To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–31

A robust statistical approach to select adequate error distributions for financial returns

J. Hambuckers[a b c *] and C. Heuchenne [a d]

[a]Chair of Statistics, Georg-August Universität Göttingen, Germany; [b]*Fonds national de la recherche scientifique (F.R.S. - FNRS), Belgium*; [c]*QuantOM, HEC Liege, University of Liege, Liege, Belgium*; [d]*Institute of Statistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

In this article, we propose a robust statistical approach to select an appropriate error distribution, in a classical multiplicative heteroscedastic model. In a first step, unlike to the traditional approach, we don't use any GARCH-type estimation of the conditional variance. Instead, we propose to use a recently developed nonparametric procedure [31]: the Local Adaptive Volatility Estimation (LAVE). The motivation for using this method is to avoid a possible model misspecification for the conditional variance. In a second step, we suggest a set of estimation and model selection procedures (Berk-Jones tests, kernel density-based selection, censored likelihood score, coverage probability) based on the so-obtained residuals. These methods enable to assess the global fit of a set of distributions as well as to focus on their behavior in the tails, giving us the capacity to map the strengths and weaknesses of the candidate distributions. A bootstrap procedure is provided to compute the rejection regions in this semiparametric context. Finally, we illustrate our methodology throughout a small simulation study and an application on three time series of daily returns (UBS stock returns, BOVESPA returns and EUR/USD exchange rates).

Keywords: error distribution, nonparametric volatility, model misspecification, goodness-of-fit, selection test, GARCH, skewed-t, NIG, hyperbolic

JEL classification: C14, C18, C46, C51

1. Introduction

Since the 2008 financial crisis, the literature faces a renewed interest in the choice of an adequate error distribution, able to capture the skewness and excess kurtosis of stochastic processes [see, among others 8, 12, 35, 37]. In this article, we propose a robust methodology to select a distribution family in a classical multiplicative heteroscedastic model. This model is defined by :

$$r_t = \sigma_t z_t, \tag{1}$$

*Corresponding author. Email: jhambuckers@ulg.ac.be

$$\sigma_t^2 = Var(r_t|\mathcal{F}_{t-1}), \tag{2}$$

$$z_t \sim F_z(\cdot), \tag{3}$$

where r_t is the daily return, σ_t^2 the conditional variance of r_t and z_t are i.i.d. random variables distributed according to a cumulative distribution function $F_z(.)$ with $E(z_t|\mathcal{F}_{t-1}) = 0$ and $E(z_t^2|\mathcal{F}_{t-1}) = 1$, $\mathcal{F}_{t-1}$ being the information set of all returns up to $t-1$.

In this context, Bollerslev [7] early emphasized the usefulness of nonnormal density functions. Later, Bai et al. [1] also noticed that the error distribution needs to exhibit strong excess kurtosis in GARCH models to ensure a theoretical unconditional kurtosis coherent with empirical evidence. Other authors [19, 22] highlighted the importance of the distribution assumption for the quality of Value-at-Risk (VaR) and expected shortfall (ES) forecasts. Meanwhile, when modeling conditional variance parametrically (e.g. with a GARCH model or one of its variants - see Francq and Zakoian [17] for a review), resulting estimators relying on the normal law may suffer from weak efficiencies. For example, Engle and Gonzalez-Rivera [15] showed that, for nonnormal data, the loss of efficiency can be up to 84% when estimating parameters with a Maximum Likelihood (ML) procedure based on the normal distribution.

These considerations lead researchers to propose alternative flexible probability density functions with heavy tails for z_t in model (1) to take these issues into account. The most common distributions are the Student's t-distribution and its generalization, the skewed-t distribution [16]. Its use in GARCH or EGARCH models considerably improves forecasts [7, 22, 28]. Another well-known family of functions is the generalized error distribution (GED).Christoffersen et al. [12] show its nice fitting characteristics on daily stock returns for different GARCH-type models. Some studies also focus on the Generalized Hyperbolic (GH) distributions, a five-parameters family of density functions, introduced first by Barndorff-Nielsen [3]. Eberlein and Keller [14] and Bingham and Kiesel [6] note the interesting goodness-of-fit (GOF) performance of these density functions for daily stock returns. More recently, Stavroyiannis et al. [37] also propose to use the Pearson type-IV distribution. Other flexible distributions not yet considered in the finance literature could be also used [see, for example 26].

Nevertheless, the coexistence of so many distributions reflects the fact that few articles concentrate on the comparison and the selection of an adequate distribution, despite the large number of available ones. Moreover, most of articles on the subject study this issue in the framework of GARCH-type models. Basically, the traditional approach consists in testing the fit of specific distribution families using GOF tests on the estimated innovations (i.e. $\hat{z}_t = r_t/\hat{\sigma}_t$) obtained using a GARCH-type estimator of the conditional variance [see, 30, for a detailed review]. But the drawback of this approach is a possible misspecification error due to the parametric variance assumption. Indeed, parametric variance models often exhibit a lack of flexibility: among others, Lamoureux and Lastrapes [29] show that GARCH models are extremely sensitive to misspecified structural breaks, Bali and Guirguis [2] point out that variance model misspecifications can cause an overestimation of the kurtosis in the estimated residuals and Jalal and Rockinger [24] emphasize the negative impact of a variance misspecification on the estimation of tail-related risk measures. Consequently, all specification and validation procedures based on these so-estimated residuals are very sensitive to the type of

variance model used. Besides, different distributional assumptions might not be rejected by classical GOF tests. In these cases, AIC, BIC or HQC criteria can help identifying the best assumption, but no formal procedure exists in our context.

This study suggest another approach, based on a two-step methodology. First, following the work of Heuchenne and Van Keilegom [23], we propose to use a nonparametric estimation of the conditional variance, instead of a classical GARCH-type estimation. This approach is a robust alternative that avoids the risk of a misspecified parametric variance. More particularly, beyond standard regression technique [32, 38], Mercurio and Spokoiny [31] developed a local constant model for the estimation of the conditional volatility (LAVE), consisting of a moving average of past squared returns over time intervals of varying lengths [25]. The advantages of this method are its ability to quickly react to jumps occurrences and its interval selection procedure independent from the true distribution of the error terms [8, 25]. Chen et al. [8] successfully applied this technique in a multiplicative model of type (1)-(3) and showed its good performance in one-day-ahead VaR forecasts.

Second, we suggest a set of estimation and model selection procedures for the error distribution, assessing both the global fit and the fit in the tails. Instead of relying on classical GOF tests like Chi-squared and Kolmogorov-Smirnov (K-S) tests (known for their lack of power), or any other single measure of the fit, we suggest to adapt four different statistics to our situation: kernel density-based selection test and Berk and Jones [5] test for an assessment of the global fit; [13] weighted likelihood scores and empirical risk level (ERL) tests to focus on the behavior in the tails. Indeed, as a model (especially in Finance) is never true, one is most often interested in finding a model that correctly fits the data rather than discovering the true model. With the proposed combination of tools, we are able to map the strengths and weaknesses of the candidate distributions. It allows researchers to decide which distribution should be preferably used in a heteroscedastic multiplicative model, according to its objective. For example, if one is more interested in ES forecasts, he could decide to favour a distribution that performs very well in the tail, even though its global fit is not the best among all tested distributions. Moreover, using several measures of the fit should allow to detect more easily possible differences between distribution, compare to a single aggregated measure (like the Anderson-Darling statistic) that averages the difference and lacks of power. The finite sample behavior of the proposed statistics are investigated in a simulation study.

Finally, we give an empirical illustration of our methodology on three daily returns time series (EUR/USD exchange rate, BOVESPA index and UBS stock) where we compare the Normal Inverse Gaussian (NIG), hyperbolic (HYP) and skewed-t distributions. These three different distribution can account for leptokurtosis and asymmetries, which make them natural candidate in a financial model of type (1) where such features have been observed.

The rest of the paper is organized as follows: in Section 2, we present the LAVE standardization technique and the different goodness-of-fit indicators used. In Section 3, we present the results of the simulation study, while Section 4 is devoted to the presentation of the empirical study. We conclude and discuss in Section 5.

2. Method

2.1 *Local Adaptive Volatility Estimation (LAVE)*

To estimate the conditional variance ($\hat{\sigma}_t$) without any risk of misspecification, we suggest to use the nonparametric LAVE technique [25, 31]. For all r_t, $t = 1, ..., n$, we compute $\hat{\sigma}_t$ using I_t previous squared returns $r^2_{t-1}, ..., r^2_{t-I_t}$:

$$\hat{\sigma}_t = (1/I_t) \sum_{i=1}^{I_t} r^2_{t-i}, \tag{4}$$

with I_t the local window length at time t. To select I_t, defined as an *interval of homogeneity*, we follow the step-by-step procedure detailed in Jeong and Kang [25], based on a power transform of r_t and a simple t-test (we have implemented this procedure in MatLab, files are available upon request to the authors). Starting from model (1)-(3), we consider that some $\gamma > 0$ exists such that,

$$|r_t|^\gamma = \sigma_t^\gamma |z_t|^\gamma = E|z_t|^\gamma \sigma_t^\gamma + \sigma_t^\gamma (|z_t|^\gamma - E|z_t|^\gamma) = \theta_t + \sigma_t^\gamma (|z_t|^\gamma - E|z_t|^\gamma), \tag{5}$$

where $\theta_t = E|z_t|^\gamma \sigma_t^\gamma$ [25]. The null hypothesis of a constant variance on I_t implies a constant trend $\theta_t = \theta_{I_t}$ for all $t \in I_t$. This trend can be approximated by the average of $|r_t|^\gamma$ over I_t :

$$\hat{\theta}_{I_t} = (1/I_t) \sum_{i=1}^{I_t} |r_{t-i}|^\gamma,$$

This estimation is used in a sequence of t-tests to select I_t as the largest interval of homogeneity. The related asymptotic theory and the detailed hypothesis tests used can be found in Jeong and Kang [25].

As explained previously, using nonparametric estimators also makes sense, since it is impossible to know the exact structure of the volatility process. Moreover, Chen et al. [8] show throughout simulations that GARCH model and the LAVE provide estimations of similar quality for various kinds of variances and nonnormal innovations. This question is beyond the scope of this paper but additional simulations are available upon demand.

2.2 *Estimation and model selection procedures*

We use the LAVE to obtain estimated innovations ($\hat{z}_t$). Then, we use both estimation and model selection procedures, to assess the quality of a single distribution and to compare two competing kind of distributions.

2.2.1 Goodness-of-fit of the whole distribution

First, we propose to use a GOF test that assesses the global fit of different density functions candidates: the Berk-Jones test (5 and more recently, 39), based on the empirical cumulative distribution function of the estimated innovations. We compute the likelihood of each estimated innovation $\hat{Z} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_n\}$, both under a tested parametric

hypothesis F_θ (an assumed parametric family under H_0) and using the empirical cdf F_n built on $\hat{Z}$. In the present situation, the B-J statistic is defined by:

$$R_{n,F_\theta} = \sup_x n^{-1} \log \left[\left(\frac{F_n(x)}{F_{\hat{\theta}}(x)} \right)^{nF_n(x)} \left(\frac{1 - F_n(x))}{1 - F_{\hat{\theta}}(x)} \right)^{n(1-F_n(x))} \right], \tag{6}$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) of θ under the assumed family F_θ. We reject the parametric hypothesis (i.e. the null hypothesis H_0 that $F = F_\theta$) if this statistic is too large.

This goodness-of-fit test provides an interesting assessment of the quality of the fit, as it does not use any bandwidth parameter. Nevertheless, the limit distribution of this statistic is only known for directly observable data. In our case, the innovations are not observable and we work with estimated residuals obtained after a nonparametric standardization. As explained by Heuchenne and Van Keilegom [23], the bootstrap is a good solution to derive the bounds of the critical region for a statistic of interest and to build hypothesis tests accordingly. Consequently, we apply the following parametric bootstrap procedure to find the critical bound of the statistic, under the null hypothesis that the innovations are F_θ distributed:

For $i = 1, ..., N$,

(1) Generate randomly n i.i.d. innovations $Z_i^* = \{z_{i,1}^*, ..., z_{i,n}^*\}$ from the parametric distribution $F_{\hat{\theta}}$.
(2) Multiply each resampled innovation by the corresponding estimated volatility $\hat{\sigma}_t, t = 1, ..., n$.
(3) We obtain $R_i^* = \{r_{i,1}^*, ..., r_{i,n}^*\}$, a particular realization of the returns sample in the bootstrap world.
(4) Estimate the conditional volatilities $\hat{\sigma}_{i,t}^*$ by LAVE, $t = 1, ..., n$.
(5) We obtain $\hat{Z}_i^* = \{\hat{z}_{i,1}^*, ..., \hat{z}_{i,n}^*\}$.

For each hypothesis to test, we obtain N resamples for each dataset leading to N realizations of R_{n,F_θ}. The null hypothesis is rejected if the statistic computed on the original sample is higher than the quantile $1-\alpha$ of these realizations (one-sided test).

Second, we propose to use a statistic relying on the *kernel density estimator* $\hat{f}(x)$ of the estimated residuals to determine which distribution displays the best fit. We compute the bandwidth using the normal rule [36]. Based on that estimated density, we compute an estimator ($KIMSE_{f_{\hat{\theta}}}$ hereunder) of the integrated mean squared error between the true and the parametrically estimated ($f_{\hat{\theta}}$) densities, defined by:

$$IMSE_{f_{\hat{\theta}}} = E[\int_{-\infty}^{\infty} (f_{\hat{\theta}}(x) - f(x))^2 dx]. \tag{7}$$

This time, we use a nonparametric bootstrap procedure to compute $KIMSE_{f_{\hat{\theta}}}$:

For $i = 1, ..., N$,

(1) Generate randomly n i.i.d. innovations from the historical distribution of the estimated innovations $\hat{Z} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_n\}$.
(2) Multiply each resampled innovation by the corresponding estimated volatility $\hat{\sigma}_t$.

(3) We obtain $R_i^* = \{r_{i,1}^*, ..., r_{i,n}^*\}$, a particular realization of the returns sample in the bootstrap world.
(4) Estimate the conditional volatilities $\hat{\sigma}_{i,t}^*$ by LAVE, $t = 1, ..., n$.
(5) We obtain $\hat{Z}_i^* = \{\hat{z}_{i,1}^*, ..., \hat{z}_{i,n}^*\}$.

Once again, we obtain N resamples for each dataset to compute $KIMSE_{f_{\hat{\theta}}}$ defined by:

$$KIMSE_{f_{\hat{\theta}}} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}(f_{\hat{\theta}_i^*}(x) - \hat{f}(x))^2 dx = \frac{1}{N}\sum_{i=1}^{N} KISE_{f_{\hat{\theta}}}^i, \tag{8}$$

where $\hat{f}(x)$ is the kernel density estimation built on the initial estimated innovations $\hat{Z}$, $f_{\hat{\theta}_i^*}$ is the parametric estimation of the error distribution and $\hat{\theta}_i^*$ are the MLE based on $\hat{Z}_i^*$, $i = 1, ..., N$. For N sufficiently large, this quantity is approximately normally distributed, given the initial sample $\hat{Z}$. Indeed, the bootstrap procedure ensures the conditional independence between the $KISE_{f_{\hat{\theta}}}^i$ for $i = 1, ..., N$. If now the goal is to compare two $IMSE_{f_{\hat{\theta}_j}}$, $j = 1, 2$, (i.e. $E[\int_{-\infty}^{\infty}(f_{\hat{\theta}_j}(x) - f(x))^2 dx]$) we can simply use the following statistic $\bar{D}$ for paired data:

$$\bar{D} = \frac{1}{N}\sum_{i=1}^{N}[KISE_{f_{\hat{\theta_1}}}^i - KISE_{f_{\hat{\theta_2}}}^i] = \frac{1}{N}\sum_{i=1}^{N} D_i, \tag{9}$$

Indeed, taking differences makes now the D_i, $i = 1, ...N$, are i.i.d. given $\hat{Z}$ and consequently:

$$\sqrt{N}\bar{D} \to N(0, \sigma_D^2), \tag{10}$$

under the null hypothesis that $IMSE_{f_{\theta_1}} = IMSE_{f_{\theta_2}}$. Using the empirical bootstrap variance of D_i, $\hat{\sigma}_D^2$, as an estimate of σ_D^2, a simple standardization gives us

$$\Delta = \bar{D}/\sqrt{\hat{\sigma}_D^2/N} \to N(0, 1), \tag{11}$$

given $\hat{Z}$ and if the observed $|\Delta| \geq \Phi^{-1}(1 - \alpha/2)$, the null hypothesis can be rejected with a test level α.

Notice that the spirit of this bootstrap procedure is different from the previous one. Indeed, here we replicate the observed data to get an estimator of $IMSE_{f_{\hat{\theta}}}$. For the B-J test, we generate data from a given parametric null hypothesis to get estimators of the critical bound of a statistic, under this hypothesis.

2.2.2 Goodness-of-fit in the tail of the distribution

As mentioned in the previous subsection, $KIMSE_{f_{\hat{\theta}}}$ and the B-J test both take into account the fit of the whole distribution. In VaR modeling, we need to focus on a specific quantile of the innovations distribution (let's say of order p) and on the fit in the tail. To measure the quality of the quantile estimation provided by the parametric method,

we first define $q_\theta(p)$, as the quantile function of the density f_θ (i.e. if a r.v. $X \sim f_\theta$, $P(X \leq q_\theta(p)) = p$). The idea is to estimate the difference between p and so-named empirical risk level (ERL) $p_{\hat{\theta}}$ given by:

$$p_{\hat{\theta}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}(\hat{z}_t \leq q_{\hat{\theta}}(p)), \tag{12}$$

where $\hat{\theta}$ are the MLE obtained from the initial sample $\hat{Z}$.

Again, we use the same bootstrap procedure as in the B-J test to obtain a bootstrap estimation of the critical bounds (α level two-sided test) for the corresponding statistic $(p_{\hat{\theta}} - p)$. Then, we are able to test if the quantile of order p of the true innovations distribution $(F_z^{-1}(p))$ is significantly different from the quantile of the same order for the assumed parametric distribution. In the latter case, the assumed parametric assumption can be rejected. Conceptually, this test can be related to the coverage test of Christoffersen [11], but applied in-sample on robust estimated innovations.

The weakness of this test is that it compares the quality of the fit of a particular distribution with respect to the true (unknown) distribution, using only a specific point of the estimated distribution. To compare the fit in the tail provided by different candidates, we need a selection test (i.e. comparing two fits) that gives a particular weight to the left tail of the distribution. Following that idea, we propose to use the selection test of Diks et al. [13], based on a weighted Kullback-Leibler Divergence (KLD). As explained in Diks et al. [13], we can test the relative accuracy of two candidate conditional distribution of the returns, g_t^1 and g_t^2, by taking the difference of their weighted KLD, at each observable r_t. This quantity can be estimated by the empirical mean $\bar{d}^{wl}$ of the weighted scores differences $d_t^{wl}, t = 1, ..., n$:

$$\bar{d}^{wl} = \frac{1}{n}\sum_{t=1}^{n} d_t^{wl} = \frac{1}{n}\sum_{t=1}^{n}(S^{wl}(\hat{g}_t^1; r_t) - S^{wl}(\hat{g}_t^2; r_t)), \tag{13}$$

with

$$S^{wl}(\hat{g}_t^j; r_t) = \mathbf{1}(r_t \in A)\log(\hat{g}_t^j(r_t)) + \mathbf{1}(r_t \in A^c)\log\left(\int_{A^c} \hat{g}_t^j(s)ds\right),\ j = 1, 2, \tag{14}$$

where $\hat{g}_t^j$ is an estimator of g_t^j,$j = 1, 2$, A is the region of interest for the fit and A^c its complement. In the empirical application, we use two different regions of interest: the 5% first observations and the 1% first observations (which are the classical test levels for VaR). The assumed conditional distributions of the returns g_t^j are linked to the distributions of the innovations f_{θ_j} through the following relationship:

$$g_t^j(r_t) = \frac{1}{\sigma_t} f_{\theta_j}(r_t/\sigma_t), j = 1, 2. \tag{15}$$

Parameters estimators for f_{θ_j} are the same as the ones used in the previous tests (thus,

MLE obtained on the whole sample of estimated innovations) and $\hat{\sigma}_t$ are computed using the LAVE. The set of d_t^{wl} is not i.i.d. but using the following statistic,

$$T = \frac{\bar{d}^{wl}}{\sqrt{\hat{\sigma}_n^2/n}}, \tag{16}$$

with $\hat{\sigma}_n^2$ being a heteroscedasticity and autocorrelation-consistent (HAC) estimator of the variance of $\sqrt{n}\bar{d}^{wl}$, we can test if $\bar{d}^{wl} = 0$. Indeed, Giacomini and White [18] demonstrate that the statistic given by equation 16 is asymptotically normally distributed assuming $\bar{d}^{wl} = 0$ and under very weak conditions (see this article and Wooldrige and White, 1988, for more details). In particular, it allows using both para- and nonparametric estimators in the computation of $\hat{g}_t^j$. For $\hat{\sigma}_n^2$, we use the same HAC estimator as in Diks et al. [13] and Giacomini and White [18]:

$$\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{G-1} a_k \hat{\gamma}_k, \tag{17}$$

where $\hat{\gamma}_k$ is the lag-k sample autocovariance of the sequence of d_t^{wl}, $a_k = 1 - k/G$, $k = 1, \cdots, G-1$, are the Bartlett weights and $G = \lfloor n^{1/4} \rfloor$ (where $\lfloor x \rfloor$ denotes the integer part of x). Given $\hat{Z}$, we test that $\bar{d}^{wl} = 0$ (i.e. the null hypothesis of an equal quality of the fits) and we reject this hypothesis if $|T| \geq \Phi^{-1}(1-\alpha/2)$ with a test level α. Moreover, if $|T| \geq \Phi^{-1}(1-\alpha/2)$ and $T < 0$ (respectively $T > 0$), then we conclude that f_{θ_1} (respectively f_{θ_2}) better fits the data.

The statistic in equation 16 has some interesting properties. First, it is a relative measure of the fit between two distributions, such that we don't need the true unknown distribution or a proxy of it. Second, this weighting scheme allows assessing the fit in the tail by controlling the impact of the central observations on the statistic: the censoring of the returns outside A allows ignoring the shape of the density function in this region. Moreover, the second term of the censored score in equation 14 avoids a possible selection bias if the tails' thickness of the compared density functions are different [13].

To summarize our approach, after the LAVE standardization, we propose to compute the four different measures of the fit presented in this section. B-J and $p_{\hat{\theta}}$ statistics are used in GOF tests to assess individually the correctness of the tested distributions while we use KIMSE and $\bar{d}^{wl}$ statistics in pairwise comparative tests to determine if some distributions have a significantly higher GOF performance than the others. In the pairwise comparison tests, we assess if one of the two distributions fits better the data than the other (if we reject the null hypothesis, we conclude that one distributions is better), using the Δ and T test statistics. Moreover, we assess both the global fit (with the B-J and Δ statistics) and the fit in the tail (with the $p_{\hat{\theta}}$ and $\bar{d}_{wl}$ statistics).

3. Practical implementation and simulations

In this section, we study the finite sample behavior of the proposed methodology. Due to the large number of observations needed and the bootstrap procedure, the computation time is quite extensive. Therefore, we only focus on three different data generating processes (DGP), combining either GARCH(1,1) or GJR-GARCH(1,1,1) [20] conditional variances with innovations distributed according to some usual parametric distributions

(we avoid the case of nested distributions, that is beyond the scope of this paper).

3.1 *Simulation set-up*

We make use of MatLab 2013a for all implementations. We use equations 1 to 3 to generate the data, with equation 2 being either a GARCH(1,1) process:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{18}$$

or a GJR-GARCH(1,1,1) process:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \mathbf{1}(r_{t-1} \leq 0)\phi r_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{19}$$

We assumed three different distributions for z_t:

$$z_t \overset{iid}{\sim} T(\nu), \tag{20}$$

$$z_t \overset{iid}{\sim} skewed - t(\lambda, \nu), \tag{21}$$

$$z_t \overset{iid}{\sim} hyperbolic(\theta_1, \theta_2, \theta_3, \theta_4), \tag{22}$$

with their two first moments equal to 0 and 1 respectively. We consider the three following DGP:

DGP1: GARCH(1,1) with $\omega = 10^{-4}, \alpha = .05$ and $\beta = .92$ combined with T-distributed z_t where $\nu = 5.4$.

DGP2: GJR-GARCH(1,1,1) with $\omega = 10^{-4}, \alpha = .05$, $\phi = .1$ and $\beta = .8$, combined with skewed-t-distributed z_t where $\lambda = 0.9$ and $\nu = 7$.

DGP3: GARCH(1,1) with $\omega = 10^{-4}, \alpha = .20$ and $\beta = .75$ combined with hyperbolic (HYP) distributed z_t where $\theta = [4.5329, 3.4371, 0.3263, -1.05]$.

Notice that all distributions parameters are chosen to exhibit leptokurtosis, and even asymmetries for the skewed-t and hyperbolic cases. Also, the sum of the variance parameters is close to unity to mimic typical GARCH parameters found on empirical data (see, e.g., Chen and Lu [10]).

For each DGP, we generate $B = 1000$ samples of size $n = 600$. For the computation of the LAVE, we set in equation 4, $m_0 = 5$ and $\gamma = .5$ [see 25]. In addition, in Jeong and Kang [25], the level of the multiple test in the LAVE computation is 0.05 (at iteration k - $(k+1)m_0$ data under study- it is divided by k to obtain the level of each separate test). We use the same procedure. To improve the computing time, I_t is bounded to a size of 200 observations. Data from time $t = 1$ to $t = 200$ are used as initial training set. We compare the true distribution (Student, skewed-t or hyperbolic) to an alternative distribution having the same number of unknown parameters (i.e. Student to GED

distributions, skewed-t and hyperbolic distributions to NIG distributions). Parameters estimates are obtained via MLE on the estimated residuals $r_t/\hat{\sigma}_t$. Based on these estimators, we compute the four different measures of the fit (Berk-Jones statistic, KIMSE, $p_{\hat{\theta}}$ and S^{wl}) for each sample and for each parametric hypothesis, using equations 6, 8, 12 and 14 proposed in Section 2. Using the Berk-Jones statistic and $p_{\hat{\theta}}$, we perform GOF tests for each tested distribution. With the KIMSE statistic and S^{wl}, we build the test statistics Δ and $\bar{d^{wl}}$ (standardized to give our statistic T) and perform pairwise comparisons between the true distribution and the tested alternative one. When the bootstrap is needed, we generate $N = 200$ resamples. Notice also that due to the small size of our samples, we compute the ERL statistics with levels of p equal to 30%, 20%, 15% and 10%. In the KLD tests, we use a region of interest A that contains respectively 20%, 10%, 5% and 1% of the data (it corresponds to censoring of 80%, 90%, 95% and 99% of the sample). We also repeat the corresponding selection tests with samples of size 1000, 2000 and 3000. The test level of all tests is set at 5%.

3.2 *Simulation results*

We observe that, for DGP 2 and DGP 3, the Berk-Jones tests exhibit low but satisfactory powers (Table 1). The tests seem a bit conservative, though, probably due to the small size of our samples (the type-I error is too low). For DGP1, the tests are not very powerful against the alternative considered. It is not surprising, as Student's t-distributions can be well approximated by GED distributions. This low level of rejections indicates that the GED distribution fits quite well the true Student's t-distribution. Overall, GOF tests based on the whole distribution can hardly reject wrong parametric hypotheses because the tested alternative distributions are very flexible. Thus, they approximate very well the true distribution.The Berk-Jones test suffers from similar drawbacks. It is one of the reasons that makes the traditional approach (based on a single GOF test) ineffective and it illustrates that combining different measures of the fit is essential.

Concerning the ERL tests (Table 1), the performance is mixed. For the second DGP, we reject the alternative hypothesis quite often for all values of p tested. However, the type-I errors are a bit too high for DGP1 and DGP3. It illustrates a weakness of this test: because we use a single point of the distribution to reject or not a parametric hypothesis, the test tends to over-reject and to not be powerful. Especially, in the case of the third DGP, these results could be attributed to the parameters estimation of the hyperbolic distribution. Indeed, as noted by [4], the likelihood functions of hyperbolic distributions are quite flat. It could cause to provide parameters estimates that fits very well the centre of the distribution at the expense of the tails. In our case, this effect could be also reinforced by the filtering process of the variance and the bootstrap procedure The results for both the ERL and the BJ tests are displayed in Figures 1 and 2.

Using the KIMSE statistic (Table 2), we are able to detect significant differences between the true distributions and the alternative ones in all DGP, in proportions ranging from 14.4% to 31.2%. Once again, for the first DGP, we detect a difference in favour of the true distribution only 14.4% of the time. These values are obviously affected by the alternative tested. Nevertheless, these tests are quite useful to detect the distribution that best fits the whole distribution of the data (obviously, the true one). Especially, it prevents us to select a wrong distribution, the proportion of times the alternative distribution is selected being always very close to 0.

The tests based on the censored likelihood score (Table 3 to 5) bring a different

perspective to the analysis. Using an uncensored statistic, this test selects quite adequately the true distribution (column $\bar{d}^{wl} < 0$), especially when the size of the sample increases (by going from 600 observations to 1000, we almost double the proportion where the correct distribution is picked up by the test, for the two first DGP). Using the censored statistics, we observe for all DGP that in the case of small-size samples (i.e. 600 observations), we select more often the true distribution compared to the test with the uncensored statistic. Thus, the censoring procedure seems to improve the detection of differences for these sample sizes. However, we select also more often the alternative distribution (column $\bar{d}^{wl} > 0$), compared to the test with the uncensored statistic (and that holds for all regions of interest and sample sizes tested). In fact, even if these results seem counter-intuitive at first, they are not that surprising, as nothing guarantees that the true distribution has the highest **censored** likelihood score: indeed, we obtain estimations of the parameters using ML techniques **based on the whole sample**. Therefore, if we use the true distribution, it tends to guarantee an estimated distribution with the lowest possible Kullback-Leibler divergence, but not one with the highest censored likelihood score. When the alternative is selected, it means that, due to the parameters estimation, this alternative has a significantly higher likelihood score in the selected tail than the true distribution with estimated parameters. A possibility to avoid these feature would have been to estimate the parameters using censored MLE. We would have had presumably lower selection ratios of the alternative, but also a less good estimation of the parameters. Figure 1 consists in the columns $\bar{d}^{wl} > 0$ for each DGP, arranged by the proportion of observation in the region of interest (A). It gives us an idea of how close the alternative distribution is from the true distribution, and in which region: for DGP1, GED appears relatively close to the Student's t-distribution for large regions of interest (e.g. 20%) but becomes less close when we go further in the tail. For DGP2, we observe the opposite effect, whereas the relative closeness between the considered distributions in DGP3 appears constant across censoring level. All this highlights that, due to the estimation of the parameters (based on a full ML procedure), the true distribution is not necessarily the one that fit best the tail of the data.

For a given region of interest, if we increase the sample size (i.e. to 2000 and 3000 observations), the selection ratios of the true distribution tend to increase (at all censoring levels). It is clearly less obvious for the alternative distribution. We also notice that for the second and third DGP, the selection ratios of the true distribution stay above the ones of the alternative, for all levels tested and all sample sizes. For the first DGP, for a region of interest consisting in 10% of the data, the selection ratio of the alternative is higher for a sample size of 600 but this effect disappears when the sample size increases.

Some could argue that working with sample sizes of 2000 or 3000 observations is unrealistic, but because we use the LAVE instead of parametric estimators of the conditional variances, we do not dread a possible parameter instability. Therefore we can make a full use of the available data (e.g., for stock returns, 10 years of data is not unusual).

Hence, in the perspective of selecting the distributions that best fit some parts of the data, these tests seem to exhibit interesting properties. In particular, this simulation study reveals the necessity to combine different measures of the fits to detect the various differences among the hypotheses tested. For instance, if the Berk-Jones tests and the ERL tests lack of power to reject the GED hypothesis, the KIMSE statistics and the censored likelihood scores could prove useful. Also, it shows that the censored likelihood scores (especially if we are far in the tail) often improve the selection of the true distri-

bution compared to the uncensored ones. Therefore, according to the objectives pursued by the modeller (e.g. VaR, ES or full density forecast), it might be interesting to map the strengths and weaknesses of the considered distribution with the different tests and then to favour the one that best fulfil our goals. Eventually, the simulations highlight the need for large samples, too.

4. Empirical illustration

In this section, we illustrate the proposed methodology on three different time series, where we test three different distributions for the innovations. Indeed, recent works emphasize the flexibility of GH subfamilies [8, 9, 34] and skewed-t distributions [22, 28]. Therefore, we will compare the NIG, the HYP (i.e. subfamilies of the GH distributions for λ equal respectively to $-1/2$ and 1) and the skewed-t [21]. We do not consider the Student-t and GED distributions, as these distributions are special cases of the other distributions [see, among others 22]. Moreover, because the filtered returns exhibit consequent asymmetries (see Table 7), it seems inadequate to consider symmetric distributions. Details concerning the GH and the skewed-t distributions can be found in Appendix B.

The goal here is to identify the most adequate distribution(s) to model the stochastic behaviour of the considered data, without bearing the risk of a misspecified parametric volatility. As it is very likely than none of the considered distribution is the true error distribution, we focus on finding the one that display the best fit, taking into account the estimation of the parameters and along the four dimensions described in the Methodology section. We compare the results of our approach with those obtained with three other measures of the fit, computed after a GARCH filtering of the data (thus, bearing the risk of a misspecification): the Kolmogorov-Smirnov statistic and the Anderson-Darling statistic (as in [27]), as well as the m-statistic proposed by [30]. The distributions of the two first statistics under the null hypothesis that the residuals stem from a particular distribution F_θ are computed with a bootstrap procedure. For the m-statistic, we use the same asymptotic result as in [30].

4.1 *Data*

We applied the proposed methodology on three different time series :

(1) Stock returns data : UBS daily returns for the period 10 June 2003 - 7 June 2013,
(2) Stock index data : BOVESPA daily returns for the period 4 January 1999 - 12 April 2012,
(3) Exchange rate data : EUR/USD daily returns for the period 15 June 2000 - 10 October 2012.

The prices have been extracted respectively from www.nasdaq.com, www.finance.yahoo.com and www.federalreserve.gov. We compute the daily log-returns from these prices ($r_t = \log(P_t/P_{t-1})$). Samples have respectively 2517, 3282 and 3215 observations. Notice also that UBS prices have been adjusted for the 2:1 stock split of 10th July 2006. A first exploratory analysis reveals also that an AR(1) (with no significant intercept) is suitable to model the conditional mean of UBS stock returns. Thus, before applying the proposed methodology, we correct this series by removing its conditional mean using the estimated AR(1) parameters. For the other time series, autocorrelations and partial autocorrelations are not significantly different from 0. We

also test for mean nonstationarity using augmented Dickey-Fuller tests with 21 lags. The unit-root hypothesis is rejected at the 99% level for all series. Finally, a graphical analysis indicates that we have series exhibiting heteroscedasticity (Figure 4) and high significant autocorrelations of the squared returns at multiple lags, indicating that equation 1 is suitable to model these returns. Graphs and detailed results of the tests can be found in Appendix A.

4.2 *Results*

4.2.1 Kolmogorov-Smirnov, Anderson-Darling and m-tests

For each sample, we remove the conditional volatility using a GARCH(1,1) model. Then, we estimate the parameters of the four parametric hypotheses and compute the various test statistics on the residuals. We use a parametric bootstrap based on the estimated parameters of the tested distributions and on the estimated GARCH(1,1) parameters to obtain the distributions of the statistics under the various null hypotheses (the details of this procedure are available upon demand to the authors). The results of the three GOF tests can be found in Table 6. We observe that a rejection occurs at the 5% test level only for three tests out of the 27 performed. When we have the asymptotic distribution of the test statistic, we also compute robust bootstrap p-values, but results stay alike. High p-values of the KS and the AD statistics suggest that the tested distributions fit correctly both the tail and the entire distribution of the random part. The more elaborate m-tests reject both NIG and skewed-t distributions for EUR/USD data. However, we do not specifically know for which reason (i.e., if it is due to asymmetry, kurtosis, etc.). We are now stuck with a set of distributions identified as equally good. We could use the p-values of the test to "rank" the distributions, but it is not possible to know if there are significant differences. Moreover, these results are subject to the correct specification of the GARCH(1,1) model. To try to circumvent these issues, we apply our approach.

4.2.2 LAVE standardization

In the LAVE computation, we set $m_0 = 5$, and $\gamma = 0.5$, as recommended in Chen et al. [8] and Jeong and Kang [25]. Figure 5 shows the estimated conditional standard deviations with this method and Figure 6 the residuals obtained after standardization. Descriptive statistics of the residuals are presented in Table 7. As expected, the kurtosis coefficients are higher than 3 and the skewness coefficients are lower than 0, indicating leptokurtosis and negative skewness. The interval where the estimated innovations take their values seems rather constant along the time, suggesting a correct standardization. Estimated parameters for all time series and for the four distributions are listed in Appendix B.

4.2.3 Fits comparisons

We use the estimated residuals to perform the tests described in Section 2. When the bootstrap is needed, we run 1000 resamples. The Berk-Jones tests do not reject any of the tested distributions (Table 8). It is not very surprising, because all distributions tested are quite flexible (they can all model asymmetries and leptokurtosis). If we stop our analysis here, it is not easy to determine if some distributions could best fit the data and we are stuck in the same situation as with the classical GOF tests. Therefore, we compute the KIMSE statistic (Table 9) and performed pairwise comparisons with the Δ statistic (Table 10). We observe that:

- the skewed-t distribution has the lowest statistic for the three series,
- significant differences are detected between NIG and skewed-t distributions, as well as between HYP and skewed-t, in favor of the skewed-t distribution (Table 10),
- no difference is detected between NIG and HYP distributions.

Hence, it seems that skewed-t distributions provide the best fits for these datasets. At the contrary, the HYP appears to have the worst fit (in term of absolute value of the KIMSE statistic).

Nevertheless, until now we only focused on the goodness-of-fit of the whole distribution. Can we also detect differences between the goodness-of-fits in the tails? The results of the tests based on the ERL statistic ($p_{\hat{\theta}}$) with 5% and 1% quantiles (typical quantiles used for VaR computations), are displayed in Table 11. Globally, the three considered distributions do not performed well. For UBS time series and $p = 5\%$, we identify the HYP and the NIG distribution as being the bests to model. For $p = 1\%$, NIG is also found to be the best distribution. For he BOVESPA, all distributions are rejected. However, the HYP (for the 5% quantile), and both the NIG and HYP (for $p = 1\%$) are found to be the closest to p. For EUR/USD, all distributions are rejected for $p = 5\%$ (HYP is the closest one) and none are rejected for $p = 1\%$ (but once again, HYP is the closest to p. Hence, it appears that both the NIG and the HYP seem to be the most adequate distributions for the 1% quantile of the UBS, BOVESPA and EUR/USD time series respectively.

The results of the tests based on the censored likelihood scores are displayed in Table 12. If we compute the scores without censoring, results are similar to the ones deduced from the KIMSE statistics (the skewed-t distribution provide the best fits). With the region of interest being the 5% tail (95% censoring), we can conclude for the UBS series that the NIG distribution provides a significantly higher score (hyperbolic and skewed-t distributions have significantly lower scores). With the region of interest being the 1% tail (99% censoring), skewed-t and NIG both appear better than the HYP distributions. For the BOVESPA time series, we can conclude that with the region of interest being the 1% tail, the skewed-t provides significant higher scores. Not enough significant differences can be detected for the scores of the EUR/USD time series, indicating that all the distributions tested provide similar goodness-of-fits.

These results differ in several ways to the ones obtained with the KS and AD statistics. While the KS and AD tests suggest that all distributions are more or less equivalent, our set of tests detects significant differences for the global fit but also for the fit in the (left) tail. The m-tests lead to different conclusions as well: they reject the skewed-t and the NIG distributions for the EUR/USD data, while our tests suggest that the skewed-t fits better the whole distribution of the data than the NIG and HYP distribution.

5. Conclusion

In this article, we contribute in two ways to the existing literature. First, we develop a statistical approach to compare the GOF of different density functions independently from a parametric variance model. We propose a method to identify and select the most appropriate error distributions in the framework of a classical multiplicative heteroscedastic model. This methodology enables a GOF analysis robust to a model misspecification, unlike traditional approaches relying on GARCH-type filtering. It also allows to use large samples without being restricted by some parameters stationarity

hypothesis. Moreover, we adapt estimation and model selection tests to this context by providing a suitable bootstrap algorithm. We also pay attention to assess not only the global fit of candidate distributions but also the fit in the (left) tail. Indeed, some of the proposed selection tests focus specifically on the left tail of the distribution and can be useful to choose an adequate distribution in the perspective of VaR or ES modeling. It would be possible to use a single statistic (like the Anderson-Darling statistic) combining both perspectives, but the risk is to be stuck with a statistic neither good to assess the global fit nor the fit in the tail. Therefore, we prefer to use a two-step procedure to distinguish the global fit from the fit in the tail. A simulation study indicates good powers of the selection tests based on the KIMSE statistic and the censored likelihood score but also highlights the need of large samples, a requirement easily met with financial time series.

Second, we illustrate our methodology in an empirical study where we compare the GOF of three different distributions (skewed-t, NIG and HYP distributions). We show, on financial time series of various kinds (stock returns, emerging market index returns and exchange rate returns), that the skewed-t distribution seems to be the best error distribution at the global level for all series. Regarding the fit in the tail, skewed-t appears very good in the 1% and 5% tails for the BOVESPA (but not for the associated quantile). However, the NIG and the HYP distributions could be more suitable if we focus respectively on the left tail or a quantile far in the tail for UBS.

More generally, both the simulations and the empirical study emphasize the necessity to combine different measures of the fit to detect possible differences between distributions. The use of the proposed tests would be a first step in selecting more properly an error distribution, that could be use later in parametric models or forecasts. Last, as interestingly pointed out by one of the reviewers, an extension of this paper would be to use the proposed statistics to combine the candidate distribution into a new piecewise function. Using model combination technique, like ensemble modeling, we could potentially combine together the various distributions in an efficient way to improve the final fit.

Acknowledgement

J. Hambuckers acknowledges the support of the Belgian National Fund for Scientific Research (F.N.R.S) with a research fellow grant.

C. Heuchenne acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet dActions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté franaise de Belgique', granted by the 'Académie universitaire Louvain.

The authors warmly thank Chris Legault and two other anonymous reviewers for the constructive and helpful comments.

References

[1] Bai, X., Russel, J., Tiao, G., 2003. Kurtosis of GARCH and stochastic volatility models with non-normal innovations. Journal of Econometrics 114, 349–360.

[2] Bali, R., Guirguis, H., 2007. Extreme observations and non-normality in ARCH and GARCH. International Review of Economics & Finance 16, 332–346.
[3] Barndorff-Nielsen, O.E., 1978. Hyperbolic distributions and distributions on hyperbolae. Scandinavian Journal of Statistics 5, 151–157.
[4] Barndorff-Nielsen, O.E., Blaesild, P., 1981. Hyperbolic distributions and ramifications: contributions to theory and application. Reidel, Dordrecht. volume 4. pp. 19–44.
[5] Berk, R., Jones, D., 1978. Relatively optimal combinations of test statistics. Scandinavian Journal of Statistics 5, 158–162.
[6] Bingham, N., Kiesel, R., 2001. Hyperbolic and semi-parametric models in finance. Disordered and Complex Systems 553, 275–280.
[7] Bollerslev, T., 1987. A conditional heteroskedasticity time series model for speculative prices and rates of returns. Review of Economics and Statistics 69, 542–547.
[8] Chen, Y., Härdle, W., Jeong, S.O., 2008. Nonparametric risk management with generalized hyperbolic distributions. Journal of the American Statistical Association 103, 910–923.
[9] Chen, Y., Härdle, W., Spokoiny, V., 2010. Risk analysis with gh distributions and independant components. Journal of Empirical Finance 17, 255–269.
[10] Chen, Y., Lu, J., 2010. Value at Risk estimation. Springer.
[11] Christoffersen, P., 1998. Evaluating interval forecasts. International Economic Review 39, 841–862.
[12] Christoffersen, P., Dorion, C., Jacobs, C., 2010. Volatility components, affine restrictions and non-normal innovations. Journal of Business and Economic Statistics 28, 483–502.
[13] Diks, C., Panchenko, V., van Dijk, D., 2011. Likelihood-based scoring rules for comparing density forecasts in tails. Journal of Econometrics 163, 215–230.
[14] Eberlein, E., Keller, U., 1995. Hyperbolic distributions in finance. Bernoulli 1, 281–299.
[15] Engle, R., Gonzalez-Rivera, G., 1991. Semiparametric arch models. Journal of Business & Economic Statistics 9, 345–359.
[16] Fernandez, C., Steel, M., 1998. On bayesian modelling of fat tails and skewness. Journal of the American Statistical Association 93, 359–371.
[17] Francq, C., Zakoian, J., 2010. GARCH Models : Structure, Statistical Inference and Financial Applications. John Wiley.
[18] Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.
[19] Giot, P., Laurent, S., 2004. Modelling daily value-at-risk using realized volatility and arch type models. Journal of Empirical Finance 11, 379–398.
[20] Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance 48, 1779–1801. doi:.
[21] Hansen, B., 1994. Autoregressive conditional density estimation. International Economic Review 35, 705–730.
[22] Hansen, P., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889.
[23] Heuchenne, C., Van Keilegom, I., 2010. Goodness-of-fit test for the error distribution in nonparametric regression. Computational Statistics & Data Analysis 54, 1942–1951.
[24] Jalal, A., Rockinger, M., 2008. Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. Journal of Empirical Finance 15, 868–877.
[25] Jeong, S.O., Kang, K.H., 2009. Nonparametric estimation of value-at-risk. Journal of Applied Statistics 36, 1225–1238.
[26] Jones, M., Pewsey, A., 2009. Sinh-arcsinh distributions. Biometrika 96, 761–780.
[27] Klar, B., Lindner, F., Meintanis, S., 2012. Specification tests for the error distribution in GARCH models. Computational Statistics and Data Analysis 56, 3587–3598.
[28] Lambert, P., Laurent, S., 2002. Modelling skewness dynamics in series of financial data using skewed location-scale distributions. Discussion Paper Université de Louvain-la-Neuve.
[29] Lamoureux, C., Lastrapes, W., 1990. Persistence in variance, structural change, and the GARCH model. Journal of Business & Economic Statistics 8, 225–234.
[30] Lejeune, B., 2009. A diagnostic m-test for distributional specification of parametric conditional heteroscedasticity models for financial data. Journal of Empirical Finance 16, 507–523.
[31] Mercurio, D., Spokoiny, V., 2004. Statistical inference for time-inhomogeneous volatility models. The Annals of Statistics 32, 577–602.
[32] Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and its Applications 9, 141142.
[33] Prause, K., 1999. The Generalized Hyperbolic model : estimation, financial derivatives, and risk measures. Ph.D. thesis.
[34] Sadefo-Kamdem, J., 2007. Var and es for linear portfolios with mixture of elliptic distributions risk

factors. Computing and Visualization in Science 4, 197–210.
[35] Scherer, M., Rachev, S., Kim, Y., Fabozzi, F., 2012. Approximation of skewed and leptokurtic return distributions. Applied Financial Economics 22, 1305–1316.
[36] Silverman, B., 1986. Density estimation. Chapman and Hall, London.
[37] Stavroyiannis, S., Makris, I., Nikolaidis, V. Zarangas, L., 2012. Econometric modeling and value-at-risk using the pearson type-iv distribution. International Review of Financial Analysis 22, 10–17.
[38] Watson, G., 1964. Smooth regression analysis. Sankhya Series 26, 359–372.
[39] Wellner, J., Koltchinskii, V., 2003. A note on the asymptotic distribution of the berk-jones type statistics under the null hypothesis. High Dimensional Probability III (T. Hoffman-Jrgensen, M. B. Marcus and J. A. Wellner, eds.) , 321–332.

Appendix A. Preliminary analysis

We check for a possible unit root in the mean of our data, that would reject the stationarity hypothesis. See the results in Table A1.

We also check for a possible conditional mean of the ARMA kind. An AR(1) model with no intercept and $\alpha = 0.1077$ seems suitable for the UBS time series. Sample autocorrelation functions (ACF) for various lags are not significant for the other time series (see Figure A.1).

The presence of heteroscedasticity is confirmed by significant ACF of the squared returns at various lag (indicating a time dependency in the variance), as shown in Figure A.2.

After the filtering of the conditional variance with the LAVE, we also check if the sample autocorrelations of the squared estimated innovations have been correctly removed. Some autocorrelations for the lags between 2 and 10 remain significantly different from zero as shown on the following graphs. Nevertheless, most of the second order time dependencies have been removed (see Figure A.3).

Appendix B. Distributions references

Generalized Hyperbolic distribution

The pdf of a GH function is given by [33] :

$$f_{GH}(x;\lambda,\alpha,\beta,\delta,\mu) = a(\lambda,\alpha,\beta,\delta)(\delta^2+(x-\mu)^2)^{(\lambda-\frac{1}{2})/2}G_{\lambda-\frac{1}{2}}(x), \tag{23}$$

$$G_{\lambda-\frac{1}{2}}(x) = K_{\lambda-\frac{1}{2}}(\alpha\sqrt{\delta^2+(x-\mu)^2})\exp(\beta(x-\mu)) \tag{24}$$

$$a(\lambda,\alpha,\beta,\delta) = \frac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\pi}\alpha^{(\lambda-1/2)}\delta^{\lambda}K_{\lambda}(\delta\sqrt{\alpha^2-\beta^2})}, \tag{25}$$

where $K_\nu(x)$ is the modified Bessel function, $\delta > 0$, $\alpha > |\beta|$ and $x \in \mathcal{R}$. For $\lambda = 1$, we obtain HYP functions, while for $\lambda = -1/2$, we obtain NIG functions (seeBarndorff-

Nielsen [3] for more details on these density functions).

Skewed-t distribution

Following the notation of Hansen [21], the pdf of a standardized skewed-t distribution is given by:

$$f_{SK}(x;\lambda,\nu) = \begin{cases} bc\left(1+\frac{1}{\nu-2}\left(\frac{bx+a}{1-\lambda}\right)^2\right)^{-(\nu+1)/2} & x < -a/b, \\ bc\left(1+\frac{1}{\nu-2}\left(\frac{bx+a}{1+\lambda}\right)^2\right)^{-(\nu+1)/2} & x \geq -a/b, \end{cases} \tag{26}$$

where $2 < \nu < \infty$ is the scale parameter and $-1 < \lambda < 1$ is the skewness parameter, for $x \in \mathcal{R}$. The constant a,b and c are given by :

$$\begin{aligned} a &= 4\lambda c\left(\frac{\nu-2}{\nu-1}\right), \\ b^2 &= 1+3\lambda^2-a^2, \\ c &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}. \end{aligned}$$

Parameters of the fitted distributions

See Tables B1 to B4 for the estimated parameters of the fitted distributions.

List of tables

DGP1	H_0^T	H_0^F	Test	Rejection of H_0^T	Rejection of H_0^F
	T	GED	$p_{\hat{\theta}}(0.1)$	6.7	8.2
			$p_{\hat{\theta}}(0.15)$	5.6	6.7
			$p_{\hat{\theta}}(0.2)$	5.1	6.2
			$p_{\hat{\theta}}(0.3)$	5.7	5.8
			BJ	4.6	3
DGP2	H_0^T	H_0^F	Test	Rejection of H_0^T	Rejection of H_0^F
	SKT	NIG	$p_{\hat{\theta}}(0.1)$	4.7	12.6
			$p_{\hat{\theta}}(0.15)$	3	20.6
			$p_{\hat{\theta}}(0.2)$	2.9	38.3
			$p_{\hat{\theta}}(0.3)$	1.6	13.9
			BJ	2.5	14
DGP3	H_0^T	H_0^F	Test	Rejection of H_0^T	Rejection of H_0^F
	HYP	NIG	$p_{\hat{\theta}}(0.1)$	6.4	9.4
			$p_{\hat{\theta}}(0.15)$	6.3	11.7
			$p_{\hat{\theta}}(0.2)$	6.1	13.5
			$p_{\hat{\theta}}(0.3)$	5.3	15.8
			BJ	4	9.6

Table 1.: Estimated type-I errors and powers of the Berk-Jones tests and the ERL tests at four different levels (30%,20%, 15% and 10%). The computed powers are associated to the wrong alternative distributions (i.e. H_0^F) specified for each DGP (e.g. for DGP1, it is the power associated to the wrong null hypothesis "f_θ is a GED distribution").

KIMSE	Sample size	DGP	$\Delta < 0$	$\Delta > 0$	No difference
	600	DGP1	14.4	0	85.6
	600	DGP2	31.2	0.1	68.7
	600	DGP3	21.2	0	78.8

Table 2.: Results for the KIMSE tests (based on the Δ statistic). $\Delta < 0$ (respectively > 0) indicates the proportion of rejection where the true distribution is selected (respectively where the alternative is selected).

DGP	f_{θ_1}	f_{θ_2}	Size	Censoring level	$\bar{d}^{wl} < 0$	$\bar{d}^{wl} > 0$	No difference
DGP1	T	GED	600	No censoring	19.7	0.3	80
				80%	29.5	21	49.5
				90%	13.7	17.7	68.6
				95%	19.2	11.6	69.2
				99%	36.9	8.6	54.5
			1000	No censoring	36.7	0.2	63.1
				80%	38.6	22.5	38.9
				90%	26.2	17.7	56.1
				95%	29.5	13.1	57.4
				99%	49.1	8.7	42.2
			2000	No censoring	70.9	0.1	29
				80%	48.8	23.9	27.3
				90%	46.2	21.6	32.2
				95%	51.2	16.9	31.9
				99%	61.3	7.1	31.6
			3000	No censoring	86.8	0	13.2
				80%	51.2	25.9	22.9
				90%	55.2	22	22.8
				95%	57.7	15.4	26.9
				99%	63.3	8.7	28

Table 3.: Rejection proportions of the Kullback-Leibler divergence (KLD) tests (based on T) for DGP1, using 80%, 90%, 95% and 99% censoring on the right. Column $\bar{d}^{wl} < 0$ indicate the proportion of samples where the true distribution (f_{θ_1} with unknown θ_1) has a significantly higher censored likelihood score, whereas column $\bar{d}^{wl} > 0$ indicate the proportion of samples where the wrong alternative distribution (f_{θ_2} with unknown θ_2) has a significantly higher likelihood score. We use samples of size 600, 1000, 2000 and 3000.

DGP	f_{θ_1}	f_{θ_2}	Size	Censoring level	$\bar{d}^{wl} < 0$	$\bar{d}^{wl} > 0$	No difference
DGP2	SKT	NIG	600	No censoring	40.3	0.3	59.4
				80%	49.8	5.2	45
				90%	55.7	13	31.3
				95%	65	14.6	20.4
				99%	65.9	24.7	9.4
			1000	No censoring	76.8	0.1	23.1
				80%	57.8	6.9	35.3
				90%	67.6	14.6	17.8
				95%	71.1	17.6	11.3
				99%	69.9	25.7	4.4
			2000	No censoring	97.2	0	2.8
				80%	64.7	7.2	28.1
				90%	73	16	11
				95%	75.8	18.8	5.4
				99%	73	23.2	3.8
			3000	No censoring	99.8	0	0.2
				80%	69	8.8	22.2
				90%	76.1	16.8	7.1
				95%	77.2	17.3	5.5
				99%	74.1	23.4	2.5

Table 4.: Rejection proportions of the Kullback-Leibler divergence (KLD) tests (based on T) for DGP2, using 80%, 90%, 95% and 99% censoring on the right. Column $\bar{d}^{wl} < 0$ indicate the proportion of samples where the true distribution (f_{θ_1} with unknown θ_1) has a significantly higher censored likelihood score, whereas column $\bar{d}^{wl} > 0$ indicate the proportion of samples where the wrong alternative distribution (f_{θ_2} with unknown θ_2) has a significantly higher likelihood score. We use samples of size 600, 1000, 2000 and 3000.

DGP	f_{θ_1}	f_{θ_2}	Size	Censoring level	$\bar{d}^{wl} < 0$	$\bar{d}^{wl} > 0$	No difference
DGP3	HYP	NIG	600	No censoring	11	0.1	88.9
				80%	44	23.1	32.9
				90%	41.9	21.9	36.2
				95%	47	24.3	28.7
				99%	53	25.4	21.6
			1000	No censoring	18.5	0.3	81.2
				80%	57.8	23.6	18.6
				90%	55.5	23.4	21.1
				95%	56.3	23.6	20.1
				99%	54.7	25.1	20.2
			2000	No censoring	29	0.3	70.7
				80%	60.2	28	11.8
				90%	60.8	24.9	14.3
				95%	61.7	23.3	15
				99%	59.9	27.2	12.9
			3000	No censoring	37.5	0.1	62.4
				80%	66.2	25.4	8.4
				90%	66.3	22.9	10.8
				95%	65.3	21.9	12.8
				99%	60.2	28.9	10.9

Table 5.: Rejection proportions of the Kullback-Leibler divergence (KLD) tests (based on T) for DGP2, using 80%, 90%, 95% and 99% censoring on the right. Column $\bar{d}^{wl} < 0$ indicate the proportion of samples where the true distribution (f_{θ_1} with unknown θ_1) has a significantly higher censored likelihood score, whereas column $\bar{d}^{wl} > 0$ indicate the proportion of samples where the wrong alternative distribution (f_{θ_2} with unknown θ_2) has a significantly higher likelihood score. We use samples of size 600, 1000, 2000 and 3000.

UBS		NIG	HYP	SKT
BS p-value	KS	0.969	0.846	0.769
	AD	0.846	0.459	0.945
	m-test	0.711	0.368	0.805
as. p-value	KS	0.99	0.9521	0.972
	AD	-	-	-
	m-test	0.649	0.746	0.882
BOVESPA		**NIG**	**HYP**	**SKT**
BS p-value	KS	0.752	0.874	0.769
	AD	0.756	0.805	0.945
	m-test	0.875	0.961	0.389
as. p-value	KS	0.837	0.956	0.972
	AD	-	-	-
	m-test	0.888	0.999	0.428
EUR/USD		**NIG**	**HYP**	**SKT**
BS p-value	KS	0.532	0.491	0.294
	AD	0.366	0.275	0.297
	m-test	0.046	0.131	0.013
as. p-value	KS	0.565	0.669	0.431
	AD	-	-	-
	m-test	0.046	0.471	0.0129

Table 6.: P-values for the Kolmogorov-Smirnov test (KS), Anderson-Darling (AD) test and m-test of [30], for the three time series considered. We compute the p-value of the KS and AD tests with a usual parametric bootstrap procedure (lines "BS p-value"), involving a GARCH standardization. We also provide the p-value of the KS test and m-test based on asymptotic results (line "as. p-value"). For the KS tests, we assume that the residuals are i.i.d. data. P-values of the m-tests are obtained from the asymptotic result presented in [30]

Descriptive statistics	UBS	BOVESPA	EUR/USD
Skewness	-0.7861	-0.6057	-0.1216
Kurtosis	10.6246	6.2410	9.3898

Table 7.: Descriptive statistics for the residuals after LAVE filtering. All series exhibit negative skewness and excess kurtosis.

$R_{n,f}$	UBS	BOVESPA	EUR/USD
NIG	0.745	0.451	0.459
HYP	0.99	0.465	0.828
SKT	0.81	0.618	0.344

Table 8.: Bootstrap p-values for the B-J test statistics for the fits with NIG, HYP and skewed-t (SKT) distributions. No rejection occurs.

KIMSE	UBS	BOVESPA	EUR/USD
NIG	0.0145	0.0080	0.0060
HYP	0.0132	0.0083	0.0075
SKT	0.0088	0.0056	0.0037

Table 9.: Values of the KIMSE statistic for the fits with NIG, HYP and skewed-t distributions. These quantities are used to compute the Δ statistics.

Δ	UBS		BOVESPA		EUR/USD	
HYP - NIG	0.219	-	0.3334	-	0.2507	-
HYP - SKT	0.0029	(SK)	0.0039	(SK)	0.0720	-
NIG - SKT	0.0002	(SK)	0.0007	(SK)	0.0034	(SK)

Table 10.: P-values of the Δ statistics for the three time series. In bold, rejection at the 5% test level. When a rejection occurs, superscripts (SH), (H), (N) or (SK) indicate which distribution has the lowest KIMSE.

p		UBS		BOVESPA		EUR/USD	
5%	NIG	0.0567	(0.31)	**0.0607**	(0.002)	**0.0632**	(0.001)
	HYP	0.0562	(0.224)	**0.059**	(0.007)	**0.0615**	(0.003)
	SKT	**0.0609**	(0.016)	**0.0621**	(0.003)	**0.0664**	(0.002)
1%	NIG	0.0132	(0.24)	**0.0094**	(0.003)	0.0142	(0.054)
	HYP	**0.0151**	(0.022)	**0.0094**	(0.001)	0.0139	(0.103)
	SKT	**0.0161**	(0.001)	**0.0108**	(0.031)	0.0143	(0.089)

Table 11.: Values of the ERL statistic at the 5% and 1% level. In parenthesis, bootstrap p-values of the test statistic. In bold, rejection at the 5% test level (two-sided test).

$\bar{d}^{wl}$	Censoring level	UBS		BOVESPA		EUR/USD	
HYP-NIG	No censoring	**-0.0040**	(0.0033)	**-0.0008**	(0.0282)	-0.0007	(0.1831)
	95%	**-0.0049**	(0.000)	-0.0002	(0.3470)	-0.0006	(0.1295)
	99%	**-0.0050**	(0.000)	-0.0003	(0.1191)	-0.0007	(0.1086)
HYP-SKT	No censoring	**-0.0078**	(0.004)	**-0.0016**	(0.0426)	-0.002	(0.1863)
	95%	**-0.0009**	(0.2938)	-0.0009	(0.0679)	-0.0021	(0.1123)
	99%	**-0.0049**	(0.000)	**-0.0014**	(0.0018)	-0.0023	(0.0891)
NIG-SKT	No censoring	**-0.0038**	(0.0097)	-0.0008	(0.0892)	-0.0013	(0.1923)
	95%	**0.0040**	(0.000)	**-0.0009**	(0.0159)	-0.0015	(0.1061)
	99%	0.0001	(0.386)	**-0.0012**	(0.000)	-0.0017	(0.0813)

Table 12.: $\bar{d}_{wl}$ statistic between HYP, NIG and skewed-t (SKT) density functions using no censored likelihood scores (first line) and censored regions up to the 5% and 1% empirical quantiles. A positive sign indicates that the first distribution of the label is the closest to the true distribution. In parenthesis, p-value of the associated test statistic T. In bold: rejection at the 5% test level (two-sided test).

Time series	Augmented DF stat	p-value
UBS	-10.2611	**0.000**
BOVESPA	-11.6946	**0.000**
EUR/USD	-11.4139	**0.000**

Table A1.: Results of the Augmented Dickey-Fuller test with 21 lags. In bold rejection of the null hypothesis of a unit root at the 1% test level.

NIG parameters	α	β	δ	μ
UBS	1.1569	-0.0943	1.1453	0.0937
BOVESPA	1.6042	-0.2607	1.5411	0.2538
EUR/USD	1.2788	0.0114	1.2787	-0.0114

Table B2.: Estimated parameters of the NIG distributions for the three time series.

HYP parameters	α	β	δ	μ
UBS	1.6770	-0.1048	0.6535	0.1041
BOVESPA	1.9999	-0.2684	1.1132	0.2615
EUR/USD	1.7099	0.0096	0.73	-0.0096

Table B3.: Estimated parameters of the HYP distributions for the three time series.

Skewed-t parameters	λ	ν
UBS	-0.0418	5.4567
BOVESPA	-0.0950	7.8023
EUR/USD	0.0053	6.4310

Table B4.: Estimated parameters of the skewed-t distributions for the three time series. α is the asymmetry parameters and ν the df

List of figures

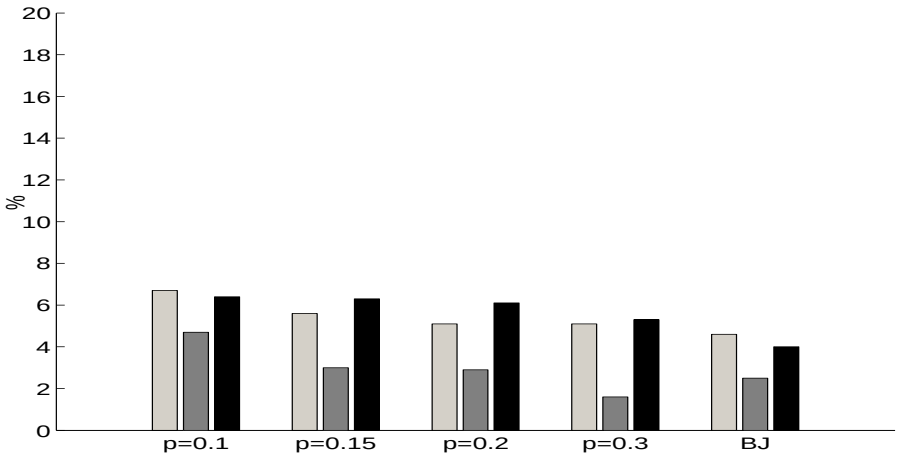

Figure 1.: Rejection rates of the true distribution (type-I error rates) for the BJ and ERL tests. For the ERL tests, the level of p is indicated on the horizontal axis. Grey: results related to DGP1. Dark grey: results related to DGP2. Black: results related to DGP3. These results can be found in Table 8.

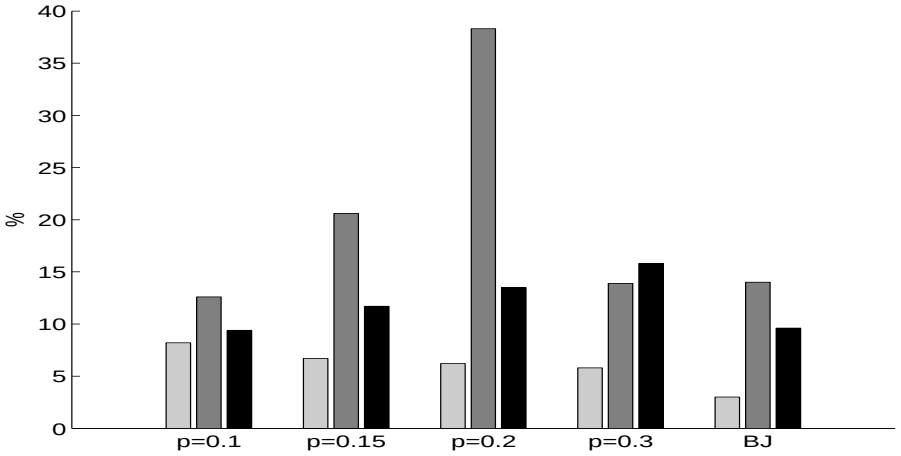

Figure 2.: Rejection rates of the alternative distribution (type-I error rates) for the BJ and ERL tests. For the ERL test, the level of p is indicated on the horizontal axis. Grey: results related to DGP1. Dark grey: results related to DGP2. Black: results related to DGP3. These results can be found in Table 8.

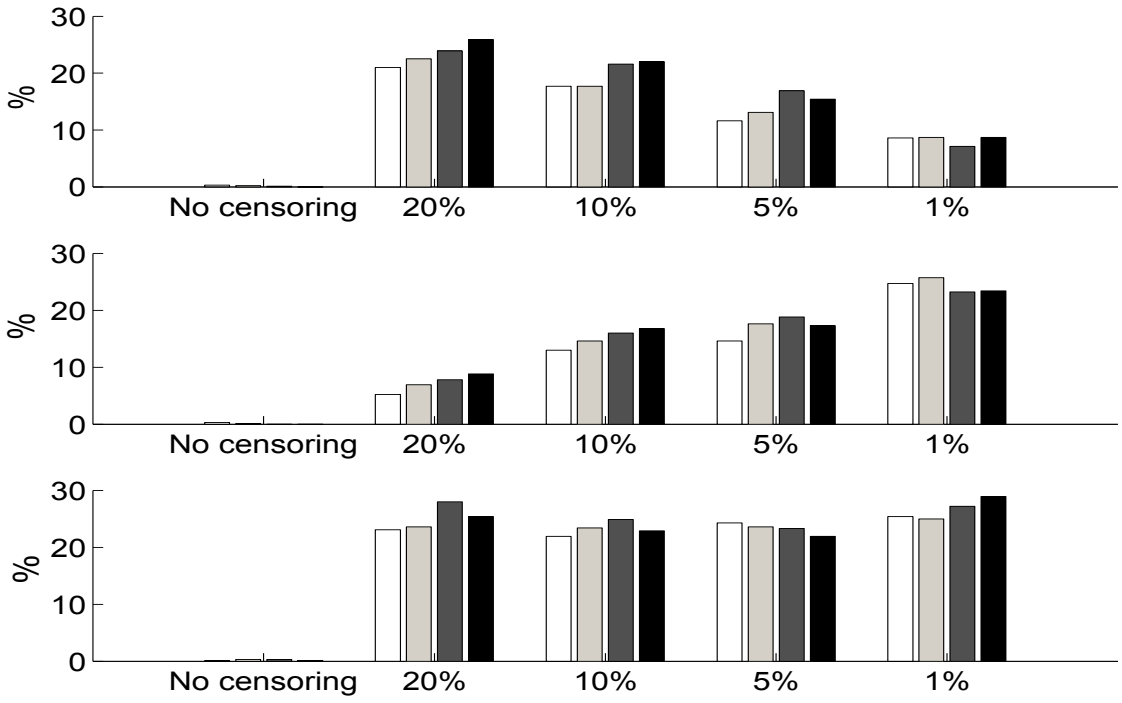

Figure 3.: Rejection rates in favour of the alternative distribution (column $\bar{d}^{wl} > 0$ in Tables 3 to 5) for DGP1 (top), DGP2 (middle) and DGP3 (bottom), as a function of the regions of interest(that correspond respectively to 80%, 90%, 95% and 99% censoring on the right). White: sample size of 600. Light grey: sample size of 1000. Grey: sample size of 2000. Black: sample size of 3000.

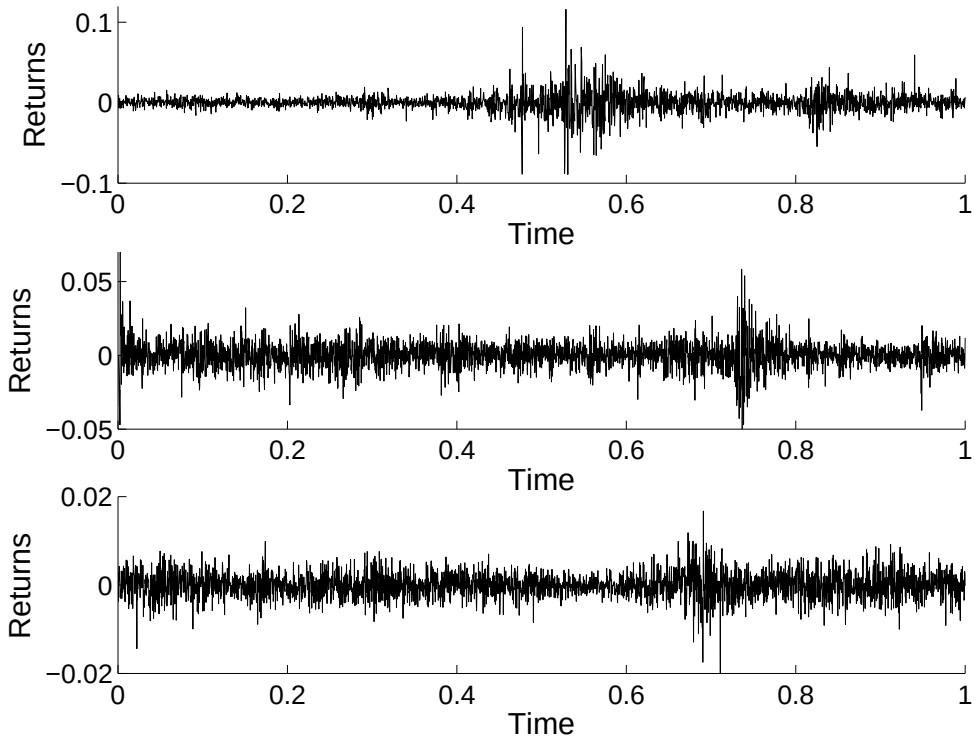

Figure 4.: Daily stock returns of the AR(1) UBS residuals, the BOVESPA and the EUR/USD time series. Notice that y axis have different scales.

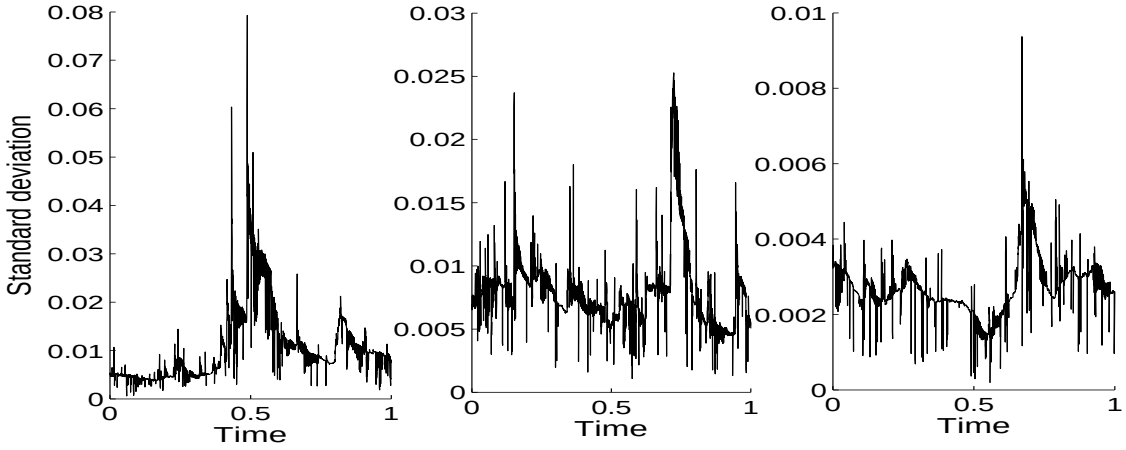

Figure 5.: Conditional standard deviation estimations of the three time series using the LAVE technique with $m_0 = 5$, $\gamma = .5$ and the 200 first observations as training set.

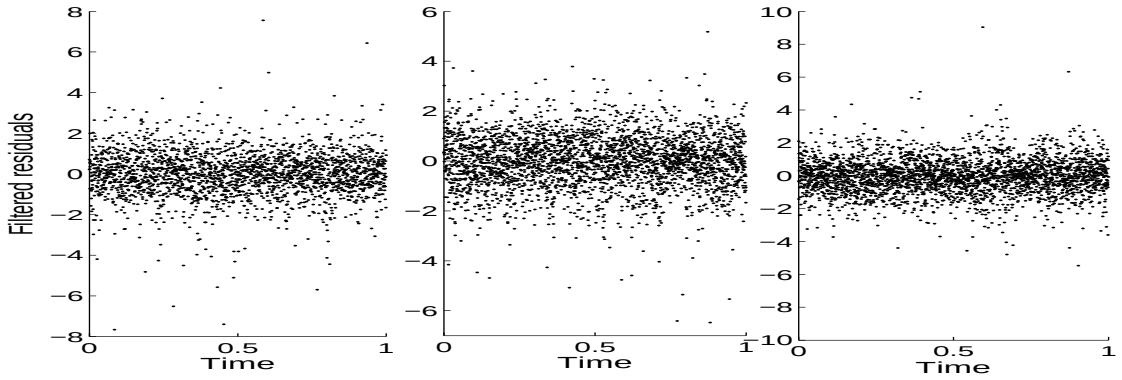

Figure 6.: Scatter plots of the daily returns after standardization of the same time series (scales of the y axis are different).

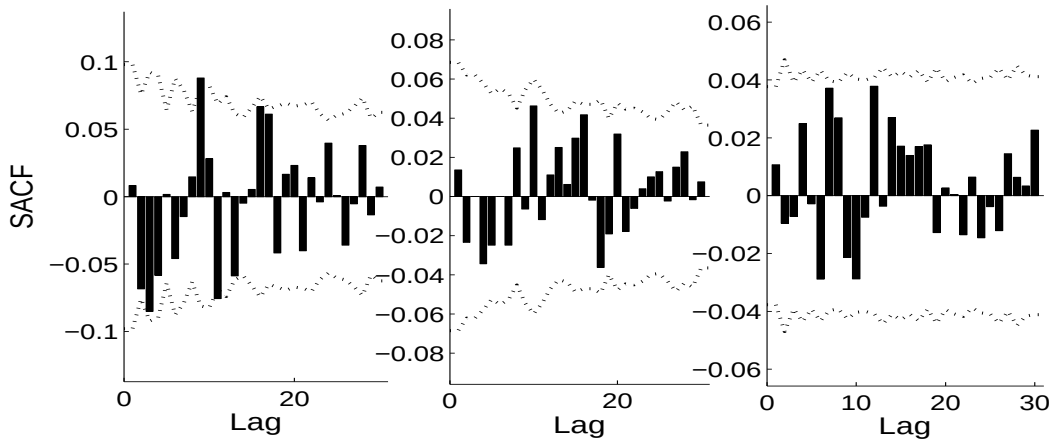

Figure A1.: Plot of the ACF with robust standard errors for AR(1) UBS errors, BOVESPA and EUR/USD time series.

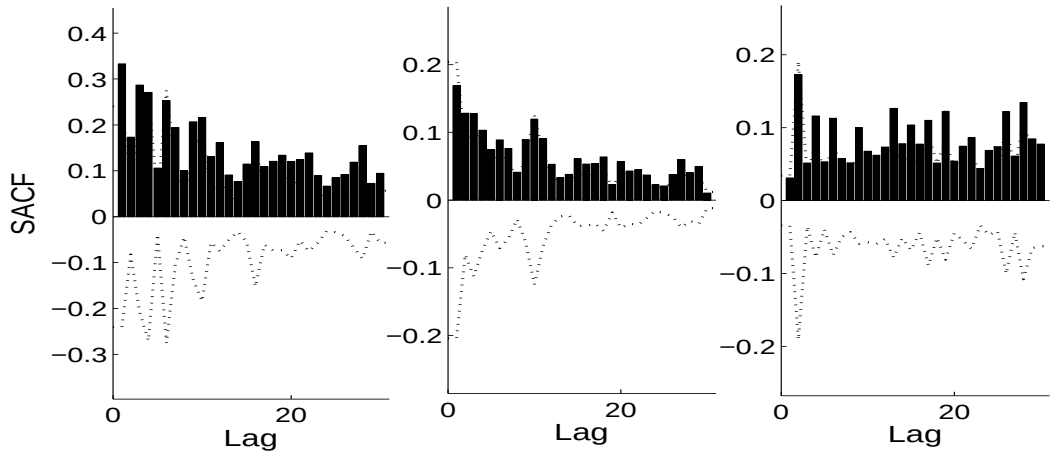

Figure A2.: Plot of the ACF with robust standard errors for squared AR(1) UBS errors, BOVESPA and EUR/USD time series.

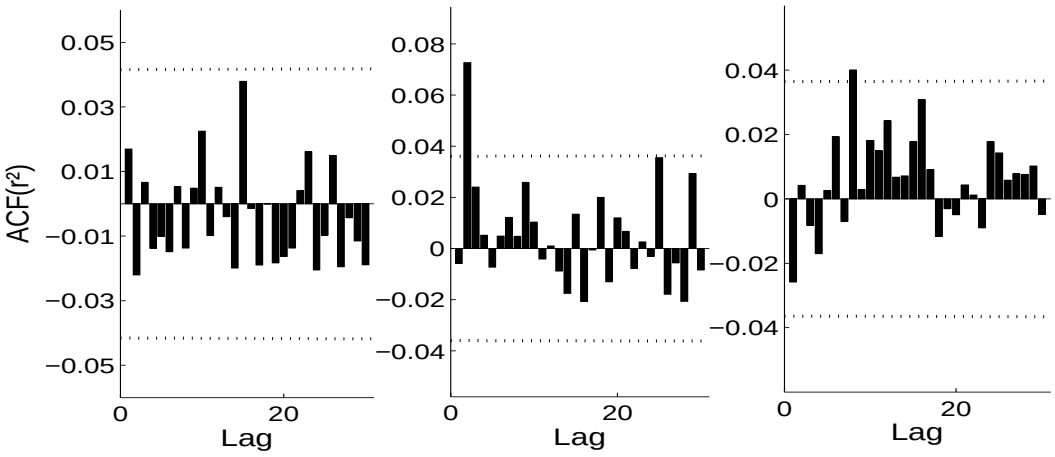

Figure A3.: Sample autocorrelations for the squared estimated innovations for the five time series tested up to lag 50. If the bar is up to the dotted line, the autocorrelation at the corresponding lag is significantly different from 0 (with a level of confidence of 95%).