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

September 20, 2005

Correspondence to:
Jurgen A. Doornik
Nuffield College
Oxford OX1 1NF
UK
email: jurgen.doornik@nuffield.ox.ac.uk

Abstract We present a new procedure for detecting multiple additive outliers in GARCH(1,1) models at unknown dates. The outlier candidates are the observations with the largest standardized residual. First, a likelihood-ratio based test determines the presence and timing of an outlier. Next, a second test determines the type of additive outlier (volatility or level). The tests are shown to be similar with respect to the GARCH parameters. Their null distribution can be easily approximated from an extreme value distribution, so that computation of p-values does not require simulation.

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 and applied to GARCH models with Student-t distributed errors.

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

Outlier Detection in GARCH Models

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. However, 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 treating 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 (2000), 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 (2000).

The new procedure for outlier detection builds on work by Doornik and Ooms (2000), which studies the impact of a dummy variable on the GARCH likelihood. In that paper, we give the conditions under which bimodality arises when adding a single-observation dummy variable to the mean equation of a GARCH(p, q) model. Interestingly, bimodality does not always happen, but tends to be more likely when there is an outlier. We also show there that adding the corresponding dummy with a lag of one period in the variance equation solves the problem of bimodality. The procedure developed below 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. Some examples to illustrate the procedure are given in §4, with a more formal description in §5. The next two sections investigate the size and power of the proposed procedure. Then in §8 we apply the procedure to the Dow Jones index, at monthly, weekly, and daily frequencies. In §9 we extend the new procedure to GARCH-t and GARCH(2,2) models. Finally, §10 concludes. Appendix B compares our procedure with those proposed by Hotta and Tsay (1998) and Franses and van Dijk (2000).

2 Additive outliers in GARCH models

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

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

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

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

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

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

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 defined as:

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

where d_t equals one when $t = s$ and zero otherwise. In (4) the outlier does not influence 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, maximum likelihood estimation is problematic because of the potential for bimodality in the likelihood.

We assume that the start-up of the GARCH(1,1) process does not depend on the parameters. The score of the log-likelihood of model (4) is given by:

$$\sum_{t=1}^{T} \frac{\partial \ell_t(\theta)}{\partial \theta} = -\sum_{t=1}^{T} \left[\frac{\varepsilon_t}{h_t} \frac{\partial \varepsilon_t}{\partial \theta} + \frac{1}{2} \frac{1}{h_t^2} \left(h_t - \varepsilon_t^2 \right) \frac{\partial h_t}{\partial \theta} \right], \tag{5}$$

with $\varepsilon_t = y_t - x_t'\zeta - d_t\gamma$. The first order condition (5) for the dummy coefficient γ can be expressed as a function of ε_s and $h_{s+1}, h_{s+2}, \ldots h_T$, since

$$\frac{\partial \varepsilon_t}{\partial \gamma} = -d_t, \; \frac{\partial \varepsilon_t^2}{\partial \gamma} = -2\varepsilon_t d_t, \quad \text{thus} \quad \frac{\partial h_t}{\partial \gamma} = -2\alpha_1 \sum_{j=1}^{t-1} \beta_1^{j-1} \varepsilon_{t-j} d_{t-j}.$$

and $d_t = 0$ for $t \neq s$ and $d_s = 1$. The score term for h_{s+1} can lead to multiple solutions for γ, depending on the GARCH parameters, on h_s and on $\varepsilon_{s+2}, h_{s+2}, \varepsilon_{s+3}, h_{s+3}, \ldots$. This type of bimodality often appears in volatile periods. Doornik and Ooms (2000) show that, when this type of bimodality in the log-likelihood occurs, the 'standard' estimate of γ that sets the residual $\widehat{\varepsilon}_s$ to zero, $\widehat{\gamma} = y_s - x_s'\widehat{\zeta}$, corresponds to a local minimum of the log-likelihood, instead of a maximum. Inference based on t-statistics in particular is compromised, motivating our decision to use likelihood-ratio based tests instead of Wald tests.

2.2 Additive volatility outliers (AVO)

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

$$\begin{aligned} y_t - x_t'\zeta - \gamma d_t &= \varepsilon_t, \quad \varepsilon_t|\mathcal{F}_{t-1} \sim N(0, h_t^*), \\ \varepsilon_t^* &= \gamma d_t + \varepsilon_t, \\ h_t^* &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^{*2} + \beta_1 h_{t-1}^*, \quad t = 1, \ldots, T, \end{aligned} \tag{6}$$

where d_t equals one when $t = s$ and zero otherwise as in (4). The log likelihood is now defined in terms of h_t^* and ε_t, where h_t^* is affected by previous outliers.

To express h_t^* in terms of the clean conditional variance h_t and a dynamic effect of the outlier, we first substitute ε_t^*:

$$h_t^* = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}^* + \alpha_1 \left(2\gamma\varepsilon_{t-1} + \gamma^2\right) d_{t-1}. \tag{7}$$

Then we find from (3):

$$h_t^* = h_t + \alpha_1 \beta_1^{t-s-1} \left(2\gamma\varepsilon_s + \gamma^2\right) I(t > s), \tag{8}$$

where $I(t > s)$ equals one when $t > s$, and zero otherwise. So the outlier has an impact on the volatility that diminishes over time, assuming $\beta_1 < 1$. In particular, when $\varepsilon_s = 0$, both a negative and a positive outlier increase volatility.

Maximum likelihood estimation (MLE) of the additive volatility outlier (AVO) model (6) is not hampered by the multiple modes for γ. The score of the log-likelihood of model (6) is given by:

$$\sum_{t=1}^{T} \frac{\partial \ell_t(\theta)}{\partial \theta} = -\sum_{t=1}^{T} \left[\frac{\varepsilon_t}{h_t^*} \frac{\partial \varepsilon_t}{\partial \theta} + \frac{1}{2} \frac{1}{h_t^{*2}} \left(h_t^* - \varepsilon_t^2\right) \frac{\partial h_t^*}{\partial \theta} \right], \tag{9}$$

with $\varepsilon_t = y_t - x_t'\zeta - d_t\gamma$.

Because the volatility equation for h_t^* is in terms of ε_t^* and not ε_t, $\partial h_t^*/\partial\gamma = 0$, since $\partial\varepsilon_t^*/\partial\gamma = 0$. The only γ solving the first order condition for MLE leads to $\widehat{\varepsilon}_s = 0$. Bimodality is not an issue, and $\widehat{\gamma} = y_s - x_s'\widehat{\zeta}$, with variance h_s^*. Detection of an outlier of type AVO simplifies to inspecting the largest standardized residual. When an outlier is found, maximum likelihood estimation of (6) is required. This option is not readily available in most current software packages, but it would be a simple extension.

3 A nesting model for generalized additive outliers (GAO)

In this section we introduce a model for generalized additive outliers that nests both the additive level and the additive volatility outlier models in GARCH processes. The first step is to introduce a lagged dummy variable in the conditional variance equation of the GARCH(p, q) model:

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

where d_t is defined as before, such that d_{t-1} equals one when $t = s + 1$ and zero otherwise. The τ parameter models the effect of an outlier on the conditional variance at time $s+1$. The polynomials in the lag operator L, $L^k x_t = x_{t-k}$, are defined as $\beta(L) = 1 - \sum_{i=1}^{p} \beta_i L^i$, and $\alpha(L) = \sum_{i=1}^{q} \alpha_i L^i$. We assume that the roots of $\beta(z) = 0$ lie outside the unit circle, and that $\beta(z)$ and $\alpha(z)$ have no common roots to ensure identification of the individual GARCH parameters. Then:

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

For the model with an additive volatility outlier, extending (6) to GARCH(p, q) processes:

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

we find, again substituting ε_t^* :

$$h_t^* = \frac{\alpha_0}{\beta(1)} + \frac{\alpha(L)}{\beta(L)}\varepsilon_t^2 + \frac{\alpha(L)}{\beta(L)}\left(2\gamma\varepsilon_t + \gamma^2\right) d_t. \tag{10}$$

In this equation for the AVO model, which extends (7), we see that the additional term multiplying $\beta(L)^{-1}d_{t-1}$ is $\alpha(L)L^{-1}(2\gamma\varepsilon_{t-1} + \gamma^2)$, while in the model with a lagged dummy in the volatility it is τ, where τ is estimated. The latter can therefore be interpreted as an unrestricted version of the AVO model.

In the second step we add the ALO dummy variable to the mean equation of the GARCH model. For this step, we again refer to Doornik and Ooms (2000), who show that, in a GARCH(p, q) model with a dummy in the mean equation and the same dummy lagged one period in the variance equation, the bimodality problem discussed in §2.1 disappears as the first order condition for γ is simplified. This motivates the adoption of the generalized additive outlier (GAO) model, which for the GARCH(1,1) case is given by:

$$\begin{aligned} y_t &= x_t'\zeta + \gamma d_t + \varepsilon_t, \\ h_t &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1 h_{t-1} + \tau d_{t-1}. \end{aligned} \tag{11}$$

In this model, the dummy variable in the mean equation for y_t sets the corresponding residual to zero when γ is estimated by maximum likelihood: $\widehat{\varepsilon}_s = 0$. Moreover, (11) nests both the AVO and ALO model, without the complexity that is created by the bimodality of the log-likelihood.

We propose to take advantage of this easy estimation in likelihood ratio tests for the presence of additive outliers. In practice the timing of the outlier, s, is often unknown. In our outlier detection

procedure we estimate a standard GARCH(1,1) model, use the largest standardized residual as the outlier candidate, then perform a likelihood-ratio type test of $\gamma = \tau = 0$ in (11). This procedure is simple enough that it can be carried out using standard GARCH software which allows for adding separate explanatory variables in the mean equation and in the variance equation, without the need for additional programming. Of course, the asymptotic distribution of the likelihood ratio test statistic is not the standard χ^2 if the timing of the outlier is unknown. We have to take account of the search for the largest outlier and approximate the distribution as we would do for an order statistic. An effective approximation is derived in §6.

If focus is only on detection of a single additive outlier, the above procedure is sufficient. It may, however, be of interest to determine whether an outlier is of type ALO or AVO. The next section gives some motivating examples before formalizing the procedure.

4 Likelihood adjustment, outlier correction and outlier classification

To illustrate the properties of the likelihood-based outlier classification, it is necessary to be able to evaluate the likelihoods of the different outlier models as a function of outlier size γ, see Figure 1 below. This is also required when a detected outlier has to be accounted for in the model. Both likelihood adjustment and outlier extensions can be implemented by a simple data transformation, which adjusts the data for the effect of the outlier.

Taking account of an additive level outlier (ALO) only involves adjusting the raw data prior to the next estimation (i.e. replacing y_t with $y_t - \widehat{\gamma} d_t$). Taking account of an additive volatility outlier (AVO) is slightly more complicated. The log-likelihood function involves both the unadjusted residuals ε_t^* that define h_t^* for $t = s+1$, and the adjusted residuals ε_t, for $t = s$. Implementing the AVO adjustment therefore requires an extension to existing GARCH code. We summarise the adjustments needed to compute the modified log-likelihoods for the different outlier models in Table 1. We call these concentrated likelihoods, although the parameters of the model other than τ, only satisfy the first order conditions for MLE at one value of γ.

Table 1: Adjustments for concentrated likelihood computation

	in volatility	in residuals	notation
ALO	$\varepsilon_t^* - \gamma d_t$	$\varepsilon_t^* - \gamma d_t$	$\ell_{\text{alo}}(\cdot\|\gamma)$
AVO	ε_t^*	$\varepsilon_t^* - \gamma d_t$	$\ell_{\text{avo}}(\cdot\|\gamma)$

$\varepsilon_t^* = y_t - x_t'\zeta$, adjustments applied to log likelihood (2).

This adjustment avoids adding parameters of dummy variables to the log-likelihood which are difficult or impossible to estimate unrestrictedly.

In the first two motivating illustrations of our procedure, we use subsets of the weekly and monthly Dow Jones returns as discussed in more detail in §8. For the weekly data we use 574 observations

covering the years 1982 to 1992. The monthly data has 420 observations for the years 1965 to 1999. In both cases, a standard GARCH(1,1) with an intercept for the mean is estimated. Also in both cases, the largest outlier is found for the first observation after the Black Monday crash of 19 October 1987. This is the outlier candidate. Using the corresponding dummy variable d_t in the mean, and d_{t-1} in the variance, the GAO model (11) is estimated next.

Table 2: Likelihood Ratio Testing for a generalized additive outlier (GAO) in Dow Jones returns

	log-likelihood	$\widehat{\gamma}$	$\widehat{\tau}$	$\widehat{\varepsilon}_s$
	Monthly data 1982–1992			
Baseline model (1)	-333.73	—	—	-4.38
GAO model (11)	-302.91	-4.39	0.08	0
Test statistic and p-value	$61.7\ [10^{-10}]$			
	Weekly data 1965–1999			
Baseline model (1)	-861.89	—	—	-9.01
GAO model (11)	-843.32	-8.98	6.09	0
Test statistic and p-value	$37.2\ [10^{-5}]$			

The s subscript refers to October 1987.

Table 2 gives the maximised log-likelihoods of the baseline GARCH(1,1) model and the GAO model. The p-values of the likelihood ratio tests treat the timing of the outlier as unknown, i.e. s is considered as estimated from the data. They are based on an extreme value approximation discussed in §6 and Appendix A. In both data sets, the outlier candidate is highly significant.

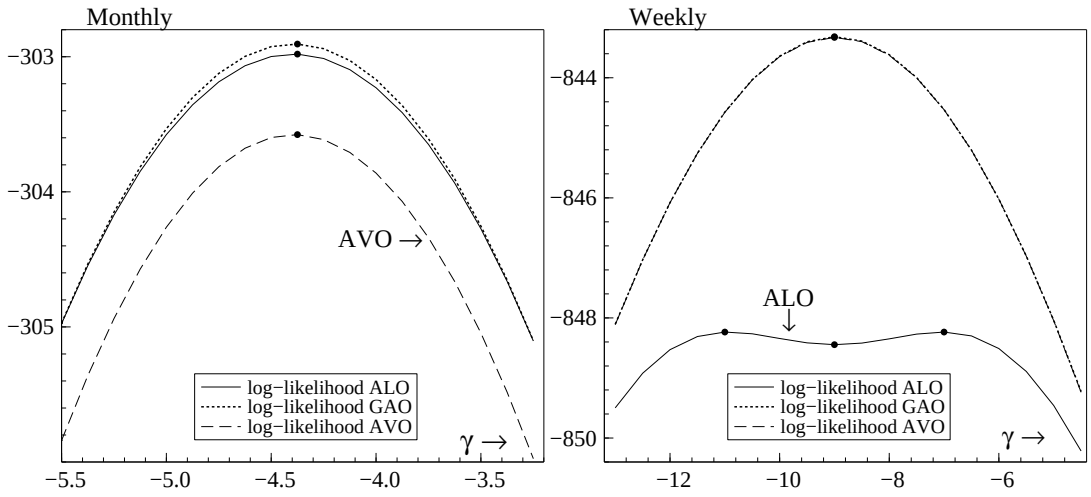

Figure 1: Likelihood grids of GAO, ALO, AVO models for GARCH(1,1) model of montly and weekly Dow Jones returns, as a function of γ, the size of the October 1987 crash, see Table (1) for ALO and AVO concentrated likelihood computation. Note: GAO and AVO indistinguishable for weekly data.

Given the presence of an outlier at $t = s$, we examine the outlier type. The decision on the type of additive outlier is based on $\ell_{\text{alo}}(\cdot|\widehat{\gamma}_{\text{gao}})$ and $\ell_{\text{avo}}(\cdot|\widehat{\gamma}_{\text{gao}})$, as discussed in Table 1. Figure 1 shows the concentrated likelihood grids as a function of the outlier size γ. The GAO model nests ALO and AVO, so always has a higher likelihood. The GAO grid can be computed from either $\ell_{\text{alo}}(\cdot|\gamma)$ or $\ell_{\text{avo}}(\cdot|\gamma)$ by adding the lagged dummy variable to the conditional variance equation and estimating τ, conditional on γ. For the monthly data, ALO is very close to GAO: there is no significant difference using a $\chi^2(1)$ test. The likelihood of ALO is higher than AVO, and the former model is preferred. For the weekly data, Figure 1 shows the ALO likelihood to be bimodal, unlike the monthly case. Here, AVO and GAO are indistinguishable, so that the AVO model is preferred.

The procedure to decide between ALO and AVO is based on the likelihoods for $\widehat{\gamma}_{\text{gao}}$ and therefore ignores the two global modes of the ALO model in case of bimodality. In practice it is possible that both ALO and AVO are significantly worse than the GAO model, although we have only encountered this very rarely. One approach to such a finding would be to adjust the GARCH likelihood in a similar manner as for ALO and AVO, so that a GAO correction can be imposed.

We conclude this section with a note on initialisation of the GARCH likelihood. In our computations for the illustration of Figure 1 we conditioned on the first observation to initialize the GARCH recursion, so that the effective sample size is 419 and 573 respectively. Then, the value of γ does not influence the likelihood of the observations prior to $t = s$. In the remainder of the paper, we use the sample mean of ε_t^2 for initialization of h_t, following the suggestion in Bollerslev (1986), which is more commonly followed in practice.

5 Detecting multiple outliers

The simplifying data adjustments and simplified likelihood-based tests are even more important when one suspects that more than one additive outlier of unknown type may be present: in that case a recursive detection procedure is required.

Based on the GAO model (11), we propose the following five step procedure to detect additive outliers in a GARCH(1,1) model:

Step 1 Estimate the baseline GARCH model (1), i.e. without any dummy variables, to obtain the log-likelihood $\widehat{\ell}_b$ and residuals ε_t^* and volatilities h_t^*.

Step 2 Find the largest standardized residual in absolute value, $\max_t |\varepsilon_t^*/h_t^*|$. Denote this observation by $t = s$. Estimate the GARCH GAO model (11) with dummy $d_t \equiv I(t = s)$ in the mean equation, and d_{t-1} in the variance equation (this can be done in most standard software packages with GARCH estimation). This gives estimates for the added parameters $\widehat{\gamma}_{\text{gao},s}$ and $\widehat{\tau}_{\text{gao},s}$ respectively, with log-likelihood $\widehat{\ell}_{\text{gao},s}$.

Step 3 If $2(\widehat{\ell}_{\text{gao},s} - \widehat{\ell}_b) < C_T^\alpha$ then terminate: no new outlier is detected. Our approximation of the asymptotic distribution of this test under the null-hypothes of no outliers suggests that $C_T \approx 5.66 + 1.88 \log T$ at a significance α of 5%. The full approximation is given in §6.

Step 4 This step implements the AVO versus ALO selection, given that an outlier was detected:

(a) If $\widehat{\tau}_{\text{gao},s} < 0$ then the outlier is of type ALO; else continue with step 4(b):

(b) Estimate the GARCH model with an ALO outlier correction of fixed size $\widehat{\gamma}_{\text{gao},s}$. The model to be estimated corresponds to $\ell_{\text{alo}}(\cdot|\widehat{\gamma}_{\text{gao},s})$ from Table 1: it is a standard GARCH model without additional dummy parameters, but with a dependent variable that is corrected for the outlier, see §4. This model gives $\widehat{\ell}_{\text{alo},s}$.

(c) Estimate the GARCH model with an AVO outlier correction of fixed size $\widehat{\gamma}_{\text{gao},s}$. The model is $\ell_{\text{avo}}(\cdot|\widehat{\gamma}_{\text{gao},s})$ from Table 1, see §4 for its implementation. This gives $\widehat{\ell}_{\text{avo},s}$.

(d) If $\widehat{\ell}_{\text{avo},s} > \widehat{\ell}_{\text{alo},s}$ the outlier is AVO, else it is ALO.

The procedure can be iterated until no further outlier is detected. Because the outlier coefficients have already been estimated at each step, we propose to use the simple data correction of Table 1 when an outlier is detected. This data adjustment procedure avoids a proliferation of parameters in the log-likelihood.

Step 4 is used to distinguish between the two types of outliers, in case one is detected. It involves two additional GARCH model estimations, which can be initialised using estimates for α_0, α_1 and β_1 from step 1, i.e. the baseline model (1) 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.

Step 4(a) uses the fact that, because $\widehat{\varepsilon}_s = 0$ for AVO: $\tau = \alpha_1\gamma^2$. Imposing $\alpha_1 > 0$ shows that a negative $\widehat{\tau}$ is incompatible with the AVO model, saving the effort of estimating the model.

6 Controlling the size of the outlier detection procedure

We use extreme value theory, see e.g. Leadbetter, Lindgren, and Rootzén (1983), and Monte Carlo simulation to determine an appropriate null distribution for the test in Step 3 of our outlier detection procedure of §5. Assuming that the single outlier test statistics are independent for all $s, s = 1, \ldots, T$, and also that the dummy variable leading to the largest statistic is selected in Step 2, one can treat the test statistic in Step 3,

$$M_T = \max_{s\in(1,\ldots,T)} LR_T^{GAO}(s) = \max_{s\in(1,\ldots,T)} 2(\widehat{\ell}_{\text{gao},s} - \widehat{\ell}_b)$$

as the maximum of a random sample of size T from a $\chi^2(2)$ distribution. Monte Carlo results in Appendix A show that the asymptotic $\chi^2(2)$ approximation for a test for a single generalized additive outlier at a known fixed time s, denoted by $LR_T^{GAO}(s)$, works well for T = 500.

Extreme value theory describes conditions under which M_T follows an extreme value limiting distribution. These conditions do not require independence of the underlying random variables $LR_T^{GAO}(s)$. In our case the limiting distribution is extreme value type I and the mean of M_T is a linear function of $\log T$ as T gets large. We use a response surface analysis of Monte Carlo experiments that leads to a good approximation of the finite sample distribution of M_T. The computation

of p-values and critical values only requires the knowledge of sample size T and does not require further simulation. Here we present the formula and the main simulation results; more detail is in Appendix A.

Steps 1–3 are simulated as described in §5 under the null hypothesis of no outlier. The Monte Carlo uses $N = 10\,000$ replications of the baseline model (1), a constant in the mean, $x_t = 1$, $\zeta = 1$ and $\alpha_1 = 0.1, \beta_1 = 0.8, \alpha_0 = 1 - \alpha_1 - \beta_1$. The sample sizes are $T = 200(100)1200, 1500, 2000, 2500$. The restrictions $0 < \widehat{\alpha}_1 + \widehat{\beta}_1 \leq 1$ and $\widehat{\alpha}_0 > 0$ are always imposed in the estimation procedure. The results, shown in the first panel of Figure 2, indicate that the mean of the test statistic increases with the sample size, in proportion to $\log T$ as $T \to \infty$, as predicted by extreme value theory. The variance, skewness and kurtosis are not very sensitive to the sample size, see the second panel of Figure 2. The last panel shows that the critical values are approximately equidistant, i.e. the critical value function, $CV(\alpha, T)$, is additively separable in two simple functions of α and T and the distances between critical values of 20% and 10% on the one hand and between 10% and 5% on the other hand, are approximately equal. This is a characteristic of a Type I extreme value distribution.

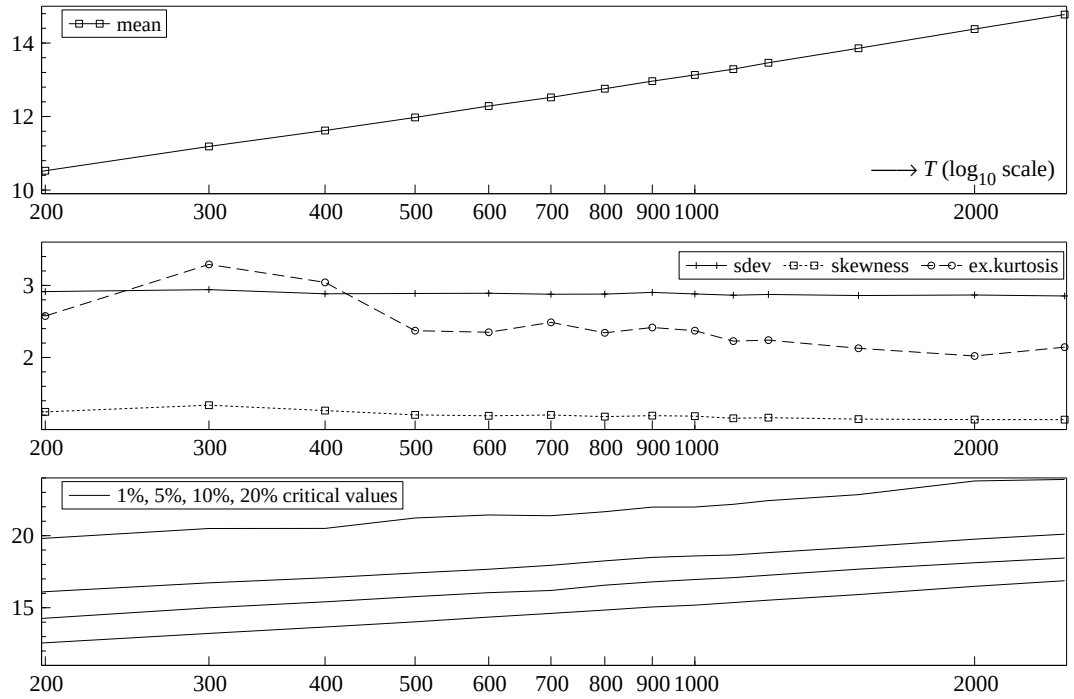

Figure 2: Simulated moments (mean, standard deviation, skewness and excess kurtosis), and critical values (1%, 5%, 10%, 20%) of the $\max_s LR_T^{GAO}(s)$ statistic under the null hypothesis.

Combining the extreme value Type I limiting distribution and the response surface analysis of Monte Carlo experiments in Appendix A, we form the following approximation for the distribution of the statistic M_T, which we denote by $\max_s LR_T^{GAO}(s)$:

$$P(\max_s LR_T^{GAO}(s) \leq x) = \exp\left\{-\exp\left[-\frac{x + 1.283 - 1.88\log T\,(1 + 12/T)}{2.223}\right]\right\}. \qquad (12)$$

To check the accuracy of this approximation, we simulate the rejection frequencies for various parameter values under the null hypothesis. Table 3 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 and β_1.

Table 3: Size of (Max-)LR-test for single generalized additive outlier at unknown time

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
		ASE	*0.006*	*0.005*	*0.003*	*0.002*

Based on $N = 4\,000$ replications.
ASE: Monte Carlo standard error of the rejection frequencies.

7 Power of the outlier detection procedure

Next, we investigate the performance of our procedure in detecting additive outliers, in selecting the type of additive outlier, and in determining the timing of the additive outlier. To investigate the power of the proposed test procedure by Monte Carlo, we select $T = 250$, and have the DGP of type AVO as in (6) 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 near the middle of the sample: $s = T/2$. The results are presented in Table 4.

The first column in Table 4 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.[2] Our procedure is also successful in detecting the type of outlier: there is no particular

[1] So we do not force γ to enter the DGP with the same sign as the drawn residual.
[2] The percentages for correct date and type are conditional on detection of an outlier.

Table 4: Size and power of outlier detection test for a generalized additive outlier in a GARCH(1,1) model

	Rejection frequencies				*Correct date*		*Correct type*	
α_1, β_1	$\gamma = 0$	-3	-4	-5	-4	-5	-4	-5
Outlier of type AVO at $T/2$								
0.1,0.8	0.054	0.23	0.53	0.84	96%	99%	77%	81%
0.3,0.5	0.050	0.20	0.53	0.83	96%	99%	76%	81%
0.5,0.3	0.048	0.20	0.52	0.83	96%	99%	76%	80%
Outlier of type ALO at $T/2$								
0.1,0.8	0.054	0.28	0.60	0.84	97%	99%	73%	75%
0.3,0.5	0.050	0.40	0.71	0.87	98%	99%	82%	84%
0.5,0.3	0.048	0.55	0.79	0.89	98%	99%	84%	85%

Based on 5% nominal rejection frequencies for $N = 4\,000$ 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.

Table 5: Samples for Dow Jones returns

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

bias towards AVO or ALO, when an outlier is detected.

While the AVO results seem independent of the GARCH parameters, the power of ALO appears to increase as α_1 increases. The likely explanation is that this corresponds to a larger volatility effect when left unmodelled, see equation (8) above. Overall, the proposed outlier detection procedure works very well, even at this small sample size where GARCH models can be somewhat harder to estimate. Essentially the same results were obtained for a sample size of 500.

8 Multiple outlier applications for Dow Jones returns

As an application of the new outlier-detection procedure we consider the returns on the Dow Jones Industrial Average index,[3] using monthly, weekly, and daily data for the period 1896, May 26, to 2001, December 5. Table 5 provides some details.

[3]The Dow Jones index data are available from www.djindexes.com.

The return data are formed by taking the first difference of the logarithms and then annualized. These returns were multiplied by the scale factor given in Table 5, selected as the integer which made the annualized average return for the daily and weekly returns as close as possible to the average for 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 months 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 last sharp fall followed the 11 September 2001 terrorist attacks on Washington and New York, which is indicated as an outlier in the daily data.

Table 6: Detected outliers in GARCH(1,1) model for monthly and weekly Dow Jones returns 1896-2001.

date	size	p-outlier	p-AVO	p-ALO	type
monthly returns: $12\Delta \log y_t^m$					
1987/10	-4.39	0.0_83	$\widehat{\tau} < 0$	0.795	ALO
1914/12	-3.58	0.00012	0.478	0.112	AVO
1940/05	-3.11	0.00022	$\widehat{\tau} < 0$	0.241	ALO
1937/09	-2.37	0.028	0.122	0.001	AVO
2001/09		0.129			—
weekly returns: $51\Delta \log y_t^w$					
1914/12/16	-16.75	0	0.0_42	0.244	ALO
1940/05/15	-7.05	0	1	0	AVO
1899/12/13	-7.14	0.0_83	0.206	0.026	AVO
1987/10/21	-8.95	0.0_53	0.287	0.010	AVO
1926/03/03	-4.84	0.00015	1	0.002	AVO
1898/05/11	7.61	0.00020	0.030	0.960	ALO
1994/03/30	-3.39	0.00075	0.120	0.536	ALO
1998/09/02		0.070			—

p-outlier is for testing no outlier against a GAO at an unknown date.
p-ALO is for testing ALO against GAO, conditional on a known outlier date.
p-AVO is for testing AVO against GAO, conditional on a known outlier date.
Notation: $0.0_45 = 0.00005$

We apply our outlier detection procedure recursively: first detect the largest outlier, then adjust for this as discussed in §4, next, detect and adjust for the subsequent outlier, until no more are found. This

approach is along the lines of Chen and Liu (1993), and therefore susceptible to the same criticism that estimates of the other model parameters, in particular α_0, are contaminated by the presence of an outlier. Robust estimation of GARCH models is possible, but rather costly and difficult to implement, see Sakata and White (1998) and therefore not yet attractive. The problem can be mitigated by applying a Student t-error distribution, see §9.2 below.

The top half of Table 6 lists the results when applying the procedure to the monthly data. The column labelled p-outlier gives the p-value of the test for a generalized additive outlier, based on the extreme value approximation. Detected are the 1987 crash, the start of the two world wars in Western Europe, as well as September 1937 (when the index dropped by 17%). The order in the table follows the order in which the outliers were detected, and we also include the first outlier with a p-value $> 5\%$. The column labelled p-AVO reports the p-value for the $\chi^2(1)$ likelihood-ratio test of the AVO restriction within the GAO model. Similarly, the next column has the test outcome for the ALO restriction. Note that AVO is rejected without further testing, when $\widehat{\tau} < 0$, according to Step 4a of the procedure. In December 1914, when the stock market reopened, both ALO and AVO are not significantly different from GAO. However, the likelihood of AVO is higher than ALO, so the former is selected.

The second part of Table 6 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 World War II is detected as an AVO. Also, the 13% fall in the second week of December 1899 is detected before the 1987 crash. Except for the final outlier in 1994, the ALO versus AVO decision is clear-cut.

In Table 7 we list the dates of outliers for the daily model, but this time in chronological order. There are more than five times as many observations as in the monthly data set, but also five times as many outliers. The procedure is found to be acceptably fast on the daily data, taking less than half an hour for nearly $30\,000$ observations (on a 800 Mhz Pentium III notebook; this includes the first estimation).

The results in this section assume that the underlying model is Gaussian GARCH(1,1), possibly 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 the GARCH model with student-t distributed errors, which is readily available in standard software, is a better description. This is explored in the next section.

9 Extensions to other models

9.1 GARCH(2,2) models

In the GARCH(p, q) case, the lag polynomial $\alpha(L)$ in (10) has q terms instead of one. The equivalent extension to the equation for h_t in the GAO model (11) would be to add the dummy variable with lags

Table 7: 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_43	1924/02/15	AVO	0.0037	1950/06/26	AVO	0.0_66
1901/05/08	AVO	0.0008	1925/11/10	AVO	0.0015	1955/09/26	AVO	0
1901/09/07	AVO	0.0208	1927/10/08	AVO	0.0325	1962/05/28	AVO	0.0039
1904/12/07	AVO	0.0_45	1929/10/28	AVO	0.0002	1982/08/17	AVO	0.0055
1907/03/14	AVO	0.0005	1933/03/15	ALO	0.0_48	1986/09/11	ALO	0.0033
1913/01/20	ALO	0.0_66	1934/07/26	ALO	0.0067	1987/10/19	AVO	0
1914/07/28	ALO	0.0_45	1939/09/05	ALO	0.0031	1989/10/13	AVO	0
1914/07/30	AVO	0.0_43	1940/05/13	AVO	0.0_63	1991/01/17	ALO	0.0158
1914/12/12	ALO	0	1943/04/09	ALO	0.0004	1991/11/15	AVO	0.0_63
1916/12/12	AVO	0.0012	1946/09/03	AVO	0.0034	1997/10/27	AVO	0.0_42
1917/02/01	ALO	0.0_61	1948/11/03	AVO	0.0_48	2000/04/14	ALO	0.0156
						2001/09/17	AVO	0.0002

1 to q as the variance equation is affected by a level outlier for q periods. As a simple alternative we do not extend the GAO model with extra lags of the dummy variable. Instead, we just apply the same procedure as for GARCH$(1,1)$, introducing only one dummy variable in the variance equation and leaving the approximation to the distribution of the test statistic in Step 3 unchanged. We evaluate our test procedure by Monte Carlo for 16 different GARCH(2,2) data generating processes both with and without additive outliers. The results in Table 8 show that the size and power are very close to that in the GARCH$(1,1)$ case. However, the procedure detects more additive level outliers in Step 4, which could be caused by the omission of the additional lagged dummies, together and the rule 4a that $\widehat{\tau} < 0$ corresponds to an ALO.

9.2 GARCH-t models and effects of outlier correction on GARCH parameter estimates

A GARCH model with Student-t distributed errors, as proposed by Bollerslev (1987), is a likely alternative for a GARCH model with additive volatility outliers. Appendix A discusses the adjustments that we made to the extreme value approximation when incorporating the standardized $t(\nu)$ distribution. As the form of the limiting extreme value distribution is nonstandard in this case and depends on the unknown ν, our approximation does not work as well as in the Gaussian model. Table 9 presents some results for the test. As expected, the actual outliers have to be considerably larger to be distinguished from the thick tail of the Student-$t(6)$ distribution.

Table 8: Size and power of the test for a generalized additive outlier at unknown time in a GARCH(2,2) model

	Rejection frequencies				*Correct date*		*Correct type*	
$\alpha_1, \alpha_2; \beta_1, \beta_2$	$\gamma = 0$	-3	-4	-5	-4	-5	-4	-5
				Outlier of type AVO at $T/2$				
$0.1, 0.1; 0.1, 0.6$	0.076	0.21	0.48	0.77	93%	98%	74%	77%
$0.1, 0.1; -0.1, 0.8$	0.061	0.16	0.41	0.74	94%	99%	60%	64%
				Outlier of type ALO at $T/2$				
$0.1, 0.1; 0.1, 0.6$	0.076	0.34	0.63	0.84	96%	98%	82%	82%
$0.1, 0.1; -0.1, 0.8$	0.061	0.32	0.63	0.85	97%	99%	85%	86%

5% nominal rejection frequencies for $N = 2\,000$, $T = 500$.

Correct date, type: % correct when an outlier was detected.

Table 9: Size and power of the test for a single generalized additive outlier at unknown time in a GARCH(1,1)-$t(6)$ model, $\alpha_1 = 0.1, \beta_1 = 0.8$

Rejection frequencies					*Correct date*			*Correct type*		
$\gamma = 0$	-5	-8	-10	-15	-8	-10	-15	-8	-10	-15
					Outlier of type AVO at $T/2$					
0.043	0.04	0.08	0.22	0.74	92%	98%	99%	77%	84%	96%
					Outlier of type ALO at $T/2$					
0.043	0.05	0.26	0.48	0.84	91%	96%	99%	90%	95%	96%

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

Correct date, type: % correct when an outlier was detected.

Empirical application to the Dow Jones industrial averages index supports the closeness of the Gaussian GARCH(1,1) model with generalized additive outliers and the GARCH(1,1)-t model. Table 10 shows that at the monthly and weekly level, the two models seem to be close substitutes, with the outlier model weakly preferred on AIC, where we treat the date and type of the outlier as known and count the sizes of the outliers as extra parameters to be estimated. At the daily level, the GARCH-t is preferred, yielding a higher log-likelihood and lower AIC than the model with outliers. Both the outlier extension and the introduction of t distributed errors significantly affect the estimates for the GARCH parameters in the weekly and daily data: $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ increase, $\widehat{\beta}_1$ decreases. The 'robust' estimation of the mean return using the Student-errors significantly increases $\widehat{\zeta}$ (the intercept) as the predominantly negative returns receive a lower weight. A similar effect was observed by Sakata and White (1998) for daily S&P 500 returns (1987/8-1991/8) when they applied robust high breakdown

estimators for the GARCH(1,1) model. The outlier correction does not have a significant impact on the estimated mean return as the percentage of outliers is very small.

Table 10: Estimated GARCH(1,1) coefficients for Dow Jones returns at various frequencies

	GARCH$(1,1)$	with outliers	GARCH$(1,1)$-$t(\nu)$	with outliers
		Monthly data: $12\Delta \log y_t^m$		
ζ	0.068 (0.015)	0.078 (0.015)	0.095 (0.015)	
α_0	0.014 (0.0040)	0.013 (0.0038)	0.017 (0.0056)	
α_1	0.114 (0.019)	0.095 (0.017)	0.102 (0.022)	
β_1	0.862 (0.021)	0.870 (0.021)	0.861 (0.027)	
α_0^*	0.582	0.377	0.459	
ν			5.357	
outliers	0	4	0	
log-lik	-1189.0	-1121.4	-1133.8	
AIC	1.889	1.782	1.803	
		Weekly data: $51\Delta \log y_t^w$		
ζ	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 (0.0078)	0.091 (0.011)	0.091 (0.010)
β_1	0.820 (0.013)	0.888 (0.0084)	0.892 (0.012)	0.894 (0.011)
α_0^*	2.036	1.437	1.644	1.645
ν			7.151	7.808
outliers	0	7	0	2
log-lik	-8372.4	-8120.2	-8162.0	-8128.3
AIC	3.090	3.000	3.013	3.000
		Daily data: $276\Delta \log y_t^d$		
ζ	0.120 (0.012)	0.124 (0.012)	0.145 (0.011)	
α_0	0.105 (0.0070)	0.069 (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	6.727	10.09	
ν			5.670	
outliers	0	34	0	
log-lik	-67539.8	-66715.8	-66476.7	
AIC	4.616	4.559	4.543	

$\alpha_0^* = \alpha_0/(1-\alpha_1-\beta_1)$. Standard errors in parentheses.

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 World War I, and AVO at the start of World War II. These are the same two leading outliers found in the normal GARCH(1,1) model. However, in terms of AIC the GARCH-t model with outliers is not an improvement over the normal GARCH(1,1) model with outliers. The effect of the outliers detection on the estimated ν is small. For the monthly data, the closest candidate outlier in the GARCH-t model was October 1987, with a p-value of 0.052. In daily data, the closest candidate was September 26, 1955, which was also the first one found in the GARCH(1,1) model, but now with p-value of 0.10 rather than zero.

10 Conclusion

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

- It is simple to implement and contains a convenient procedure to compute p-values for tests, without the need for simulation.
- It is likelihood-based and associated tests are asymptotically similar with respect to the GARCH parameters α_1 and β_1.
- Simple nested tests distinguish between Additive Level Outliers and Additive Volatility Outliers.
- The procedure can be extended to other types of GARCH models such as EGARCH, etc.

Our applications on monthly, weekly and daily Dow Jones returns show that the test procedure also works well in practice. We compare estimates of our outlier model with a GARCH-t model, also possibly affected by outliers. The GARCH-t model without outliers is to be preferred over the normal GARCH with outliers for the daily Dow Jones returns.

Other practical aspects of the procedure could be examined. Although the in-sample fit of a GARCH-t and normal-GARCH with outliers for the monthly Dow Jones returns 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. Conclusions regarding leverage effects in the form of asymmetric volatility could be also different: the outlier detection, for the data considered, predominantly removes negative shocks.

The proposed method could become a useful addition to the toolkit of empirical volatility modellers. The first-step outlier test can serve as a mis-specification test for the model. Next, the iterated procedure can be used as a robustification of the model (with too many outliers suggesting that the model is inadequate). Finally, the detected outliers can complement value-at-risk estimations: in large samples, the distribution of outliers is informative in itself, otherwise the estimates may require their absence.

Acknowledgements

We wish to thank Siem Jan Koopman for helpful discussions and suggestions. We'd also like to thank Dayong Zhang for reporting some errors in earlier versions of the paper. Financial support from the UK ESRC (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 the first version. The computations were performed using the Ox programming language (Doornik, 2001).

References

Abraham, B. and N. Yatawara (1988). A score test for detection of time series outliers. *Journal of Time Series Analysis 9*(2), 109–119.

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.

Franses, P. H. and D. van Dijk (2000). Outlier detection in GARCH models. Econometric Institute Report EI-9926/RV, 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.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer-Verlag, New York, Heidelberg, Berlin.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics, Third Edition*. McGraw-Hill.

Sakata, S. and H. White (1998). High breakdown point conditional dispersion estimation with application to s&p 500 daily returns volatility. *Econometrica 66*, 529–567.

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.

Tsay, R. S. (2002). *Analysis of Financial Time Series*. New York: John Wiley & Sons.

A Approximating the distribution of the $\max_s LR_T^{GAO}(s)$ test

This appendix describes the details of the experiments leading to the approximation for the $\max_s LR_T^{GAO}(s)$ test in the normal case described in §6. We also present adjustments to the approximation for the case of Student-t errors, that we discuss in §9. In order to simplify the presentation we denote $LR^{GAO}(s)$ by X_s and $\max_s LR_T^{GAO}(s)$ by M_T.

The single likelihood-ratio test for a Generalized Additive Outlier at a known time $t = s$, denoted by $LR^{GAO}(s)$, involves two parameters that are well identified under the null, giving the test statistic an asymptotic $\chi^2(2) \equiv \exp(1/2)$ distribution. The effectiveness of this asymptotic approximation for a sample size of 500 is illustrated below.

As we effectively do T such tests we wish to approximate the distribution of the maximum: $M_T = \max(X_1, \ldots, X_T)$. Assuming independently and identically distributed X_s the cumulative distribution function F_{M_T} of M_T is given by

$$F_{M_T}(x) = \{F_X(x)\}^T = \left\{1 - e^{-\frac{1}{2}x}\right\}^T.$$

Using

$$\frac{1}{T}\log F_{M_T}(x) = \log\left(1 - e^{-\frac{1}{2}x}\right) \approx -e^{-\frac{1}{2}x},$$

when x is large, gives

$$F_{M_T}(x) \approx \exp\left\{-Te^{-\frac{1}{2}x}\right\} = \exp\left\{-\exp\left(-\frac{x - 2\log T}{2}\right)\right\},$$

such that for large x and large T, M_T has a Type I extreme value limiting distribution. Our approximations are based on this distribution type. Leadbetter, Lindgren, and Rootzén (1983, Chapters 1,3) show that Type I extreme value (or Gumbel-) limiting distributions apply much more generally. The X_s need not be exponential and independent, although these are the cases where the asymptotic theory works well, also in moderately sized samples.

In general, when

$$F_{M_T}(x) = \exp\left\{-\exp\left(-\frac{x - a_T}{b}\right)\right\}, \tag{13}$$

the expectation and variance of M_T are given by $\mathsf{E}[M_T] \equiv m_T = a_T + \delta b$, where $\delta \approx 0.577216$, and $\mathsf{V}[M_T] = b^2\pi^2/6$, see e.g. Mood, Graybill, and Boes (1974, Appendix B). Critical values at significance level α can therefore be computed as

$$C_T^\alpha = -b\log(-\log(1-\alpha)) + a_T. \tag{14}$$

Although the X_s are not independently distributed in our case, we can use the extreme value distribution (13) as the limiting distribution. The X_s are not fat tailed and they are short memory under the null hypothesis of no outliers, so the required distributional mixing conditions for a Type I extreme value distribution are met, see Leadbetter, Lindgren, and Rootzén (1983, Ch. 3). The general theory allows the variance of the approximating distribution, and therefore b, to depend on T. This does not apply to our statistic.

Simulating the distribution of M_T for increasing sample sizes T, as reported in §6, we observe that the simulated standard deviation, $V[W_T]^{1/2}$, of the test statistic is close to constant. Its asymptotic value is 2.851 with standard error 0.008, found from a regression on a constant, T^{-1} and T^{-2}. This results in $\widehat{b} = 2.223$.

After some experimentation, we found that the means from the Monte Carlo experiment in §6, depicted as 14 observations in the upper panel of Fig. 2, are very well described by the following regression:

$$\widehat{m}_{T_i} = \underset{(0.0013)}{1.880}\ \log(T_i) + \underset{(1.2)}{22.7}\ \log(T_i)/T_i, \qquad i = 1, \ldots, 14,$$

where heteroscedasticity consistent standard errors (HCSE) are given in parentheses. The residual normality test of this regression insignificant, but there is significant heteroscedasticity. The intercept is insignificant. The resulting response surface for m_T, b as a function of sample size T is:

$$\begin{array}{lcl} a_T & = & m_T - \delta b \approx 1.88\log T(1 + 12/T) - 1.283, \\ b & \approx & 2.223. \end{array} \tag{15}$$

Figure 3 shows how well the approximation (14), (15) works when applied to a selection of critical values.

Because the number of Monte Carlo experiments is quite big, we have a large number of draws from the (hypothesized) extreme value distribution. This in turns leads to accurate estimates of the mean and standard deviation, for which we adopted the method of moments to determine the parameters of the limiting distribution. We could instead have used the disaggregated data along the lines of, e.g., Tsay, 2002, §7.5.2.1, to directly estimate a_T and b_T. This would give very similar outcomes for the resulting critical values, but with better estimates of the overall parameter uncertainty in the approximation at varying levels of T.

Abraham and Yatawara (1988) (AY88) use a similar extreme value approximation for the maximum of a sequence of $\chi^2(2)$ distributed LM tests for time series model outliers. They do not fit equations for the moments of the extreme value distribution, but instead adjust the critical value approximation (14), with fixed $b = 2$ and derive a constant term in the critical value equation, $\log(\theta)$, using Monte Carlo Simulations.

$$(AY88) : C_T^{\alpha} = -2\log(-\log(1-\alpha)) + 2\log(T) + \log(\theta) \tag{16}$$

with T the number of (dependent) outlier tests. They estimated an extremal index $\theta = 0.8$. The term $\log(\theta)$ corrects the critical values for the dependence of the test statistics, see Leadbetter, Lindgren, and Rootzén (1983, p.67) for a formal definition. $\theta = 1$ for (asymptotically) independent statistics. Abraham and Yatawara (1988) also note that applying the test with estimated parameters for the time series model, rather than using known parameters markedly decreases the empirical critical values for the test. A similar effect may explain that the coefficient of $\log(T)$ is lower than two in our approximation formula (15). The specification (16) would work badly in our case.

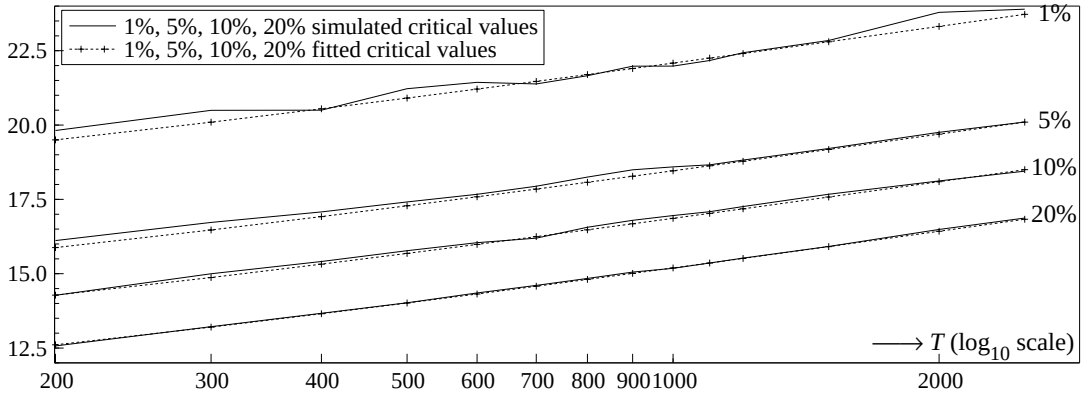

Figure 3: Simulated and fitted critical values (1%, 5%, 10%, 20%) of the $\max_s LR_T^{GAO}(s)$ test statistic under the null hypothesis.

Next, we turn to the case with a Student-t error term. We first note that M_T of a sample of $t(\nu)$-distributed variables has a type II extreme value (Fréchet) limiting distribution:

$$F_{M_T}(x) = \exp\left\{-x^{-\nu}\right\}, \tag{17}$$

where ν, the tail index, determines the shape of the distribution. The k-th moments of M_T are now given by $\mathsf{E}[M_T^k] = \Gamma(1 - k/\theta)$, where Γ is the gamma function. In this case ν equals the degrees of freedom of the Student distribution, see Mood, Graybill, and Boes (1974, §6.5.3, example 12).

As the LR^{GAO} test involves the test for a single outlier in a GARCH-t model, one may perhaps expect that the type II behaviour also applies here. It may also be that the type I approximation is still reasonable for common values of ν .

In order to investigate this issue we compare the distributions of the $LR^{GAO}(s)$ test for a fixed s in the cases of normal errors and Student-t errors using simulation. Figure 4 presents QQ plots of the simulation results for the design given in §6 for $T = 500$ and $N = 10000$, testing for an outlier at the middle of the sample: $s = T/2$. The solid line in Figure 4 makes clear that the distribution for the LR^{GAO} test for an outlier at a known point in a GARCH model with normal errors is indeed close to $\chi^2(2)$.

Note that for Student-$t(6)$ errors, the distribution of LR^{GAO} is considerably more spread towards the right tail. At first sight, this may indicate that a type I extreme value distribution does not apply here. However, if we simulate the critical values of the $\max_s LR_T^{GAO}(s)$ test under Student-t errors, the distribution of the

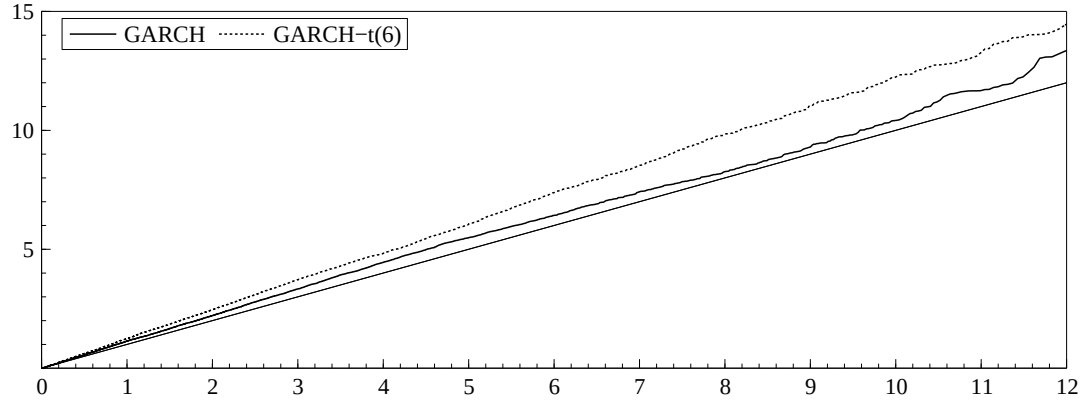

Figure 4: QQ plot against a $\chi^2(2)$ reference distribution of the LR(2) test for an outlier in the middle of the sample: normal GARCH(1,1) (solid line) versus GARCH-t(1,1) with $t(6)$ errors. $T = 500$, $N = 10\,000$.

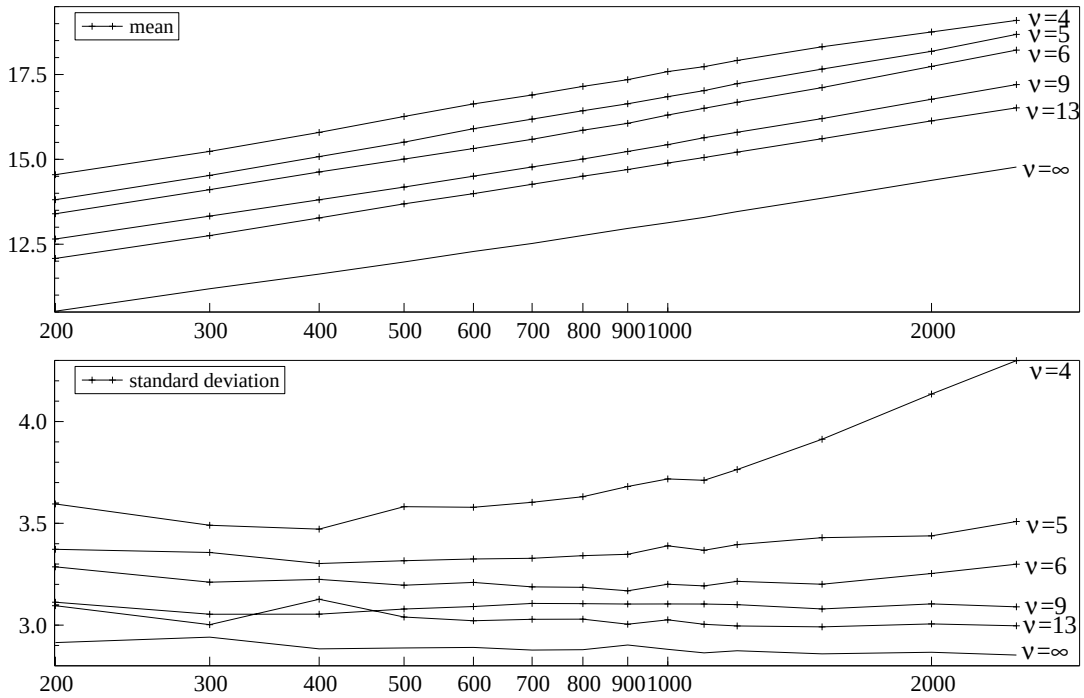

Figure 5: Simulated moments (mean, standard deviation), of the $\max_s LR_T^{GAO}(s)$ statistic in a GARCH-$t(\nu)$(1,1) under the null hypothesis, for $\nu = 4, 5, 6, 9, 13$ and ∞ (normal).

test shifts with ν, but the distance between critical values at 5% and 10% and between 10% and 20% for a specific ν remain very close, as in Figure 2. Figure 6 presents simulated critical values for Student-t errors. The differences in critical values for a type II extreme value distribution are determined by $[-\log(1-\alpha)]^{-1/\nu}$ which should not lead to an equal spacing between 5% and 10 % and 10% and 20% critical values. This is an indication that that the type II approximation would not work well here.

Instead of using a type II approximation, we adapt the type I extreme value approximation under normal errors to the Student-$t(\nu)$ case by allowing m_T and b to depend on ν. Based on GARCH(1,1)-$t(\nu)$ Monte Carlo 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:

$$\begin{array}{lcl} m(T,\nu) & \approx & m_T + 11\nu^{-1} + 0.25 m_T \nu^{-1/2}, \\ b(\nu) & \approx & b + 12\nu^{-2}, \end{array} \tag{18}$$

with m_T and b given in (15). We did not allow b to depend on T, although the simulations show the variance to be somewhat u-shaped for ν ranging from 4 to 6. The response surface for the mean fits remarkably well. The approximation to the critical values is satisfactory, see Figure 6, except when $\nu = 4$, and to a lesser extent for $\nu = 5$ at 1%.

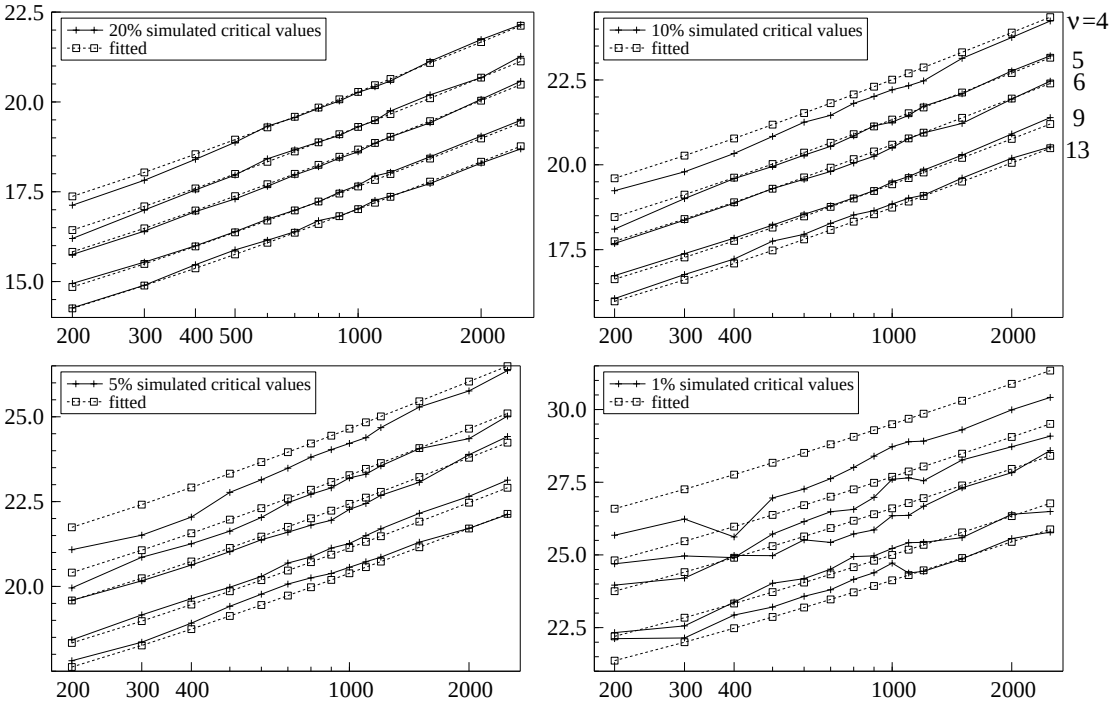

Figure 6: Simulated and fitted critical values (1%, 5%, 10%, 20%) of the test statistic for GAO in GARCH-t(ν) under the null hypothesis.

B Alternative outlier detection procedures for GARCH(1,1) models

We discuss to alternative approaches. Hotta and Tsay (1998) built a procedure based on LM tests. Franses and van Dijk (2000) suggested a procedure based on regressions.

B.1 Additive volatility outliers

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

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

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

B.2 Additive level outliers

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

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

$t < J \leq T$ 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 (2000) suggest the following procedure for detecting additive level outliers in GARCH(1,1) models. Using the 'variance innovations' $u_t = \varepsilon_t^2 - h_t$ and $u_t^* = \varepsilon_t^{*2} - h_t^*$ they rewrite (8) as (so this is under the impact of a neglected outlier):

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

where $\phi = 2\gamma\varepsilon_s + \gamma^2$ is the direct impact of the outlier on the sequence of variance innovations. The ϕ parameter is estimated by regression of $\widehat{u}_t^* = \widehat{\varepsilon}_t^{*2} - \widehat{h}_t^*$ on $\left\{ I(t=s) - \widehat{\alpha}_1 \widehat{\beta}_1^{t-s-1} I(t>s) \right\}$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained in the baseline GARCH(1,1) model. From this they solve for γ:

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

The largest $\widehat{\gamma}_s$ exceeding a certain critical value is used to remove the outlier from the data. An approximation for the critical value is offered for certain significance levels. If an outlier is found, at t_0 say, the procedure is repeated for $y_t - \widehat{\gamma_{t_0}} I(t=t_0)$ 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 (2000) select the smallest solution (in absolute value). Although this provides a unique choice for γ, their regression method for the variance innovations often suggests the existence of multiple solutions for γ, even when these are not indicated by the log-likelihood. See, e.g., the left hand side likelihood grid in Figure 1.

Both the LM based approach and the regression procedure are rather complex, and suffer from non-similarity with respect to the GARCH parameters, so that new simulations are needed to compute p-values in each empirical application.

B.3 Simulation comparison

Next, we contrast our procedure to these alternative methods, denoted FD for the regression procedure of Franses and van Dijk (2000), and HT for the LM test based approach of Hotta and Tsay (1998). The results are in Table 11.[4] The main findings are that FD, although not designed to test for AVO, it will have some power against it; FD has lower power than HT when the outlier is of type ALO, probably because HT actually uses two tests (a more appropriate comparison would be with LM_{ALO} only). HT and our procedure have similar power, but the latter is much better at dating the outlier. Surprisingly, HT is worse at dating for the larger outliers as the LM tests lose their optimal power properties for distant alternatives. In addition, our procedure is more successful in classifying the outlier.

Table 11: Size and power of outlier detection tests for a single outlier in a GARCH(1,1) model

		Rejection frequencies			*Correct date*		*Correct type*	
	α_1, β_1	$\gamma = 0$	-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 LM approach of Hotta and Tsay (1998); FD is regression method of Franses and van Dijk (2000). For further notes: see Table 4.

B.4 Application Comparison for the Dow Jones returns

Table 12 lists the results when applying the three procedures to the monthly Dow Jones returns. The order in the table is that in which the outliers were detected, and we also include the first outlier with a p-value $> 5\%$.

The procedure of Hotta and Tsay (1998) finds the same outliers as our method, with two additional ones. Note that Hotta and Tsay (1998) use 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 (2000)'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 rather different from the other methods, as the multiple solution for γ suggested by their regression for the variance innovations did not arise in the other methods.

This could also explain the subsequent differences 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.

[4] To compute the rejection frequency, we used the extreme value approximation (12) for 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 $\max(3, \kappa_\epsilon)$. 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 12: 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	0.0_83	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$^{\text{AVO}}$	p-LM$^{\text{ALO}}$
1987/10	ALO	-4.38	0.0_74	0
1914/12	ALO	-3.58	0.0_45	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 (2000)				
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: $0.0_45 = 0.00005$

Table 13 gives the results for the weekly data. Four out of the seven outliers that are found by both our procedure and HT are now of a different type. The HT procedure detects two more outliers albeit at p-values that are not very low.

The application comparison shows two clear benefits of our new procedure: 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 less than half an hour for nearly $30\,000$ observations (on a 800 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 is scope for implementing this more efficiently.

Table 13: 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$^{\text{AVO}}$	p-LM$^{\text{ALO}}$	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.