Analysis of financial time series: a new robust approach based on the forward search

Lisa Crosato

Department of Economics, Management and Statistics (DEMS), University of Milano - Bicocca, 26100 Milano, Italy

Luigi Grossi*

Department of Economics, University of Verona, Via Cantarane, 24, 37129 Verona, Italy

Abstract

The main purpose of the paper is to study the effect of extreme observations on the estimates of GARCH(1,1) models. To this end the Forward Search (FS) approach has been extended to time series. As the FS was originally developed for independent data, the extension to dependent data raises issues which have been addressed in the earlier literature by using methods for dealing with missing observations. In the present paper a new Weighted Forward Search (WFS) approach is introduced. It is based on a weighting system of each unit and overcomes the issue of missing data. A WFS test is suggested and calibrated by an extensive Monte Carlo simulation to detect highly influential observations in GARCH(1,1) models. The size and the power of the WFS test are usually similar and in some cases better than those of other outlier detection methods suggested in the literature for GARCH models. After the definition of the test, a robust WFS estimator of GARCH(1,1) coefficients is introduced. The Monte Carlo experiment shows that the bias of the estimator is low even when a strong contamination is introduced into the time series. Finally, the application of the WFS test and estimator to several financial time series of the NYSE reveals the robustness of the method to the presence of extreme returns.

JEL codes: C13, C15, C58, C63

Keywords: Extreme observations, GARCH models, Outliers, Robust statistics, Weighted Forward Search

*Corresponding author

Email addresses: lisa.crosato@unimib.it (Lisa Crosato), luigi.grossi@univr.it (Luigi Grossi)

Abbreviations: FS = Forward Search, WFS = Weighted Forward Search, CDS = Clean Data Set

July 22, 2016

1. Introduction

Return series of financial assets are characterized by high order dependence, volatility clustering and high kurtosis. ARCH and GARCH models, since their introduction by Engle (1982) and Bollerslev (1986), have become increasingly popular in parameterizing these characteristics. In order to capture excess kurtosis of the data, GARCH models have been also extended to t-Student errors (Bollerslev, 1987), but empirical evidence documents that estimated residuals still exhibit excess kurtosis, often due to the presence of extreme outliers (Bali and Guirguis, 2007). Outliers are extreme observations that affect the estimation of parameters (Van Dijk *et al.*, 1999; Galeano and Tsay, 2010), the tests of conditional homoscedasticity (Carnero *et al.*, 2007; Grossi and Laurini, 2009) as well as the out-of-sample volatility forecasts (Chen and Liu, 1993a; Franses and Ghijsels, 1999; Catalan and Trivez, 2007; Charles, 2008).

Some contributions proposing robust estimators for ARCH and GARCH models have been recently published. A robust estimator for GARCH(p, q) models has been introduced by Muler and Yohai (2008). A new estimator for heavy-tailed and asymmetric GARCH models, based on the negligibly trimming QML criterion has been suggested and discussed by Hill (2015). Hung (2014) uses a robust Kalman filter to forecast conditional volatility, which is used to improve the robust performance of GARCH models. A M-estimator for multivariate GARCH models with t-Student distribution has been proposed by Boudt and Croux (2010).

Several proposals to detect outliers in GARCH models can be found in the literature. The method developed by Hotta and Tsay (2012), based on a Lagrange multiplier test, suffers from the masking effect, which occurs when the presence of one influential observation masks the presence of other outliers. The methods suggested by Franses and Ghijsels (1999), Franses and Van Dijk (2000), Charles and Darnè (2005) and Doornik and Ooms (2005) are instead extensions of the procedure which was introduced for ARIMA models by Chen and Liu (1993b). This procedure is iterative and works on single deletion diagnostics which are based on coefficients estimated assuming outlier-free data. The result is that, when multiple outliers are present in the data set, deletion diagnostics can be badly biased by the presence of other outliers. Alternative procedures for outlier detection in

GARCH models were proposed by Bilen and Huzurbazar (2002) and Grané and Veiga (2010), both based on wavelets. According to Grané and Veiga's results, the main advantage of the wavelet-based procedures is that they avoid the masking effect and, in particular, noticeably lower the detection rate of false outliers. The method introduced by Laurent *et al.* (2014), based on a semi-parametric statistical test for outlier detection, does not suffer from the masking effect. However, the large number of detected outliers raises concerns about the size of the test. Grané and Veiga (2014) compare some of the robust estimators and tests for the GARCH outliers described above, finding a prevalence of the method proposed by Grané and Veiga (2010).

This paper has two purposes. The first is to introduce a new outlier test for GARCH(1,1) models, assessing its size and power. The second is to define a robust estimator for the same class of models. Both are achieved by extending the Forward Search (FS) technique (Atkinson and Riani, 2000) to GARCH(1,1) models. The FS is an efficient and robust method originally developed for linear models to unmask multiple outliers and to measure their effect on model estimates. As previously said, some methods have been developed to identify observations which strongly influence GARCH model estimates. These methods, like all backward methods, are affected by the masking effect which prevents the detection of multiple outliers. The FS overcomes the drawback of single deletion methods because it monitors outlier diagnostics starting from an initial subset free from outliers (the Clean Data Set: CDS hereafter).

One of the distinctive characters of the FS is that it ranks the observations according to their degree of accordance with the model. Since time series data possess by definition a temporal ordering, there is a conflict between these two ranking criteria.

Another distinctive feature of the FS is that many models should be iteratively estimated based on different subsets of observations. When data are independent, the method does not raise any issues as it is possible to select subsets of observations without constraints. In time series this is not possible. We cannot select any possible subset of the original time series, but only patches of consecutive observations to respect the time order of units. A natural way to deal with the problem is to consider observations not belonging to the subset as missing data and estimate the parameters with estimators suitable for treating missing observations (Riani, 2004). An alternative solution, suggested by Grossi (2004), is to replace observations outside the subset with data simulated using robust parameters. The two solutions, although empirically effective, are not optimal because they rely on methods which cannot be always applied, as in the case of GARCH models (see next section for details). To fill these gaps, we suggest a new robust procedure where observations outside the estimation subsets are down-weighted instead of being treated as missing or replaced by simulated data. The weighting function is defined within the iterative FS, so that no arbitrary elements and choices are introduced. In this way, all observations are used for estimating parameters and maintained in the original time position. The theoretical background is mimicked from the classic FS, but could be considered a generalization of the parent method. As will be discussed throughout the paper, the classic FS can be considered a special case of the new procedure which will be called Weighted Forward Search (WFS hereafter).

The paper is structured as follows. In section 2 the general GARCH(1,1) model is presented and the main issues related to the extension of the FS are discussed. The new WFS procedure is introduced in section 3. The new method is organized into two main steps which are discussed in two different subsections: the choice of the initial subset free from outliers (section 3.1) and the weighting scheme applied to each unit (3.2). A typical output of the WFS is shown in section 3.3. The WFS test and the WFS estimator are introduced in section 4. As the WFS test is crucial for the definition of WFS estimator an extensive Monte Carlo simulation is reported in section 4.1 and a comparison of the size and power of alternative outlier tests is presented in section 4.2. An application of the WFS test and estimator to financial time series is carried out in section 5. Section 6 reports final remarks and conclusions.

2. GARCH models and the forward search

In this section we provide some mathematics related to the basic GARCH(1,1) model. As the procedure applied in this paper is a novelty in regard to the GARCH family models, we start from the simplest specification; but the procedure we suggest could be easily extended to more complex models.

Let r_t, with $t = 1, \ldots, T$ denote an observed time series of returns: $r_t = log\left(p_t/p_{t-1}\right)$ where p_t is the price of a security at time t.

The GARCH(1,1) model can be simply described as:

$$r_t = \mu + \epsilon_t, \epsilon_t|F_{t-1} \sim N\left(0, \sigma_t^2\right) \text{ and } \sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \qquad (1)$$

with $\alpha_0 > 0, \alpha_1 \geq 0, \beta \geq 0, \alpha_1 + \beta < 1$.

When we try to extend the FS to time series, we face the problem of time ordering. As well known, observations outside the CDS in the classic FS applied to independent data are simply considered as missing values (Atkinson and Riani, 2000, pp. 22-24). Handling missing values when dealing with dependent data, such as time series, is a much more difficult task (Penzer and Shea, 1999; Johansen and Nielsen, 2016). A family of robust estimators for autoregressive processes with missing values has been introduced by Kharin and Voloshko (2011). In the framework of ARMA models, the Kalman filter could be conveniently applied to estimate the coefficients with missing values (Gómez *et al.*, 1999; Proietti, 2008). This approach is adopted by Riani (2004). In a first step he uses a state space representation of an ARMA model. In the second step, when missing values occur, the Kalman filter equations are not updated and the innovations are set equal to zero. This is a very natural way to deal with missing values in time series. Unfortunately, this procedure cannot be easily extended to other models. For instance, the state space representation of GARCH models is not easy to obtain, as indicated in Penzer (2007) and in Ossandón and Bahamonde (2011). It is certainly easier to obtain a state space representation for a simple ARCH model. However, once the state space formulation is obtained, the Kalman filter approach in Jones (1980) cannot be applied to compute the likelihood of an ARCH process with missing values. Bondon and Bahamonde (2012) suggested a least square estimation with missing values for ARCH models, but the method cannot be extended to GARCH models.

Bearing these open issues in mind, Grossi (2004) proposes extending the FS to GARCH models by avoiding the parameter estimation with missing values. In his procedure, Grossi suggests replacing the observations not belonging to the CDS with values simulated from a stochastic process whose parameters have been estimated at the previous step of the search. Although this method proves quite effective, it is not completely consistent, particularly at the first steps of the FS, when observations are simulated using parameters estimated on a very small number of observations. Moreover, the method is very computationally demanding and could be infeasible for large datasets.

3. A weighted forward search for GARCH models

In this paper we suggest filling the gap in the literature by introducing a modified version of the FS to avoid the issues related to the estimation of

GARCH parameters with missing observations.

The FS makes it possible to inspect the role played by each observation in the parameter estimation procedure. It is based on robust and efficient estimators and consists of three steps: the first concerns the choice of an initial subset free from outliers called Clean Data Set (CDS); the second refers to the way in which we progress in the FS; and the third relates to monitoring some relevant statistics during the search's progress. As said in the previous section, the difficulty of applying the FS to time series is due to the conflict between the ordering of the data introduced by the search and the natural temporal ordering of the data. The solution suggested in this paper is to move from the classic FS where observations could have just two weights (0 for potential outliers and 1 for remaining observations), to a weighted version of the FS where each observation receives a weight between 0 and 1. The weighting system depends on the degree of agreement with the model at the previous step of the FS. On the other hand, observations that would have been included in the search are given weight 1. In the next subsections, we describe the method in detail following the three steps of the classic FS.

3.1. Choice of the initial subset

The FS is always initialized on a subset of observations free from outliers which must be selected among a set of possible combinations of equal size. When units are assumed to be independent, the initial CDS of generic size m is chosen among all possible m-sized combinations of units which can be obtained starting from a set of n observations. For time series, the usual procedure proposed for independent data cannot be applied for several reasons which depend on the temporal dependence of the data. In particular, when GARCH models are used for forecasting purposes, future volatility is predicted iteratively and the estimation is based on past observations. Moreover, the log-likelihood function of GARCH models is estimated iteratively and the initialization is based on the first observations of the time series.

A method to deal with this issue is to transfer the idea of block sampling (Heagerty and Lumley, 2000) into the FS framework. According to this assumption, which seems particularly sensible in the case of stationary time series like financial returns, we can identify subgroups of contiguous units which maintain the same dependence structure of the original time series.

A procedure for the selection of the initial CDS based on the idea of block sampling was originally proposed by Riani (2004) for the estimation of

ARIMA parameters when time series present missing values and by Grossi (2004) to extend the classic FS procedure to GARCH(1,1) models.

The approach applied in the present paper, based on the idea of block-sampling, follows the article by Grossi (2004). The main features and symbols are recalled in the next paragraphs, while details can be found in the original paper.

Let T be the size of a time series of financial returns r_t, and b a number of initial observations. The main feature of the block sampling is the splitting of the remaining $T-b$ observations into a number f of subsets of contiguous units. To simplify the notation, we assume, without loss of generality, that the size of each subset is $g = (T-b)/f$, where g is assumed to be integer.

The generic h-th subset, with $h = 1, 2, \ldots, f$ subset $S_h^{(g+b)}$ is then made up of the units $r_1, r_2, \ldots, r_b, r_{b+1+(h-1)g}, \ldots, r_{b+hg}$.

The criterion used to select the best initial subset is the same as applied in the case of independent data: that is, the minimization of the median of squared residuals (Least Median of Squares estimator, Rousseeuw and Leroy, 1987).

For GARCH(1,1) models, to introduce the selection procedure, we need to define the standardized residuals as

$$\widetilde{e}_{t,S_h^{(g+b)}} = e_{t,S_h^{(g+b)}} / s_{t,S_h^{(g+b)}} \tag{2}$$

where e_t and s_t, are the estimates of ε_t and σ_t, respectively (see equation (1)). The estimated residuals are based on the MLE of the GARCH coefficients obtained using only the observations included in $S_h^{(g+b)}$ (see Grossi, 2004, for details on the log-likelihood function).

Thus, the best initial subset is given by the observations which minimize the median of squared standardized residuals, that is

$$\underset{h}{min} \left[\widetilde{e}^2_{[med],S_h^{(g+b)}} \right] \tag{3}$$

where $\widetilde{e}^2_{[med],S_h^{(g+b)}}$ is the j-th ordered residual estimated on observations in $S_h^{(g+b)}$, among $t = b+1, \ldots, T$ and $med = [(T-b)/2]$.

The choice of the initial subset is influenced by two factors. First, the possible presence of outliers among the b initial observations of the original time series. Second, the choice of the number of subsets of size $g + b$. The first issue can be resolved with backward forecasts of the b initial observations based on the remaining $T-b$ units. A solution to the second problem is given by a heuristic rule which finds the optimal value as $g = \sqrt{T}$. For a detailed discussion of these issues, see Grossi (2004).

3.2. Weighting observations during the forward search

At each step of the classic FS, the size of the initial subset is increased by adding new observations to the CDS and, sometimes, removing others. Therefore, the observations augmenting the size of the CDS at a given step contribute to the estimates, while the others are excluded from the estimation process until the subsequent step when a new ranking on all units is defined. One of the main contributions of this paper, which characterizes the Weighted FS (WFS) for time series, is the introduction of a new approach. At each step of the search, estimation is carried out on all observations but not all of them contribute with their observed value: observations are weighted to account for their degree of outlyingnes[2].

In particular, moving from step k to step $k+1$ of the search

- all units, but the first b, are sorted according to their degree of agreement with the parameters estimated at the previous step of the search. The degree of agreement is measured by squared standardized residuals defined in equation (2), obtained from estimates of step k. Thus, at each step of the algorithm, the data are ordered by the WFS, as in the case of the classic FS;

- the first $g+k$ observations in the ranking defined at the previous step are given weight 1; the remaining observations are given a weight which is proportional to the corresponding value of the complementary cumulative distribution function of the squared standardized residuals defined on the whole sample.

In this way, each observation that would not have joined the CDS according to the classic FS, is down-weighted by the probability that the corresponding or a larger residual may be observed. Weights range from 0 to 1, so that the closer the weight to 0 the higher is the likelihood of the observation

[2]Other robust estimators are based on weighting schemes (Hill and Prokhorov, 2016). However, in this paper weights are computed considering all observations, including very extreme ones. In our approach, weights are adapted to the contamination pattern of the time series and, thanks to the test which will be introduced in the next section, they are not influenced by the presence of very large observations.

being an outlier, while the closer the weight to 1 the stronger is the degree of agreement of the observation with the estimated model. Observations with weight 1 form the CDS.

With this approach it is possible to achieve two goals, which cannot be pursued with the classical FS:

1. the temporal structure of the time series is respected, filling the time gaps created by the forward ordering;
2. all observations can be ordered according to their degree of agreement with the model estimated at the previous step.

Thus, the autocorrelation structure of the data is maintained, since volatility clustering will be accounted for by heavier weights, while the influence of outliers will be watered out by smaller weights.

The details of the procedure can be summarized as follows. In order to obtain stable estimates of GARCH parameters, the first b observations are always considered with their original value. Let $m = b + g$ be the size of the subsample chosen at the first step. Going from step 1 to step 2 of the search, all $T - b$ observations are then ordered according to their squared standardized residuals $\widetilde{e}^2_{t,S^m}$ for $t = b+1,\ldots,T$, so that each observation obtains a forward ordering given by squared residuals.

At each step $j = (b+g),\ldots,T$, the WFS assigns to each observation r_t a weight, say $w_{t,j}$, which is defined as:

- $w_{t,j} = 1, t = 1,\ldots,m-b+1$, for the observation $r_1,\ldots,r_{m-b+1}$ belonging to the CDS

- $w_{t,j} = 1 - F_{res}\left(\widetilde{e}^2_{t,S^m}, 1\right), t = m-b+2,\ldots,T.$

where $F_{res}(\cdot)$ is the squared standardized residuals distribution function[3]. In practice, at each step, standardized residuals are tested to be χ^2 distributed. If the p-value of the Kolmogorov-Smirnov test exceeds the critical value 0.05, we use tabulated values of the $\chi^2(1)$ distribution; otherwise a kernel density estimation is used.

The weighted observations are then re-ordered according to time, so that the temporal structure of the time series is recovered. The weighted ordered

[3]It is important to note that weights are obtained by exploiting the FS ordering.

series is used to estimate the GARCH parameters. Finally, we move from step 2 to step 3 in the same way, until all observations enter the CDS with their original value, that is, with unit weight. As the WFS proceeds, the net around outliers becomes tighter and their value is down-weighted until the end of the search.

Note that the outlier decontamination process begins with the initialization of conditional variance σ_t^2, given the importance of choosing a suitable number of initial observations to estimate the initial instances of the conditional variance process in GARCH estimation (Pelagatti *et al.*, 2009). In our context, the conditional variance initialization is even more important since outliers entering the variance process at the earlier steps could have a ripple effect on the whole GARCH estimation. We adopt here a forward variance initialization approach, since it turned out to produce the steadier estimates throughout the search if compared with other approaches. More precisely, we use only the observations belonging to the CDS in order to assign an initial value to the conditional variance not influenced by outliers.

It is very important to stress that the classic FS is a particular case of the WFS, where the weights could only assume two values: zero when the observation does not belong to the CDS; or one when the observation is in the CDS. Thus, the weighted forward approach could be considered a more general procedure which, of course, maintains the diagnostic properties of the classic FS. When the WFS is applied, it is possible, as in the classic FS, to measure the influence on estimates and on trajectories of residuals of each observation at the time it enters the CDS and receives weight 1. Note that weighted data at step $k+1$ are conditional to weighted data at step k, and that coefficients are estimated using the data set composed of all observations, whatever the weight is. On the contrary, residuals used to order observations are based on the coefficients estimated on the original $T-b$ units[4].

Furthermore, by including all observations at each step, we obtain a more stable pattern in the output of the search. If the estimates were based only on observations with weight 1 and on the units filling the time gaps, on moving from step k to step $k+1$ we would not know the number of observations

[4]At step k the CDS is made up of $b+g+k-1$ observations maintaining their original value while the remaining units are down-weighted. Overall, a WFS over T observations counts $T+1-(b+g)$ steps.

involved in the estimation process, which would depend on the time distance between units with weight 1.

3.3. Weighted forward output

The second stage of the WFS illustrated in the previous section is repeated until all units are included with their original value; therefore, until the CDS coincides with the original time series. The output of the search is mainly graphical as in the classic FS: separate plots with coefficient estimates, t-statistics and residuals can be reported for the last steps of the search.

A simple example of the output is reported in Figure 1 where residuals (top right panel), coefficient estimates (second row panels) and t-statistics (third row panels) are reported. All statistics have been obtained applying the WFS to a simulated GARCH(1,1) series of length 500 with $\alpha_0 = 0.01, \alpha_1 = 0.07, \beta = 0.9$ (top left panel), during the last 10% steps of the search. The series were contaminated by 3 Additive Outliers (AO) of different magnitude following the framework of Carnero *et al.* (2007). Hence, observations contaminated by AO are as follows:

$$r_t^* = r_t + \omega \tag{4}$$

with $\omega = 5\sigma_r$, $\omega = 10\sigma_r$ and $\omega = 15\sigma_r$, where σ_r indicates the standard deviation of the time series before contamination.

Figure 1 about here

Most of the time, the graphical output of the FS provides the researcher with quite clear indications on which observations should be considered as outliers and, therefore, should be removed or corrected (Cerioli *et al.*, 2014). In the FS plots sudden jumps of the trajectories indicate that one or more influential observations entered the CDS (Riani *et al.*, 2015).

However, this approach inherently leaves a certain degree of subjectivity to the researcher. Consider for instance the example in Figure 1, which is applied to a trajectory generated by a GARCH(1,1) process[5]: it is evident

[5]Of course, this running example is only one of the possible infinite trajectories gen-

that the introduction of the second last observation causes a shift downward in the estimated value of β (second row, right panel), which further drops after the last observation joins the CDS. A similar pattern can be observed for α_0 and, though not easy to detect, for α_1.

The presence of at least two outliers is confirmed by the analysis of the WFS standardized residuals of the same series (last 50 steps of the search, see the top right panel). The residual trajectories of two observations, $t = 54$ and $t = 425$, markedly depart with respect to the others, and a third one ($t = 192$) also departs from the main group, suggesting a third outlier could be present. Note that the third last observation causes a mini-break in the plot of estimates (second row, in particular for α_1 and β) when it enters the estimation process with its exact value. The last two t-statistics (third row in Figure 1) go down for all three parameters, while the third last drops for α_0, stays in the range of preceding values for α_1 and goes up for β. Accordingly, it seems somewhat troublesome to decide whether observation $t = 95$ is influential.

In order to limit the subjective choices linked to the graphical visualization, we add to the classical FS output a new WFS test which can be used to mark an observation as an outlier with a probability level.

4. A weighted forward test for outliers

The introduction of a WFS test is crucial for two reasons:

1. the chances of arbitrary selecting one observation as outlier is reduced because, given a significant level of probability, it is always possible to say whether a unit can be considered an outlier;
2. once the number of outliers has been defined, the forward plots of coefficient estimates can be cut automatically: at this point, WFS robust estimates of the coefficients are obtained because they are not influenced by the detected extreme values (see section 4.3).

The null hypothesis of the new WFS test is as follows: given a single observation generated by a GARCH(1,1) model, this observation is not an outlier.

erated by the GARCH(1,1) process with that set of parameters. However, this example precisely mimics the situation that we encounter when we analyze one, and only one, real time series.

The WFS test for the presence of outliers is defined using simulated envelopes generated according to Atkinson and Riani (2000) through the following steps:

1. perform the WFS on a number, say "$n.sim$", of series of size T simulated by a set of GARCH(1,1) processes. The set of processes is defined by different combinations of parameters commonly observed when financial returns are modeled $\alpha_0 = 0.01$ and $\alpha_1 + \beta = 0.97$ (see Table 1). The total number of simulated trajectories is $n.sim = 10,000$.

Table 1 about here

2. detect the observation in the CDS that gives the largest standardized residual in absolute value

$$r_{i,t}^{max,j} = \underset{r_t \in CDS_i^j}{argmax} \left(|\widetilde{e}_{i,t}^{j}| \right)$$

   where
   $j = 1, 2, \ldots, 10000$ is the series index and
   $i = n - (p \times T - 1), n - (p \times T), \ldots, n - 1, n$ is the sequence of the steps over the last $p \times T$ steps, p is a given percentage of T and n is the total number of steps in the WFS;
3. the bounds of the outlier detection interval are given by the αth and $(1 - \alpha)$th percentiles, for each of the last $p \times T$ steps, over

$$\widetilde{e}\left(r_{i,t}^{max,j}\right), \quad \text{where} \quad i = n - (p \times T - 1), n - (p \times T), \ldots, n - 1, n.$$

Observation r_t is declared to be an outlier if the corresponding standardized residual trajectory crosses the outlier detection interval for a fixed number of times (number of exceedances, $n.ex$) at least.

The final simulated envelopes, which are used as outlier detection intervals, have been obtained as the average of the intervals based on different sample sizes: that is, $T = 250, 500, 1000$, which are typical sample sizes observed when financial returns are analyzed[6]. Note that, by construction, the WFS test is independent of the parameters of the process.

[6]The ideal situation happens when the test is based on simulated time series of the same

4.1. Power and size of the test: a simulation study

The outlier test that we propose can be calibrated according to the level of significance α and the number of exceedances, $n.ex$. In this tuning procedure, we must consider the usual trade-off which exists between the maximization of the power of the test and the need to keep a reasonably low size.

The Monte Carlo study that we perform in this section is based on $N = 1000$ trajectories of size $T = 250, 500, 1000$ simulated by a GARCH(1,1) process. The coefficients of the process are set to $\alpha_0 = 0.01, \alpha_1 = 0.07, \beta = 0.9$, and the series are contaminated according to three different patterns:

- one single outlier of magnitude $\omega = 5\sigma_r, \omega = 10\sigma_r, \omega = 15\sigma_r$;
- three outliers, one for each of the above magnitudes;
- ten outliers: two of magnitude $\omega = 5\sigma_r$, four of magnitude $\omega = 10\sigma_r$ and four of magnitude $\omega = 15\sigma_r$.

All outliers are placed randomly along the series.

The power of the WFS test, which is run over each simulated series, is calculated as the mean of the percentage of correctly detected outliers over all replications, while the size is calculated as the mean of the percentage of erroneously detected outliers over all uncontaminated replications.

The power and size curves of the WFS test related to the number of exceedances expressed as a percentage of the final steps of the search are shown in Figure 2, Figure 3 and Figure 4.

Figure 2 about here

Figure 3 about here

size as the series under test. However, replication of many WFSs is quite time consuming. In addition, we have run several tests based on different sizes over the same time series and obtained negligible differences in the results.

Figure 4 about here

Three curves are represented in each panel: solid curves correspond to the $\alpha = 1\%$ intervals, dashed curves to $\alpha = 5\%$ intervals, and dotted curves to the $\alpha = 10\%$ intervals. Obviously, we expect that the larger the number of exceedances required to define outliers, the smaller the power and the size of the test. Nonetheless, a quite constant pattern emerges from most figures for either power or size depending on the magnitude and number of outliers and on the sample size.

Three general points must be stressed. First, regardless of the sample size, the power curves for the single largest outliers are basically equal to 100% whatever the number of exceedances and the level of significance (see second and third panel in the first row of Figure 2 - 4). Second, the $\alpha = 1\%$ significance level interval (solid curves) is the worst performer in terms of power (and the best one in terms of size), while differences in power (and size) among the 5% and 10% intervals are small. For this reason we shall focus the following comments on the 5% (dashed) curves. Third, power curves for the largest sample size ($T = 1000$) show a constant pattern up to 95% of the exceedances in the smallest outlier case, as well as in the multiple outlier cases.

Considering the remaining cases we observe that:

- $\omega = 5\sigma_r$ (top left panel of Figure 2 - 4): the power of the test does not show substantial changes when the sample size decreases and remains over 95%;

- three outliers (bottom left panel of Figure 2 - 4): as T decreases the power curves become slightly steeper, so that the loss of power is negligible using a smaller number of exceedances in proportion to the interval length. When $T = 500$ the loss of power with respect to $T = 1000$ is negligible, while for $T = 250$ the power can still be pushed above 95% using up to 70% of the exceedances;

- ten outliers (bottom central panel of Figure 2 - 4): the power still remains quite high, although it decreases more rapidly as the exceedances increase. For $T = 500$ a power of 95% is achieved with a number of exceedances up to 80% of the interval length. For $T = 250$ this limit falls to 40% of the interval length;

- size: (bottom right panel of Figure 2 - 4) empirical sizes are very low compared to the significance levels throughout the whole range of exceedances for any sample size.

In summary, the WFS test shows high levels of power combined with low sizes in all situations, the small and multiple outlier cases included. Importantly, decreasing sample size implies quite moderate losses in power and size of the test. From a practical point of view, the 5% test offers a good balance between power and size. At that level, the number of exceedances, if lower than a given threshold in proportion to the interval length, has no relevant impact on power. In particular, when the presence of multiple outliers is suspected in short series, the number of exceedances should reach at most one third of the considered final steps. However, the flexibility of the test allows power and size to be adjusted according to the specific series at hand.

4.2. Comparison with other methodologies

In this section, we compare the performance of the WFS procedure with other outlier detection methods for GARCH models. Recently, Grané and Veiga (2010) measured the performance of four different outlier tests, including the one proposed in their paper, calculating:

1. the percentage of outliers correctly detected over the total number of outliers in the simulated series (power of the test);
2. the average number of false outliers detected on the contaminated series.

Results are then compared over different sample sizes and different types of contaminations (see Table 2). The benchmark for power levels is the test suggested by Franses and Ghijsels (1999), which performs best in all cases, while the benchmark for the average number of false outliers is the test by Grané and Veiga (2010), which achieves an extraordinarily low average number of false outliers.

We calculate the same measures by running the WFS test for different levels of significance ($\alpha = 1\%, 5\%, 10\%$) and sample sizes ($T = 500$ and $T = 1000$), using the outlier detection interval based on 10% of the sample size and fixing the number of exceedances to 60% of the interval length[7]. We further add the results of the WFS test in the case of 10 outliers and a sample size ($T = 250$) smaller than those considered by Grané and Veiga (2010).

Table 2 about here

Three levels of significance of the WFS test are displayed to show its flexibility. Observe that in the case of a single outlier of magnitude 5σ, the rate of outlier detection of the WFS test is very close to the rate achieved by Franses and Ghijsels (1999) at 10% level, although a larger number of false outliers (8.88) is obtained. This number can be reduced to 0.40 maintaining a good power (88.0%) at the 1% level. For one outlier of size 10σ or 15σ, all the WFS tests have the maximum power (99.9 to 100%) with less than 0.5 false outliers per series (1% test). With three outliers, our test outperforms the others in terms of power, for all levels of significance (94.2% for the 1% test versus the 92.4% of F&G). The average number of false outliers per series (0.33) is the second lowest at the 1% level. The ten-outlier case confirms that the power of the WFS test is not significantly weakened by the increase in the number of outlying observations (it remains over 93%). At the same time, the average number of false outliers decreases as well as its variability. In fact, we have a smaller average number of false outliers with respect to F&G and a smaller variability. Their method produces 6 false outliers on average with a standard deviation of 10; while the WFS test detects at most 5.55 false outliers with a standard deviation of 1.62.

Indeed, this result points to a scant impact of both the masking effect and the swamping effect on the performance of the WFS test.

Results for $T = 250$ and $T = 500$ highlight this main strength that the WFS inherits directly from the classical FS method. In reducing the sample size, the WFS improves both power and the average number of false outliers, while Grané and Veiga's method loses in power, and that of Franses and Ghijsels increases the average number of false outliers in 4 out of 5 cases.

[7]Note that these parameters have been selected taking into account the results of the Monte Carlo study reported in section 4.1

Concluding, the WFS test's performance is generally in line with the other tests and is sometimes even better, with particular reference to the multiple outlier and short series cases. A plus of our test is that it is possible to tailor the level of significance to the user's preferences, according to the size and power which are needed. For example, if a preliminary observation of the time series does not suggest clear evidence of outliers, it is possible to use a higher level of significance (maximizing the power), risking a little more on the false outlier side. On the contrary, if one is willing to minimize the probability of erroneously declaring an observation as an outlier to less than one, the optimal level of significance drops to 1%.

While the WFS is in all occurrences on top of the table as far as the power is concerned, in order to reduce the average number of false outliers substantially one should push the number of exceedances up to 100%. For example, by placing a small, single outlier in a series of length $T = 1000$ we can reach 0.02 average outliers per series with a power of about 73.7% (complete results for the whole range of exceedances are available on request).

It is very important to stress that this test does not depend on the set of true parameters characterizing the time series generating process, since the outlier detection interval is based on a wide range of parameters; nor does it depend on the size of the time series under test because it was obtained by averaging detection intervals of different lengths.

4.3. The WFS estimator

It is well known that Maximum Likelihood (ML) and Generalized Least Squares (GLS) are not robust estimators of the parameters for a GARCH(1,1) model except for very large samples. Although the Quasi-maximum Likelihood (QML) estimator based on the Student likelihood is more robust than the classic estimators, it is still affected by outliers, particularly in the case of the coefficient β (Sakata and White, 1998; Mendes, 2000; Carnero *et al.*, 2007). In this section we introduce a new robust estiamtor of GARCH(1,1) models called the Weighted Forward Search Estimator (WFSE) and assess its robustness to the presence of outliers.

Furthermore, we control for the impact of false outlier detection on the same estimator, in order to obtain some indications on the balancing of size and power when testing for outliers in GARCH(1,1) models.

Let $\theta = (\alpha_0, \alpha_1, \beta)$ be the vector of parameters of a GARCH(1,1) model as in Section 2 and $\hat{\theta}_i$ be the MLE of θ at the i-th step of the search, $i = 1, \ldots, n$, with $n = T + 1 - (b + g)$. We define the WFSE of θ as:

$$WFSE\,(\theta) := \hat{\theta}_{n-n.out} \tag{5}$$

where $n.out$ is the number of outliers detected by means of the WFS outlier test. From this definition it follows that the WFS estimates derive automatically from the outlier identification process and that there is no need for further corrections of outliers.

As an example, consider the WFSE applied to the same trajectory simulated by a GARCH(1,1) process that we have already analyzed in Figure 1. The WFS test at the 5% level of significance detects exactly the three outliers with which the series was contaminated (top right panel of Figure 5). Accordingly, the WFS estimates of the three coefficients are identified by the vertical lines cutting the sequence of ML estimates before the three outliers enter with their original value (second row of Figure 5). As can be seen, the WFSE automatically corrects for outliers: once they are identified they are downweighted as seen in section 3.2 in order to achieve a robust estimate. Estimates of the three GARCH(1,1) parameters are close to the true values indicated by the horizontal dashed lines.

We now move to study the performance of the WFS test and estimator over 1000 trajectories simulated by the same GARCH(1,1) process with the same set of parameters ($\alpha_0 = 0.01, \alpha_1 = 0.07, \beta = 0.9$). In order to verify the robustness of the WFSE we compare the distribution of the ML estimates obtained over uncontaminated series with the distribution of the WFS estimates on the same series after contamination with three outliers of size $\omega = 5\sigma_r, 10\sigma_r, 15\sigma_r$. The WFS estimates are plotted at $7, 6, 5 \ldots, 1, 0$ steps before the end of the search (see Figure 6, Figure 7 and Figure 8) under the hypothesis that the WFS test has previously identified $7, 6, 5 \ldots, 1, 0$ outliers in all replications.

Figure 5 about here

Starting from coefficient α_0 (Figure 6), we can see that when the WFS test detects all the three outliers (bottom left panel, $n-3$ step of the search) there is a substantial overlap between the two distributions. Going back to 8 steps to the end of the search (top left panel) we still observe quite similar distributions, while moving forward we observe that if the test misses one outlier (the smaller, $n-2$ step) the WFSE distribution (grey) slightly separates from the MLE distribution on uncontaminated data, and it is completely biased if the test misses the whole set of outliers (bottom right panel, end of the WFS). Note that at the end of the search the WFS estimates coincide with the ML estimates over contaminated data.

Figure 6 about here

We can extend the above conclusions to coefficients α_1 and β (Figure 7 and Figure 8), although the overlap between the two distributions in the best hypothesis is not as clear as in the α_0 case, particularly for α_1.

Figure 7 about here

Figure 8 about here

Thus far the sequence of estimates has been artificially cut at the right point $(n-3)$. In Figure 9 we see instead the comparison between MLE over uncontaminated data and the WFS estimates automatically determined by the algorithm. Again, the overlap between the distributions is quite good, and, although coefficients α_1 and β show a larger variability with respect to α_0, the correction applied by the WFS estimator to the contaminated series appears to have made a pretty good clean up.

Figure 9 about here

5. Application to financial time series

The WFS test and estimator were applied to analyze a set of securities quoted on the New York Stock Exchange (NYSE) and the main stock exchange index. The securities were selected to create a quite diversified portfolio covering some of the main industries of the NYSE (see Table 3). Daily prices were downloaded from the Bloomberg platform for a sampling period extending from the beginning of 2006 to the end of 2015. We moved to weekly series of 522 observations selecting the intermediate day of each week, so that weekly log-returns measure the relative change of prices with respect to the previous Wednesday[8].

Table 3 about here

The plots of the weekly log-returns for four securities (Abbott Laboratories, Kimberly-Clark, S&P500 and Wal-Mart Stores)[9] are reported in Figure 10.

Figure 10 about here

Extreme returns detected by the WFS test are denoted by red circles, while the corresponding dates are shown in Table 4, third column.

Table 4 about here

[8] Weekly frequency makes it possible to analyze long time periods with not too long time series. On the other hand the presence of extreme returns is more probable than in the case of daily returns.

[9] The same plots are available for the remaining companies analyzed and can be obtained from the authors upon request.

The number of detected observations ranges from a minimum of 6 (Sysco corp.) to a maximum of 14 (HP). It is interesting to note that some dates are detected as extremely influential in many time series. For instance, observations 144 (8th October 2008) is pointed out in 8 out of 10 series. Indeed, on October 6, 2008 the Dow-Jones index closed below 10,000 for the first time since 2004. This was the lowest minimum of the market after September 14, when the Lehman Brothers announced the largest bankruptcy filing in U.S. history at that time. Another example is given by observation index 292, corresponding to August 10th 2011 (i.e. the return with respect to the previous Wednesday August the 3rd 2011) which is detected as an influential observation in 7 series. This has been one of the most critical times in the world financial crisis. The United States credit rating was downgraded by Standard & Poors from AAA to AA+ on 6 August 2011 for the first time since 1941 due to the slow economic growth in the US. The European Central Bank was expected to start buying Spanish and Italian government bonds in order to save the Euro, and there was fear of contagion to other European countries. As a consequence, many stock exchanges around the world, NYSE included, experienced large losses which have been detected by the WFS test. A more recent event which affected financial returns was the stock market sell-off, which began in the United States on August 18, 2015, when the Dow Jones Industrial Average fell 33 points, triggered by concerns that China was not doing enough to stabilise its economy. The downward effect on the US financial market has been detected in 5 series as a big negative return recorded on 26th August 2015 (index number 503).

The detection procedure can be visualized by looking at Figure 11, where the WFS trajectories of standardized GARCH residuals are reported for the usual selection of four companies.

Figure 11 about here

In each panel the dashed bold lines identify a 95% confidence region obtained as simulated envelopes described in section 4, and the outlying trajectories discovered according to WFS test are the colored bold lines. The numbers which appears at the right side of each panel correspond to the red circles drawn on the log-return plots (Figure 10). From the plots reported in Figure 11 is possible to see clear examples of the masking effect which has a negative impact on the ability to detect influential observations when backward detection methods are applied. As said earlier, the masking effect occurs when one or more outliers are masked by the presence of other outliers in the same dataset. In this case, the MLE residuals of influential observations are not particularly high, but, on the contrary, tend to be very close to or even lower than the residuals of "normal" units. This is exactly what happens, for instance, to observations 384 and 270 in the ABT panel, and to observations 436 in the SP500 panel. MLE residuals of these observations at the end of the search are mixed with residuals of other units inside the confidence regions. This means that detection methods based on the observation of residuals calculated on the whole sample, but even detection methods based on the deletion of few observations (backward methods), would not be able to correctly identify influential observations which are instead easily detected by observing the WFS trajectories.

Looking at Table 4 (last two columns) it is possible to compare the robust and non-robust estimates of the GARCH coefficients (α_0, α_1 and β) for the four securities shown in Figure 10. The comparison gives an idea of the correction obtained when the robust estimator is used. For example, the WFS estimate for Abbott is around 0.9, while the MLE is approximately 0.77; for Walmart the estimate moves from 0.97 to 0.84.

As it is well known (Hwang and Pereira, 2006), GARCH estimates has proven to be very unstable when the sample size is small. The WFSE has shown to be robust even to the reduction of sample size. As an exercise, we have reduced the sample size to 50% of the original time series length, considering only the last five years, to compare the estimated GARCH coefficients obtained both by the MLE and the WFSE. The results are quite interesting because they reveal that the instability of the MLE is due to very few large returns which strongly affect the estimation of the GARCH coefficients with a particular emphasis to the coefficient β. For example, the WFS estimate for Abbott is around 0.65, while the MLE is approximately 0.1; for Kimberly-Clark the estimate moves from 0.65 to 0.06.

Table 5 about here

Finally, Table 5 reports for each series the t-stats of the GARCH coefficients obtained as ratios of the average values of estimates and standard errors, before (tstatsB) and after (tstatsO) the step identified by the WFS test. The average value has been computed either as the arithmetic mean or as the median. The extent of the difference between the two types of t-statistics gives an idea of the influence of extreme observations on the significance of coefficients. When the robust version is computed the t-statistics tend to increase, particularly when the median is used. In some cases (Abbott, Sysco, Walmert), considering the a significance level of 1%, the corresponding estimates of α_1 become not significant when influential observations are included.

6. Concluding remarks

This paper has proposed a new robust estimator of the GARCH(1,1) model based on the generalization of the FS procedure (Atkinson and Riani, 2000) to the case of time series. The extension of the FS to time series has been suggested in previous papers with reference to ARMA models (Riani, 2004) and to GARCH models (Grossi, 2004). The main issue discussed in the literature is how to deal with missing values generated during the FS. The solutions have so far consisted in estimators based on the Kalman filter, which enables estimation of ARMA coefficients with missing observations, or in replacing missing values with simulated data. Both approaches have the advantage of maintaining the temporal order of the units even when a subset of the initial sample is used for estimation purposes. However, they can be considered sensible solutions to cope with the problem of missing data in particular cases, which cannot be generalized to different classes of time series models. Moreover, previous works have focused on monitoring the effect of extreme observations on the estimation output, while the problem of finding a robust estimator has been neglected.

We have tried to fill these gaps through a new approach based on a weighting system of single units which leads to a generalized method called Weighted Forward Search (WFS). The classic FS can therefore be considered as a special case of the WFS where observations can be weighted using just two values: zero when the observation does not belong to the Clean Data Set; one when the observation has not yet joined the subset. The main advantage of the WFS is that all observations are used to estimate parameters at each step of the search, but their impact is given by a weight which is always inside the interval $(0, 1)$. Consequently, this method might easily be extended to any type of time series model.

Our methodology was developed and applied through three main steps.

First, we have introduced a WFS test based on simulated envelopes (Atkinson, 1994). The test was calibrated by means of an extensive Monte Carlo experiment, resulting in a set of simulated confidence regions ready to be effectively applied to detect outliers in time series with a length which is usually observed in daily or weekly financial prices. The size and power of the WFS test was assessed through another Monte Carlo simulation and then compared to other detection procedures proposed in the literature. The results are promising since the performance of the WFS test is on average on the same level as that of the best methods, with reference to both size and power. In particular, in the presence of multiple outliers the WFS shows the highest power with a competitive number of false outliers.

Second, a WFS estimator based on the number of exceedances of the WFS trajectories of residuals with respect to the simulated regions was defined. The weighted estimates obtained by downweighting the outlying observations were identified as the robust WFS estimates. Furthermore, the bias of the robust estimator was studied, either with uncontaminated or contaminated time series, revealing its good performance.

Finally, the application of the robust WFS test and estimator to several time series of returns computed on the NYSE confirmed that the suggested approach relies on the main bulk of observations, while the ML estimator is usually badly biased by a few extreme observations which strongly influence the GARCH estimates and their significance.

Further research will be devoted to study the theoretical properties of the WFS estimator. It is well known that the robustification of the estimators is obtained at the cost of a lower efficiency (Rousseeuw and Leroy, 1987). The FS is a very flexible procedure which combines the properties of robust methods with the high efficiency of MLE. However, the use of the WFS test could affect the efficiency of the robust WFS estimator. Moreover, additional simulations should be carried out to observe the performance of the robust estimator when larger sample sizes are considered. This is particularly interesting when high-frequency data are analyzed. Finally, it would be very interesting to study how the application of the WFS estimator could improve the forecasting performance of conditional volatility models.

Acknowledgements

The authors wish to thank all the participants at the final conference of Project MIUR (2012) in Benevento, at Sco Meeting (2013) in Milan, and at the meeting on Robust Statistics at the Joint Research Centre of the European Commission in Ispra (2014) where preliminary versions of the present work were presented. All comments and suggestions have been considered and have contributed to improving the quality of the manuscript. This work was supported by the project MIUR 2012 - MISURA and by the Department of Economics (University of Verona), grant FAR/2015. The usual disclaimer applies.

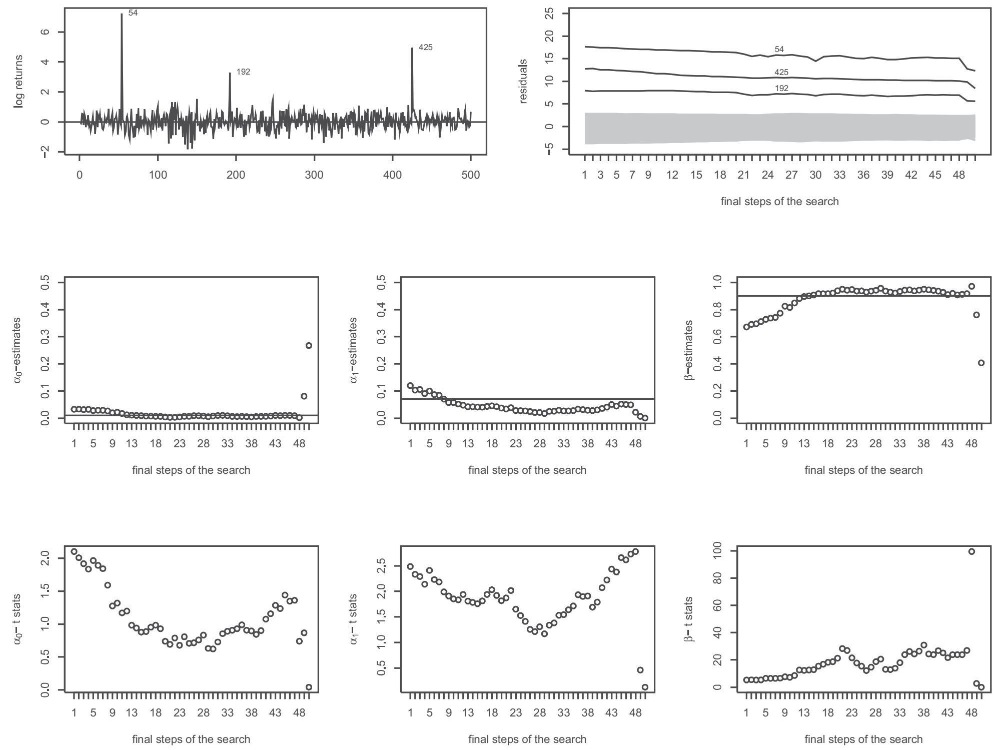

Figure 1: GARCH(1,1) simulated series with three AO of size 5σ, 10σ and 15σ (top left panel), standardized residuals along the WFS (top right panel, last 10% steps of the WFS) and coefficient estimates (second row, last 10% steps of the WFS). Third row: t-statistics.

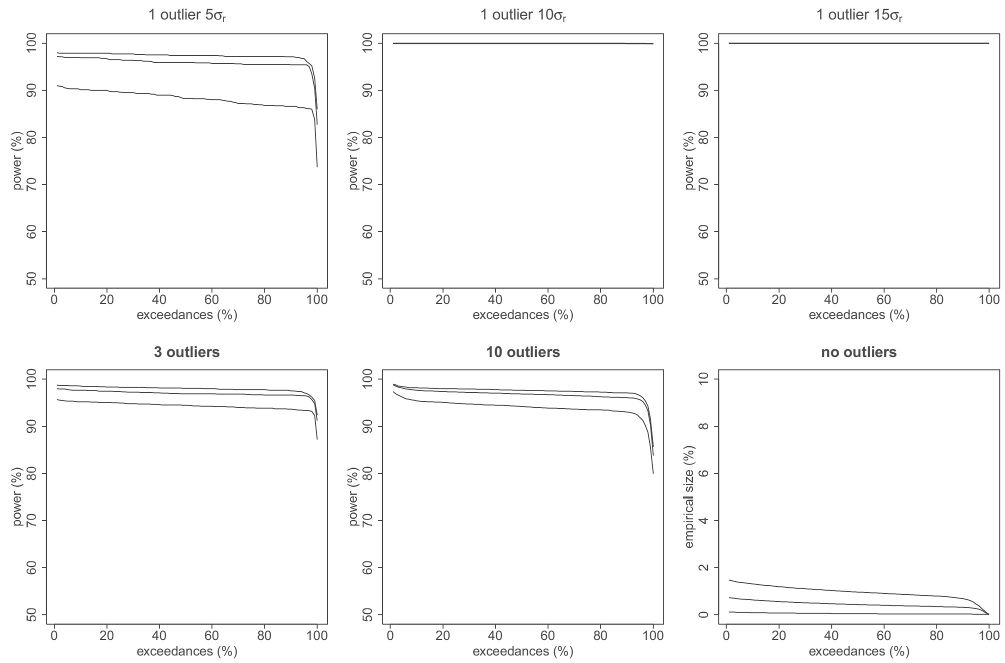

Figure 2: Power and size curves for the WFS test according to different significance levels (solid line $\alpha = 1\%$, dashed line $\alpha = 5\%$, dotted line $\alpha = 10\%$) and number of exceedances as a percentage of the final steps of the search (on the horizontal axis). The test is executed on N=1000 GARCH(1,1) trajectories of size T=1000 contaminated by outliers of different magnitude.

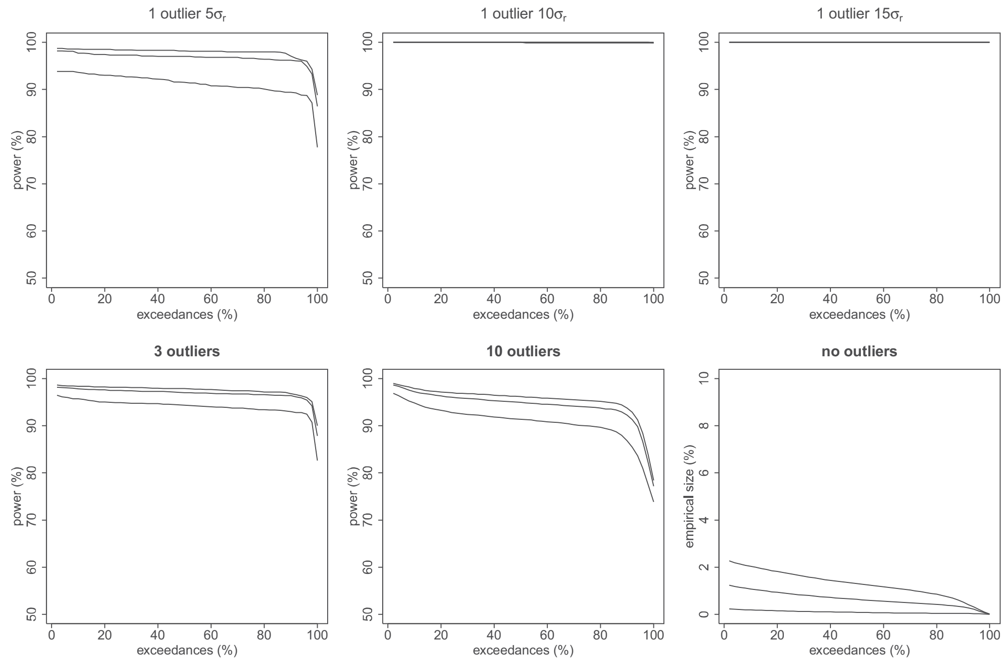

Figure 3: Power and size curves for the WFS test according to different significance levels (solid line $\alpha = 1\%$, dashed line $\alpha = 5\%$, dotted line $\alpha = 10\%$) and number of exceedances as a percentage of the final steps of the search (on the horizontal axis). The test is executed on N=1000 GARCH(1,1) trajectories of size T=500 contaminated by outliers of different magnitude.

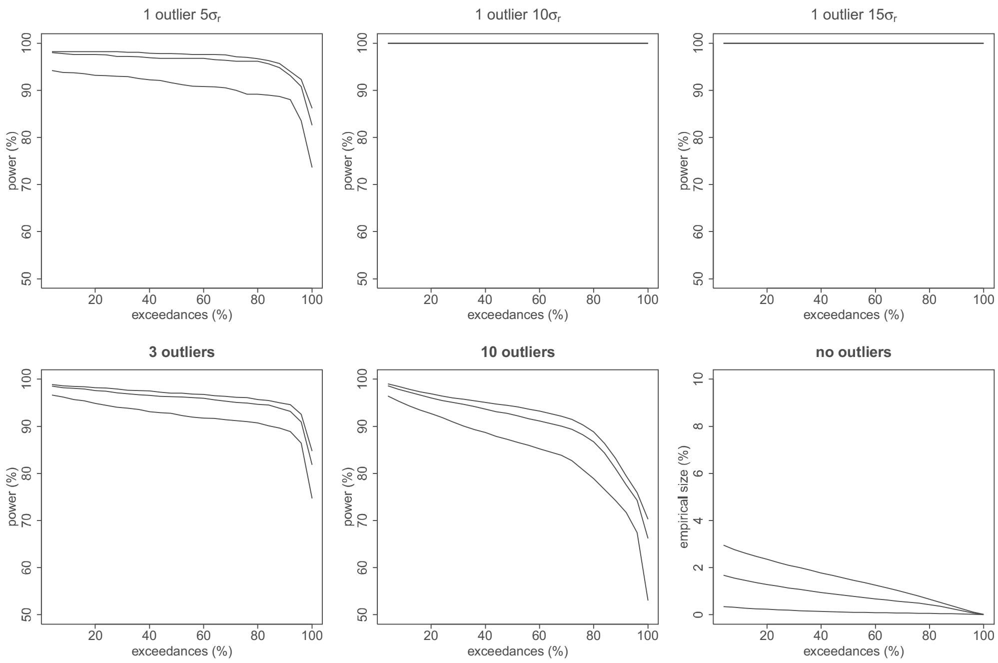

Figure 4: Power and size curves for the WFS test according to different significance levels (solid line $\alpha = 1\%$, dashed line $\alpha = 5\%$, dotted line $\alpha = 10\%$) and number of exceedances as a percentage of the final steps of the search (on the horizontal axis). The test is executed on N=1000 GARCH(1,1) trajectories of size T=250 contaminated by outliers of different magnitude.

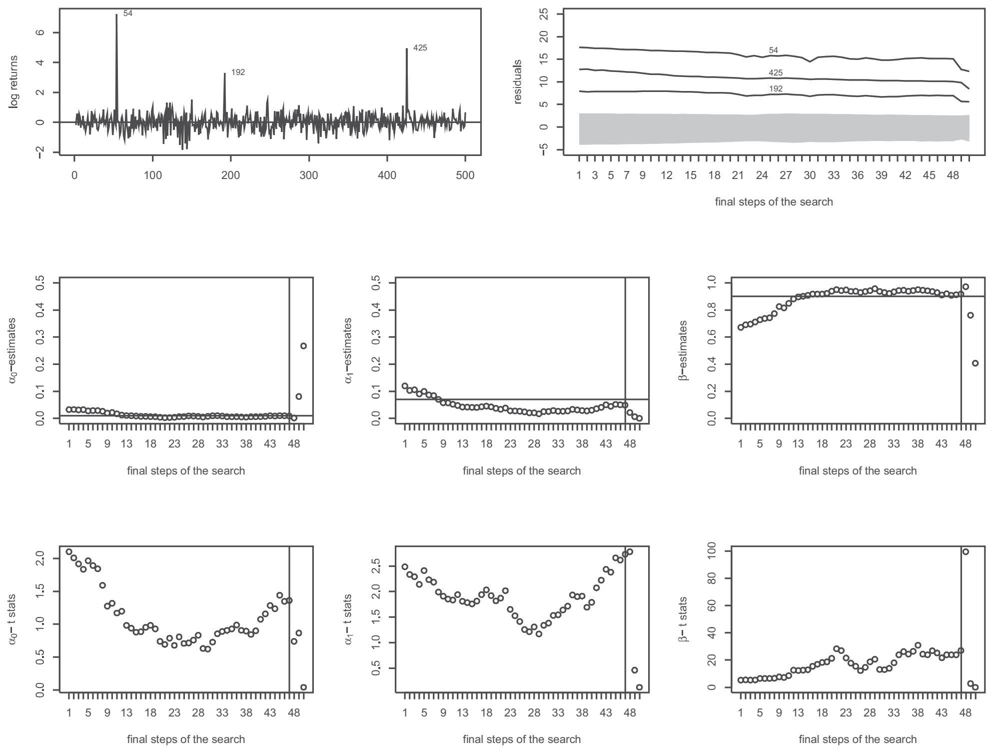

Figure 5: The WFS process of estimation on the series of Figure 1. Top right panel: standardized residuals and outlier detection interval (dashed lines). Second row: coefficient estimates; horizontal lines are the true coefficient values and vertical lines cut the plot into the WFS estimates. Third row: t-statistics.

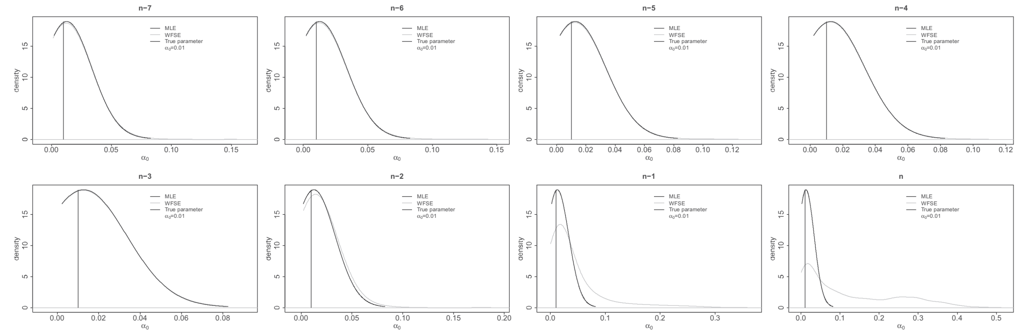

Figure 6: Density of α_0 ML estimates on uncontaminated series (black) vs WFS estimates on the same series contaminated by three outliers (grey), last 8 steps of the FS. The WFS estimates are obtained by cutting the sequence of estimates at $7, 6, \ldots, 1, 0$ steps before the end of the search for all the series.

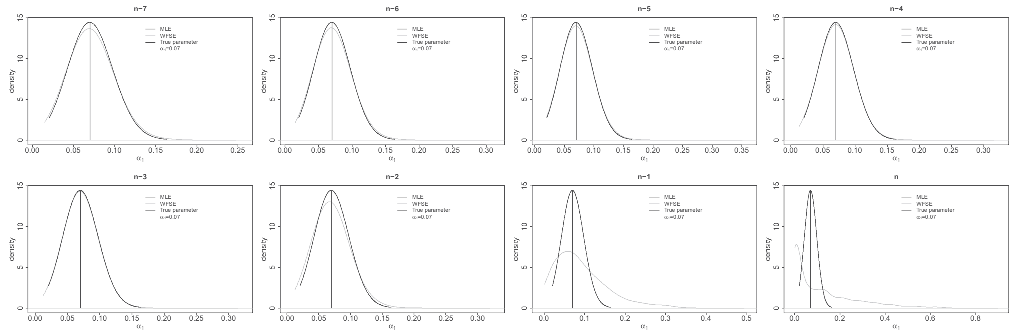

Figure 7: Density of α_1 ML estimates on uncontaminated series (black) and WFS estimates (grey) on the same series contaminated by three outliers, last 8 steps of the FS. The WFS estimates are obtained by cutting the sequence of estimates at $7, 6, \ldots, 1, 0$ steps before the end of the search for all the series.

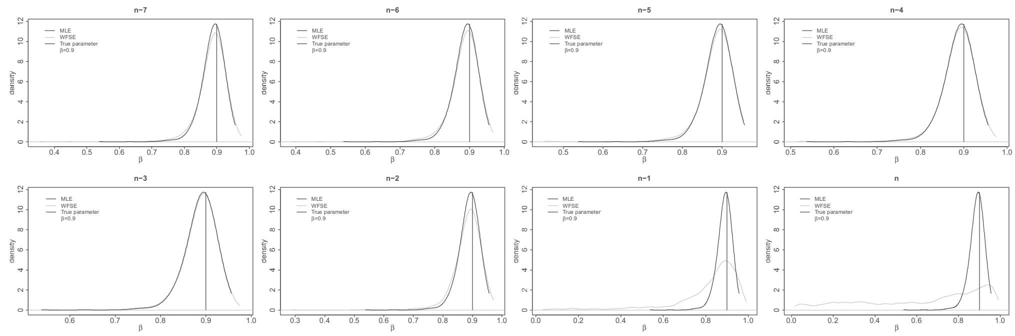

Figure 8: Density of β ML estimates on uncontaminated series (black) vs WFS estimates on the same series contaminated by three outliers (grey), last 8 steps of the FS. The WFS estimates are obtained by cutting the sequence of estimates at $7, 6, \ldots, 1, 0$ steps before the end of the search for all the series.

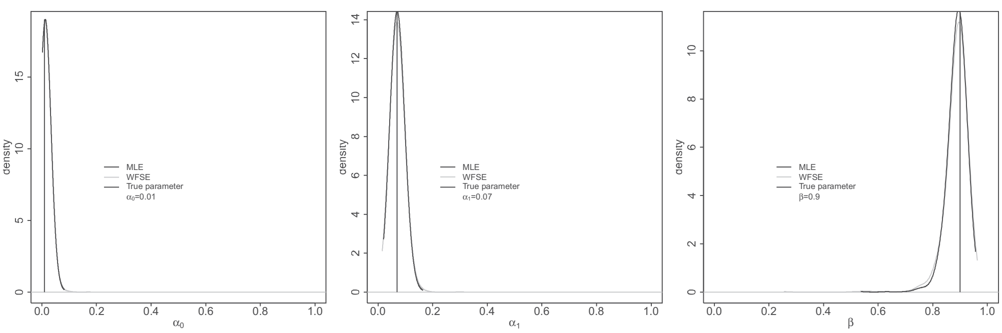

Figure 9: ML estimates on uncontaminated series (black) vs WFS estimates on the same series contaminated by three outliers.

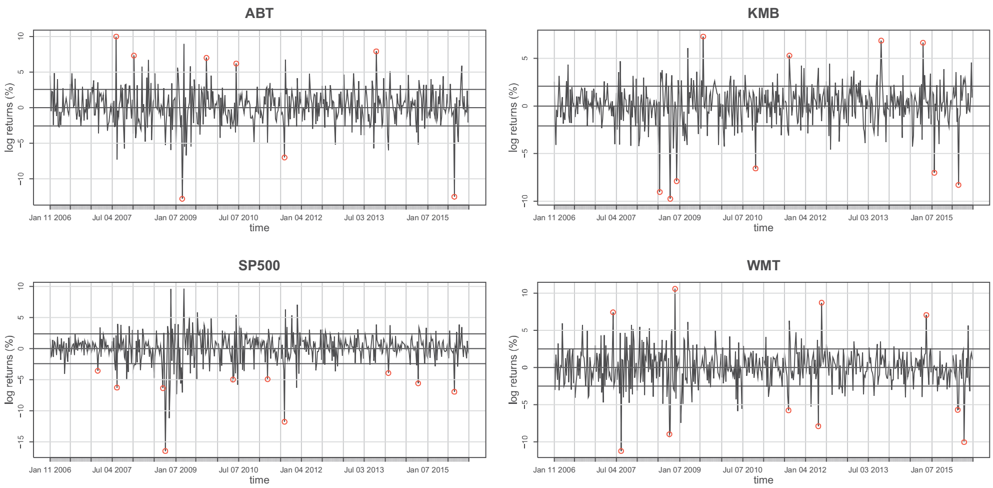

Figure 10: Log-returns of weekly prices collected on four companies quoted on the NYSE. Sample period is 05/01/2011 - 31/12/2015. Outliers detected by the WFS test are identified by red circles. The corresponding dates are reported in Table 4, fourth column.

e

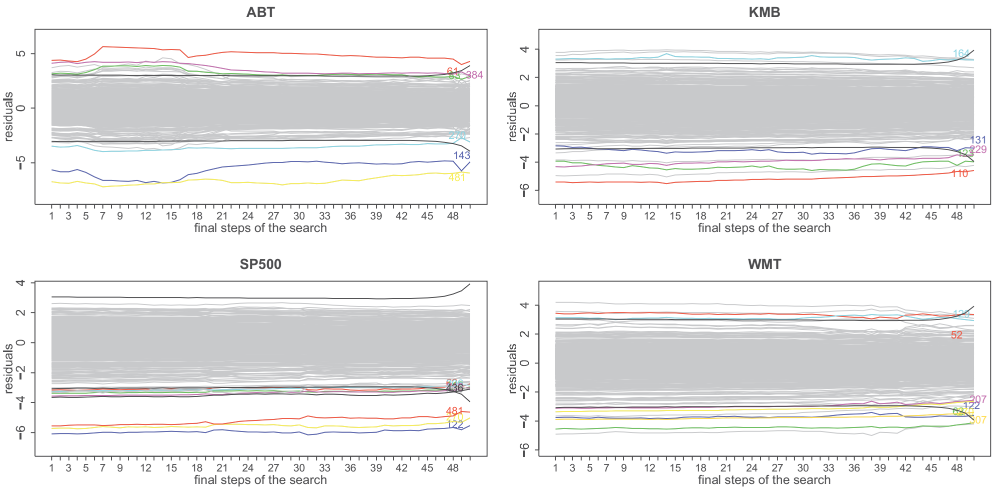

Figure 11: Trajectories of standardized residuals along the final 10% steps of the FS for selected companies. Dashed bold lines represent the 95% confidence regions defined by simulated envelopes. Outlying trajectories are identified by colored bold lines. Unit indexes are reported on the right.

α_0	α_1	β
0.01	0.02	0.95
0.01	0.07	0.9
0.01	0.12	0.85
0.01	0.17	0.8
0.01	0.27	0.7
0.01	0.37	0.6
0.01	0.47	0.5
0.01	0.57	0.4
0.01	0.67	0.3
0.01	0.77	0.2
0.01	0.87	0.1

Table 1: Combinations of parameters used to generate trajectories from GARCH(1,1) models ($\alpha_1 + \beta = 0.97$).

sample size	T	WFS ($\alpha = 1\%$)	WFS ($\alpha = 5\%$) $\theta = (0.01, 0.07, 0.9)$	WFS ($\alpha = 10\%$)	G&V	B&H	D&O $\theta = (0.0126, 0.0757, 0.9152)$	F&G
Percentage* of correct detection of additive outliers								
1 outlier of size 5sd	*250*	92.2	96.9	97.9				
	500	90.8	96.8	98.1	66.3	93.4	80.2	96.7
	1000	88.0	95.7	97.4	66	90.2	76.1	98
1 outlier of size 10sd	*250*	100	100	100				
	500	100	100.0	100	98.4	100	86.7	100
	1000	99.9	100	100	98.9	99.9	92	100
1 outlier of size 15sd	*250*	100	100	100				
	500	100	100.0	100	99.0	100	91.3	100
	1000	100	100	100	99.7	100	93.7	100
3 outliers	*250*	93.1	96.6	97.5				
	500	94.0	96.9	97.7	63.3	91.8	68.5	90.9
	1000	94.2	96.9	98.0	71.4	92	80.8	92.4
10 outliers	*250*	88.7	93.6	95.0				
	500	90.9	94.5	95.9				
	1000	93.9	96.7	97.6				
Average number of false additive outliers (standard deviation in parenthesis)								
1 outlier of size 5sd	*250*	0.32 (0.63)	2.13 (1.85)	3.99 (1.85)				
	500	0.29 (0.57)	2.57 (1.76)	5.55 (1.76)	0.02 (0.15)	1.96 (7.99)	1.11 (0.54)	0.63 (1.36)
	1000	0.40 (0.63)	3.87 (2.16)	8.88 (2.16)	0.05 (0.22)	1.91 (1.58)	1.03 (0.20)	1.23 (1.73)
1 outlier of size 10sd	*250*	0.30 (0.58)	1.95 (1.67)	3.66 (1.67)				
	500	0.43 (0.76)	3.19 (2.22)	6.53 (2.22)	0.03 (0.20)	3.85 (19.25)	1.21 (0.92)	3.44 (7.75)
	1000	0.42 (0.68)	3.91 (2.13)	8.79 (2.13)	0.03 (0.16)	2.21 (1.91)	1.05 (0.47)	1.67 (3.79)
1 outlier of size 15sd	*250*	0.29 (0.63)	1.86 (1.64)	3.56 (1.64)				
	500	0.40 (0.70)	3.08 (2.23)	6.37 (2.23)	0.04 (0.20)	5.07 (22.16)	1.38 (2.08)	10.21 (12.84)
	1000	0.38 (0.63)	3.76 (2.13)	8.54 (2.13)	0.04 (0.21)	4.35 (27.28)	1.11 (0.76)	6.11 (10.15)
3 outliers^{+}	*250*	0.19 (0.52)	1.43 (1.50)	2.73 (1.50)				
	500	0.35 (0.68)	2.70 (2.06)	5.63 (2.06)	0.03 (0.19)	5.00 (8.84)	1.22 (1.04)	7.46 (11.50)
	1000	0.33 (0.60)	3.36 (2.00)	7.81 (2.00)	0.04 (0.24)	7.34 (41.28)	1.11 (0.88)	6.23 (10.06)
10 outliers^{++}	*250*	0.06 (0.25)	0.44 (0.78)	0.94 (0.78)				
	500	0.17 (0.46)	1.44 (1.54)	3.12 (1.54)				
	1000	0.21 (0.48)	2.30 (1.62)	5.55 (1.62)				

* Over 1000 replications of size T for GARCH(1,1) with errors following a normal distribution
($^{+}$) 1 outlier of size 5σ, 1 outlier of size 10σ, 1 outlier of size 15σ
($^{++}$) outliers of size 5σ, 4 outliers of size 10σ and 4 outliers of size 15σ

Table 2: Comparison of the WFS test with outlier tests by Grané and Veiga (2010) - G&V, Bilen and Huzurbazar (2002) - B&H, Doornik and Ooms (2005) - D&O and Franses and Ghijsels (1999) - F&G.

Ticker	Company	Description	Industry classification (SIC)
ABT	Abbott Laboratories	It discovers, develops, manufactures and sells health care products	Surgical and Medical Instruments and Apparatus
APA	Apache Corp.	It is an independent energy company that explores, develops and produces natural gas, crude oil and natural gas liquids	Crude Petroleum and Natural Gas
BAC	Bank of America	Through its subsidiaries, it provides banking and non-banking financial services and products throughout the United States and in selected international markets	National Commercial Banks
HPQ	HP	It is a provider of products, technologies, software, solutions and services to individual consumers, small- and medium-sized businesses including customers in the government, health and education sectors.	Electronic Computers
KMB	Kimberly-Clark	It is engaged in the manufacturing and marketing of products made from natural or synthetic fibers using technologies in fibers, nonwovens and absorbency.	Sanitary Paper Products
MCD	Mc Donald's	It franchises and operates McDonald's restaurants in the food service industry.	Restaurants, Licensed
SYY	Sysco Corp.	Through its subsidiaries and divisions, it is engaged in the distribution of food and related products to the foodservice or food-away-from-home industry.	Other Foods, Wholesale
UNP	Union Pacific Corp	It is a rail transporting company. Its operating company is Union Pacific Railroad Company. It links 23 states in the western two-thirds of the country by rail.	Railroad Rolling Stock Industry
WMT	Wal-Mart Stores Inc.	It operates retail stores in various formats under various banners. Its operations comprise of three reportable business segments, Walmart U.S., Walmart International and Sam's Club	Other Retail Stores
SP500	S&P500		

Table 3: List of the analyzed time series. Nine companies of the NYSE and the main index (S&P500) have been selected. The industry representation criteria has been applied to cover most of the main industries of the U.S. market. Description and SIC classification have been obtained from the NYSE website (https://www.nyse.com/index)

Ticker	Company	Obs. position	Date	MLE α_0, α_1, β	WFSE α_0, α_1, β
ABT	Abbott Laboratories	83	08/08/2007	1.1 0.06 0.77	0.3 0.04 0.9
		105	09/01/2008		
		165	04/03/2009		
		195	30/09/2009		
		232	16/06/2010		
		292	10/08/2011		
		406	16/10/2013		
		503	26/08/2015		
KMB	Kimberly - Clark	132	16/07/2008	0.14 0.03 0.94	0.13 0.02 0.94
		145	15/10/2008		
		153	10/12/2008		
		186	29/07/2009		
		251	27/10/2010		
		293	17/08/2011		
		407	23/10/2013		
		459	22/10/2014		
		473	28/01/2015		
		503	26/08/2015		
WMT	Walmart	74	06/06/2007	0.44 0.09 0.84	0.05 0.02 0.97
		84	15/08/2007		
		144	08/10/2008		
		151	26/11/2008		
		292	10/08/2011		
		329	25/04/2012		
		333	23/05/2012		
		463	19/11/2014		
		502	19/08/2015		
		510	14/10/2015		
SP500	S&P500 index	60	28/02/2007	0.52 0.24 0.68	0.19 0.13 0.82
		84	15/08/2007		
		141	17/09/2008		
		144	08/10/2008		
		228	19/05/2010		
		271	16/03/2011		
		292	10/08/2011		
		421	29/01/2014		
		458	15/10/2014		
		503	26/08/2015		

Table 4: Influential observations detected in each series of prices of companies quoted on the NYSE. Unit index, date, MLE and WFS estimates of GARCH coefficients are reported.

Ticker	Company	TstatsB Mean			TstatsO Mean			TstatsB Median			TstatsO Median		
		α_0	α_1	β	α_0	α_1	β	α_0	α_1	β	α_0	α_1	β
ABT	Abbott Laboratories	1.62	2.29	14.12	1.64	1.93	5.75	1.86	2.49	16.38	1.46	1.77	8.96
APA	Apache Corp.	1.78	3.17	22.68	1.91	3.36	23.09	1.88	3.19	23.08	1.87	3.4	24.96
BAC	Bank of America	1.97	6.49	40.95	2.16	5.89	33.41	1.99	6.55	41.28	2.31	5.91	34.07
HPQ	HP	1.7	2.9	20.12	2.22	3.07	16.09	1.67	3.02	20.76	2.15	2.97	15.01
KMB	Kimberly - Clark	1.23	1.7	27.13	1.13	1.56	18.73	1.23	1.69	28.56	1.35	1.74	21.15
MCD	Mc Donald's	1.27	2.28	22.37	0.97	1.53	9.94	1.05	2.19	35.95	0.9	1.6	8.63
SYY	Sysco Corp.	1.22	2.49	35.76	1.47	2.24	25.12	1.17	2.4	35.58	1.38	2.1	23.05
UNP	Union Pacific Corp.	1.56	3.31	37.69	1.76	3.39	29.33	1.61	3.51	38.6	1.77	3.36	28.03
WMT	Walmart	1.29	2.54	34.14	1.52	2.24	17.27	1.35	2.67	32.62	1.26	2.15	15.75
SP500	S&P500 index	2.77	3.97	14.48	2.74	3.88	14.28	2.75	4.06	14.00	2.51	3.77	14.33

Table 5: Comparison of tstatsB and tstatsO t-statistics computed in the last steps of the search. the two types of t-statistics are based on different sets of observations. tstatsB are obtained considering observations Before the step which separates influential observations from the remaining units. tstatsO are based on the Extreme Observations identified by the WFS test. In the columns labeled as "Mean" the numerator of the ratio is given by the arithmetic mean of the estimates, the denominator is given by the arithmetic mean of the standard errors. In the columns labeled as "Median" the arithmetic mean is replaced by the median.

References

Atkinson A. and Riani M. (2000) *Robust diagnostic regression analysis*, Springer, New York.

Atkinson A.C. (1994) Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, 89, 428, 1329–1339.

Bali R. and Guirguis H. (2007) Extreme observations and non-normality in ARCH and GARCH, *International Review of Economics and Finance*, 16, 3, 332 – 346.

Bilen C. and Huzurbazar S. (2002) Wavelet-based detection of outliers in time series, *Journal of Computational and Graphical Statistics*, 11, 2, 311–327.

Bollerslev T. (1986) Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, 31, 3, 307–327.

Bollerslev T. (1987) A conditionally heteroskedastic time series model for speculative prices and rates of return, *The Review of Economics and Statistics*, 69, 3, 542–547.

Bondon P. and Bahamonde N. (2012) Least squares estimation of ARCH models with missing observations, *Journal of Time Series Analysis*, 33, 6, 880–891.

Boudt K. and Croux C. (2010) Robust m-estimation of multivariate GARCH models, *Computational Statistics and Data Analysis*, 54, 11, 2459 – 2469, the Fifth Special Issue on Computational Econometrics.

Carnero M., Peña D. and Ruiz E. (2007) Effects of outliers on the identification and estimation of GARCH models, *Journal of Time Series Analysis*, 28, 4, 471–497.

Catalan B. and Trivez F.J. (2007) Forecasting volatility in GARCH models with additive outliers, *Quantitative Finance*, 7, 6, 591–596.

Cerioli A., Farcomeni A. and Riani M. (2014) Strong consistency and robustness of the forward search estimator of multivariate location and scatter, *Journal of Multivariate Analysis*, 126, 167 – 183.

Charles A. (2008) Forecasting volatility with outliers in GARCH models, *Journal of Forecasting*, 27, 7, 551–565.

Charles A. and Darnè O. (2005) Outliers and GARCH models in financial data, *Economics Letters*, 86, 3, 347 – 352.

Chen C. and Liu L.M. (1993a) Forecasting time series with outliers, *Journal of Forecasting*, 12, 1, 13–35.

Chen C. and Liu L.M. (1993b) Joint estimation of model parameters and outlier effects in time series, *Journal of the American Statistical Association*, 88, 421, 284–297.

Doornik J.A. and Ooms M. (2005) Outlier Detection in GARCH Models, Economics Papers 2005-W24, Economics Group, Nuffield College, University of Oxford.

Engle R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, 50, 4, 987–1007.

Franses P.H. and Ghijsels H. (1999) Additive outliers, GARCH and forecasting volatility, *International Journal of Forecasting*, 15, 1, 1–9.

Franses P.H. and Van Dijk D. (2000) *Non-linear time series models in empirical finance*, Cambridge University Press, Cambridge.

Galeano P. and Tsay R.S. (2010) Shifts in individual parameters of a GARCH model, *Journal of Financial Econometrics*, 8, 1, 122–153.

Gómez V., Maraval A. and Peña D. (1999) Missing observations in ARIMA models: Skipping approach versus additive outlier approach, *Journal of Econometrics*, 88, 2, 341 – 363.

Grané A. and Veiga H. (2010) Wavelet-based detection of outliers in financial time series, *Computational Statistics & Data Analysis*, 54, 11, 2580–2593.

Grané A. and Veiga H. (2014) Outliers, GARCH-type models and risk measures: A comparison of several approaches, *Journal of Empirical Finance*, 26, 26–40.

Grossi L. (2004) Analyzing financial time series through robust estimators, *Studies in Nonlinear Dynamics & Econometrics*, 8, 2, article 3.

Grossi L. and Laurini F. (2009) A robust forward weighted Lagrange multiplier test for conditional heteroscedasticity, *Computational Statistics & Data Analysis*, 53, 6, 2251–2263.

Heagerty P.J. and Lumley T. (2000) Window subsampling of estimating functions with application to regression models, *Journal of the American Statistical Association*, 95, 449, 197–211.

Hill J.B. (2015) Robust estimation and inference for heavy tailed GARCH, *Bernoulli*, 21, 3, 1629–1669.

Hill J.B. and Prokhorov A. (2016) GEL estimation for heavy-tailed GARCH models with robust empirical likelihood inference, *Journal of Econometrics*, 190, 1, 18 – 45.

Hotta L.K. and Tsay R. (2012) Outliers in GARCH processes, in: *Economic time series: Modeling and Seasonality*, Bell W.R. H.S. and McElroy T., eds., Chapman and Hall, 337–358.

Hung J.C. (2014) Robust Kalman filter based on a fuzzy GARCH model to forecast volatility using particle swarm optimization, *Soft Computing*, 19, 10, 2861–2869.

Hwang S. and Pereira P.L.V. (2006) Small sample properties of GARCH estimates and persistence, *The European Journal of Finance*, 12, 6-7, 473–494.

Johansen S. and Nielsen B. (2016) Asymptotic theory of outlier detection algorithms for linear time series regression models, *Scandinavian Journal of Statistics*, 43, 2, 321–348.

Jones R.H. (1980) Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics*, 22, 3, 389–395.

Kharin Y.S. and Voloshko V.A. (2011) Robust estimation of {AR} coefficients under simultaneously influencing outliers and missing values, *Journal of Statistical Planning and Inference*, 141, 9, 3276 – 3288.

Laurent S., Lecourt C. and Palm F.C. (2014) Testing for jumps in conditionally Gaussian ARMA-GARCH models, a robust approach, *Computational Statistics & Data Analysis*, In press, DOI: 10.1016/j.csda.2014.05.015.

Mendes B.V.D.M. (2000) Assessing the bias of maximum likelihood estimates of contaminated GARCH models, *Journal of Statistical Computation and Simulation*, 67, 4, 359–376.

Muler N. and Yohai V.J. (2008) Robust estimates for GARCH models, *Journal of Statistical Planning and Inference*, 138, 10, 2918 – 2940.

Ossandón S. and Bahamonde N. (2011) On the nonlinear estimation of GARCH models using an extended Kalman filter, in: *Lecture Notes in Engineering and Computer Science. Proceedings of The World Congress on Engineering*, WCE 2011, 6-8 July, 2011, London, U.K., 148–151.

Pelagatti M., Lisi F. *et al.* (2009) Variance initialisation in GARCH estimation, in: *Complex data modeling and computationally intensive statistical methods for estimation and prediction*, S.Co 2009, 14-16 September, 2009, Milan, Italy, Maggioli Editore, 323–328.

Penzer J. (2007) Inference for GARCH and related models with incomplete data, Working Paper, LSE.

Penzer J. and Shea B. (1999) Finite sample prediction and interpolation for ARIMA models with missing data, *Journal of Forecasting*, 18, 6, 411–419.

Proietti T. (2008) Missing data in time series: A note on the equivalence of the dummy variable and the skipping approaches, *Statistics & Probability Letters*, 78, 3, 257–264.

Riani M. (2004) Extensions of the forward search to time series, *Studies in Nonlinear Dynamics & Econometrics*, 8, 2, article 2.

Riani M., Perrotta D. and Cerioli A. (2015) The forward search for very large datasets, *Journal of Statistical Software*, 67, 1, 1–20.

Rousseeuw P. and Leroy A. (1987) *Robust regression and outlier detection*, Wiley, New York.

Sakata S. and White H. (1998) High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility, *Econometrica*, 66, 3, 529–567.

Van Dijk D., Franses P.H. and Lucas A. (1999) Testing for ARCH in the presence of additive outliers, *Journal of Applied Econometrics*, 14, 5, 539–562.