Robust Forecasting of Value-at-Risk and Expected Shortfall with Markov-Switching GARCH Models

Jean S. M. Diniz and Luiz K. Hotta

University of Campinas

Abstract: The study of volatility is important in several fields of the financial area and GARCH models have been frequently used in the literature, due to their ability to capture some of the stylised facts of financial time series. However, in some financial time series there is a change of structure in the volatility dynamics, in which cases, GARCH models tend to be inadequate, while Markov-switching GARCH models can be used. On the other hand, outliers are often present in empirical data. This paper shows that the usual estimators can be strongly affected by outliers and proposes an estimator that is more robust to outliers. Simulations show that the proposed robust estimator outperforms the traditional estimators in forecasting the volatility and VaR.

Key Words: Robust statistical methods, Volatility models, Hidden Markov models, Outliers, Value at risk.

1 Introduction

Volatility is one of the most intensely studied aspects in financial econometrics in recent decades. From an empirical point of view, several authors have observed that financial time series present some stylised facts (Francq and Zakoian, 2019, Section 1.3, pg. 7). In this scenario, Engle (1982) proposed a conditionally heteroscedastic autoregressive (ARCH) model, but like in several series, when estimating ARCH models, high orders were observed , Bollerslev (1986) presented an extension of the ARCH model, the Generalised ARCH (GARCH) model.

The most popular GARCH estimator is the quasi-maximum likelihood (QML). Francq and Zakoian (2019) (Section 7, pg. 175) showed that, under certain conditions, these estimators have excellent asymptotic properties, such as consistency. However, this estimator is not robust to outliers.

Two approaches are frequently used to deal with outliers: the first is to detect outliers and treat them in some appropriate way; and the second is to use estimators robust to outliers. Hotta and Tsay (2012), based on ARCH(p) and GARCH(1,1) models, studied the impact of outliers on maximum likelihood estimation as well as proposing a method for detecting outliers. Several authors have presented robust methods to estimate GARCH models. For example, Sakata and White (1998) used S-estimators; Muler and Yohai (2008) presented an estimator based on M-estimators combined with a robust filter mechanism to reduce the effect of outliers on conditional variances, and Boudt et al. (2013) proposed a modification of the objective function of the M-Estimator and the mechanism used by Muler and Yohai (2008). In the presence of outliers these estimators have a significant gain over the non-robust estimators, as shown by Carnero et al. (2012) for the estimator of Muler and Yohai (2008). Additionally Trucíos et al. (2017) presented a modification for robust prediction of volatility in GARCH models.

On the other hand, financial time series often have structural breaks over time, usually associated with political events and/ or financial crises, such as the subprime crisis in 2008-9. Thus, especially in long series, it is not suitable to assume that the same parameters of the GARCH model explain the series for the whole period. In the presence of structural breaks, such models become inappropriate, Hillebrand (2005) pointed out that this adverse effect is independent of the estimation method. For instance, Hwang and Valls Pereira (2008) showed that the estimated persistence of conditional volatility in large samples can be excessive, especially in the presence of structural breaks in the ARCH and GARCH parameters.

Hamilton (1989) and Cai (1994) proposed the conditional heteroscedastic autoregressive model with Markov-switching (MS-ARCH) regime. By allowing switching between different regimes, this model is able to capture more complex dynamic patterns. Because the regimes are not observable, in the maximum likelihood estimation of the model, one needs to infer the probability of being in a given regime in time. For this, the mechanism used is the Hamilton (1989) filter.

When extending this method to GARCH Markov-switching (MS-GARCH) models, it was observed that the maximum-likelihood estimation was unfeasible because it depended on all possible path combinations. This problem can be circumvented via Bayesian estimation, as shown by Das et al. (2004) and Bauwens et al. (2010). In the traditionl approach, Gray (1996) was the first author who dealt with the path dependency problem making maximum likelihood estimation possible. Later, Haas et al. (2004) presented a new approach to MS-GARCH models, with a new interpretation of the model presented by Gray (1996).

MS-GARCH models have become popular in the literature because in many cases they have a better performance than single-regime GARCH models. For instance, Ardia et al. (2018) found that MS-GARCH models provide better predictions of VaR and expected shortfall compared to GARCH models with a single-regime in several cases. Furthermore, they observed that the improvements were more pronounced when the Markov switching mechanism was applied to simple specifications, such as the GARCH-Normal model.

However, like other usual filters, Hamilton (1989)'s filter performs poorly in the presence of outliers as shown by Calvet et al. (2015). Thus, the objective of this work is to present robust alternatives to maximum-likelihood estimators of MS-GARCH models. To do this, we suggest some modifications, for instance the use of robust filters presented by Petrus (1999) and Calvet et al. (2015), always having in mind the robustness objectives, (Huber and Ronchetti, 2009, Section 1.2, pg. 5), namely that the estimator must have a reasonably good efficiency in the assumed model; it should be robust in the sense that small deviations from the model's assumptions hurt performance only marginally; and slightly larger deviations from the model do not cause a catastrophe. In this sense, other modifications in the estimation methods of MS-GARCH models are suggested.

As far as we know, there is no work in the literature dealing with the robust estimation MS-GARCH models. The goal of the paper is to show the effect of outliers on the estimation of MS-GARCH models and then to propose a robust alternative. In Section 2, the GARCH model and non-robust and robust estimation methods are presented, in particular the robust methods of Muler and Yohai (2008) and Boudt et al. (2013). In Section 3, MS-GARCH models are presented along with their properties, interpretation and advantages in relation to GARCH models in certain scenarios. Among the different specifications of MS-GARCH models, we present the model of Haas et al. (2004). In addition, we describe an alternative robust estimation that consists of adapting the techniques used by Boudt et al. (2013) to the case of regime-switching models. Section 4 presents a Monte Carlo simulation study to evaluate the performance of the estimators of the MS-GARCH models. We evaluate both the quasi-likelihood estimation through Hamilton (1989)'s filter and robust alternatives. In this section we focus on evaluating the accuracy of the parameter estimates and the ability of the filter to identify the regimes. We observe that, in general, the QML estimators using the t-Student (QML-t) and the robust Hamilton filter perform reasonably well in identifying regimes, unlike the QML using the normal distribution (QML-n). On the other hand, when estimating the conditional variance parameters, we note that the estimators tend to estimate some parameters well, but not others. Finally, in Section 5 the main conclusions found in the simulations and applications in relation to the considered estimators are discussed.

2 GARCH models

According to Bollerslev (1986), y_t follows a GARCH(p,q) model if:

$$\begin{aligned} y_t &= \sqrt{h_t} z_t, \\ h_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \end{aligned} \tag{1}$$

where $\{z_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with mean 0 and variance 1; and $\alpha_0 > 0$, $\alpha_i \geq 0$, and $\beta_j \geq 0$ are the sufficient conditions for h_t to be positive. Stationary conditions can be found, for instance, in Francq and Zakoian (2019) (Section 2.2.2) and in the Supplementary Materials.

2.1 Estimators

Various methods exist to estimate GARCH models. For instance, Baillie and Chung (2001) estimated the GARCH(1,1) model using a minimum distance estimator for the squared process autocorrelations and Pérez-Cruz et al. (2003) used the support vector machine method. Aknouche and Guerbyenne (2006) develop three recursive online algorithms based on a two-stage least squares scheme, while (Ardia, 2008, Section 3) presented methods for Bayesian estimation. From a frequentist approach, the most common estimation is by QML.

2.1.1 Quasi-maximum likelihood estimator

Denote $\boldsymbol{y}_T = \{y_1, y_2, \ldots, y_T\}$, $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, with $\boldsymbol{\alpha} = (\alpha_0.\alpha_1, \ldots, \alpha_p)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, so that it belongs to the parametric space of the form $\Theta \subset (0. + \infty) \times [0.\infty)^{p+q}$ - often the space in which the GARCH model is stationary. To obtain the likelihood of the GARCH(p,q) model defined in (1), it is necessary to specify the distribution of $\{z_t\}$. The name quasi-likelihood comes from the fact that one distribution is used, even when knowing that this distribution is not correct. For the normal case, given the initial values, for example, $\boldsymbol{\phi}_0 = \{y_0.\ldots, y_{1-p},\ h_0, ..., h_{1-q}\}$, the quasi-likelihood can be expressed as:

$$L(\boldsymbol{\theta}|\boldsymbol{y}_T, \boldsymbol{\phi}_0) = \prod_{t=1}^{T} f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}, \boldsymbol{\phi}_0) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi h_t}} \exp\left\{-\frac{y_t^2}{2h_t}\right\}, \tag{2}$$

where h_t can be obtained recursively through (1). Thus, the QML estimate $\hat{\boldsymbol{\theta}}$, when we assume the standard normal distribution (QML-n), is given by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\boldsymbol{y}_T, \boldsymbol{\phi}_0). \tag{3}$$

Or by maximising the log-likelihood:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y_T}, \boldsymbol{\phi}_0) = \log(L(\boldsymbol{\theta}|\boldsymbol{y}_T, \boldsymbol{\phi}_0)) \propto -\frac{1}{2}\sum_{t=1}^{T}\left[\ln(h_t) + \frac{y_t^2}{h_t}\right]. \tag{4}$$

Under certain conditions, the QML estimator is consistent and the distribution converges (Francq and Zakoian, 2019, Seção 7.1.1), i.e.,

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N\left(0, (\kappa_z - 1)J^{-1}\right), \quad \text{where}$$
$$J := E\left(\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{y_t}, \boldsymbol{\phi}_0)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}\right).$$

The choice of initial values in finite samples can produce different values of the estimates, especially when the sample size is small. Pelagatti and Lisi

(2009) showed the accuracy of the QML-n estimates for the GARCH(1,1) model under different initial value choices.

2.2 Robust estimation of GARCH models

There are numerous methods to robustify estimators of GARCH models, focusing mainly on additive outliers. We highlight Basu et al. (1998); Sakata and White (1998); Park (2002); Lee and Song (2009); Hotta and Trucíos (2018) and Crosato and Grossi (2019). However, Muler and Yohai (2008) pointed out that the estimates of the robust methods cited above may not be robust in the presence of a fraction of outliers in the sample, since they are based on conditional variance, which is very sensitive to outliers. Therefore, they presented a modification of the M-Estimators and a modified specification of the conditional variance of the GARCH model to decrease the effect of outliers.

2.2.1 BIP-GARCH model

Muler and Yohai (2008) proposed a mechanism similar to robust filtering to estimate the conditional variance. This mechanism, called BIP (bounded innovation propagation), restricts the propagation of the outliers' effect present at time $t-1$ in the conditional variance of time t. For the GARCH(p,q) model h_t is calculated as:

$$\begin{aligned} h_t^* &= \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{t-i}^* r_k\left(\frac{y_{t-i}^2}{h_{t-i}^*}\right) + \sum_{i=1}^{q} \beta_j h_{t-i}^*, \\ r_k(u) &= \begin{cases} u, & \text{if } u \le k \\ k^*, & \text{if } u > k, \end{cases} \end{aligned} \tag{5}$$

with $k^* = k$. However, Boudt et al. (2013) observed in their simulations that when using the BIP specification in (5), the estimates of the parameters of the GARCH model showed small biases. So they introduced a correction factor c_{k_δ}, to guarantee that the expectation of $c_{k_\delta} h_{t-1}^{(k)} r_{k_\delta}(y_{t-1}^2/h_{t-1}^{(k)})$, conditional on $\mathcal{F}_{t-2}$, is equal to the conditional expectation of y_{t-1}^2, when there is no outlier. $\mathcal{F}_{t-2}$ is the set of all available information up to time $t-2$. Equation (5) becomes

$$\begin{aligned} h_t^* &= \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{t-i}^* c_{k_\delta} r_{k_\delta}\left(\frac{y_{t-i}^2}{h_{t-i}^*}\right) + \sum_{i=1}^{q} \beta_j h_{t-i}^*, \\ r_{k_\delta}(u) &= \begin{cases} u, & \text{if } u \le k_\delta \\ k_\delta^* & \text{if } u > k_\delta. \end{cases} \\ c_{k_\delta} &= \frac{E[W]}{E[r_{k_\delta}(W)W]} = \frac{1}{F_{\chi_3^2}(k_\delta) + (1-\delta)k_\delta}, \end{aligned} \tag{6}$$

where $k_\delta^* = k_\delta$, and W is a chi-square random variable with one degree of freedom, $F_{\chi_3^2}(.)$ is the cumulative distribution function of the chi-square distribution with 3 degrees of freedom and k_δ is the δ quantile of the chi-square distribution with one degree of freedom. The notation k_δ was introduced to clarify the meaning of the threshold, from which one suspects that the observed value is affected by an outlier. Table 1 of the article by Boudt et al. (2013) presents the values of c_{k_δ}.

Other authors have proposed different values of k^* and k_δ^* in (5) and (6), respectively. Carnero et al. (2012) proposed to use $k^* = 1$. i.e., when there is an atypical observation, it is replaced by $E[y_{t-1}^2] = h_{t-1}$. For cases where another value for k_δ is used in (6), when $u > k_\delta$, say k_{δ^*}, we have that the value of c_{k_δ} is given by $c_{k_\delta} = [F_{\chi_3^2}(k_\delta) + k_{\delta^*}(1 - F_W(k_\delta))]^{-1}$. For the case of Carnero et al. (2012), where it is replaced by 1, we have that $\delta^* = 0.6827$.

The constant δ works as a trade-off between robustness and efficiency, if the chosen value is too high h_t^*, tends to coincide with h_t, while for small values h_t^* becomes more robust to outliers.

Boudt et al. (2013) proposed to use variance target, that is, α_0 is reparameterized as $\hat{h}(1 - \alpha_1 - \beta_1)$ - with $\hat{h}$ being a consistent estimator of the unconditional variance h is estimated before the other parameters. Without outliers $\hat{h}$ can be obtained simply by the sample variance, but this estimator is greatly affected by the presence of outliers. A robust estimator for the unconditional variance is the reweighted variance estimator proposed by Boudt et al. (2013). The reweighted variance estimator, when $z_t \sim N(0.1)$, is given by:

$$
\begin{aligned}
\hat{h} &= 1.318 \frac{\sum_{i=1}^{T} y_t J_t}{\sum_{i=1}^{T} J_t} \\
J_t &= \mathbb{I}\left[\frac{y_t^2}{\mathrm{MAD}_t^2} \leq \chi_1^2(95\%)\right],
\end{aligned}
$$

such that $\mathbb{I}(.)$ is the indicator function and the MAD of a sequence of observations $y_1, \ldots, y_T$ is defined as $1.486 \times \mathrm{median}_i(|y_i - \mathrm{median}_j(y_j)|)$, where 1.486 is a correction factor to ensure that the MAD is a consistent scale estimator of the normal distribution.

Finally, Boudt et al. (2013) estimated the parameters of the BIP-GARCH model using an estimator that minimises an objective function ρ evaluated for the logarithm of the squared standardised observations. This is,

$$
\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{T} \sum_{t=1}^{T} \rho\left(\ln \frac{y_t^2}{h_t}\right). \tag{7}
$$

Based on the comparison of several ρ candidate functions, Boudt et al. (2013) recommended one associated with the t-Student density with 4 degrees of freedom. For the most general case, associated with the t-Student density with ν degrees of freedom we have

$$
\rho(z) = -z + \sigma_\nu \rho_{t_\nu}(\exp(z)), \tag{8}
$$

$$
\rho_{t_\nu}(u) = (1 + \nu) \log\left(1 + \frac{u}{\nu - 2}\right), \tag{9}
$$

$$\sigma_\nu = \frac{1}{E[\rho'_{t_\nu}(W)W]}. \tag{10}$$

2.3 Volatility prediction

For the sake of simplicity and notation, here we will consider the GARCH(1,1) model, given by (1), to illustrate how volatility forecast τ steps forward is given. Suppose the model has been estimated over the time period $t = 1, 2, \ldots, T$. The τ steps ahead volatility forecast at T, given the information up to this time, denoted by $h_{T+\tau|T}$, with least mean squared error, is given by

$$h_{T+\tau|T} = \alpha_0 + \alpha_1 y^2_{T+\tau-1|T} + \beta_1 h_{T+\tau-1|T} \tag{11}$$

where, similarly, $y^2_{T+\tau|T}$ is the forecast of y_t^2 τ-steps ahead, conditional on $\mathcal{F}_T$; that is, $y^2_{T+\tau|T} \equiv E[y^2_{T+\tau}|\mathcal{F}_T]$. $\boldsymbol{\theta} = (\alpha_0.\alpha_1, \beta_1)'$ can be estimated by (3). As with parameter estimation, the presence of outliers affects the volatility forecast. To deal with this, an alternative to the forecast τ steps ahead of the conditional variance, as presented by Boudt et al. (2013), is given by:

$$h_{T+\tau|T} = \alpha_0 + \alpha_1 c_{k_\delta} h_{T+\tau-1|T} r_{k_\delta}(y^2_{T+\tau-1|T}/h_{T+\tau-1|T}) + \beta_1 h_{T+\tau-1|T}, \tag{12}$$

where $r_{k_\delta}(.)$ is the function defined in (6). And, as mentioned by Boudt et al. (2013), in the presence of outliers, y_t^2 is not observed, and is replaced by h_t when using the r_{k_δ} mechanism. Thus, the effect of these outliers on future variations is unlimited in (11), but limited in (12). Additionally, Carnero et al. (2012) observed that, in the presence of outliers, non-robust estimators of the parameters of the GARCH model generally lead to biases in the estimated of volatilities.

3 Markov-switching GARCH models

Lamoureux and Lastrapes (1990), Horváth et al. (2006) and other authors have provided empirical evidence of structural changes of stock returns. In this scenario, as cited by Gray (1996), many popular models produce poor results when fitted to the data. Lamoureux and Lastrapes (1990), Hillebrand (2005), Hwang and Valls Pereira (2008) and other authors have shown that in the presence of structural breaks in the parameters of the ARCH and GARCH processes, when adjusting the GARCH model, the estimated persistence, even in large samples, may be excessive. Furthermore, Hillebrand (2005) stated that this negative effect is independent of the estimation method of the GARCH model. Additionally, some authors have observed this fact empirically, such as Mikosch and Stărică (2004), who estimated a GARCH model in a sample that exhibits structural changes in its conditional variance and obtained persistence practically equal to that of Billio et al. (2016).

To deal with this situation, Cai (1994) and Hamilton and Susmel (1994) proposed MS-ARCH models. The idea of the MS approach to modelling heteroscedasticity is that the distribution of the ARCH process depends on an unobserved state. The MS-ARCH model presented by Cai (1994), for example, is given by:

$$
\begin{aligned}
y_t &= \sqrt{h_t} z_t, \\
h_t &= \alpha_0^{(\Delta_t)} + \sum_{i=1}^{p} \alpha_i y_{t-i}^2,
\end{aligned}
$$

where Δ_t is a homogeneous ergodic Markov chain on a finite set $S = \{1, \ldots, K\}$.

However, the extension of this specification to MS-GARCH was not feasible. Due to the dependence of the model path, the likelihood function for a sample size T requires integration over all K^T possible combinations up to the T-th observation, which is impracticable.

Gray (1996) was the first author who dealt with the problem of path dependency. First, he observed that the conditional density of y_t is a mixture of normals with mixture weight $(p_{t-1}(\Delta_t = j))$ varying over time, i.e

$$
f(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{K} p_{t-1}(\Delta_t = j) f_N(y_t|\mathcal{F}_{t-1}, \Delta = k)
$$

where $f_N(.)$ represents the density of a normal distribution. In order to facilitate the notation, consider that $h_t^{(k)}$ is the conditional variance when $\Delta_t = k$, that is,

$$
h_t^{(k)} = \alpha_0^{(k)} + \alpha_1^{(k)} y_{t-1}^2 + \beta_1^{(k)} h_{t-1}^{(k)}. \tag{13}
$$

Hence, the conditional variance of y_{t-1} given $\mathcal{F}_{t-2}$ is given by $h_{t-1} = \sum_{k=1}^{K} p_{t-2}(\Delta_{t-1} = k) h_{t-1}^{(k)}$. Gray (1996) replaced $h_{t-1}^{(k)}$ with h_{t-1} in Equation (13). Thus, the specific conditional variance of each regime is given by:

$$
h_t^{(k)} = \alpha_0^{(k)} + \alpha_1^{(k)} y_{t-1}^2 + \beta_1^{(k)} h_{t-1}. \tag{14}
$$

and now h_{t-1} is not path dependent, so the likelihood estimation is tractable. Abramson and Cohen (2007) showed some inferential results of this model.

Haas et al. (2004) pointed out that the interpretation of the model proposed by Gray (1996) is problematic. As they mentioned, when one specifies a model with regime switching, the interest is to capture the dynamics in the conditional variance in periods of high and low volatility throughout the series. Therefore, we are interested in the relationship of the parameters of the GARCH model corresponding to the variance of its regime $h_t^{(j)}$, similar to the GARCH model with a single regime described in Section 2. However, when using the conditional variance (14), $\beta_1^{(k)}$ cannot be interpreted as being the inertia of $h_t^{(k)}$, since h_t is defined as the average of the conditional variances of all K regimes with respect to the probabilities of being in the $k-$th regime. And since $\beta \neq 0$, $h_t^{(k)}$ will be affected by shocks as h_t varies even if $\alpha_1^{(k)}$ is equal to zero.

In view of this, Haas et al. (2004) presented a new approach to MS-GARCH models with a clear interpretation in relation to the conditional variance of the regime and with feasible estimation.

3.1 Haas, Mittnik and Paolella (2004)'model

Haas et al. (2004) presented a new approach to univariate to MS-GARCH models with a clear interpretation in relation to the conditional variance of the regime and feasible estimation. The MS(K)-GARCH(1,1) model proposed by Haas et al. (2004) can be expressed as:

$$y_t = \sqrt{h_t^{(\Delta_t)}} z_t, \tag{15}$$

$$h_t^{(k)} = \alpha_0^{(k)} + \alpha_1^{(k)} y_{t-1}^2 + \beta_1^{(k)} h_{t-1}^{(k)}, \tag{16}$$

where $h_t^{(k)}$ are the conditional variances of regime k. Note that, in this definition, the term accompanying the parameter $\beta_1^{(k)}$, that is, $h_{t-1}^{(k)}$, is what differentiates this model from the one proposed by Gray (1996). $\alpha_0^{(k)} > 0$, $\alpha_1^{(k)} \geq 0$, $\beta_1^{(k)} \geq 0$ for all $k = 1, \ldots, K$ are the sufficient conditions for $h_t^{(k)}$ to be positive; $\{z_t\}$ is a sequence of independent random variables with mean 0 and variance 1, commonly chosen as N(0.1). Finally, Δ_t is a Markov chain with finite state space $\mathcal{S} = \{1, \ldots, K\}$ and a transition matrix $K \times K$ $\mathbf{P} := [p_{ij}]$, where $p_{ij} = P(\Delta_t = j | \Delta_{t-1} = i)$, with the sum of each row equal to 1 and all elements of the matrix p_{ij} greater than zero. Assuming that the Markov chain is irreducible and aperiodic, the stationary distribution $\boldsymbol{\pi}_\infty = (\pi_\infty^{(1)}, \pi_\infty^{(2)}, \ldots, \pi_\infty^{(K)})'$ can be obtained by $\boldsymbol{\pi}_\infty = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{v}$, where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_k - \boldsymbol{P} \\ \mathbf{1}_k' \end{bmatrix} \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} \mathbf{0}_k \\ \mathbf{1}_k \end{bmatrix}, \tag{17}$$

$\boldsymbol{I}_k$ is the $k \times k$ identity matrix, and $\mathbf{0}_k$ and $\mathbf{1}_k$ are $k \times 1$ vectors of zeros and ones, respectively.

Note that, in this model, the conditional variance $h_t^{(k)}$ depends only on the respective parameters of its regime k, in contrast to the one given by Gray (1996) in (14). Therefore, if $\max\{\beta_1^{(1)}, \ldots, \beta_1^{(K)}\} < 1$, like the GARCH models with a single regime, we have that $\alpha_1^{(k)}$ represents the impact of a shock at $t-1$ on $h_t^{(k)}$; $\beta_1^{(k)}$ represents the memory in $h_t^{(k)}$; and $\alpha_1^{(k)}(1-\beta_1^{(k)})^{-1}$ reflects the total impact of a unit shock on future variations of the regime k.

To study the second-order stationarity of the model, first, consider the $K^2 \times K^2$ matrix $\boldsymbol{M} = [\boldsymbol{M}_{ij}]$, where $\boldsymbol{M}_{ij} = p_{ij}(\boldsymbol{\beta} + \boldsymbol{\alpha}_1 \boldsymbol{e}_k')$ are $K \times K$ matrices, $\boldsymbol{\beta} = \text{diag}(\beta_1^{(1)}, \beta_1^{(2)}, \ldots \beta_1^{(K)})'$, $\boldsymbol{\alpha_1} = (\alpha_1^{(1)}, \alpha_1^{(2)}, ..., \alpha_1^{(K)})'$, p_{ij} denotes the transition probabilities and $\boldsymbol{e}_k'$ is a $K \times 1$ vector that contains zeros, except in the $k-$th element. According to Haas et al. (2004), the process is weakly stationary if and only if $\rho(\boldsymbol{M}) < 1$, where $\rho(.)$ is the largest eigenvalue in modulus of the matrix $\boldsymbol{M}$. Although, the strict stationarity condition is complicated, Liu (2006) presented the conditions for strict stationarity. Haas et al. (2004)

stated that, a necessary condition for the weak stationarity of the MS(K)-GARCH(1,1) process is that, for all k, $\beta_1^{(k)} < 1$, while Abramson and Cohen (2007) presented the conditions of strict and second-order stationarity for MS(K)-GARCH(p,q). More information on the stationarity and properties of the model can be found in the Supplementary Material.

3.2 Non-robust estimators

Here we are interested in maximum likelihood estimator, more specifically in the model defined by Haas et al. (2004). For this, we aim to maximize:

$$L(\boldsymbol{\theta}|\boldsymbol{y}_T) = \prod_{t=1}^{T} f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \prod_{t=1}^{T}\sum_{k=1}^{K} \left[f(y_t|\Delta_t = k, \mathcal{F}_{t-1}, \boldsymbol{\theta}) P(\Delta_t = k|\mathcal{F}_{t-1})\right], \tag{18}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, p_{11}, \ldots, p_{kk})'$ and $f(.)$ is the conditional density of y_t. However, as the regime is unobserved, we need to infer the probability of being in time t in a given regime $P(\Delta_t = k|\mathcal{F}_{t-1})$. For this, the mechanism used here is the Hamilton (1989) filter, which is a nonlinear iterative filter to estimate state probabilities of an autoregressive model with Markov-switching regime.

3.2.1 Estimation via Hamilton filter

Define $\boldsymbol{\xi}_t$ as the vector of conditional probabilities of being in each regime at time t, that is $\boldsymbol{\xi}_t = (P(\Delta_t = 1|\mathcal{F}_{t-1}), \ldots, P(\Delta_t = K|\mathcal{F}_{t-1}))'$ and $\boldsymbol{h}_t$ as the vector of the conditional variances of each regime, that is, $\boldsymbol{h}_t = (h_t^{(1)}, \ldots, h_t^{(K)})'$. So, given the initial values $(y_0.\boldsymbol{h}_0, \boldsymbol{\xi_0})'$ (or $(y_1, \boldsymbol{h}_1, \boldsymbol{\xi_1})'$), the estimation of $\boldsymbol{\xi}_t$ by the Hamilton filter proceeds as follows:

1) Given the vector of conditional probabilities $\boldsymbol{\xi}_{t-1}$:

$$\xi_{i,t-1} = P(\Delta_{t-1} = i|\mathcal{F}_{t-1}). \tag{19}$$

2) Calculate the conditional variances of each regime:

$$h_t^{(k)} = \alpha_0^{(k)} + \alpha_1^{(k)} y_{t-1}^2 + \beta_1^{(k)} h_{t-1}^{(k)}. \tag{20}$$

Under the normality assumption, calculate the conditional densities of each regime.

$$\eta_{it} = f(y_t|\Delta_t = i, \mathcal{F}_{t-1}) = (2\pi h_t^{(i)})^{-1/2} \exp\left\{-y_t^2/(2h_t^{(i)})\right\}. \tag{21}$$

The normality assumption is not the only option. For instance, Paolella et al. (2012) used the t-Student distribution. And, if the distribution of innovations is not normal, but leptokurtic, the use of normality in the regime can seriously affect the identification of the state of the regime and, consequently, also the estimation of the parameters for each regime. That is, the model with normal distribution in regimes will tend to detect regime changes very frequently due to an atypical observation (Paolella et al., 2012).

3) Calculate the conditional density of the t-th observation as:

$$f(y_t|\mathcal{F}_{t-1}) = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} P(\Delta_{t-1} = i|\mathcal{F}_{t-1})\eta_{jt} = \mathbf{1}'(\boldsymbol{P}\boldsymbol{\xi}_{t-1} \odot \boldsymbol{\eta}_t), \tag{22}$$

where $\odot$ represents the Hadamard product and $\boldsymbol{P}$ is the transition matrix.

4) Update the probabilities of being in each regime as:

$$\boldsymbol{\xi}_t = \frac{\boldsymbol{P}\boldsymbol{\xi}_{t-1} \odot \boldsymbol{\eta}_t}{\mathbf{1}'(\boldsymbol{P}\boldsymbol{\xi}_{t-1} \odot \boldsymbol{\eta}_t)}. \tag{23}$$

As a result of iterating from 1) to 4), we can compute the conditional quasi-likelihood function over the entire sample. Thus, the estimate of $\boldsymbol{\theta}$ is obtained by maximizing the likelihood or log-likelihood through some numerical method. In such a way that:

$$\hat{\boldsymbol{\theta}} \ \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=2}^{T} \log f(y_t|\mathcal{F}_{t-1}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=2}^{T} \log(\mathbf{1}'(\boldsymbol{P}\boldsymbol{\xi}_{t-1} \odot \boldsymbol{\eta}_{jt})). \tag{24}$$

As with single-regime models, the choice of initial values can influence parameter estimates. For the initial values of $\boldsymbol{\xi}_1$, the estimated stationary distribution of the Markov chain is often used (17), or else considered to be equiprobable, or considered to be parameters and estimated by QML. For the values of $\boldsymbol{h}_1$, we can assume that they are a function of the parameters, such that for the regime k we have that $h_1^k = \alpha_{0k}/(1 - \alpha_{1k} - \beta_{1k})$.

3.2.2 Notes on quasi-maximum likelihood estimation

As of the writing, we are unaware of any asymptotic properties, such as consistency and asymptotic distribution of the QML estimator for the MS-GARCH model that follows the specification proposed by Haas et al. (2004).

Paolella et al. (2012) pointed out that, similar to the GARCH mixing models, log-likelihood maximisation (24) is often performed by an optimisation method based on the Hessian matrix with numerically determined derivatives. Because there may be several plausible local maxima of the likelihood function, only by using various initial values one can obtain the possible the global maximum. For a slightly different model, Augustyniak (2014) also highlighted the impact of initial values on parameter estimates and proposed an algorithm that can reduce the sensitivity to initial values. Another point is that, due to the nature of the likelihood function as a mixture, it may have singularities (infinite likelihood values).

Andreou and Ghysels (2009) showed different statistical methods to look for the number of regimes. However, Paolella et al. (2012) stated that these standard tests and also tests based on likelihood ratios do not perform well, while information criteria such as AIC and BIC are reasonably good in this context. On the other hand, in empirical problems, taking into account interpretability, the number of regimes chosen is usually 2, since with $K > 2$, at least one of the regimes tends to have a very small stationary probability, so its parameters are subject to large estimation errors (Paolella et al., 2012).

3.3 Robust estimation and prediction

We propose a robust estimator that uses three modifications of the one presented previously: i) the BIP method given by Equation (6) in Section 2.2 to minimise the effect of outliers on the conditional variance estimates of each regime; ii) the use of the t-Student distribution with 4 degrees of freedom in the conditional densities; and iii) the objective function proposed by Boudt et al. (2013) for the univariate case.

The motivation of using the BIP is the same as for the single-regime case. The use of the t-Student distribution with 4 degrees of freedom in the conditional densities, is based on the techniques used in robust filters. Different authors have mentioned that the usual filters, as well as Hamilton's, are not robust. Even in the mildest cases they have unsuitable performance. Petrus (1999), Zou et al. (2000), and Calvet et al. (2015) showed the poor performance of several filters such as least squares and Bayesian, for example, in the presence of outliers.

To understand the representation of outliers in filters, we follow Calvet et al. (2015). Let $y_t \in \mathbb{R}$ be the observation with conditional density $f(y_t|\Delta_t, \mathcal{F}_{t-1})$. As in (23), by Bayes' rule, the probabilities of being in each regime inferred by the filter satisfy:

$$\lambda(\Delta_t|y_t, \mathcal{F}_{t-1}) = \frac{f(y_t|\Delta_t, \mathcal{F}_{t-1})\lambda(\Delta_t|\mathcal{F}_{t-1})}{f(y_t|\mathcal{F}_{t-1})}, \tag{25}$$

where $f(y_t|\mathcal{F}_{t-1}) = \int f(y_t|\Delta_t, \mathcal{F}_{t-1})\lambda(\Delta_t|\mathcal{F}_{t-1})d\Delta_t$. Thus, let a family of observation densities contaminated by outliers be denoted by $f_{cont}(.|\Delta_t, \mathcal{F}_{t-1}, \omega)$, parameterised by ω, where ω belongs to an interval $\mathcal{D}$ on the real line containing zero. If $\omega = 0$, the contaminated density coincides with the uncontaminated observation density, that is, $f_{cont}(y_t|\Delta_t, \mathcal{F}_{t-1}, \omega = 0) = f(y_t|\Delta_t, \mathcal{F}_{t-1})$ for all Δ_t and $\mathcal{F}_{t-1}$. When contamination is not taken into account, the naive application of the Bayes rule leads to formula (25).

To robustify the filter, Calvet et al. (2015) and Petrus (1999) used robust densities. Calvet et al. (2015) used one based on the limitation of the derivative of the log density and Petrus (1999) employed one based on Huber's objective function, $\rho(u) = u^2 I(|u| \leq k) + k|u|I(|u| < k)$. Analogously to Boudt et al. (2013), we use a t-Student distribution with 4 degrees of freedom.

Therefore, the robust estimation proposal will modify the estimation presented in Subsection 3.2.1 in such a way that, in step 2) we will use the BIP mechanism to minimise the effect of outliers on the conditional variance estimates of each regime and we will use the t-distribution with 4 degrees of freedom to calculate of the conditional density of each regime; and, finally, instead of maximising the log-likelihood as in (24), we will use a modification of it combined with the function ρ given in (8). Thus, the estimate of $\boldsymbol{\theta}$ is obtained by minimising:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=1}^{T} \log \left[\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} P(\Delta_{t-1} = i|\mathcal{F}_{t-1}) \frac{1}{\sqrt{h_t^{(j)}}} \left(1 + \frac{y_t^2}{2h_t^{(j)}}\right)^{-\frac{(1+4)\sigma_4}{2}} \right]. \tag{26}$$

where σ_4 is given by (10).

To estimate the one-step ahead volatility, we propose to use the robust BIP mechanism given by Equation (6). For more than one-step-ahead prediction we can use the procedure of Paolella et al. (2012), which is given in the Supplementary Material, because from the second-step we use the expected values, which are not affected by outliers. Using the one-step-ahead conditional distribution of each state, we can use numerical methods or path simulations to estimate the one-step-ahead VaR. For two or more steps ahead we have to use path simulations.

For the robust method the volatility prediction can be found by simulating paths of the returns using the estimated robust model.

3.4 Expected shortfall estimation

Consider that a model predict that the return on time t will come from a mixture of K distributions with density $f_k(.)$, each one with weight $p_k = P(\Delta_t = k)$, $k = 1, ..., K$, For the t-th observation, the VaR at the α risk level, denoted by VaR_α , is given by

$$P[y_t < -VaR_\alpha] = \sum_{k=1}^{K} P(\Delta_t = k) F_k(-VaR_\alpha)$$

where $F_k(.)$ is the distribution function of the k-th regime.

The expected shortfall at the α risk level, denoted by ES_α is defined as $-E(y_t|y(t) < -VaR_\alpha)$, i.e.,

$$ES_\alpha \;=\; -E(y_t|y(t) < -Var_\alpha) = -\alpha^{-1} \sum_{k=1}^{K} p_k \int_{-\infty}^{-VaR_\alpha} z f_k(z) dz$$

For the Gaussian case, where the distribution for each regime is $N(0, h^{(k)})$, the ES is given by

$$\begin{aligned} ES_\alpha &= -\sum_{k=1}^{K} p_k \frac{1}{\sqrt{2\pi h_t^{(k)}}} \int_{-\infty}^{-VaR_\alpha} z \exp\left(\frac{-z^2}{2h_k^{(t)}}\right) dz \\ &= -\alpha^{-1} \sum_{k=1}^{K} p_k \sqrt{h_t^{(k)}} \psi(\mathbf{\Phi}^{-1}(-VaR_\alpha / \sqrt{h_t^{(k)}}) \end{aligned} \tag{27}$$

where $\mathbf{\Phi}(.)$ and $\psi(.)$ are the distribution and probability density function of the standard normal distribution, respectively.

For the t-Studennt case where the degree of freedom of each regime is given by ν_k, the ES is given by

$$ES_\alpha = -\alpha^{-1} \sum_{k=1}^{K} p_k \frac{\nu_k + \left(\frac{VaR_\alpha}{\sqrt{h_t^{(k)}}} \sqrt{\nu_k/(\nu_k - 2)}\right)^2}{\nu_k - 1} t_{\nu_k} \left(\frac{VaR_\alpha}{\sqrt{h_t^{(k)}}} \sqrt{\nu_k/(\nu_k - 2)}\right), \tag{28}$$

Note that for the robust method, besides using the robust estimates we use the BIP mechanism given by Equation (6) to estimate the variances for each regime. The estimation of the ES for two or more steps-ahead can be done by simulating paths.

4 Simulation

In this Section we discuss the performance of the estimators of the MS-GARCH models in the presence or absence of additive outliers. We use GARCH(1,1) models for both regimes. As cited by Francq and Zakoian (2019) (Section 8.5, page 235), the reason for choosing GARCH(1,1) models is that they are by far the most widely used by professionals who want to estimate the volatility of daily returns, a practice motivated by the common belief that GARCH(1,1) is sufficient to capture the properties of financial series and higher order models can be unnecessarily complicated.

4.1 Data generator process

We consider a sample size of $T = 3000$ and 1000 Monte Carlo replications. Given the choices $p_{11} = 0.96$ and $p_{22} = 0.98$, the stationary distribution of the Markov chain is $\boldsymbol{\pi}_\infty = (1/3, 2/3)'$ such that, the average numbers of observations in each regimen are 1000 and 2000. The observed series y_t^*, with additive outliers, is defined as:

$$y_t^* = y_t + \text{sign}(y_t) I_A(t)\, d_t \sqrt{h_t^{(\Delta_t)}},$$

where y_t is given by (15), sign(.) represents the sign function, $I_A(t)$ is an indicator variable representing the position of the outliers, with A being the set of observations affected by outliers, d_t is a constant belonging to $\mathbb{R}_+$ indicating the size of the outliers in terms of conditional volatility, and $\sqrt{h_t^{(\Delta_t)}}$ is defined as (16).

Some authors use isolated and consecutive outliers at strategic points, such as at the beginning, middle and end of the series, to assess their respective impacts. As in Muler and Yohai (2008) and Boudt et al. (2013), we evaluate the estimators when a fraction of observations is contaminated. The fraction, say ϵ, are given by $\epsilon = 1\%$, $\epsilon = 5\%$, $\epsilon = 10\%$ and $\epsilon = 0\%$ when there are no outliers. The positions of the outliers are randomly sampled without replacement. Finally, we consider two values for d, 3 and 5. The data generating process is:

$$y_t = \sqrt{h_t^{(\Delta_t)}} z_t \tag{29}$$

$$h_t^{(1)} = 2.0 + 0.10\, y_{t-1}^2 + 0.6\, h_{t-1}^{(1)} \tag{30}$$

$$h_t^{(2)} = 0.3 + 0.35\, y_{t-1}^2 + 0.2\, h_{t-1}^{(2)}, \tag{31}$$

where $z_t \sim N(0, 1)$. In each replication, a burn-in of 500 observations is considered. Although the models are not exactly the same, we use for the simulations the same parameter values of the generating process presented by Bauwens et al. (2010), Augustyniak (2014) and Billio et al. (2016).

The unconditional variance of the first regime is 10 times greater than that of the second. That is, we have $\alpha_0^{(1)}/(1 - \alpha_1^{(1)} - \beta_1^{(1)}) = 20/3$ and $\alpha_0^{(2)}/(1 - \alpha_1^{(2)} - \beta_1^{(2)}) = 2/3$.

4.2 Estimators

A robust estimator i) must have a reasonably good efficiency (optimal or almost optimal) if the assumed model is true; ii) must be robust, in the sense that small deviations from the model's assumptions should only slightly affect the estimates; and iii) must not have catastrophic performance with slightly larger deviations from the model. We consider three estimators in the simulation: QML-n and QML-t estimators defined in Section 3.2, and the robust (Rob) estimator defined in Section 3.3. A modification was introduced in the function $r_{k_{\delta,1}}(u)$, defined in Equation (6), so that the estimators produce better results in estimating the parameters. In the literature, some authors propose replacing the value of $k_{\delta,1}$ when $x > k_{\delta,1}$, like Muler and Yohai (2008) and Boudt et al. (2013), for example. However, others suggest using 1 (expected value) when $x > k_{\delta,1}$, such as Carnero et al. (2012). This value corresponds to approximately $k_{0.69,1}$. We use the midpoint between $\delta = 0.95$ and 0.69. Thus, when $x > k_{\delta,1}$, we replace it with $k_{0.82,1}$.

4.3 Initial values

As observed by Paolella et al. (2012), there may be several "plausible" local maxima of the MS-GARCH model's likelihood function. An alternative way to deal with this is to use several initial values to find the global maximum. Thus, in our simulation, for each replication in which we evaluate the performance of the estimators, we take into account several different initial values and for each we evaluate the likelihood function. For the three initial values that generate the highest likelihood (if it is a loss function, the lowest), we estimate the parameters giving us three parameter estimates generated by three different initial values. Among these three values, we use the following mechanism to determine which estimate corresponds to the global maximum. The mechanism is based on the following idea: When the 3 estimates converge to the same place, that is, the estimated parameters are all the same and the final likelihood found is the same, we select the respective value; otherwise, when the 3 estimates do not converge to the same place, we select the one with the highest final likelihood, or the average of the 3 estimates if there is no difference between the likelihoods.

Here, we consider the parameters and the final likelihood to be different if there is a difference greater than 10^{-4} either in the parameters or in the final likelihood, except for the intercept of the high volatility regime, which will be different if there is a difference greater than 10^{-2}.

Those cases where the three estimates do not converge to the same place are exceptions. For instance, we generate 1000 Monte Carlo replications of the process defined in (31) and generated the above mechanism using maximum likelihood estimation. Of these replications, only 27 (2.7%) did not converge to the same place, and of these cases, only 14 (1.4%) converged to the same likelihood value, but with different parameter values.

The performance of the Hamilton filter is analysed in Section 4.4 and the precision in estimating the model parameters is examined in Section 4.5.

4.4 Analysis of the filter performance

Figure 1 presents the proportion of correct classifications, in which the data generating process is given by Equation (31), considering (a) only the observations affected by outliers and (b) all the observations. We set a threshold of 0.50 to classify whether an observation is in one regime or not. As expected, the filters perform better in the high volatility regime and, in this regime, the performance tends to improve with greater contamination and larger outlier size, since the effect of outliers is to classify the observation as generated by the high volatility regime.

For the case of the low volatility regime, with 1% contamination, when using the QML-n estimator, the percentage of correct classification is less than 14% of cases, even with outliers with size equal to 3, in contrast to the QML-t and Robust estimators, when the hit rate is greater than 75% when $d = 3$ and greater than 48% when $d = 5$. When the degree of contamination increases to 5%, the performance of the estimators deteriorates. The performances of the QML-t and robust estimators are still reasonable for $d = 3$ with more than 60% correct classification, but the percentage drops to less than 17% when $d = 5$. Finally, when the proportion of outliers is 10% and $d = 3$, the robust estimator performs similarly with contamination of 5% and better than QM-t. However, when $d = 5$ all perform poorly.

Generally, the three estimators perform well in the high volatility regime. In the low volatility regime, the filter with the QML-n estimator performs poorly, and when $d = 3$, the filter with the robust estimator performs reasonably well, better than with the QML-t estimator. When the size of the outlier is $d = 5$, the performance of the filter used the three estimators is very poor when the contamination is greater than or equal to 5%

Considering all observations, for the low volatility regime, all estimators perform well, but, in general, the robust estimator performs slightly better than QML estimators. As for the high volatility regime, we note the superior performance of the QML-t and robust estimator in relation to the QML-n estimators, except for the case in which the proportion of outliers is 10% and $d = 5$.

4.5 Model, VaR and ES estimation

The bias and RMSE of the parameter model estimates of the Monte Carlo simulation are given in the Supplementary Material, while Figures 2 and 3 present the boxplots for the parameter estimates of the MS-GARCH model for the high and low volatility regime, respectively.

As expected, in the absence of outliers ($\epsilon = 0\%$), the QML-n estimator performs better than the others. And as the presence of outliers increases, we expected this relationship to reverse, which was confirmed in the simulations. For the parameters of the variance equations of the MS-GARCH model, we observe that, in terms of RMSE, as the fraction and size of the *outliers* increase, the performance of the estimators deteriorates. Also, in general the

QML-n estimator performs worse than the QML-t and robust estimators. We also observe that the robust estimator stands out in most cases in estimating these parameters.

In the high volatility regime (Figure 2), the performance of the QML-n estimator, of 5% with $d = 3$ and 1% with $d = 5$ deteriorates substantially. When $d = 3$ the robust estimator performs better than the QML-t. When $d = 5$, for 5% the estimates of parameters β_1, p_{11} and p_{22} deteriorate for all estimators. When $d = 5$, for the parameters α_0 and α_1, there is no advantage regarding the estimators.

For the low volatility regime (Figure 3), the robust and QML-t estimators perform better than the QML-n, except for the parameter α_1 when there is 1% contamination by outliers. Furthermore, the robust estimator is better than the QML-t estimator.

Finally, the QML-t estimator obtained the best results when estimating the probabilities of the Markov transition matrix. The RMSE of the QML-t estimator is half that of the robust estimator, except when the proportion of outliers is 10%. For 5% contamination, the performance of the QML-n estimator is very poor. There is also a larger variability in the QML-n estimates for both regimes for different windows.

The literature contains some results of Monte Carlo simulations in relation to the estimation of the parameters of the MS-GARCH model. Augustyniak (2014) presented the simulation results using the specification given by Gray (1996); Billio et al. (2016) showed the simulation results of estimating the parameters via Bayesian methods. In both cases, there was difficulty in estimating the parameters $\alpha_0^{(1)}$, $\beta_1^{(1)}$ and $\beta_1^{(2)}$, even in the absence of *outliers*. In our simulations, we also noticed that in general, these parameters were the most problematic. Also, the BIP mechanism tended to work better when the percentage of outliers was higher. Furthermore, the QML-t estimator performed well in estimating the Markov chain transition matrix probabilities.

Tables 1 presents the performance of the VaR and ES prediction under different outlier contamination for different methods. We present the mean absolute prediction error (MAPE) and the mean absolute percentage prediction error (MPPE) for VaR and ES.

The robust estimator presented a good performance, except for 10% contamination. When there is no contamination the QML-n presented the best performance, as expected, while the robust estimator presented a close performance. Even when there is no or few outliers, the robust estimator presented a better performance than the QML-t.

5 Final remarks

The objective of this paper was to evaluate the performance of some MS-GARCH model estimators in the presence of additive outliers. Since there are no studies of robust techniques to estimate these models, this work contributes to performance simulation studies for different outlier contamination scenarios. The suggested method adapts the robust estimation methods presented by Boudt et al. (2013) for the case of Markov regime switching.

The main conclusions from the simulation results are: i) the QML estimator using the normal distribution as the conditional density of each regime performs catastrophically as the proportion and size of outliers increases (the estimator performed poorly in terms of filter classification capacity and estimation accuracy); and ii) using the QML estimation with t-Student distribution with four degrees of freedom as the conditional density of each regime, we observed a reasonable performance of the filter classification capacity and in the estimation of some parameters of the volatility equations. As for the robust estimation, it also had reasonable performance in the filter classification capacity and in the estimation of some parameters of the volatility equations.

As future works we can mention: Application to real data; study the BIP in case of regime change; consider possible changes in the function $r_{k_{\delta,1}}(u)$, defined in Equation (6), including different functional forms; study of the behaviour of robust techniques in relation to different filters and in relation to Bayesian estimation; consider the feasibility of putting transition probabilities as time variants, as, for example, in Diebold et al. (1994); Psaradakis and Sola (2021); Pouzo et al. (2022) and Wang et al. (2022); and compare with the indicator saturation method developed by Santos et al. (2008) and applied, for example, in Bauwens and Sucarrat (2010) and Pretis et al. (2018).

References

Abramson, A. and Cohen, I. (2007). On the stationarity of Markov-switching GARCH processes. *Econometric Theory*, 23(3):485–500.

Aknouche, A. and Guerbyenne, H. (2006). Recursive estimation of GARCH models. *Communications in Statistics-Simulation and Computation*, 35(4):925–938.

Andreou, E. and Ghysels, E. (2009). *Structural Breaks in Financial Time Series*, pages 839–870. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ardia, D. (2008). *Financial Risk Management with Bayesian Estimation of GARCH Models*. Springer, Berlin, Heidelberg, 1st edition.

Ardia, D., Bluteau, K., Boudt, K., and Catania, L. (2018). Forecasting risk with Markov-switching GARCH models: A large-scale performance study. *International Journal of Forecasting*, 34(4):733–747.

Augustyniak, M. (2014). Maximum likelihood estimation of the Markov-switching GARCH model. *Computational Statistics & Data Analysis*, 76:61–75.

Baillie, R. T. and Chung, H. (2001). Estimation of GARCH models from the autocorrelations of the squares of a process. *Journal of Time Series Analysis*, 22(6):631–650.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Bauwens, L., Preminger, A., and Rombouts, J. V. (2010). Theory and inference for a Markov switching GARCH model. *The Econometrics Journal*, 13(2):218–244.

Bauwens, L. and Sucarrat, G. (2010). General-to-specific modelling of exchange rate volatility: A forecast evaluation. *International Journal of Forecasting*, 26(4):885–907.

Billio, M., Casarin, R., and Osuntuyi, A. (2016). Efficient Gibbs sampling for Markov switching GARCH models. *Computational Statistics & Data Analysis*, 100:37–57.

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

Boudt, K., Danielsson, J., and Laurent, S. (2013). Robust forecasting of dynamic conditional correlation GARCH models. *International Journal of Forecasting*, 29(2):244–257.

Cai, J. (1994). A Markov model of switching-regime Arch. *Journal of Business & Economic Statistics*, 12(3):309–316.

Calvet, L. E., Czellar, V., and Ronchetti, E. (2015). Robust filtering. *Journal of the American Statistical Association*, 110(512):1591–1606.

Carnero, M. A., Peña, D., and Ruiz, E. (2012). Estimating GARCH volatility in the presence of outliers. *Economics Letters*, 114(1):86–90.

Crosato, L. and Grossi, L. (2019). Correcting outliers in GARCH models: a weighted forward approach. *Statistical Papers*, 60(6):1939–1970.

Das, D., Yoo, B. H., et al. (2004). A Bayesian MCMC algorithm for Markov switching GARCH models. *Econometric Society 2004 North American Summer Meetings*, pages 1–16.

Diebold, F. X., Lee, J.-H., and Weinbach, G. C. (1994). Regime switching with time-varying transition probabilities. In Hargreaves, C., editor, *Nostationary Time Series Analysis and Cointegration*, page 1998. Oxford University Press.

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

Francq, C. and Zakoian, J.-M. (2019). *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62.

Haas, M., Mittnik, S., and Paolella, M. S. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, 2(4):493–530.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1-2):307–333.

Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *Journal of Econometrics*, 129(1-2):121–138.

Horváth, L., Kokoszka, P., and Zhang, A. (2006). Monitoring constancy of variance in conditionally heteroskedastic time series. *Econometric Theory*, 22(3):373–402.

Hotta, L. K. and Trucíos, C. (2018). Inference in (M) GARCH models in the presence of additive outliers: Specification, estimation, and prediction. In Lavor, C. and Gomes Neto, F. d. A. M., editors, *Advances in Mathematics and Applications*, pages 179–202.

Hotta, L. K. and Tsay, R. S. (2012). Outliers in GARCH processes. In Bell, W. R., Holan, S. H., and McElroy, T. S., editors, *Economic Time Series: Modeling and Seasonality*, pages 337–358.

Huber, P. J. and Ronchetti (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Hwang, S. and Valls Pereira, P. L. (2008). The effects of structural breaks in ARCH and GARCH parameters on persistence of GARCH models. *Communications in Statistics—Simulation and Computation*, 37(3):571–578.

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business & Economic Statistics*, 8(2):225–234.

Lee, S. and Song, J. (2009). Minimum density power divergence estimator for GARCH models. *Test*, 18(2):316–341.

Liu, J.-C. (2006). Stationarity of a Markov-switching GARCH model. *Journal of Financial Econometrics*, 4(4):573–593.

Mikosch, T. and Stărică, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, 86(1):378–390.

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

Paolella, M. S., Haas, M., Bauwens, L., Hafner, C. M., and Laurent, S. (2012). Mixture and regime-switching GARCH models. *Wiley Handbooks in Financial Engineering and Econometrics*, (3):71–102.

Park, B.-J. (2002). An outlier robust GARCH model and forecasting volatility of exchange rate returns. *Journal of Forecasting*, 21(5):381–393.

Pelagatti, M. and Lisi, F. (2009). Variance initialisation in GARCH estimation. In *S. Co. 2009. Sixth Conference. Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction*, page 323.

Pérez-Cruz, F., Afonso-Rodriguez, J. A., Giner, J., et al. (2003). Estimating GARCH models using support vector machines*. *Quantitative Finance*, 3(3):163–172.

Petrus, P. (1999). Robust Huber adaptive filter. *IEEE Transactions on Signal Processing*, 47(4):1129–1133.

Pouzo, D., Psaradakis, Z., and Sola, M. (2022). Maximum likelihood estimation in Markov regime-switching models with covariate-dependent transition probabilities. *Econometrica*, 90(4):1681–1710.

Pretis, F., Reade, J. J., and Sucarrat, G. (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software*, 86(3):1–44.

Psaradakis, Z. and Sola, M. (2021). Markov-switching models with state-dependent time-varying transition probabilities. *Econometrics and Statistics*.

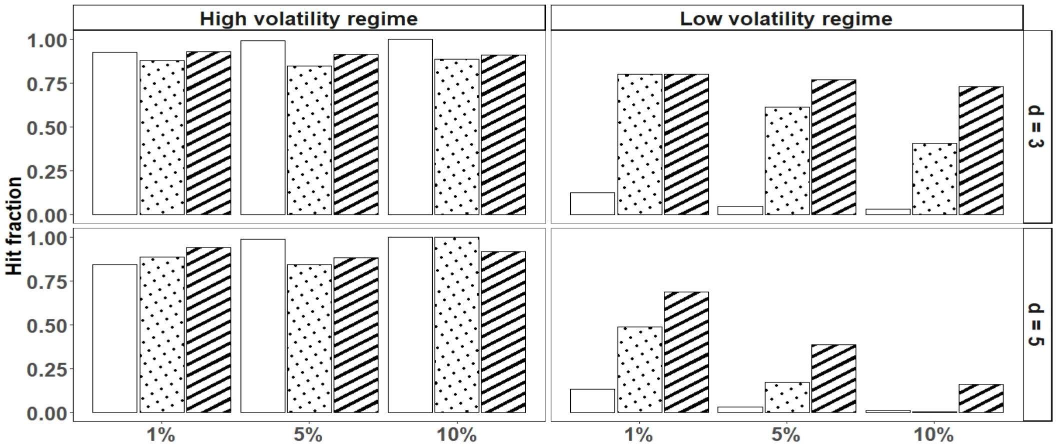
Sakata, S. and White, H. (1998). High breakdown point conditional dispersion estimation with application to s&p 500 daily returns volatility. *Econometrica*, pages 529–567.

Santos, C., Hendry, D. F., and Johansen, S. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23(2):317–335.

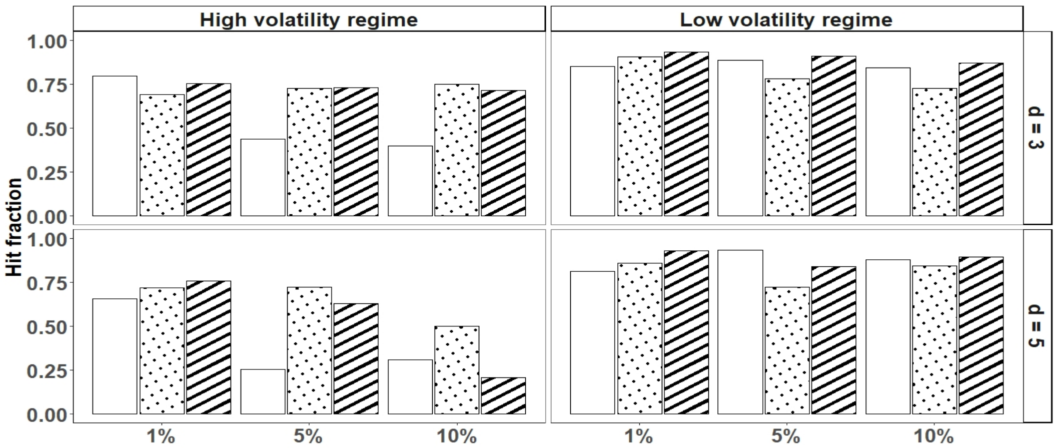
Trucíos, C., Hotta, L. K., and Ruiz, E. (2017). Robust bootstrap forecast densities for GARCH returns and volatilities. *Journal of Statistical Computation and Simulation*, 87(16):3152–3174.

Wang, J., Ma, F., Liang, C., and Chen, Z. (2022). Volatility forecasting revisited using Markov-switching with time-varying probability transition. *International Journal of Finance & Economics*, 27(1):1387–1400.

Zou, Y., Chan, S.-C., and Ng, T.-S. (2000). Least mean m-estimate algorithms for robust adaptive filtering in impulse noise. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(12):1564–1569.

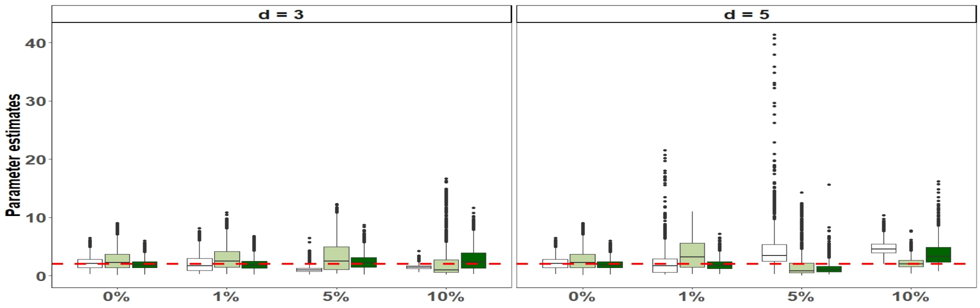


(a) contaminated observations

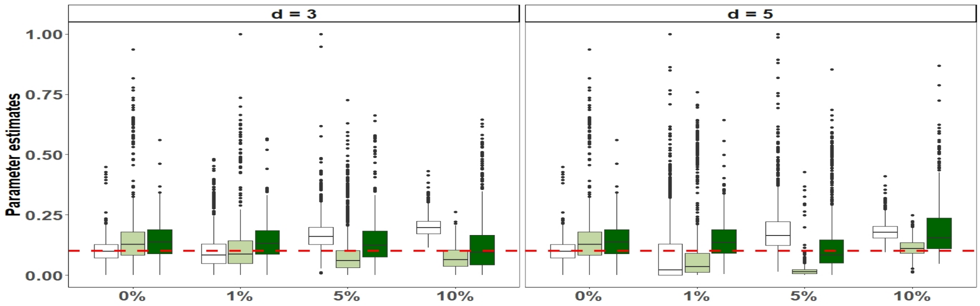


(b) all observations

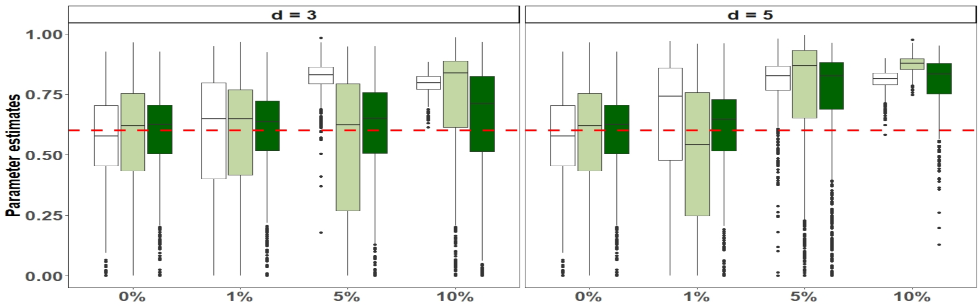
Figure 1: Proportion of correct classification of the different estimators for the MS-GARCH model in the (a) contaminated observations (b) all observations. The horizontal axis gives the percentage of outliers. The empty bars denote the QML-n estimator, the bars with dots represent the QML-t estimator and those with diagonal lines denote the robust estimator. The graphs on the left indicate when the correct regime has high volatility, that is, regime 1, and those on the right indicate regime of low volatility, regime 2. The upper graphs show outliers of size 3 times the conditional standard deviation, while the lower ones show 5 times the conditional standard deviation.



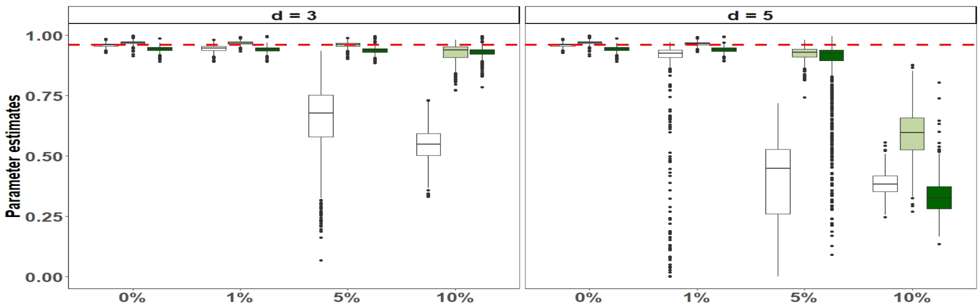
(a) $\alpha_0^{(1)}$



(b) $\alpha_1^{(1)}$

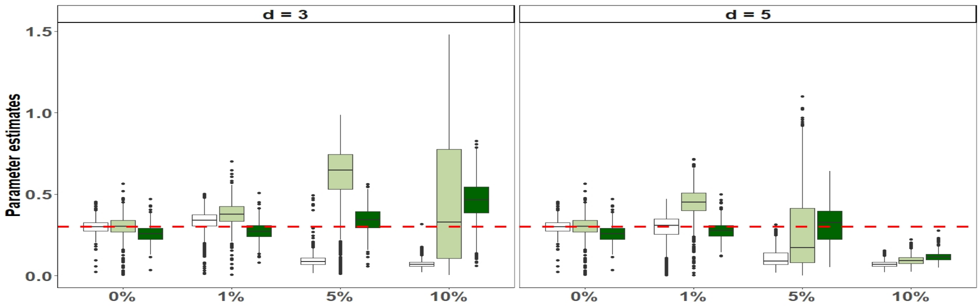


(c) $\beta_1^{(1)}$

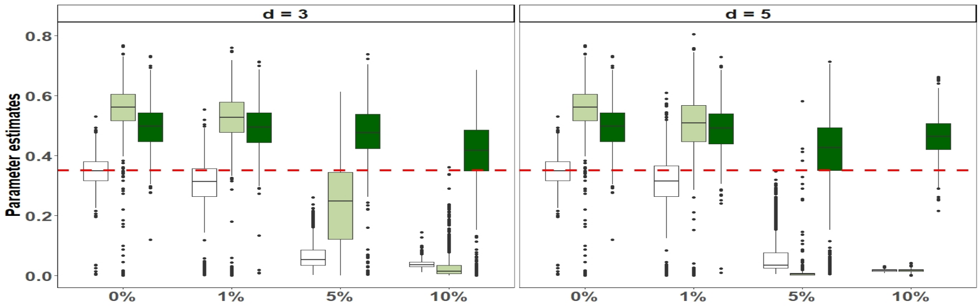


(d) p_{11}

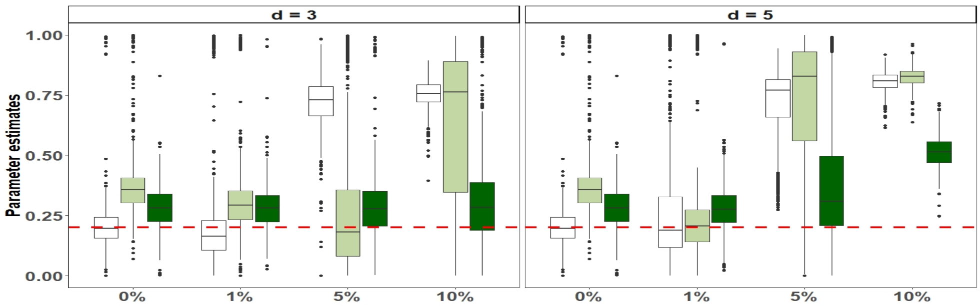
Figure 2: Boxplot of parameter estimates for the high volatility regime of the Monte Carlo replications of the MS-GARCH(1,1) model. The x-axis represents the percentage of outliers and the y-axis represents the parameter estimate values. The left frame contains the values of the sizes of outliers with $d = 3$ and in the right $d = 5$. The QML-n estimator is in white, the QML-t estimator is in light green and and the robust estimator is in dark green. The red dashed line represents the actual value of the parameter. The parameters in each graph are: (a) $\alpha_0^{(2)}$, (b) $\alpha_1^{(2)}$, (c) $\beta_1^{(2)}$ and (d) p_{22}.



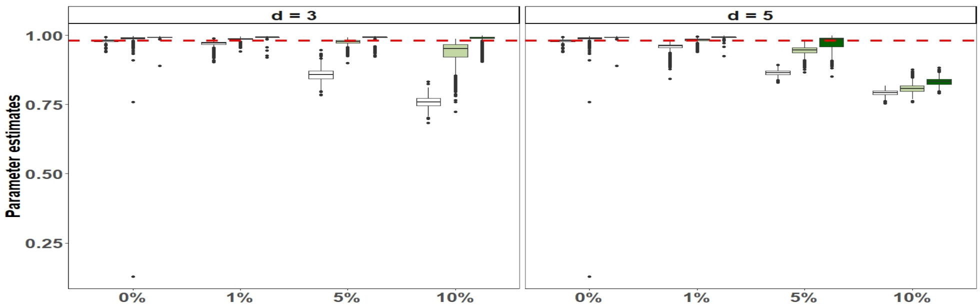
(a) $\alpha_0^{(2)}$



(b) $\alpha_1^{(2)}$



(c) $\beta_1^{(2)}$



(d) p_{22}

Figure 3: Boxplot of the estimates of the parameters of the low volatility regime of the Monte Carlo replications of the MS-GARCH(1,1) model. The x-axis represents the percentage of outliers and the y-axis represents the parameter estimate values. The left frame contains the values of the sizes of outliers with $d = 3$ and in the right $d = 5$. The QML-n estimator is in white, the QML-t estimator is in light green and and the robust estimator is in dark green. The red dashed line represents the actual value of the parameter. The parameters in each graph are: (a) $\alpha_0^{(2)}$, (b) $\alpha_1^{(2)}$, (c) $\beta_1^{(2)}$ and (d) p_{22}.

Table 1: Mean absolute prediction error (MAPE) and the mean absolute percentage prediction error (MPPE) for VaR and ES. In bold are those that presented the best results within each group.

Contami nation	α	Method	VaR MAPE	VaR MPPE	ES MAPE	ES MPPE	% Hits
$\varepsilon = 0\%$	1%	QML-N	**0.557**	**0.219**	**0.740**	**0.237**	**1.1**
		QML-t	1.365	0.487	2.500	0.655	0.3
		Rob	0.936	0.341	1.792	0.474	0.8
	2.5%	QML-N	**0.396**	**0.173**	**0.570**	**0.214**	**2.5**
		QML-t	0.734	0.322	1.568	0.515	1.6
		Rob	0.533	0.221	1.078	0.360	2.7
$\varepsilon = 1\%$ d = 3	1%	QML-N	**0.823**	**0.344**	**1.130**	**0.377**	0.7
		QML-t	1.522	0.558	2.773	0.745	0.2
		Rob	1.022	0.373	1.927	0.511	**0.8**
	2.5%	QML-N	**0.544**	0.252	**0.841**	**0.333**	2.3
		QML-t	0.823	0.374	1.749	0.590	1.5
		Rob	0.569	**0.242**	1.174	0.393	**2.5**
$\varepsilon = 1\%$ d = 5	1%	QML-N	1.266	0.532	**1.857**	0.619	0.5
		QML-t	1.641	0.616	2.991	0.823	0.3
		Rob	**1.032**	**0.379**	1.942	**0.518**	**0.8**
	2.5%	QML-N	0.713	0.312	1.278	0.505	1.9
		QML-t	0.882	0.411	1.887	0.652	1.4
		Rob	**0.580**	**0.248**	**1.186**	**0.399**	**2.3**
$\varepsilon = 5\%$ d = 3	1%	QML-N	2.433	1.058	2.954	0.946	0
		QML-t	2.267	0.860	3.917	1.097	0.2
		Rob	**1.411**	**0.518**	**2.543**	**0.682**	**0.6**
	2.5%	QML-N	1.61	0.832	2.349	0.943	0.7
		QML-t	1.301	0.609	2.561	0.898	0.9
		Rob	**0.776**	**0.349**	**1.609**	**0.543**	**1.7**
$\varepsilon = 5\%$ d = 5	1%	QML-N	5.005	2.135	6.511	2.051	0
		QML-t	2.852	1.148	4.818	1.425	0
		Rob	**1.734**	**0.675**	**3.100**	**0.879**	**0.3**
	2.5%	QML-N	2.913	1.540blz	4.909	1.943	0
		QML-t	1.651	0.813	3.184	1.182	0.8
		Rob	**0.959**	**0.455**	**1.971**	**0.705**	**1.3**
$\varepsilon = 10\%$ d = 3	1%	QML-N	3.425	1.461	4.065	1.273	0
		QML-t	3.212	1.241	5.287	1.508	0.1
		Rob	**2.034**	**0.759**	**3.513**	**0.964**	**0.3**
	2.5%	QML-N	2.436	1.286	3.347	1.322	0.2
		QML-t	1.950	0.926	3.567	1.275	0.8
		Rob	**1.174**	**0.542**	**2.295**	**0.791**	**1.2**
$\varepsilon = 10\%$ d = 5	1%	QML-N	7.027	2.923	8.685	2.684	0
		QML-t	5.914	2.397	9.379	2.794	0
		Rob	**4.867**	**2.095**	**8.062**	**2.512**	0
	2.5%	QML-N	4.801	2.473	6.949	2.690	0
		QML-t	3.545	1.794	6.464	2.421	0
		Rob	**2.666**	**1.449**	**5.365**	**2.120**	0