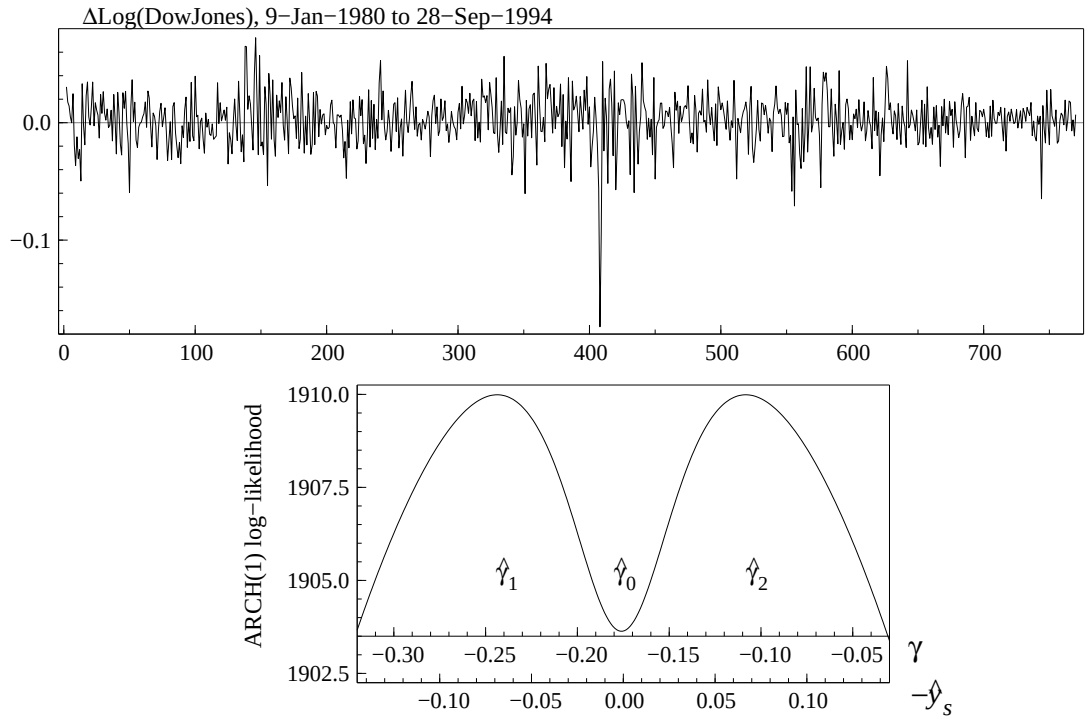

Figure 1: Log-returns on Dow–Jones index (top), with likelihood grid for the dummy parameter γ, corresponding to the 1987 crash (bottom); $-\widehat{y}_s = -(y_s - \gamma)$.

the maximum likelihood estimates, also see equation (2) below. The bottom panel of Figure 1 plots the log-likelihood values as a function of γ, with the remaining coefficients kept fixed at $\widehat{c}, \widehat{\alpha}_0, \widehat{\alpha}_1$. The figure shows a pronounced bimodal shape of the likelihood, with a local minimum at $\widehat{\gamma}_0$, and two maxima at $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ ($\widehat{\gamma} = \widehat{\gamma}_1 = \widehat{\gamma}_2$). The corresponding interpolated value is given on the lower horizontal axis of the bottom graph. Quite surprisingly, adding an ARCH term to a regression model with a dummy variable clearly changes the role of that variable. Table 1 provides details on the two maxima and single minimum.

Table 1: Extremes of the ARCH(1) likelihood from Figure 1b.

y_s	γ	$\widehat{y}_s(\gamma)$	$\widehat{\varepsilon}_s(\gamma)$	
-0.174	-0.244	-0.242	0.068	left mode, $\widehat{\gamma}_1$
-0.174	-0.176	-0.174	0	local minimum, $\widehat{\gamma}_0$
-0.174	-0.108	-0.106	-0.068	right mode, $\widehat{\gamma}_2$

Even in the simplest ARCH model the estimate for a missing observation does not always corre-

spond to its conditional expectation given the other observations. For interpolation in this case, $\widehat{y}_s$ is not determined by its expectation. An exceptional return, implicit in $\widehat{\gamma}_1$ or $\widehat{\gamma}_2$, can be more likely than an average return, implicit in $\widehat{\gamma}_0$. We provide an analytical explanation below.

2.1 GARCH models with a dummy variable in the mean

The following proposition explains the effect of the dummy variable for the GARCH(p, q) model.

Proposition 1 *Consider the GARCH(p, q) 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 $t = s, 1 < s < T$, and $d_t = 0$ otherwise. Define*

$$G_s = \frac{1}{\alpha_1}\left[h_{s+1} - \alpha_0 - \sum_{i=2}^{q}\alpha_i\varepsilon_{s+1-i}^2 - \sum_{i=1}^{p}\beta_i h_{s+1-i}\right].$$

(a) When $\widehat{G}_s = 0$ the log-likelihood $\ell(\theta)$ has a unique maximum *for γ:*

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

with $\widehat{\varepsilon}_s = 0$.

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

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

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

The role of G_s and the properties of the likelihood are discussed in the next section.

Proposition 1 indicates that the dummy variable does not always lead to multimodality. In the first case, $\widehat{\gamma} = y_s - x_s'\widehat{\zeta}$, and the dummy plays a similar role as in the linear regression model without GARCH errors. However, when G_s is positive at the maximum, there are two identical modes. The value of G_s depends on the parameter values and on past and future residuals. In an ARCH(1) model we can consider G_s^* (defined in (11) below) as a function of the parameters (i.e. not just evaluated at the values corresponding to the maximum). The next section then shows that negative G_s^* leads to one maximum, and positive to two. Also, G_s^* depends only on the residuals immediately after and before the time of the impulse and both $\partial G_s^*/\partial\varepsilon_{s-1}^2 > 0$ and $\partial G_s^*/\partial\varepsilon_{s+1}^2 > 0$. Proposition 1 states 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 two different t-values. The estimation procedure may pick either maximum, but deciding significance by looking at the t-value is problematic. Note that a dummy at the end of the sample cannot lead to multimodality.

When a dummy is included as regressor, standard econometric software may find the local minimum instead of one of the maxima: if the starting value for the dummy parameter (often determined by a prior regression) corresponds to the local minimum, the derivative is zero. Then, during subsequent iterations, the dummy coefficient will not move, and convergence is to the local minimum. This will show up when the standard error is computed, because the variance matrix is negative definite.

Bimodality leads to two residuals: $\widehat{\varepsilon}_{1,s} = \widehat{G}_s^{1/2}$ and $\widehat{\varepsilon}_{2,s} = -\widehat{G}_s^{1/2}$ corresponding to two $\widehat{y}_s$: $\widehat{y}_{1,s} = y_s - \widehat{G}_s^{1/2}$, $\widehat{y}_{2,s} = y_s + \widehat{G}_s^{1/2}$. In Table 1, the solution $\widehat{y}_{2,s}$ might be more appealing from an economic point of view, but this does not follow from the statistical model. Diagnostic tests based on the residuals (or standardized residuals: there is only one value for h_s) will have different outcomes, unless only the squared values are used.

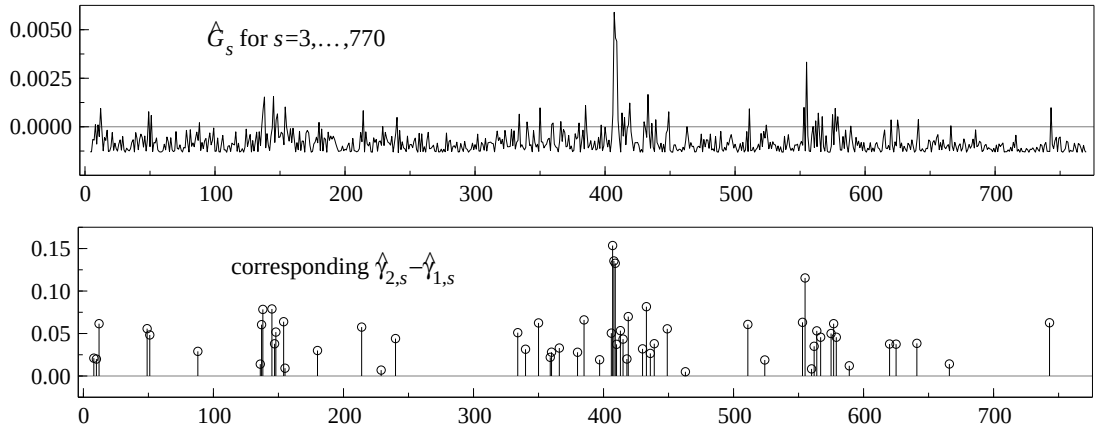

Figure 2: ARCH(1) model for growth rates of Dow-Jones with moving dummy variable: $\widehat{G}_s^*$ (top), $\widehat{\gamma}_{2,s} - \widehat{\gamma}_{1,s}$ (bottom).

To assess the relevance of Proposition 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 $t = s,\ s = 3, \ldots, 770$). Figure 2a plots the value of $\widehat{G}_s$ for the 768 estimated ARCH(1) models, with positive values indicating multiple maxima. There are 59 cases with $\widehat{G}_s > 0$, and correspondingly with two solutions for γ; the second graph displays the difference $\widehat{\gamma}_{2,s} - \widehat{\gamma}_{1,s} = 2\widehat{G}_s^{1/2}$. For the cases without multimodality there is only one estimate of γ and $\widehat{y}_s = \widehat{c} + \widehat{\gamma}$.

In Figure 3 we only consider the cases which have multimodality. The top graph shows the t-values when $\widehat{G}_s > 0$. Using a critical value of two, there are several cases with one t statistic

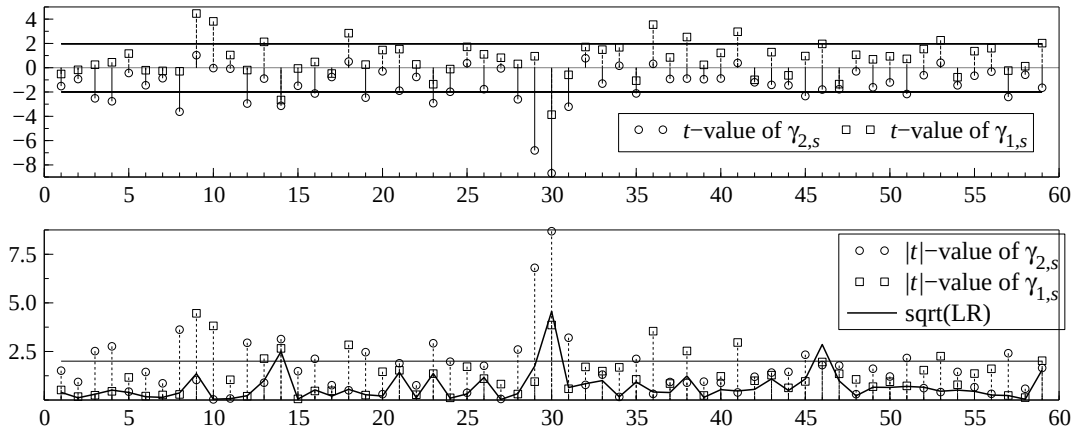

Figure 3: t-values of left and right mode (top), and absolute t-values (squares and circles) with square root of likelihood-ratio test (continuous line). Both only for cases with multimodality.

insignificant, and the other significant. In a few cases (the last, for example), the left mode has nearly significant negative value, and the right mode a significantly positive value if a critical value of 2 is used. The second panel shows the square root of the likelihood-ratio test, which has one degree of freedom, together with the absolute values of the t statistics. The LR test has only three of the displayed observations significant. Interestingly, it follows very closely the lowest of the absolute t-values, suggesting that the smallest $|\widehat{\gamma}|$ should be selected. Unfortunately, in practice it will be unknown which of the two t-values is found, unless the modeller is aware of the problem.

2.2 Proof of Proposition 1

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

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

Assuming that the start-up of the recursive process does not depend on the parameters, the score is given by:

$$\frac{\partial \ell_t(\theta)}{\partial \theta} = -\frac{\varepsilon_t}{h_t} \frac{\partial \varepsilon_t}{\partial \theta} - \frac{1}{2} \frac{1}{h_t^2} \left(h_t - \varepsilon_t^2 \right) \frac{\partial h_t}{\partial \theta}, \tag{3}$$

with $\varepsilon_t = y_t - x_t'\zeta - d_t\gamma$. It is convenient to use the ARCH(∞) form. Define the lag polynomials $\beta(L) = 1 - \sum_{i=1}^{p} \beta_i L^i$, and $\alpha(L) = \sum_{i=1}^{q} \alpha_i L^i$, such that

$$h_t = \beta(L)^{-1} \left(\alpha_0 + \alpha(L)\varepsilon_t^2 \right) = \alpha_0^* + \sum_{j=1}^{\infty} \delta_j \varepsilon_{t-j}^2. \tag{4}$$

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 to ensure identification of the individual GARCH parameters. As discussed in detail in §3.1, nonnegativity of the δ_is will ensure that h_t is always positive when $\alpha_0 > 0$.

The main example is the GARCH(1,1) model with $0 \leq \beta_1 < 1$, $\alpha_1 \neq \alpha_0\beta_1$:

$$h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1 h_{t-1},$$

which can be written as

$$h_t = \alpha_0^* + \alpha_1 \sum_{j=1}^{t} \beta_1^{j-1}\varepsilon_{t-j}^2, \tag{5}$$

given ε_0 and h_0, where $\alpha_0^* = \alpha_0(1-\beta_1^t)/(1-\beta_1) + \beta_1^t h_0$, which does not depend on γ. In the ARCH(∞) representation (4) of the GARCH(1,1) case: $\delta_1 = \alpha_1$, $\delta_2 = \alpha_1\beta_1$, $\delta_3 = \alpha_1\beta_1^2$, ...

The first order conditions (3) for γ can be expressed as a function of ε_s and h_t:

$$\frac{\partial \varepsilon_t}{\partial \gamma} = -d_t, \ \frac{\partial \varepsilon_t^2}{\partial \gamma} = -2\varepsilon_t d_t, \quad \text{thus} \quad \frac{\partial h_t}{\partial \gamma} = -2\sum_{j=1}^{t-1} \delta_j \varepsilon_{t-j} d_{t-j}.$$

Since $d_t = 0$ for $t \neq s$ and $d_s = 1$:

$$\frac{\partial h_t}{\partial \gamma} = -2\delta_{t-s}\varepsilon_s \quad \text{for } t > s,$$

and zero otherwise. The full score with respect to γ is:

$$\frac{\partial \ell(\theta)}{\partial \gamma} = \frac{\varepsilon_s}{h_s} + \varepsilon_s \sum_{t=s+1}^{T} \delta_{t-s} \frac{1}{h_t^2}\left(h_t - \varepsilon_t^2\right) = \frac{\varepsilon_s}{h_s}\left[1 + \frac{\delta_1 h_s}{h_{s+1}^2}\left(h_{s+1} - \varepsilon_{s+1}^2\right) + \kappa_s\right], \tag{6}$$

where $\kappa_s = h_s \sum_{t=s+2}^{T} \delta_{t-s} h_t^{-2}(h_t - \varepsilon_t^2)$. In (6), ε_s is a function of γ; h_{s+1} depends on ε_s^2, and is therefore also a function of γ, as are all h_t for $t \geq s+1$, and therefore κ_s. Define

$$Q_1(h_{s+1}) \equiv (1+\kappa_s)\, h_{s+1}^2 + \delta_1 h_s h_{s+1} - \delta_1 h_s \varepsilon_{s+1}^2, \tag{7}$$

so that maximizing the log-likelihood w.r.t. γ requires solving:

$$\frac{\partial \ell(\theta)}{\partial \gamma} = \frac{\varepsilon_s}{h_s h_{s+1}^2} Q_1(h_{s+1}) = 0. \tag{8}$$

In order to prove that a solution leads to a minimum or maximum we need the second derivative of the log-likelihood with respect to γ:

$$\frac{\partial^2 \ell(\theta)}{(\partial\gamma)^2} = -\frac{1+\kappa_s}{h_s} - \frac{\delta_1}{h_{s+1}^2}\left(h_{s+1} - \varepsilon_{s+1}^2\right) - 2\varepsilon_s^2 \sum_{t=s+1}^{T} \delta_{t-s}^2 \left(\frac{2\varepsilon_t^2 - h_t}{h_t^3}\right). \tag{9}$$

Two situations can attain when solving (8):

- $\widehat{\varepsilon}_s = 0$ is the only solution of (8), corresponding to $\widehat{\gamma} = y_s - x_s'\zeta$.

  For $\widehat{\varepsilon}_s = 0$, the second derivative matrix at the solution is block diagonal with respect to γ, because all terms in the derivative of (6) w.r.t. the GARCH parameters contain a factor ε_s. The last term in (9) drops out when $\widehat{\varepsilon}_s = 0$. The first two terms add up to $Q_1(h_{s+1})$ divided by $-h_s h_{s+1}^2$. Since $Q_1(h_{s+1})$ is monotonically increasing for positive values of h_{s+1} when $1 + \kappa_s > 0$, we can infer that it must be positive for $\widehat{\varepsilon}_s = 0$ to be the only solution. This makes the Hessian element negative, as required for a maximum.

  Although $1 + kappa$ could be negative, we have not seen any cases where $Q_1 < 0$ at $\widehat{\gamma}$.

- There is a $\widetilde{h}_{s+1}$ such that $Q_1(\widetilde{h}_{s+1}) = 0$.

  Two additional solutions to (8) can then be derived from:

$$\widetilde{\varepsilon}_s^2 = \frac{1}{\alpha_1}\left[\widetilde{h}_{s+1} - \alpha_0 - \sum_{i=2}^{q}\alpha_i\varepsilon_{s+1-i}^2 - \sum_{i=1}^{p}\beta_i h_{s+1-i}\right] \equiv G_s\left(\widetilde{h}_{s+1}\right) \equiv \widetilde{G}_s.$$

  This is now positive, and the additional two solutions are

$$\widetilde{\gamma} = y_s - x_s'\zeta \pm \widetilde{G}_s^{1/2}.$$

  In that case, the log-likelihood and its derivatives are identical for both values.

  Now $\widehat{\varepsilon}_s = 0$ leads to a negative $Q_1(h_{s+1})$. This creates a positive diagonal element in the Hessian, violating the conditions for a maximum.

A necessary condition for bimodality is that the solution to $Q_1(h_{s+1}) = 0$:

$$h_{s+1}^* = \frac{h_s\delta_1}{2\left(1+\kappa_s\right)}\left[-1 \pm \left(1 + \frac{4\varepsilon_{s+1}^2\left(1+\kappa_s\right)}{\delta_1 h_s}\right)^{1/2}\right] \tag{10}$$

is positive. In addition, the implied ε_s^{*2} must be non-negative. Therefore, when $1 + \kappa_s$ is positive, the negative solution can be ignored. When $1 + \kappa_s$ is negative, there are two solutions h_{s+1}^*. However, only one of these corresponds to $\widetilde{h}_{s+1}$. □

The expression for G_s merits further discussion.

In the ARCH(1) model we have $\delta_t = 0$ for $t > 1$, so that $\kappa_s = 0$ for all s. Now (7) can be solved expicitly for γ. It is zero when

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

is positive (the negative solution can be discarded). Then

$$G_s^* = \frac{h_s}{2}\left[-1 + \left(1 + \frac{4\varepsilon_{s+1}^2}{\alpha_1 h_s}\right)^{1/2}\right] - \frac{\alpha_1}{\alpha_0}, \tag{11}$$

and it is easy to see that G_s^* depends positively on both ε_{s+1}^2 and ε_{s-1}^2. Dummies in a volatile period can lead to multimodality.

In the GARCH(1,1) model, we can no longer solve (7) analytically. We can only derive some properties that a solution will have. In particular, knowing $\widehat{\kappa}_s$, there will be two modes if (10) has a positive solution, which can not be ruled out, in particular when ε_{s+1} is (also) large. The fact that dummies shortly before a volatile period can lead to multimodality is illustrated in our empirical application in the next subsection. In practice, if estimation of the model with a dummy yields $\widehat{\varepsilon}_s = 0$, then this is either a local minimum or a global maximum, which can be verified be inspecting the second derivative. Otherwise,

$$\widehat{G}_s = \frac{1}{\alpha_1}\left[\widehat{h}_{s+1} - \alpha_0 - \beta_1 h_s\right]$$

is positive and there are two global maxima.

2.3 Dummy variables in EGARCH models

The proof in §2.2 makes it clear that multimodality may occur in GARCH(p, q) models, especially when a sequence of large squared standardized residuals is present and a dummy is introduced in the preceding period. A similar effect could be expected for the EGARCH model (Nelson, 1991), although not necessarily symmetric multimodality.

Figure 4 shows the likelihood grid when specifying the example of Figure 1 as GARCH and EGARCH. The EGARCH(p, q) parameterisation for the conditional variance reads:

$$\log h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \left\{\vartheta_1 \xi_{t-i} + \vartheta_2 \left(|\xi_{t-i}| - E|\xi_t|\right)\right\} + \sum_{i=1}^{p} \beta_i \log h_{t-i}, \tag{12}$$

with $\alpha_1 = 1$. The main added flexibility of the EGARCH model derives from the asymmetry term $\vartheta_1 \xi_{t-i}$, which usually implies larger effects on h_t from negative ξ_{t-1} than from positive ξ_{t_1}. As before, we plot the likelihood grid as a function of γ, with the other parameters fixed at their values found at the global maximum.

Both plots in Figure 4 exhibit bimodality. For EGARCH, the two maxima are at different likelihood values, owing to the asymmetry term. When imposing $\vartheta_1 = 0$ in (12) both modes are at the same likelihood value. Because of the appearance of absolute value in (12), the local minimum is at a point where the likelihood is non-differentiable. Unless the iterative estimation procedure starts at

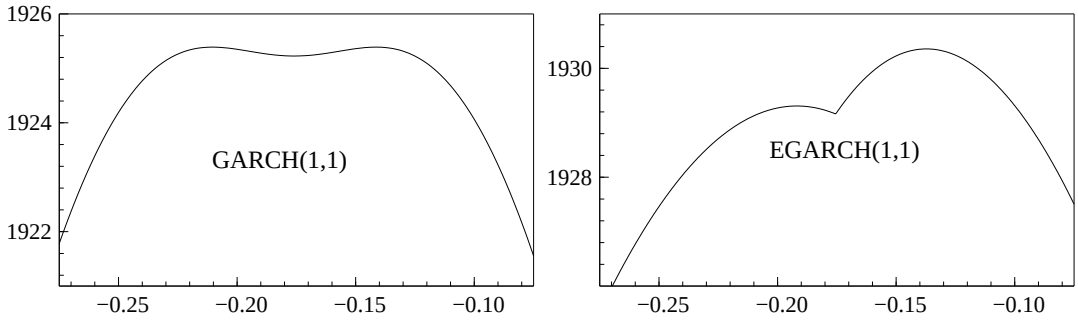

Figure 4: Likelihood grid for the dummy parameter corresponding to the 1987 crash, GARCH(1,1) (left) and EGARCH(1,1) (right).

the local minimum, this non-differentiability will not cause problems in practice. However, now it matters whether the local or the global maximum is found.

Figure 5 plots $\widehat{\gamma}_{2,s} - \widehat{\gamma}_{1,s}$ for the GARCH(1,1) and EGARCH(1,1) models. Now there are about 30 cases with two modes in the likelihood. Note that the values for $\widehat{\gamma}_{2,s} - \widehat{\gamma}_{1,s} = 2\widehat{G}_s^{1/2}$ for the GARCH model are much larger in the four weeks before the 1987 crash than in the week of the crash itself. Effects of shifting and extending the dummy are considered in the next subsection.

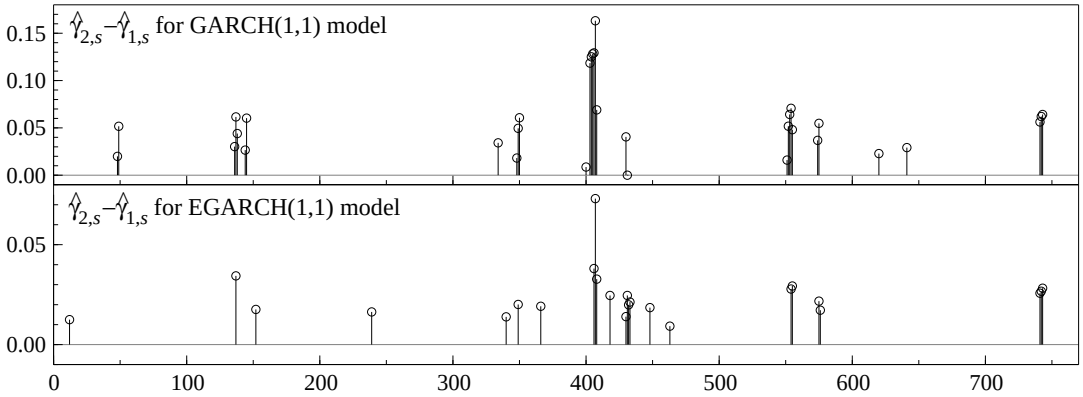

Figure 5: Estimates of $\widehat{\gamma}_{2,s} - \widehat{\gamma}_{1,s}$ for the GARCH(1,1) model (top) and the EGARCH(1,1) model (bottom).

For completeness, we remark that GARCH models with Student-t errors can also exhibit multi-modality, although we had to try another data set (UK quarterly inflation for 1955Q1 – 2000Q4) to find some examples.

2.4 Extended dummy variables in GARCH models

Up to this point, all dummy variables had only one non-zero observation. Here, we consider a dummy variable that is unity for j consecutive observations: $s, \ldots, s+j-1$. The score (6) for γ in the GARCH(p, q) model becomes:

$$\frac{\partial \ell(\theta)}{\partial \gamma} = \sum_{k=0}^{j-1} \frac{\varepsilon_{s+k}}{h_{s+k}} \left[1 + \frac{\delta_1 h_{s+k}}{h_{s+k+1}^2} \left(h_{s+k+1} - \varepsilon_{s+k+1}^2 \right) + \kappa_{s+k} \right]. \tag{13}$$

For example, when $j = 2$, the dummy variable is unity for two observations in a row. Then (13) is zero when $\widehat{\varepsilon}_s = \widehat{\varepsilon}_{s+1} = 0$, but also has a solution when $G_s > 0$ and $G_{s+1} > 0$. In general, any dummy variable that picks out observations with positive G_s will have two modes. In such a situation there may even be more than two modes.

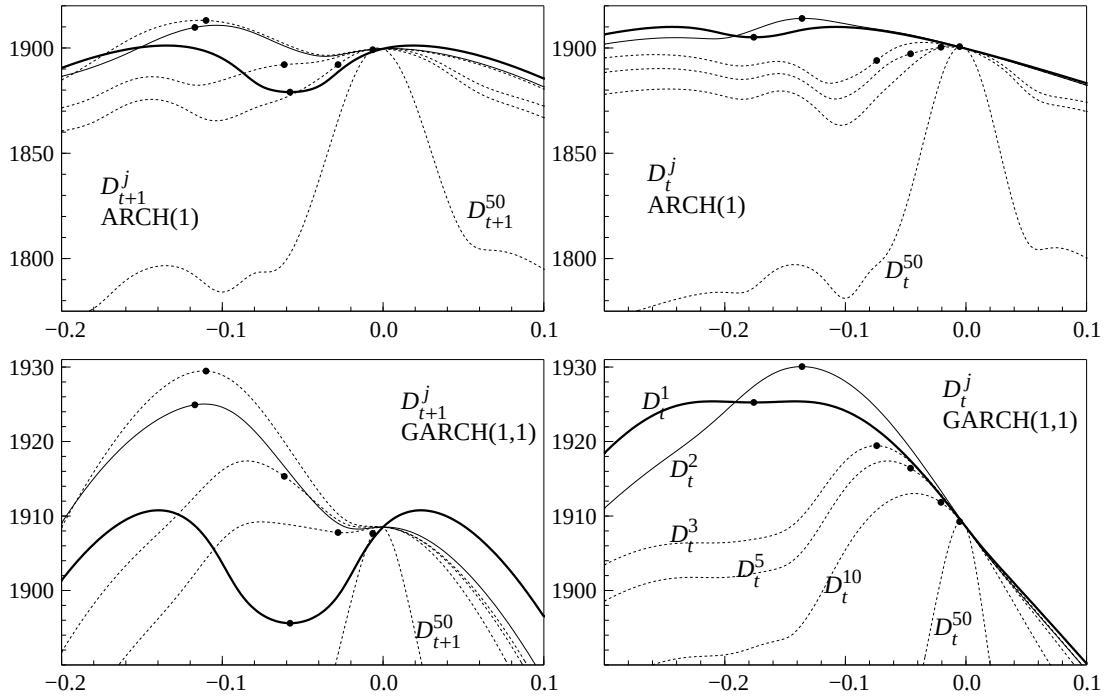

Figure 6: Concentrated likelihood grid for the coefficient of D_t^j. D_{t+1}^j starts in the week before the 1987 crash (left graphs); D_t^j starts in the week of the crash (right graphs). Dummy continues for j periods, $j = 1, 2, 3, 5, 10, 50$. The thick solid lines are for one-period dummies ($j = 1$). The thin solid lines are for the two period dummies ($j = 2$). The dashed lines for $j = 3, 5, 10, 50$ are below each other within each panel.

To investigate the case with extended dummies, define s as the week of the 1987 crash (this is as

before; remember that we use weekly data). We now construct the following variable:

$$D_t^j = \begin{cases} 1 & \text{for } t = s, \ldots, s+j-1, \\ 0 & \text{otherwise,} \end{cases}$$

We also use this dummy variable with a lead of one period, D_{t+1}^j, which now starts one observation prior to the crash.

Figure 6 show the effect of shifting and extending the dummy on the ARCH(1) and GARCH(1,1) profile likelihoods for γ. The thick solid lines in the right graphs are for D_t^1 and correspond to the dummy variable d_t used in Figures 1 and 4. In Figures 1 and 4 we fixed the remaining parameters in the construction of the likelihood grids, but Figure 6 plots the concentrated log-likelihoods: for a range of values for γ, construct $y_t^* = y_t - \gamma D_t^j$ and estimate an ARCH(1) model (top two panels) and a GARCH(1,1) model (bottom two panels) for y_t^*. We expect the GARCH parameter estimates to depend more strongly on the dummy parameter when the dummy extends over a longer period, and we therefore choose to re-estimate the GARCH parameters in this case. In all cases, a constant is the only other regressor in the mean. The circles in the bottom two graphs match the solution that would be found if the coefficient of the dummy variable is fixed by a prior regression of y_t on a constant and D_t^j, i.e. the solution corresponding to $\sum_{t=s}^{s+j-1} \widehat{\varepsilon}_t = 0$.

The symmetric bimodalities of the thick solid lines correspond to the single period dummies. By comparing left and right figures, one observes that the bimodality is more pronounced when the single dummy is added in the week before the crash. For the GARCH model, the multimodality largely disappears as the dummy is extended. Note that the model with the three period dummy, attains the highest likelihood of all models where the dummy starts one period before the crash: D_{t+1}^3. When the dummy starts with the crash, the two period dummy has the highest likelihoods: D_t^2. The profile likelihoods for the ARCH(1) models reveal many cases with multiple modes once the dummy is unity over two or more periods.

2.5 GARCH with a dummy variable in the conditional variance

Proposition 2 *Consider the GARCH(p, q) regression model specified with a lagged variance dummy as follows:*

$$\begin{aligned} y_t &= x_t'\zeta + \gamma d_t + \varepsilon_t, \\ \beta(L)h_t &= \alpha_0 + \alpha(L)\varepsilon_t^2 + \tau d_{t-1}, \end{aligned}$$

where $d_t = 1$ when $t = s, 1 < s < T$, and $d_t = 0$ otherwise. This combination of dummy variables does not induce multimodality in the log-likelihood function.

Proof

We extend the proof in §2.2. The crucial term in the likelihood function, h_{s+1}, is now a function of both γ and τ. The ARCH(∞) equation (4) is extended to

$$h_t = \alpha_0^* + \sum_{j=1}^{t} \delta_j \varepsilon_{t-j}^2 + \tau \sum_{j=1}^{t-1} \phi_j d_{t-j}, \tag{14}$$

where $\phi(L) = \phi_1 L + \ldots = L\left[\beta(L)\right]^{-1}$. In particular ϕ_1 = 1. So

$$\frac{\partial h_t(\gamma,\tau)}{\partial \tau} = \phi_{t-s} \quad \text{for } t > s,$$

and zero otherwise. $\partial \ell(\theta)/\partial\gamma$ was given in equation (6) and does not change by the introduction of the variance dummy. From (3), we find the full score with respect to τ as:

$$\frac{\partial \ell(\theta)}{\partial \tau} = -\frac{1}{2} \sum_{t=s+1}^{T} \phi_{t-s} \frac{1}{h_t^2}\left(h_t - \varepsilon_t^2\right) = -\frac{1}{2}\left[\frac{1}{h_{s+1}^2}\left(h_{s+1} - \varepsilon_{s+1}^2\right) + \lambda_s\right],$$

where $\lambda_s = \sum_{t=s+2}^{T} \phi_{t-s} h_t^{-2}(h_t - \varepsilon_t^2)$. This leads to a second quadratic equation for $h_{s+1}(\gamma,\tau)$:

$$Q_2(h_{s+1}) \equiv \lambda_s h_{s+1}^2 + h_{s+1} - \varepsilon_{s+1}^2 = 0. \tag{15}$$

Although $\lambda_s = \lambda_s(\gamma,\tau)$ (unless an ARCH(1) model is considered), an additional solution would solve $Q_2 = 0$, which can be expressed in terms of $\widehat{h}_{s+1}$. Q_2 has a positive real solution if $|\lambda_s| > 0$ and $\lambda_s > -\frac{1}{4}\varepsilon_{s+1}^{-2}$:

$$\begin{array}{rcll} \widehat{h}_{s+1} & = & (2\lambda_s)^{-1}\left[-1 + \left(1 + 4\varepsilon_{s+1}^2\lambda_s\right)^{1/2}\right], & \lambda_s > 0 \\ \widehat{h}_{s+1} & = & (2|\lambda_s|)^{-1}\left[1 \pm \left(1 + 4\varepsilon_{s+1}^2\lambda_s\right)^{1/2}\right], & -\frac{1}{4}\varepsilon_{s+1}^{-2} < \lambda_s < 0 \end{array} \tag{16}$$

while $\widehat{h}_{s+1} = \varepsilon_{s+1}^2$ is the solution if λ_s = 0.

The multimodality result for γ only extends to the current model if the root of Q_2 simultaneously solves $Q_1(h_{s+1}) = 0$, given in (7), because the extra stationary points of $\ell(\theta)$ were caused by solutions to $Q_1(h_{s+1}) = 0$. Otherwise, multimodality is avoided, and $\widehat{\varepsilon}_s = 0$ in (6) provides the single solution for γ.

If $\lambda_s = 0$, the solution to (15) simplifies to $\widehat{h}_{s+1} = \widehat{\varepsilon}_{s+1}^2$. Substituting $\widehat{h}_{s+1}$ into (7) shows that $Q_1(h_{s+1}) = 0$ then requires $\varepsilon_{s+1}^2 = h_{s+1} = 0$, which is not a feasible solution to the maximum likelihood problem as the log-likelihood becomes minus infinity for $h_{s+1} = 0$. If $|\lambda_s| > 0$ substitution of (16) into (7) shows this is not a solution either, unless $\varepsilon_{s+1}^2 = h_{s+1} = 0$, which again can be ruled out. □

Proposition 2 shows that adding the corresponding dummy with one lag to the variance equation provides a way to avoid the bimodality in the GARCH(p, q) model that was detected in Proposition 1. It is instructive to see what happens for $p = 0$. The ARCH(q) model has $\lambda_s = 0$, and therefore $\widehat{h}_{s+1} = \widehat{\varepsilon}^2_{s+1}$. A straightforward solution for τ follows:

$$\widehat{\tau} = \widehat{\varepsilon}^2_{s+1} - \alpha_0 - \alpha(L)\widehat{\varepsilon}^2_{s+1}.$$

Finally, consider another relative timing of the two dummies. If the dummy enters both the mean and variance without lag, the solution to $Q_2(\cdot) = 0$ applies to h_s instead of h_{s+1}, which does not immediately interfere with the first order conditions for γ, so bimodality remains a potential issue. If $G_s < 0$ and $p = 0$ the first order condtions for γ and τ lead to $\widehat{\varepsilon}^2_s = \widehat{h}_s = 0$ and a log-likelihood of minus infinity results.

3 Multimodality without dummy variables

We have shown how the introduction of dummy variables, which is regularly done in practice, can cause multimodallity in the GARCH likelihood. However, adding dummy variables may not be the only cause of multimodality. The objective in this section is to investigate the incidence of multiple modes without a regression part for the mean. As modes may occur at unreasonable values for the GARCH parameters, we first discuss restrictions on the GARCH parameter space in §3.1. Implementation details will also be provided. Next, §3.2 discusses the effects of the restrictions on the number and type of modes found in samples from simulated GARCH(2,2) processes, and in an empirical data set concerning daily British Pound/US Dollar exchange rates.

3.1 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.[1] The restrictions are imposed in

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

the ARCH(∞) form, which was introduced earlier in equation (4) in connection with multimodality caused by dummy variables. The parameter restrictions may have an impact in that context as well, but in the remaining part of this paper we focus on multimodality in the absence of dummy variables. Nelson and Cao (1992) require $\alpha_0^* = \alpha_0/\beta(1) > 0$ and $\delta_i \geq 0\ \forall i$. This implies 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 2 we refine the Nelson and Cao (1992) conditions for the GARCH($2, q$) case, i.e. for $p = 2$, by removing redundant conditions. Table 2 summarizes the restrictions for low-order GARCH models. The conditions on the roots when $p = 2$, as given in Table 2, can also be expressed as $\beta_2 + \beta_1 < 1,\ \beta_1^2 + 4\beta_2 \geq 0$. In the original formulation, 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 $\alpha_0 > 0$.[2] In addition, Appendix 2 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	$\alpha_2 + (\rho_1 + \rho_2)\alpha_1 \geq 0,$ and $\alpha_2 + \rho_1\alpha_1 > 0.$

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.[3] For example, the restrictions imply $h_t \geq \alpha_0^*$. And, when the initial δ_i are positive and dominate the coefficients at higher lags, the probability of obtaining a negative conditional variance becomes essentially zero.

Because the Nelson and Cao (1992) constraints are very complex for higher order models, we now suggest another set of constraints. These relax the positivity restrictions in a different way, and are easier to implement and interpret. They are based on the ARMA representation for the variance

[2] 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.

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

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

$$\varepsilon_t^2 = \alpha_0 + \sum_{i=1}^{m}(\alpha_i + \beta_i)\varepsilon_{t-i}^2 - \sum_{i=1}^{p}\beta_i u_{t-i} + u_t, \tag{17}$$

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

Taking unconditional expectations of (17), we can ensure positivity and invertibility by the conditions:

$$\begin{aligned} \alpha_0 &> 0, \\ \alpha_i + \beta_i &\geq 0, \qquad \text{for } i = 1, \ldots, m. \\ 0 < \textstyle\sum_{i=1}^{m} \alpha_i + \beta_i &< 1, \end{aligned} \tag{18}$$

where, as before, $m = \max(p, q)$. 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 coefficients in the ARMA representation (17) are:

$$\varepsilon_t^2 = (\alpha + \beta)(L)^{-1}\left(\alpha_0 + \beta(L)u\right) = \alpha_0^{**} + \sum_{i=0}^{\infty}\gamma_i u_{t-i}, \tag{19}$$

where $\beta(L) = 1 - \sum_{i=1}^{p}\beta_i L^i$, $(\alpha + \beta)(L) = 1 - \sum_{i=1}^{m}(\alpha_i + \beta_i)L^i$, and $\gamma_0 = 1$. The γ_i coefficients show the IGARCH boundary: if they remain constant after an initial period, then $\sum_{i=1}^{m}\alpha_i + \beta_i = 1$.

Table 3: Types of GARCH parameter restriction.

UNR	Unrestricted, except for: $\alpha_0 > 0$;
N&C	Positive conditional variance: conditions (DO1)–(DO4), see Appendix 2;
UV	Positive and finite unconditional variance: restrictions (18), see Appendix 3;
POS	All coefficients positive: $\alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0$, also see Appendix 1.

Table 4: GARCH processes A–D.

Process	α_1	α_2	β_1	β_2	ρ_1	ρ_2	$\sum \alpha_i + \beta_i$
A	0.10	0	0.85	0	0.85	0	0.95
B	0.10	0.10	0.10	0.65	0.85777	−0.75777	0.95
C	0.10	0.10	−0.10	0.85	−0.97331	0.87331	0.95
D	0.35	−0.20	0.70	0.10	0.82170	−0.12170	0.95

Table 3 summarizes the parameter restrictions that are considered here. The relevant appendices show how these restrictions can be implemented through parameter transformations. Then, restricted estimation can be implemented as an extension to unrestricted estimation, using the Jacobian of the transformation (which can be computed analytically or numerically).

To compare the impact of these restrictions, we use a GARCH(1,1) and three GARCH(2,2) processes, see Table 4. Processes C and D are not allowed when all coefficients are forced to be non-negative (POS). Process D is not allowed by parametrization UV, because $\alpha_2 + \beta_2 < 0$, but is fine for N&C. Process C is just allowed by UV, but not by N&C because the largest absolute root is negative.

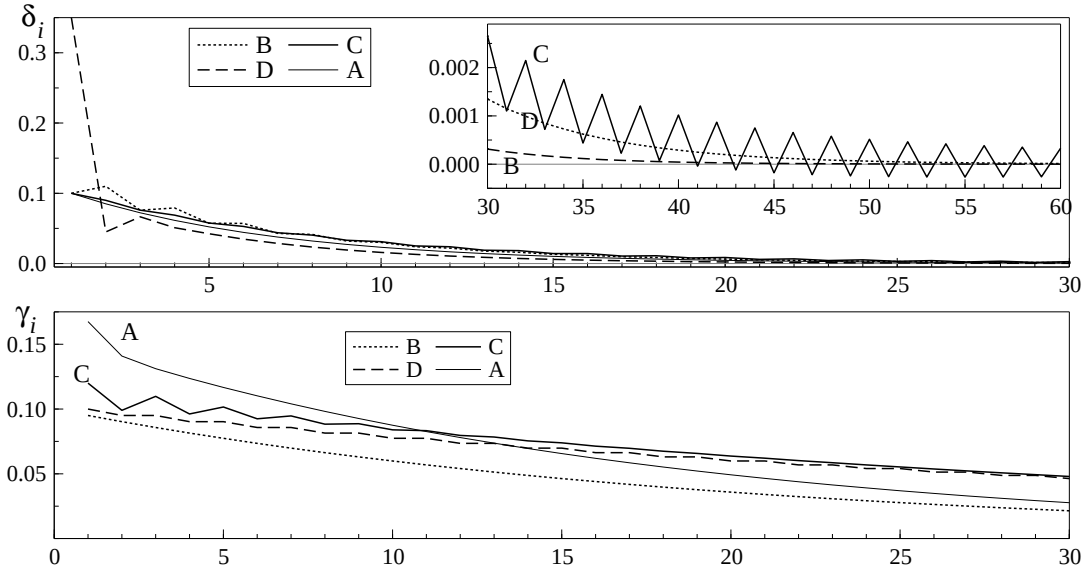

Figure 7: Coefficients δ_i (top) and γ_i (bottom) for GARCH(2,2) processes B,C,D and GARCH(1,1) process A.

Figure 7a plots the coefficients δ_i from (4) for the three GARCH(2,2) processes. The section from lag 30 to 60 is shown as a separate inset. Process B starts with a zig-zag pattern, but becomes smooth as it approaches zero. Process C, on the other hand, only starts to really zig-zag as the lag-length gets beyond 15. The fact that it moves around zero, while getting smaller, is not allowed by N&C-type restrictions. Moving process C onto the N&C boundary ($\beta_1 = 0, \beta_2 = 0.75$) makes the coefficients behave like a step-function, with increasingly smaller steps as they approach zero. A feature of D is that δ_2 is smaller than δ_3. Figure 7b plots the γ_i coefficients from (19), omitting γ_0 ($= 1$).

3.2 Searching for multiple modes

This section presents some evidence on the possible occurrence of multimodality when the mean only consists of a constant term. We consider the four types of parameter restrictions UNR, N&C, UV and POS, as discussed in the previous section. Implementation of N&C is explained at the end of Appendix 2; for implementation details of UV see Appendix 3. The choice of parameter 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 $\widehat{\theta}$ (say). We then re-estimate with $\widehat{\theta} + \epsilon$ as starting values, with ϵ drawn from the standard normal distribution.[4] In case restrictions are imposed, the transformed estimates from the first estimation are randomized (there are no restrictions in the transformed space, see the Appendices) to provide starting values for subsequent estimations. This automatically 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.

We start by considering a single sample of 1000 replications for GARCH processes A–D. For each process, this is generated from the same random normal sequence, and 250 initial observations are discarded. Table 5 gives the maximum values that were found after this search. For each process, the same log-likelihood was found when estimating a GARCH(1,1) model. The table lists the improvement in log-likelihood when moving from GARCH(1,1) to GARCH(2,2): $\widehat{\ell}_{2,2} - \widehat{\ell}_{1,1}$, and from GARCH(2,2) to GARCH(3,3): $\widehat{\ell}_{3,3} - \widehat{\ell}_{2,2}$. A single star indicates that the likelihood-ratio is significant at 5% on a $\chi^2(2)$ test, while two stars indicates significance at 1%.

A notable feature, which we also found in other samples, is that overparametrized unrestricted estimation finds maxima at 'strange' parameter values. These maxima tend to be considerably better, therefore likely to be accepted on LR or AIC criteria.

For N&C there are two cases in Table 5 where the more general model has a lower log-likelihood. This indicates a local maximum, because the more restricted model with $\widehat{\alpha}_3 = \widehat{\beta}_3 = 0$ would be better. In the random search for GARCH(2,2) maxima on the process A data, the overall maximum was found 90% of the time, and the local in the remaining 10%. For the GARCH(3,3) estimates of process A, the overall maximum was only found in 3% of the cases, the next best in 1.5%, and the

[4] We could have considered using the estimated variance for the normal distribution. However, there is no guarantee that a local optimum would provide a good estimate of the variance. Moreover, this would not allow parameters with low 'standard errors' to move very much.

Table 5: Changes in likelihood values $\widehat{\ell}_{p,p}$ at located maxima for GARCH(p,p) models, $p = 1, 2, 3$ for a single replication from processes A–D.

	UNR	N&C	UV	POS
		process A		
$\widehat{\ell}_{2,2} - \widehat{\ell}_{1,1}$	$3.0678*$	1.1473	0.4637	0.4290
$\widehat{\ell}_{3,3} - \widehat{\ell}_{2,2}$	$9.2421**$	0.1022	0.6982	0.7296
		process B		
$\widehat{\ell}_{2,2} - \widehat{\ell}_{1,1}$	2.6383	1.6220	0.9074	0.8567
$\widehat{\ell}_{3,3} - \widehat{\ell}_{2,2}$	$8.5432**$	-0.4132	0.3014	0.3521
		process C		
$\widehat{\ell}_{2,2} - \widehat{\ell}_{1,1}$	2.3391	1.5305	0.6626	0.6122
$\widehat{\ell}_{3,3} - \widehat{\ell}_{2,2}$	$7.6686**$	-0.2400	0.6279	0.6783
		process D		
$\widehat{\ell}_{2,2} - \widehat{\ell}_{1,1}$	$6.1170**$	$6.1170**$	$6.1170**$	$4.4192*$
$\widehat{\ell}_{3,3} - \widehat{\ell}_{2,2}$	0.7556	0.7556	0.3146	1.2327

$\widehat{\ell}_{p,p}$is the log-likelihood for GARCH(p,p) model.

worst in 95%. Note that, when searching, the most common solution was randomized (i.e. that found from default starting values). In general, our experience was that the global maxima of unrestricted estimation can be hard to find. For the restricted parameterizations, on the other hand, the most commonly found solution is also usually the best.

Figure 8a shows the coefficients δ_i for all unrestricted GARCH(1,1) and GARCH(2,2) estimates that converged when using the realization from the GARCH(1,1) process A. The corresponding GARCH(3,3) results are in Figure 8b. The different patterns are quite surprising.

Figure 9 shows the coefficients γ_i for N&C and UV. In this case UV and POS are identical, except that UV found a small number (about 1%) of local solutions on the IGARCH boundary. The corresponding figures for the realization of processes B and C look very similar.

Finally, we look at selecting a GARCH(p,q) model for the British pound to US dollar daily exchange rate.[5] The sample has 2915 observations (7-Jun-1973 to 28-Jan-1985), and is similar to some

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

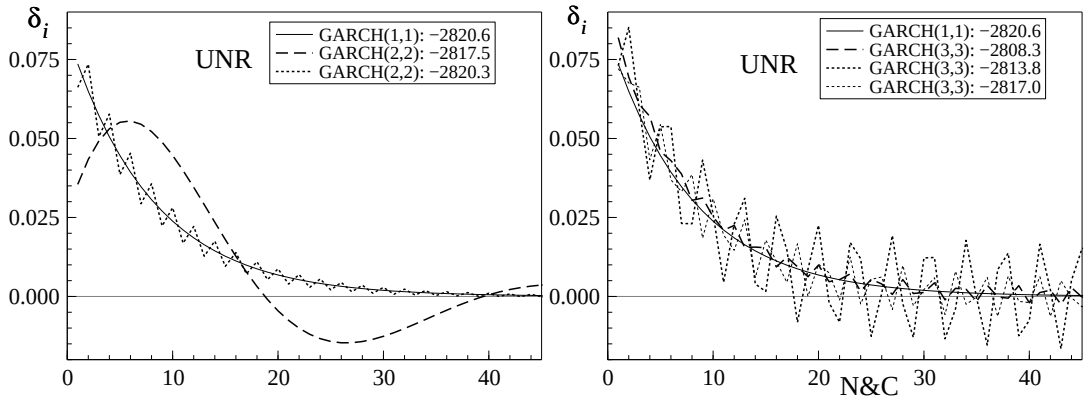

Figure 8: Coefficients δ_i unrestricted GARCH(2,2) (left) and GARCH(3,3) (right) estimates, for all (local) maxima. Data is from process A.

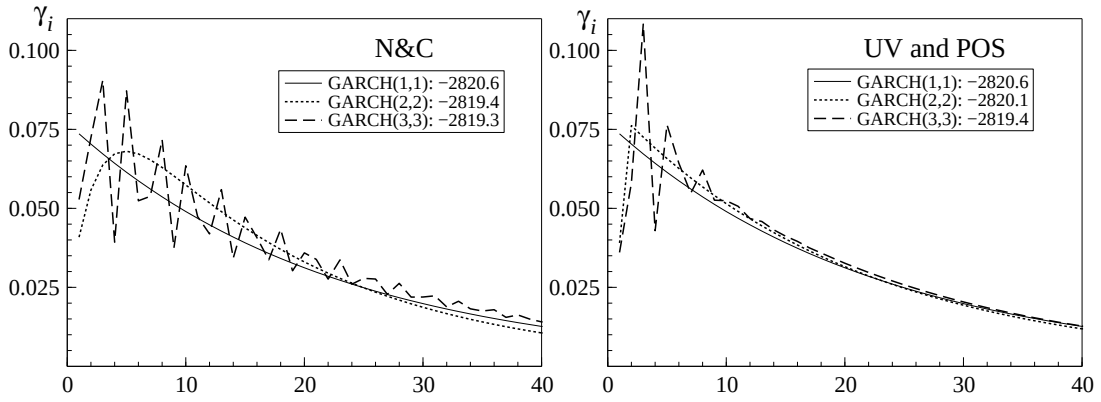

Figure 9: Coefficients γ_i of GARCH(2,2) and GARCH(3,3) estimates for N&C, UV and POS. Data is from process A.

of the estimations in Nelson and Cao (1992). For GARCH(1,1), GARCH(2,1), and GARCH(1,2) we found no multimodality, but for higher order models, we did find multiple solutions, even with such a large sample size. Table 6 lists some of the maxima that were found for selected GARCH models. The columns labelled 'robustness' give the percentage of time the same solution was found when using it as a starting point for a randomized search. Again this is based on 250 successful estimations. A low robustness value could indicate that it is difficult to locate that particular mode.

In the unrestricted case in particular, the random search delivered considerably higher likelihoods. The same happened with GARCH(3,3) estimation for the N&C case. For the other cases, the solutions are very close in terms of the log-likelihood. Testing down the lag length is problematic when there are many local maxima: it can easily happen that a sequence of nested hypotheses is not nested in terms of

likelihood values (as happened for the GARCH(3,3) estimates under UV). None of our GARCH(1,1) estimates, either on the artificial processes or actual data, revealed multiple modes.

Table 7 reports the models that are selected on the AIC criterion. The last column is for the 'global' maximum (although we cannot rule out that even better solutions exist). Each parameterization selects a different model: the estimated GARCH(3,3) for the unrectricted case is quite different from the Nelson&Cao restrictions. The column labelled 'Robust' only considers those modes which were found at least 60% of the time when re-estimating from that solution with randomization. This yields a different GARCH(3,3) model for unrestricted estimation, and a GARCH(3,1) instead of GARCH(3,3) for N&C. In the remaining two cases the solution does not change: all modes are very robust.

Figure 10 expresses the models in terms of the estimated coefficients γ_i. Note that the UV model is IGARCH, and the best unrestricted model goes beyond that with a sum of GARCH parameters equal to 1.005.

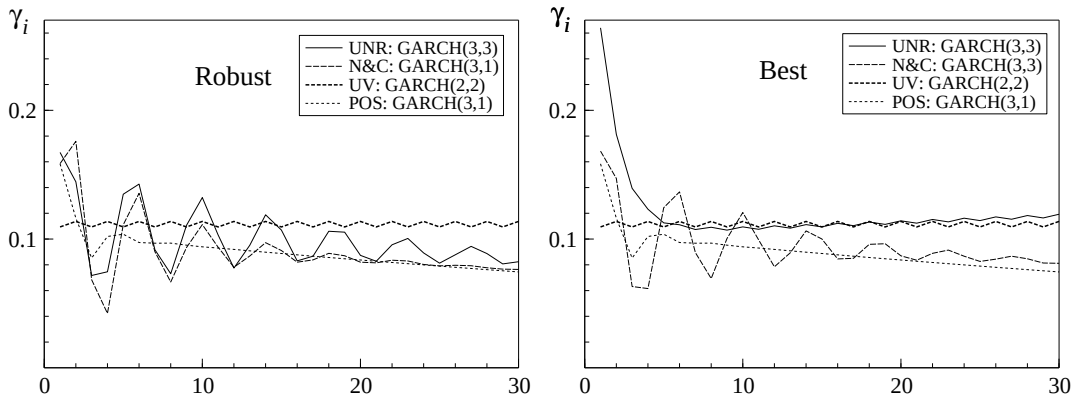

Figure 10: Coefficients γ_i for models corresponding to Table 7.

4 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. Interestingly, whether this happens depends on the data, but it is likely when correcting for large outliers. We believe that this curiosity, while of empirical relevance, has not yet been explicitly noted in the literature.

This finding has important consequences for empirical modelling. First, a t-test on the coefficient of a dummy variable cannot be used in GARCH regression models. When there are two maxima, 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. Asymptotic likelihood

Table 6: GARCH(3,3), GARCH(2,3), GARCH(2,2) likelihood values at located maxima for £/$ returns ($T = 2915$). And the percentage of occurrence based on 250 model estimates from random starting values.

UNR		N&C		UV		POS	
loglik	robustness	loglik	robustness	loglik	robustness	loglik	robustness
			GARCH$(3,3)$				
-2093.7	4.0%	-2128.0	47.6%	-2141.3	84.4%	-2142.3	98.0%
-2109.5	8.4%	-2130.8	84.8%	-2144.1	99.6%		
-2123.7	65.2%	-2139.1	96.8%	-2145.1	55.6%		
-2134.5	3.6%						
-2138.4	60.8%						
-2140.9	39.6%						
			GARCH$(2,3)$				
-2095.9	6.8%	-2141.3	100%	-2139.0	96.4%	-2142.6	74.0%
-2112.7	3.2%			-2142.6	94.4%	-2143.9	77.2%
-2141.3	81.0%						
			GARCH$(2,2)$				
-2113.1	7.6%	-2142.3	100%	-2139.0	94.4%	-2142.6	72.8%
-2134.8	1.2%			-2142.6	92.8%	-2143.9	70.4%
-2142.6	92.4%			-2144.9	99.6%		

Robustness is the percentage of estimates that found same mode in randomization.
loglik is the log-likelihood; see Table 3 for UNR, N&C, UV, POS.

theory is affected by this violation of the regularity conditions. Secondly, all model statistics which involve the value of the dummy are affected. We also noted that with only dummies as regressors, standard software may find a local minimum of the likelihood. Finally, we showed that adding the dummy with one lag in the conditional variance equation avoided the multimodality. We use this result in Doornik and Ooms (2002) to develop a procedure for outlier detection in GARCH models.

Next, 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 proposed a simpler alternative which allows imposition of the IGARCH boundary, while also being more general than forcing all coeffi-

Table 7: GARCH model for £/$ returns selected by AIC, for GARCH($p \leq 3, q \leq 3$).

	Robust	Best
UNR: unrestricted	$(3,3)$	$(3,3)$
N&C: positive conditional variance	$(3,1)$	$(3,3)$
UV: positive and finite unconditional variance	$(2,2)$	$(2,2)$
POS: all coefficients positive	$(3,1)$	$(3,1)$

Robust is outcome with robustness $> 60\%$.

Best is outcome using best solution.

cients to be positive. This seems to behave as well in our applications, albeit with a somewhat higher incidence of boundary solutions.

We have shown that multimodality of the GARCH likelihood is of practical relevance. It is likely that 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 on 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 GARCH coefficients greater than one). In light of this, it is important that restrictions are imposed on the parameter space.

Acknowledgements

We wish to thank Peter Boswijk, Bruce McCullough, Neil Shephard and David Hendry for helpful discussions and suggestions, as well as seminar attendees at the 2000 World Congress of the Econometric Society, 9th ESTE, 2001 meeting of the Society for Computational Economics and Warwick University. Financial support from the UK Economic and Social Research Council (grant R000237500) is gratefully acknowledged by JAD. All computations were done using Ox (Doornik, 2001) versions 2.11 – 3.2.

Appendix 1 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. In this paper we follow the suggestion of Bollerslev (1986) to use the mean of the squared residuals:

$$\varepsilon_i^2 = h_i = T^{-1} \sum_{t=1}^{T} \varepsilon_t^2, \text{ for } i \leq 0. \tag{20}$$

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 $\phi_0, \ldots, \phi_{q+p}$ be the parameters used in estimation, then $\alpha_0, \alpha_1, \ldots, \beta_p = e^{\phi_0}, \ldots, e^{\phi_{q+p}}$ will ensure that all coefficients are positive. The Jacobian matrix of this transformation is $\mathrm{dg}(\alpha_0, \alpha_1, \ldots, \beta_p)$. More general formulations are discussed in §3.1, and below.

3. Which maximization technique to use;

   We have found BFGS (see e.g. Fletcher, 1987 or Gill, Murray, and Wright, 1981) to be the most successful numerical maximization method. This corresponds with the consensus view in the numerical analysis literature that BFGS is the preferred quasi–Newton method, see e.g. Fletcher (1987, p.71) and Nocedal and Wright (1999, p.197). BFGS avoids the need for second derivatives. It 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). BFGS may be somewhat slower than some other methods, but we believe that robustness (i.e. success in convergence) is more important.

4. How to compute starting values for the parameters;

   We use the ARMA parameterization of the variance process from (17) and apply the method of Galbraith and Zinde-Walsh (1997), developed for estimation of ARMA models, to the squared data, after removing regression effects in the mean. If necessary, the resulting parameter values are adjusted to enforce the unconditional variance to exist.

5. Whether to use numerical or analytical derivatives;

   All estimates in this paper use analytical derivatives, except when imposing all positive or Nelson&Cao-type restrictions, and for EGARCH-type models. When the Hessian matrix is required for the variance-covariance matrix this is also computed numerically.

6. Which estimate of the variance-covariance matrix to use.

   A comparison of various estimators is given in Fiorentini, Calzolari, and Panattoni, 1996.

Appendix 2 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 $i = 1, 2, \ldots$:

$$\begin{aligned} \delta_i &= 0, && i < 1, \\ \delta_i &= \textstyle\sum_{j=1}^{p} \beta_j \delta_{i-j} + \alpha_i, && i \le q, \\ \delta_i &= \textstyle\sum_{j=1}^{p} \beta_j \delta_{i-j}, && i > q. \end{aligned} \tag{21}$$

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:

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

NC Theorem 2 gives these conditions as:

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

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

Proposition 3 *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:

$$\begin{aligned} &\rho_1\alpha_1 + \alpha_2 > 0, && \text{(NC4)},\\ &(\rho_1 + \rho_2)\alpha_1 + \alpha_2 > 0, && \text{from } \delta_q \text{ in (21)}. \end{aligned}$$

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

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

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:[6]

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

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 $\phi_0, \phi_1 \ldots, \phi_q, \psi_1, \psi_2$ be the unrestricted parameters. Then:

(a) $\alpha_0 = \exp(\phi_0)$,
(b) $\rho_1 = \frac{\exp(\psi_1)}{1+\exp(\psi_1)},\ \rho_2 = \rho_1\frac{1-\exp(\psi_2)}{1+\exp(\psi_2)}$,
(c) $\beta_1 = \rho_1 + \rho_2,\ \beta_2 = -\rho_1\rho_2$,
(d) $\alpha_i = \delta_i - \sum_{j=1}^{p}\beta_j\delta_{i-j}$ using $\delta_i = \exp(\phi_i)$ for $1 \le i \le q-1,\ \delta_i = 0$ for $i < 1$,
(e) $\alpha_q = -\frac{1}{2}(x+y) + \frac{1}{2}\left[(x-y)^2 + 4\exp(\phi_q)\right]^{1/2},\quad x = \sum_{j=1}^{p}\beta_j\delta_{q-j},\ y = \sum_{j=1}^{q-1}\rho_1^{q-j}\alpha_j$.

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

As NC point out, starting the h_t recursion with the sample mean (20) will ensure positive conditional variance. This is not necessarily the case when using other methods to initialize pre-sample values of h_t.

Appendix 3 Positive and finite unconditional variance

Estimation under restrictions (18) 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 $m = 3$:

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

[6]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.

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 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. *Journal of 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, pp. 2959–3038. Amsterdam: North-Holland.

Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox* (4th ed.). London: Timberlake Consultants Press.

Doornik, J. A. and M. Ooms (2002). Outlier detection in GARCH models. mimeo, Nuffield College.

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 inflation. *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.

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

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.

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

Gómez, V., A. Maravall, and Peña (1999). Missing observations in ARIMA models: Skipping approach versus additive outlier approach. *Journal of Econometrics 88*, 341–363.

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.

Marriott, J. M. and A. F. M. Smith (1992). Reparameterization aspects of numerical Bayesian methodology for autoregressivce moving-average models. *Journal of Time Series Analysis 13*, 327–343.

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.

Nocedal, J. and S. J. Wright (1999). *Numerical Optimization*. New York: Springer-Verlag.

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