

Outlier Detection in GARCH Models

Jurgen A. Doornik*

Nuffield College, University of Oxford, Oxford OX1 1NF, UK

Marius Ooms

*Department of Economics, Free University of Amsterdam
1081 HV Amsterdam, The Netherlands*

February 12, 2002

Abstract We present a new procedure for detecting additive outliers in GARCH(1,1) models. The outlier candidate is the observation with the largest standardized residual. First, a likelihood-ratio based tests determines the presence of an outlier. Next, a second LR test determines the type of outlier (volatility or level). This procedure is shown to be approximately independent from the GARCH parameters, with a null distribution that can be easily approximated. The procedure outperforms alternative methods, especially when it comes to determining the date of the outlier. We apply the method to returns of the Dow Jones index, using monthly, weekly, and daily data. The procedure is extended to cover GARCH models with Student- t distributed errors.

Keywords: Dummy variable, GARCH, GARCH-t, Outlier detection.

JEL classification: C22

1 Introduction

Financial data typically show volatility clustering and so-called thick tails. The ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models were designed to capture these features. But, when estimating a GARCH model with normal errors, there are frequently more outliers than expected. Two approaches come readily to mind to address this issue: using a distribution with fatter tails, such as the Student- t distribution, or to treat the outliers as being generated separately, and using dummy variables to remove them. Here we are concerned with the latter, and discuss methods for outlier detection in GARCH models.

The focus in this paper is on additive outliers, for which we shall follow the classification of Hotta and Tsay (1998). They distinguish between additive outliers that only affect the level, but leave the variance unaffected, and those that also affect the conditional variance. We label the first type ‘ALO’, and the second ‘AVO’. Like Hotta and Tsay (1998) and Franses and van Dijk (1999), our approach is inspired by Chen and Liu (1993), who discuss outlier detection in standard time-series models. Our approach, however, is based on likelihood-ratio tests, instead of lagrange-multiplier tests, which leads to much simpler procedures than either Hotta and Tsay (1998) or Franses and van Dijk (1999).

The new procedure for outlier detection builds on previous work (Doornik and Ooms, 2000), which studies the impact of a dummy variable on the GARCH likelihood. In that paper, we give the

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

conditions under which bimodality arises when adding a single-observation dummy variable to the mean of a GARCH(p, q) model. Interestingly, bimodality does not always happen, but tends to be more likely when there are outliers. We also show that adding the corresponding dummy with a lag of one period in the variance equation solves the problem of bimodality. The procedure developed in this paper is based upon this observation.

The organization of this paper is as follows. In §2 we review the two types of additive outliers introduced by Hotta and Tsay (1998). We then propose a nesting model for additive outliers in §3 and use this as the basis for a new likelihood-based detection procedure in §4. The next two sections investigate the size and power of the new procedure. Then in §7 we apply the procedure to the Dow Jones index, at monthly, weekly, and daily frequencies. In §8 we extend the new procedure to GARCH-t and GARCH(2,2) models. Finally, §9 concludes.

2 Additive outliers in GARCH models

The GARCH(p, q) regression model with normally distributed errors is defined as:

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

\mathcal{F}_t is the filtration up to time t . In practice, x_t may only consist of the constant term. Recent surveys include Bollerslev, Engle, and Nelson (1994), Shephard (1996), and Gouriéroux (1997). The log-likelihood of (1) is given by:

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

For a GARCH(1,1) model with $0 \leq \beta_1 < 1$, which is the main focus, we can write

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

as

$$h_t = \alpha_0^* + \alpha_1 \sum_{j=1}^{t-1} \beta_1^{j-1} \varepsilon_{t-j}^2, \quad (3)$$

given ε_0 and h_0 , where $\alpha_0^* = \alpha_0(1 - \beta_1^t)/(1 - \beta_1) + \beta_1^t h_0$.

2.1 Additive level outliers (ALO)

The GARCH(1,1) model with an additive level outlier is:

$$\begin{aligned} y_t - x_t' \zeta - \gamma d_t &= \varepsilon_t, \quad \varepsilon_t | \mathcal{F}_{t-1} \sim N(0, h_t), \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \quad t = 1, \dots, T, \end{aligned} \quad (4)$$

where d_t equals one when $t = s$ and zero otherwise. In (4) the outlier is removed from the lagged disturbances that enter the conditional variance. The occasion could be a market correction that does not influence volatility, an institutional change, or even a rogue trade.

Model (4) is a standard GARCH model with a dummy variable as regressor. Although this data generation process is well-defined, estimation is problematic because of the potential for bimodality in the likelihood. Doornik and Ooms (2000) show that, when bimodality occurs, the estimate of γ that sets the residual to zero, $\hat{\gamma} = y_s - x'_s \hat{\zeta}$, corresponds to a local minimum of the log-likelihood. Inference based on t -statistics in particular can be difficult.

2.2 Additive volatility outliers (AVO)

The GARCH(1,1) model for an additive volatility outlier is:

$$\begin{aligned} y_t - x'_t \hat{\zeta} &= \varepsilon_t^*, & \varepsilon_t^* &= \gamma d_t + \varepsilon_t, & \varepsilon_t | \mathcal{F}_{t-1} &\sim N(0, h_t^*), \\ h_t^* &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^{*2} + \beta_1 h_{t-1}^*, & t &= 1, \dots, T, \end{aligned} \quad (5)$$

where d_t equals one when $t = s$ and zero otherwise. The likelihood is now defined in terms of h_t^* and ε_t . Substituting ε_t :

$$h_t^* = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}^* + \alpha_1 (2\gamma \varepsilon_{t-1} + \gamma^2) d_{t-1}. \quad (6)$$

Expressing (6) in terms of h_t we find from (3):

$$h_t^* = h_t + \alpha_1 \beta_1^{t-s-1} (2\gamma \varepsilon_s + \gamma^2) I(t > s), \quad (7)$$

where $I(t > s)$ equals one when $t > s$, and zero otherwise.

In (5), the outlier has an impact on the volatility that diminishes (assuming $\beta_1 < 1$). In particular, when $\varepsilon_s = 0$, both a negative and a positive outlier increase volatility.

MLE of AVO is straightforward. Because the volatility is in terms of ε_t^* , and not ε_t , $\partial h_t^* / \partial \gamma = 0$, and therefore $\hat{\varepsilon}_s = 0$. Bimodality is not an issue, and $\hat{\gamma} = y_s - x'_s \hat{\zeta}$, with variance h_s^* . Detection of an outlier of type AVO simplifies to finding large standardized residuals. As noted in Chen and Liu (1993), estimation of the other model parameters is contaminated by the presence of an outlier.

When an outlier is found, full estimation of (5) is required, which is not readily possible in most current software packages (but is a simple extension).

3 A nesting model for additive outliers

In a general GARCH(p, q) model, using the lag operator L , and assuming that $1 - \beta(L)$ is invertible, we start by introducing a lagged dummy in the variance:

$$h_t = \alpha_0 + \alpha(L) \varepsilon_t^2 + \beta(L) h_t + \tau d_{t-1},$$

which can be written as:

$$h_t = \frac{\alpha_0}{1 - \beta(1)} + \frac{\alpha(L)}{1 - \beta(L)} \varepsilon_t^2 + \frac{\tau L}{1 - \beta(L)} d_t I(t > s),$$

where $I(t > s)$ is one when $t > s$ and zero otherwise.

For the model with an additive volatility outlier, extending (5):

$$\begin{aligned} \varepsilon_t^* &= \gamma d_t + \varepsilon_t, \\ h_t^* &= \alpha_0 + \alpha(L) \varepsilon_t^{*2} + \beta(L) h_t^*, \end{aligned}$$

we find :

$$h_t^* = \frac{\alpha_0}{1 - \beta(1)} + \frac{\alpha(L)}{1 - \beta(L)} \varepsilon_t^2 + \frac{\alpha(L)}{1 - \beta(L)} (2\gamma\varepsilon_t + \gamma^2) d_t I(t > s). \quad (8)$$

In the AVO case the additional term multiplying $d_{t-1}I(t > s)[1 - \beta(L)]^{-1}$ is $\alpha(L)L^{-1}(2\gamma\varepsilon_t + \gamma^2)$, while in the model with a lagged dummy in the volatility it is τ , where τ is estimated. The latter can be interpreted as an unrestricted version of the AVO model. Next, we refer to Doornik and Ooms (2000), who show that, in a GARCH(p, q) model with a dummy in the mean and the same dummy lagged one period in the variance, the bimodality disappears. Moreover, the dummy in the mean now sets the corresponding residual to zero.

So, in a GARCH(1,1) model, the outlier candidate is the largest standardized residual when the alternative is given by:

$$\begin{aligned} y_t &= x_t' \zeta + \gamma d_t + \varepsilon_t, \\ h_t &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} + \tau d_{t-1}. \end{aligned} \quad (9)$$

Therefore, (9) nests both the AVO and ALO case, without the complexity that is created by the bimodality of the log-likelihood. We propose to use this fact for a likelihood ratio test: estimate a standard GARCH(1,1) model, use the largest standardized residual as the outlier candidate, then perform a likelihood-ratio type test relative to (9). This procedure is simple enough that it can be carried out using standard GARCH software, without the need for additional programming.

If an outlier is detected, it will be of interest to test if it is of type AVO or ALO. If the outlier is AVO, $\hat{\tau}$ provides an estimate of $\alpha_1(2\gamma\varepsilon_s + \gamma^2)$ for the GARCH($p, 1$) model. This provides one estimate of $\hat{\gamma}$, because we only use the smallest (in absolute value) solution of the quadratic term. This solution allows for possible bimodality in the log-likelihood. The alternative solution is the residual from the original model. Estimating the original model twice, with the γ candidates subtracted from the dependent variable, gives two further likelihood values. The best of these can be compared with that of the model with two dummies, and, if the former cannot be rejected in favour of the latter, we may conclude that a level outlier was found.

4 The outlier detection procedure

Based on the previous section, we propose the following procedure to detect an outlier in a GARCH(1,1) model:

Step 1 Estimate the baseline GARCH model to obtain log-likelihood $\hat{\ell}_b$ and residuals ε_t^* .

Step 2 Find the largest (in absolute value) standardized residual $\max_t |\varepsilon_t^*/h_t^*|$, at $t = s$, say. Estimate the extended GARCH model with dummy $d_t \equiv I(t = s)$ in the mean, and d_{t-1} in the variance. This gives estimates for the added parameters $\hat{\gamma}_m$ and $\hat{\tau}_m$ respectively, with log-likelihood $\hat{\ell}_m$. The estimated GARCH parameters are $\hat{\alpha}_{1,m}, \hat{\beta}_{1,m}$.

Step 3 If $2(\hat{\ell}_m - \hat{\ell}_b) < C_T^\alpha$ then terminate: no further outliers are present. Our simulations suggest that $C_T \approx 5.66 + 1.88 \log T$ at 5%. The full approximation is given in the next section.

The next two steps implement the AVO versus ALO test, given that an outlier was detected:

Step 4 Estimate the baseline GARCH model for $y_t - \hat{\gamma}_m d_t$, giving $\hat{\ell}_0$.

The next stage solves $(2\gamma\varepsilon_s^* - \gamma^2) = \tau/\alpha_1$ for γ , using $\varepsilon_s^* = \hat{\gamma}_m$ and $\hat{\tau}_m, \hat{\alpha}_{1,m}$. If $\hat{\gamma}_m^2 - \hat{\tau}/\hat{\alpha}_{1,m} > 0$ compute:

$$\hat{\gamma}_1 = \begin{cases} \hat{\gamma}_m - (\hat{\gamma}_m^2 - \hat{\tau}/\hat{\alpha}_{1,m})^{1/2} & \text{if } \hat{\gamma}_m \geq 0, \\ \hat{\gamma}_m + (\hat{\gamma}_m^2 - \hat{\tau}/\hat{\alpha}_{1,m})^{1/2} & \text{if } \hat{\gamma}_m < 0. \end{cases}$$

else set $\hat{\gamma}_1 = 0$. Estimate the baseline GARCH model for $y_t - \hat{\gamma}_1 d_t$, giving $\hat{\ell}_1$. If $\hat{\ell}_0 \geq \hat{\ell}_1$ set $\hat{\gamma}_2 = \hat{\gamma}_m$ and $\hat{\ell}_2 = \hat{\ell}_0$, else set $\hat{\gamma}_2 = \hat{\gamma}_1$ and $\hat{\ell}_2 = \hat{\ell}_1$.

Step 5 If $2(\hat{\ell}_m - \hat{\ell}_2) \leq 3.84$ then outlier is of type ALO, with estimated coefficient $\hat{\gamma}_2$. Else the outlier is of type AVO, with estimated coefficient $\hat{\gamma}_m$.

Step 4 is used to distinguish between the two types of outliers, in case one is detected. It involves two additional GARCH estimations, which can be started using estimates from step 1 (i.e. the baseline model without any outlier effects). The same can be done in step 2, so that the additional overhead of the three maximum likelihood estimations is small.

The procedure can be iterated until no further outlier is detected. Because the outlier coefficients have already been estimated at each step, we propose a simple data correction when an outlier is detected. An additive level outlier involves adjusting the raw data prior to the next estimation ($y_t \leftarrow y_t - \hat{\gamma}_2 d_t$), whereas the volatility outlier uses the unadjusted residuals in the h_t , but the adjusted values in the log-likelihood function. The latter will require a simple extension to the GARCH code. This data adjustment procedure avoids a proliferation of parameters in the log-likelihood. moreover, an ALO cannot be corrected by adding a dummy variable for estimation, because of the bimodality issue. A new type of regressor could be created for AVO, but we found that estimation of the standard errors appears unreliable.

5 Size of the outlier detection procedure

To determine an appropriate distribution for the test procedure we first we simulate steps 1–3 as described in the previous section under the null hypothesis of no outlier. The Monte Carlo uses $M = 10\,000$ replications, a constant in the mean, and $\alpha_1 = 0.1, \beta_1 = 0.8, \alpha_0 = 1 - \alpha_1 - \beta_1$. The sample sizes are $T = 200(100)1200, 1500, 2000, 2500$. In the estimation procedure, we always impose $0 < \hat{\alpha}_1 + \hat{\beta}_1 \leq 1$ and $\hat{\alpha}_0 > 0$. The results indicate that the mean of the test statistic is increasing with the sample size, but less so as T gets large. The variance, skewness and kurtosis are not very sensitive to the sample size.

If all the tests are independent, the distribution could be described by the maximum of a random sample of size T from a $\chi^2(2)$ distribution. Some casual experimentation indicates that the $\chi^2(2)$ works reasonably well for a single test: 10% and 5% rejection frequencies are 0.126 and 0.066 at $T = 500, M = 1000$ for the given design. The maximum of a $\chi^2(2)$ sample can be approximated by an extreme value distribution, which is combined in Appendix A with the simulation results to form the approximation for the LR statistic of Step 3:

$$P(Y \leq y) = \exp \left\{ - \exp \left[- \frac{y - 1.283 - 1.88 \log T (1 + 12/T)}{2.223} \right] \right\}. \quad (10)$$

To check the accuracy of the approximation, we simulate the rejection frequencies for various parameter values under the null hypothesis. Table 1 lists the empirical size, showing that the procedure works well enough for practical use. The table also illustrates that the approximation works well for a range of GARCH parameters, indicating that the test is asymptotically similar with respect to α_1, β_1 .

Table 1: Size of test for single outlier using extreme value approximation

α_1	β_1	T	20%	10%	5%	1%
0.6	0.2	500	0.184	0.091	0.046	0.013
0.4	0.2	500	0.189	0.093	0.045	0.012
0.2	0.4	500	0.191	0.094	0.048	0.011
0.2	0.6	500	0.194	0.094	0.048	0.009
0.05	0.9	500	0.204	0.108	0.056	0.015
0.1	0.8	250	0.191	0.102	0.055	0.012
0.1	0.8	500	0.191	0.097	0.049	0.013
0.1	0.8	1000	0.195	0.100	0.056	0.011
0.1	0.8	2500	0.199	0.097	0.050	0.012
<i>ASE</i>			<i>0.006</i>	<i>0.005</i>	<i>0.003</i>	<i>0.002</i>

Based on $M = 4\,000$ replications.

ASE: Monte Carlo standard error of the rejection frequencies.

Table 2: Size and power of proposed outlier detection test for a single outlier in a GARCH(1,1) model

α_1, β_1	<i>Rejection frequencies</i>				<i>Correct date</i>		<i>Correct type</i>	
	$\gamma = 0$	-3	-4	-5	-4	-5	-4	-5
Outlier of type AVO at $T/2$								
0.1,0.8	0.051	0.22	0.52	0.83	98%	99%	48%	53%
0.3,0.5	0.047	0.21	0.51	0.82	98%	99%	58%	64%
0.5,0.3	0.040	0.21	0.51	0.82	98%	99%	62%	68%
Outlier of type ALO at $T/2$								
0.1,0.8	0.051	0.27	0.57	0.82	98%	99%	89%	91%
0.3,0.5	0.047	0.38	0.68	0.85	98%	99%	93%	93%
0.5,0.3	0.040	0.53	0.77	0.87	99%	100%	92%	92%

Based on 5% nominal rejection frequencies for $M = 4\,000$ and $T = 250$.

Correct date: % with the correct date when an outlier was detected.

Correct type: % with the correct outlier type when an outlier was detected.

100% correct date is: > 99.5% correct.

6 Power of the outlier detection procedure

To investigate the power of the proposed test procedure, we select $T = 250$, and have the DGP of type AVO as in (5) as well as of type ALO as in (4).¹ The DGP parameters are set as $\alpha_0 = 1 - \alpha_1 - \beta_1$, with $\gamma = -3, -4, -5$. The outlier enters at $T/2$. The first column Table 2 gives the GARCH design parameters. The next four columns give the rejection frequencies at a 5% significance level. The results for $\gamma = 0$ correspond to the size of the test, confirming a level close to 5%. The remainder shows that the proposed procedure has satisfactory power to detect the outlier, regardless of the type of outlier. It is also remarkably good at detecting the date (i.e. the location) of the outlier, which, of course, is an important aspect of any detection procedure. (Note that the correct date and type

¹So we do not force γ to enter the DGP with the same sign as the drawn residual.

percentages are conditional on detection of an outlier.)

Our procedure is not so good at detecting the type of outlier: it is somewhat biased against AVO, when an outlier is detected. When the outlier is of type AVO, not much more than half of the outliers are detected as such; for ALO-type outliers, the success rate increases to over 90%. There is some evidence that, as α_1 gets larger, there is an increased success rate of finding AVO outliers when they are true. Decreasing the critical value is not a solution: we would detect more AVO when it is true, but also when ALO is the correct type. We found similar results to Table 2 when the outlier was dated at $T/4$ or $3T/4$.

7 Some applications

We apply the new outlier detection procedure to the returns on the Dow Jones Industrial Average index,² using monthly, weekly, and daily data for the period 1896, May 26, to 2001, December 5:

frequency	index at	no. of observations	scale
daily	close of trade	29269	276
weekly	midweek (or nearest day before)	5422	51
monthly	end of month	1264	12

The return data is formed by taking the first difference of the logarithms. This was multiplied by the scale factor given above: selected as the integer which made the annualized average for the daily and weekly returns as close as possible to the monthly data.

Visual inspection of the daily returns shows the largest drop in 1914, followed closely by 1987. In 1914, the exchange was closed for four and a half month following the outbreak of World War I. So there is a long period of missing data in 1914 (during that period, grey trading continued outside the exchange). The year 1929 is characterized by boom and bust, followed by a period of long decline, and is historically the period with the highest volatility. October 1987 saw the largest one-day drop in the index, but it took less than two years to reach the pre-crash levels again. The most recent sharp fall followed the 11 September 2001 terrorist attacks on Washington and New York. The results below indicate that this tragedy is only visible as an outlier in the daily data.

The top half of Table 3 lists the results when applying the new procedure to the monthly data. Detected are the 1987 crash, the start of the two world wars, as well as September 1937 (in which the index dropped by 17%). The order in the table is that in which the outliers were detected, and we also include the first outlier that was rejected.

The second part of Table 3 gives the results for the weekly data. We see more AVO outliers, as expected. At different frequencies, the pattern of outliers will also be different: a brief crash or rally within a month can be hidden by only looking at the end-of-month data. The world wars are now the largest outliers, and WWII has become AVO. Also, the fall of December 1899 (13% down in second week) is detected before the 1987 crash. The new procedure has two clear benefits: it is a nested procedure, avoiding the need to have to compare p -values of two separate tests, possibly at different dates. It is also easy to compute p -values at the second stage, allowing for better classification in ALO and AVO.

In Table 4 we list the dates of outliers for the daily model in chronological order. There are more than five times the number of observations than in the monthly data set, but also five times as many outliers. The new procedure is found to be acceptably fast on the daily data, taking about half an hour for nearly 30 000 observations (on a 700 Mhz Pentium III notebook; this includes the first estimation).

²The Dow Jones index data are available from www.djindexes.com.

Table 3: Detected outliers in GARCH(1,1) model for monthly and weekly Dow Jones returns.

date	type	size	p -outlier	p -ALO
monthly returns: $12\Delta \log y_t^m$				
1987/10	ALO	-4.38	0.0 ₈ 3	0.795
1914/12	ALO	-3.58	0.00012	0.112
1940/05	ALO	-3.11	0.00018	0.251
1937/09	AVO	-2.37	0.036	0.002
2001/09	—		0.139	
weekly returns: $51\Delta \log y_t^w$				
1914/12/16	ALO	0	0.244	-16.75
1940/05/15	AVO	0	0	-7.05
1899/12/13	AVO	0.0 ₈ 3	0.026	-7.14
1987/10/21	AVO	0.0 ₅ 3	0.010	-8.95
1926/03/03	AVO	0.00015	0.002	-4.84
1898/05/11	ALO	0.00020	0.960	7.61
1994/03/30	ALO	0.00075	0.536	-3.39
1998/09/02	—	0.070		

p -ALO is for testing ALO, when an outlier is detected.
 Notation: 0.0₄5 = 0.00005

Table 4: Detected outliers using the new procedure in GARCH(1,1) model for daily Dow Jones returns: $276\Delta \log y_t^d$

date	type	p -outlier	date	type	p -outlier	date	type	p -outlier
1899/12/08	AVO	0.0 ₄ 3	1924/02/15	ALO	0.0037	1950/06/26	AVO	0.0 ₆ 6
1901/05/08	AVO	0.0008	1925/11/10	ALO	0.0016	1955/09/26	AVO	0
1901/09/07	AVO	0.0212	1927/10/08	ALO	0.0329	1962/05/28	AVO	0.0039
1904/12/07	AVO	0.0 ₄ 6	1929/10/28	AVO	0.0002	1982/08/17	AVO	0.0056
1907/03/14	AVO	0.0005	1933/03/15	ALO	0.0 ₄ 8	1986/09/11	ALO	0.0034
1913/01/20	ALO	0.0 ₆ 6	1934/07/26	ALO	0.0068	1987/10/19	AVO	0
1914/07/28	ALO	0.0 ₄ 5	1939/09/05	ALO	0.0031	1989/10/13	AVO	0
1914/07/30	AVO	0.0 ₄ 3	1940/05/13	AVO	0.0 ₆ 3	1991/01/17	ALO	0.0160
1914/12/12	ALO	0	1943/04/09	ALO	0.0004	1991/11/15	ALO	0.0 ₆ 3
1916/12/12	AVO	0.0013	1946/09/03	AVO	0.0034	1997/10/27	AVO	0.0 ₄ 3
1917/02/01	ALO	0.0 ₆ 1	1948/11/03	AVO	0.0 ₄ 8	2000/04/14	ALO	0.0157
						2001/09/17	AVO	0.0002

The results assume that the underlying model is GARCH(1,1), contaminated with outliers. Outliers only exist with reference to a model, and using the wrong model could lead to the detection of too many outliers. Especially for the daily data, it may be that a GARCH model with student- t distributed errors is a better description. This is explored in the next section.

8 Extensions to other models

8.1 GARCH(2,2) models

In the GARCH(p, q) case, (8) has q additional terms, and the equivalent extension would be to add the dummy with lags 1 to q in the variance. The first lagged residual has non-zero expectation under the alternative hypothesis, but further lags do not, so a restriction may have to be imposed. As a simple alternative we just apply the same procedure as for GARCH(1, 1), leaving the approximation to the distribution unchanged. Table 5 shows that the resulting size and power are very close to that in the GARCH(1, 1) case.

Table 5: Size and power of proposed outlier detection test for a single outlier in a GARCH(2,2) model

$\alpha_1, \alpha_2; \beta_1, \beta_2$	<i>Rejection frequencies</i>			
	$\gamma = 0$	-3	-4	-5
Outlier of type AVO at $T/2$				
0.1,0.1;0.1,0.6	0.07	0.21	0.49	0.79
0.1,0.1;-0.1,0.8	0.07	0.18	0.45	0.75
Outlier of type ALO at $T/2$				
0.1,0.1;0.1,0.6	0.07	0.35	0.63	0.83
0.1,0.1;-0.1,0.8	0.07	0.33	0.63	0.85

5% nominal rejection frequencies for $M = 2\,000, T = 250$.

8.2 GARCH- t models

A GARCH model with Student- t distributed errors, as proposed by Bollerslev (1987), is a likely alternative for a GARCH model with outliers. Appendix A gives the adjustments that need to be made to the extreme value approximation when incorporating the standardized $t(\nu)$ distribution. Table 6 presents some results for the test. The actual outliers have to be considerably larger to be detected among the thicker tail of the Student- t distribution.

Table 6: Size and power of proposed outlier detection test for a single outlier in a GARCH(1,1)- $t(6)$ model

	<i>Rejection frequencies, $\alpha_1 = 0.1, \beta_1 = 0.8$</i>				
	$\gamma = 0$	-5	-8	-10	-15
Outlier of type AVO at $T/2$	0.046	0.04	0.08	0.22	0.76
Outlier of type ALO at $T/2$	0.046	0.05	0.26	0.50	0.84

Based on 5% nominal rejection frequencies for $M = 2\,000$ and $T = 1\,000$.

Empirical application to the Dow Jones industrial averages index, supports the closeness of the outlier-corrected GARCH(1,1) and the GARCH(1,1)- t model. Table 7 shows that at the monthly and weekly level, the two models seem to be close substitutes, with the outlier-corrected model weakly preferred on AIC. At the daily level, the GARCH- t is preferred, yielding a higher log-likelihood and lower AIC than the model with outliers.

For each frequency we also applied the GARCH- t outlier test to the GARCH- t models. Only for the weekly data were outliers detected: ALO when the market reopened after WWI, and AVO at the start of WWII. These are the same two leading outliers found in the GARCH(1,1) model. However, in terms of AIC it is not an improvement over the normal GARCH(1,1) model with outliers. For the monthly data, the candidate was 1987/10, with p -value of 0.052. In daily data, the candidate was 1955/09/26, also the first found in the GARCH(1,1) model, but now with p -value of 0.10 rather than zero.

Table 7: Estimated GARCH(1,1) coefficients

	GARCH(1,1)	with outliers	GARCH(1,1)- $t(\nu)$	with outliers
Monthly data: $12\Delta \log y_t^m$				
c	0.068 (0.015)	0.079 (0.015)	0.095 (0.015)	
α_0	0.014 (0.0040)	0.013 (0.0039)	0.017 (0.0056)	
α_1	0.114 (0.019)	0.102 (0.018)	0.102 (0.022)	
β_1	0.862 (0.021)	0.866 (0.021)	0.861 (0.027)	
α_0^*	0.582	0.404	0.459	
ν			5.357	
outliers	0	4	0	
log-lik	-1189.0	-1122.3	-1133.8	
AIC	1.889	1.790	1.803	
Weekly data: $51\Delta \log y_t^w$				
c	0.100 (0.013)	0.089 (0.013)	0.111 (0.013)	0.110 (0.013)
α_0	0.063 (0.0077)	0.023 (0.0039)	0.027 (0.0057)	0.025 (0.0052)
α_1	0.149 (0.012)	0.095 (0.0079)	0.091 (0.011)	0.090 (0.010)
β_1	0.820 (0.013)	0.888 (0.0084)	0.892 (0.012)	0.894 (0.012)
α_0^*	2.036	1.437	1.644	1.604
ν			7.151	7.839
outliers	0	7	0	2
log-lik	-8372.4	-8120.2	-8162.0	-8128.0
AIC	3.090	3.000	3.013	3.000
Daily data: $276\Delta \log y_t^d$				
c	0.120 (0.012)	0.123 (0.012)	0.145 (0.011)	
α_0	0.105 (0.0070)	0.070 (0.0052)	0.082 (0.0085)	
α_1	0.094 (0.0032)	0.072 (0.0026)	0.080 (0.0041)	
β_1	0.896 (0.0032)	0.918 (0.0027)	0.912 (0.0043)	
α_0^*	10.69	6.852	10.09	
ν			5.670	
outliers	0	34	0	
log-lik	-67539.8	-66693.4	-66476.7	
AIC	4.616	4.562	4.543	

$$\alpha_0^* = \alpha_0 / (1 - \alpha_1 - \beta_1)$$

9 Conclusion

We introduced a new detection procedure for outliers in GARCH models. This procedure has several advantages over existing procedures:

- simple to implement,
- likelihood-based, with asymptotical similarity with respect to α_1, β_1 ,
- convenient procedure to compute p -values, without the need for simulation,
- nested test to distinguish between ALO and AVO, avoiding the need to have to compare p -values of two separate tests, possibly at different dates.

Our applications show that the test procedure also works well in practice. Although the in-sample fit of a GARCH-t and normal-GARCH with outliers may be quite similar, the forecasted volatility will be quite different. It may be that the former is preferred in practice, for example for value-at-risk estimations.

Acknowledgements

We wish to thank Siem Jan Koopman for helpful discussions and suggestions. Financial support from the UK Economic and Social Research Council (grant R000237500) is gratefully acknowledged by JAD. JAD would also like to thank the Graduate School of Business at Stanford University for their hospitality while writing this paper. The computations were performed using the Ox programming language (Doornik, 2001).

A Approximating the distribution of the outlier test

A single likelihood-ratio test involves two parameters, giving the test statistic X_i an asymptotic $\chi^2(2) \equiv \exp(1/2)$ distribution. We do T such tests, and wish to find the distribution of the maximum: $Y = \max(X_1, \dots, X_T)$. Assuming independently and identically distributed X_i :

$$F_Y(y) = \{F_X(y)\}^T = \left\{1 - e^{-\frac{1}{2}y}\right\}^T.$$

Using

$$\frac{1}{T} \log F_Y(y) = \log \left(1 - e^{-\frac{1}{2}y}\right) \approx -e^{-\frac{1}{2}y},$$

when y is large, gives

$$F_Y(y) \approx \exp \left\{-T e^{-\frac{1}{2}y}\right\},$$

which is an extreme value distribution: $Y \sim EV(2 \log T, 2)$. In general, when $W \sim EV(a, b)$, then:

$$F_W(w) = \exp \left\{-\exp \left(-\frac{w-a}{b}\right)\right\}, \quad (11)$$

$E[W] \equiv m = a + \gamma b$, where $\gamma \approx 0.577216$, and $V[W] = b^2 \pi^2 / 6$. Critical values at significance level α can be computed as

$$C_T^\alpha = -b \log(-\log(1 - \alpha)) + a. \quad (12)$$

Although the sequence of LR tests in the GARCH(1,1) case is not independently distributed, we use the extreme value distribution (11) as the approximating distribution. This is supported by the simulations of §5: the 20%, 10%, 1%, 1% simulated critical values as a function of sample size have a fixed distance. The simulated standard deviation of the test statistic is close to constant. Its asymptotic value is 2.85, found from a regression on a constant, T^{-1} and T^{-2} ; at small samples (T up to 1000) it is about 2.9. The mean is well described by $\log T$, although with a coefficient less than two. The resulting response surface for m, b as a function of T is:

$$\begin{aligned} m &\approx 1.88 \log T (1 + 12/T), \\ b &\approx 2.223, \end{aligned} \quad (13)$$

where $a = m - \gamma b$.

Based on GARCH(1,1)- $t(\nu)$ simulations for $\nu = 4, 5, 6, 9, 13$, the following adjustments can be used to approximate the distributions for the outlier test in the GARCH(1,1)- $t(\nu)$ model:

$$\begin{aligned} m(\nu) &\approx m + 11\nu^{-1} + 0.25m\nu^{-1/2}, \\ b(\nu) &\approx b + 12\nu^{-2}. \end{aligned} \quad (14)$$

We assumed that the variance is approximately constant, although the simulations show it to be somewhat u-shaped for ν ranging from 4 to 6. The response surface for the mean fits remarkably well.

B Alternative outlier detection procedures

B.1 Additive volatility outliers

Hotta and Tsay (1998) propose an LM test on the largest standardized residual:

$$LM^{AVO} = \max_{1 < t < T} \frac{\hat{\varepsilon}_t^2}{\hat{h}_t}.$$

This is approximately distributed as the maximum of a random sample of size $T - 2$ from a $\chi^2(1)$ distribution.³

³They actually use the maximum from a sample of size $T - 20$, using 20 observations to initialize the GARCH recursion. This, however, is a rather non-standard method for estimating GARCH models. The most commonly used procedure is to use the mean of the squared residuals, see Fiorentini, Calzolari, and Panattoni (1996).

B.2 Additive level outliers

Hotta and Tsay (1998) propose an LM test for the ALO case:

$$\text{LM}^{\text{ALO}} = \max_{1 < t < T} \frac{\hat{\varepsilon}_t^2}{\hat{h}_t} \frac{\left\{ 1 + \hat{\alpha}_1 \hat{h}_t \sum_{j=t+1}^J \hat{\beta}_1^{j-(t+1)} \hat{h}_j^{-2} (\hat{h}_j - \hat{\varepsilon}_j^2) \right\}^2}{1 + 2\hat{\alpha}_1^2 \hat{h}_t^2 \sum_{j=t+1}^J \hat{\beta}_1^{2[j-(t+1)]} \hat{h}_j^{-2}}.$$

$t < J \leq T$ is a truncation parameter that is introduced to avoid ‘swamping’. The distribution of LM^{ALO} depends on the choice of J , and the true values of α_1 and β_1 , requiring simulation for every test. Finally, they suggest, when both LM^{ALO} and LM^{AVO} are significant, to adopt the one with the most significant value. The p -values of LM^{ALO} can only be obtained by simulation, which can hinder the decision between outlier types: if the AVO test has a very small p -value, many replications are required to decide whether the ALO test has an even smaller p -value or not. Moreover, there is no guarantee that the candidate outliers for both tests occur at the same observation.

Franses and van Dijk (1999) suggest the following procedure for detecting additive level outliers in GARCH(1,1) models. Using $u_t = \varepsilon_s^2 - h_t$ they rewrite (7) as (so this is under the impact of a neglected outlier):

$$u_t^* = \phi \{ I(t = s) - \alpha_1 \beta_1^{t-s-1} I(t > s) \} + u_t,$$

where $\phi = 2\gamma\varepsilon_s + \gamma^2$. From this they estimate $\hat{\phi}$ by OLS, and:

$$\hat{\gamma}_s = \begin{cases} 0 & \text{if } \varepsilon_s^{*2} - \hat{\phi} < 0, \\ \varepsilon_s^* - (\varepsilon_s^{*2} - \hat{\phi})^{1/2} & \text{if } \varepsilon_s^{*2} - \hat{\phi} \geq 0 \text{ and } \varepsilon_s^* \geq 0, \\ \varepsilon_s^* + (\varepsilon_s^{*2} - \hat{\phi})^{1/2} & \text{if } \varepsilon_s^{*2} - \hat{\phi} \geq 0 \text{ and } \varepsilon_s^* < 0. \end{cases}$$

The largest $\hat{\gamma}_s$ exceeding a certain critical value is used to remove the outlier from the data (an approximation is offered for certain significance levels). If one is found, at t_0 say, the procedure is repeated for $y_t - \hat{\gamma}_{t_0} I(t = t_0)$ until no further outliers are detected. This procedure could be combined with LM^{AVO} along the lines suggested by Hotta and Tsay (1998) (i.e. selecting the outcome with the smallest p -value). In both cases, the assumption is that the outlier is of the same sign as the observed residual. In addition, Franses and van Dijk (1999) select the smallest solution (in absolute value). Although a unique solution is found, this method does impose bimodality on the likelihood, even when it is not present. An will be illustrated in the empirical application in §7.

Both procedures are rather complex, and suffer from non-similarity with respect to the GARCH parameters.

B.3 Simulation comparison

Next, we contrast our procedure to these alternative methods, denoted FD for Franses and van Dijk (1999), and HT for Hotta and Tsay (1998). The results are in Table 8.⁴ FD is not designed to test for AVO, but will have some power against it; FD has lower power than HT when the outlier is of type ALO, but that is probably because HT actually uses two tests (a more appropriate comparison would be with LM^{ALO} only). HT and the new procedure have similar power, but the latter is much better at dating the outlier (surprisingly, HT is worse at dating for the larger outliers). In addition, the new procedure is more successful in classifying the outlier.

B.4 Application comparison

Table 9 lists the results when applying the three procedures to the monthly data. The order in the table is that in which the outliers were detected, and we also include the first outlier that was rejected.

⁴To compute the rejection frequency, we used the extreme value approximation (10) for the our procedure. For HT we used simulation based on 1000 replications and $J = 3$. For FD we use the given critical value approximation, except that we replace κ_ε with $\max(3, \kappa_\varepsilon)$. This is not a good solution, though, e.g. when $\alpha = 0.6$ and $\beta = 0.2$, we would use the value 3, but simulations find a size of 20% in that case.

Table 8: Size and power of outlier detection tests for a single outlier in a GARCH(1,1) model

	α_1, β_1	Rejection frequencies			Correct date		Correct type	
		$\gamma = 0$	-4	-5	-4	-5	-4	-5
Outlier of type AVO at $T/2$								
HT	0.1,0.8	0.047	0.55	0.84	97%	96%	50%	39%
HT	0.3,0.5	0.044	0.54	0.85	82%	75%	54%	48%
HT	0.5,0.3	0.045	0.54	0.85	70%	66%	60%	56%
Outlier of type ALO at $T/2$								
FD	0.1,0.8	0.050	0.45	0.73	91%	97%		
FD	0.3,0.5	0.042	0.30	0.55	82%	91%		
FD	0.5,0.3	0.075	0.27	0.50	72%	85%		
HT	0.1,0.8	0.047	0.58	0.82	97%	96%	75%	80%
HT	0.3,0.5	0.044	0.69	0.85	88%	78%	75%	81%
HT	0.5,0.3	0.045	0.76	0.86	72%	56%	60%	75%

HT is Hotta and Tsay (1998); FD is Franses and van Dijk (1999).
Further notes: see Table 2.

Table 9: Detected outliers in GARCH(1,1) model for monthly Dow Jones returns: $12\Delta \log y_t^m$

new procedure					
date	type	size	p -outlier	p -ALO	
1987/10	ALO	-4.38	0.083	0.795	
1914/12	ALO	-3.58	0.00012	0.112	
1940/05	ALO	-3.11	0.00018	0.251	
1937/09	AVO	-2.37	0.036	0.002	
2001/09	—		0.139		
Hotta and Tsay (1998)					
date	type	size	p -LM ^{AVO}	p -LM ^{ALO}	
1987/10	ALO	-4.38	0.074	0	
1914/12	ALO	-3.58	0.045	0	
1940/05	ALO	-3.11	0.008	0.002	
1899/12	ALO	-2.49	0.039*	0.025	
1937/09	AVO	-2.38	0.0457	0.046*	
1990/08	ALO	-1.79	0.052*	0.043	
2001/09	—		0.053	0.069**	
Franses and van Dijk (1999)					
date	type	size			
1987/10	ALO	-3.78			
1932/08	ALO	+3.44			
1940/05	ALO	-2.54			
1914/12	ALO	-2.76			

p -ALO is for testing ALO, when an outlier is detected.
* at date of subsequent outlier candidate; ** at 1907/3.
Notation: 0.045 = 0.00005

The procedure of Hotta and Tsay (1998) finds the same outliers, with two additional ones. This procedure

uses simulation to determine p -values for the ALO test. For large outliers, the result is a p -value of zero, because it would be too time consuming to find accurate values (we use 1000 replications and $J = 3$). In our implementation, ALO is selected over AVO in that situation.

Franses and van Dijk (1999)'s procedure only detects ALO, which is less of a problem with monthly data, nonetheless giving quite different results. This method was the only to detect a positive outlier in the monthly data: August 1932 saw a large upswing in the index. The size of the first detected outlier is different from the other methods, corresponding to the imposition of bimodality on the likelihood, even when it is not present. This could explain the subsequent difference in the detection path. For the weekly and daily results we exclude this method, because it would need to be combined with an AVO detection (adding LM^{AVO} is simple, but does require simulation to determine p -values). The consequence of only correcting for ALO in the weekly returns is that about twice as many outliers are found, often close to each other. This illustrates the advantages of implementing volatility outliers.

Table 10: Detected outliers in GARCH(1,1) model for weekly Dow Jones returns: $51\Delta \log y_t^w$

new procedure				
date	type	p -outlier	p -ALO	size
1914/12/16	ALO	0	0.244	-16.75
1940/05/15	AVO	0	0	-7.05
1899/12/13	AVO	0.0 ₈ 3	0.026	-7.14
1987/10/21	AVO	0.0 ₅ 3	0.010	-8.95
1926/03/03	AVO	0.00015	0.002	-4.84
1898/05/11	ALO	0.00020	0.960	7.61
1994/03/30	ALO	0.00075	0.536	-3.39
1998/09/02	—	0.070		
Hotta and Tsay (1998)				
date	type	p - LM^{AVO}	p - LM^{ALO}	size
1914/12/16	AVO	0	0**	-16.75
1940/05/15	AVO	0	0	-7.05
1899/12/13	ALO	0.0 ₉ 3	0	-7.14
1987/10/21	ALO	0.0 ₆ 5	0	-8.95
1898/05/11	ALO	0.0 ₄ 3*	0	7.61
1994/03/30	ALO	0.0 ₄ 3*	0	-3.39
1926/03/03	ALO	0.0 ₄ 2	0	-4.84
1998/09/02	ALO	0.030	0.018	-4.73
1929/10/30	AVO	0.043	0.061*	-8.67
1927/10/19	—	0.115	0.059	

* at subsequent outlier candidate.

** at previous observation: 1914/7/29.

Table 10 gives the results for the weekly data. The new procedure has two clear benefits: it is a nested procedure, avoiding the need to have to compare p -values of two separate tests, possibly at different dates. It is also easy to compute p -values at the second stage, allowing for better classification in ALO and AVO.

The new procedure is found to be considerably faster on the daily data, taking about half an hour for nearly 30 000 observations (on a 700 Mhz Pentium III notebook; this includes the first estimation). HT takes two and a half hours, requiring simulation, and FD more than seven hours. FD requires nearly 30 000 regressions for each test, but there may be scope for implementing this more efficiently.

References

- Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 51, 307–327.
- Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics* 69, 542–47.
- 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.
- Chen, C. and L. M. Liu (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* 88, 284–297.
- Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox* (4th ed.). London: Timberlake Consultants Press.
- Doornik, J. A. and M. Ooms (2000). Multimodality in the GARCH regression model. mimeo, Nuffield College.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.
- Fiorentini, G., G. Calzolari, and L. Panattoni (1996). Analytic derivatives and the computation of GARCH estimates. *Journal of Applied Econometrics* 11, 399–417.
- Franses, P. H. and D. van Dijk (1999). Outlier detection in the GARCH(1,1) model. Econometric Institute Report EI-9926/A, Erasmus University Rotterdam.
- Gourieroux, C. (1997). *ARCH Models and Financial Applications*. New York: Springer Verlag.
- Hotta, L. K. and R. S. Tsay (1998). Outliers in GARCH processes. mimeo, IMECC, Brazil and University of Chicago.
- 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.