To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–23

# A new methodological approach for error distributions selection in Finance.

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

[a] *QuantOM, HEC Management School, University of Liège, Liège, Belgium*; [b] *Fonds national de la recherche scientifique (F.R.S. - FNRS), Bruxelles, Belgium*; [c] 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 methodology to select the most appropriate error distribution candidate, 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 [30]: 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 given distribution as well as to focus on its behavior in the tails. Finally, we illustrate our methodology on three time series (UBS stock returns, BOVESPA returns and EUR/USD exchange rates).

**Keywords:** error distribution, nonparametric, misspecification, goodness-of-fit, selection test, Value-at-Risk, GARCH

*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, 34, 36]. In this article, we propose a robust methodology to select a distribution family in a classical multiplicative heteroscedastic model. This model is defined by :

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

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

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

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

where $r_t$ is the daily return, $\sigma_t^2$ 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 law and its generalization, the skewed-t distribution [16]. Its use in GARCH or EGARCH models considerably improves forecasts [7, 22, 27]. 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. [36] also propose to use the Pearson type-IV distribution.

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, 29, 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 [28] 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 [31, 37], Mercurio and Spokoiny [30] 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. 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), skewed-t and the Sinh-arcsinh distributions [26].

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, 30]. For all $r_t$, $t = 1, ..., n$, we compute $\hat{\sigma}_t$ using $I_t$ previous squared returns $r_{t-1}^2, ..., r_{t-I_t}^2$:

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

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, 38), 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 $F_\theta$ (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:

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

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) of $\theta$ under the assumed family $F_\theta$. We reject the parametric hypothesis 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 $F_\theta$ 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 $R_{n,F_\theta}$. 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* of the estimated residuals to determine which distribution displays the best fit. We compute the bandwidth using the normal rule [35]. Based on that estimated density, we compute an estimator of the integrated mean squared error between the true and the parametrically estimated ($f_{\hat{\theta}}$) densities ($KIMSE_{f_{\hat{\theta}}}$ hereunder). We use a nonparametric bootstrap procedure this time to estimate this quantity:

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:

$$KIMSE_{f_{\hat{\theta}}} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}(f_{\hat{\theta}_i^*}(x) - \hat{f}(x))^2 dx = \frac{1}{N}\sum_{i=1}^{N} KISE_{f_{\hat{\theta}}}^i, \tag{7}$$

where $\hat{f}(x)$ 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. $E[\int\limits_{-\infty}^{\infty}(f_{\hat{\theta}_j}(x) - f(x))^2 dx]$) we can simply use the following statistic $\bar{D}$ for paired data:

$$\bar{D} = \frac{1}{N}\sum_{i=1}^{N}[KISE_{f_{\hat{\theta}_1}}^i - KISE_{f_{\hat{\theta}_2}}^i] = \frac{1}{N}\sum_{i=1}^{N} D_i, \tag{8}$$

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

$$\sqrt{N}\bar{D} \to N(0, \sigma_D^2), \tag{9}$$

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 $\sigma_D^2$, a simple standardization gives us $\Delta = \bar{D}/\sqrt{\hat{\sigma}_D^2/N} \to N(0,1)$ given $\hat{Z}$ and if the observed $|\Delta| \geq \Phi^{-1}(1-\alpha/2)$, the null hypothesis can be rejected with a test level $\alpha$.

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 $f_\theta$ (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:

$$p_{\hat{\theta}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}(\hat{z}_t \leq q_{\hat{\theta}}(p)), \tag{10}$$

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 ($\alpha$ 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 $d_t^{wl}, t = 1, ..., n$:

$$\bar{d^{wl}} = \frac{1}{n}\sum_{t=1}^{n} d_t^{wl} = \frac{1}{n}\sum_{t=1}^{n}(S^{wl}(\hat{g}_t^1; r_t) - S^{wl}(\hat{g}_t^2; r_t)), \tag{11}$$

with

$$S^{wl}(\hat{g}_t^j; r_t) = \mathbf{1}(r_t \in A)\log(\hat{g}_t^j(r_t)) + \mathbf{1}(r_t \in A^c)\log\left(\int_{A^c} \hat{g}_t^j(s)ds\right), \; j = 1, 2, \tag{12}$$

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. We propose to 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 $f_{\theta_j}$ through the following relationship:

$$g_t^j(r_t) = \frac{1}{\sigma_t} f_{\theta_j}(r_t/\sigma_t), j = 1, 2. \tag{13}$$

Parameters estimators for $f_{\theta_j}$ 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,

$$T = \frac{\bar{d^{wl}}}{\sqrt{\hat{\sigma}_n^2/n}}, \tag{14}$$

with $\hat{\sigma}_n^2$ being a heteroscedasticity and autocorrelation-consistent (HAC) estimator of the variance of $\sqrt{n}\bar{d^{wl}}$, Giacomini and White [18] demonstrate that this statistic is asymptotically normally distributed under very weak conditions (see this article and Wooldrige and White, 1988, for more details). In particular, it allows using both para- and non-parametric 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]:

$$\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{G-1} a_k \hat{\gamma}_k, \tag{15}$$

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 $x$).

This statistic 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 (13) 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 statistics presented in this section. B-J and $p_{\hat{\theta}}$ statistics enable to reject irrelevant distributions while we use KIMSE and $\bar{d}^{wl}$ statistics in comparative tests to determine if some distributions have a significantly higher GOF performance than the others. Moreover, we assess both the global fit (with the B-J and the KIMSE statistics) and the fit in the tail (with $p_{\hat{\theta}}$ 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.

### 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:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{16}$$

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

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \mathbf{1}(r_{t-1} \leq 0)\phi r_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{17}$$

We assumed three different distributions for $z_t$:

$$z_t \overset{iid}{\sim} T(\nu), \tag{18}$$

$$z_t \overset{iid}{\sim} skewed - t(\lambda, \nu), \tag{19}$$

$$z_t \overset{iid}{\sim} hyperbolic(\theta_1, \theta_2, \theta_3, \theta_4), \tag{20}$$

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 (5), $m_0 = 5$ and $\gamma = .5$ [see 25]. In addition, in this last paper, the level of the multiple test 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). 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-t, skewed-t or hyperbolic) to an alternative distribution having the same number of unknown parameters (i.e. Student-t 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, $\bar{d^{wl}}$ and $p_{\hat{\theta}}$) for each sample and for each parametric hypothesis, using equations (6), (7), (10) and (11) proposed in Section 2. 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 censoring scores at 20%, 10%, 5% and 1% levels. 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 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 distributions can be well approximated by GED distributions. This low level of rejections indicates that the GED distribution fits quite well the true Student distribution.

Concerning the ERL tests, 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.

Using the KIMSE statistic (Table 2), we are able to detect significant differences between the true distributions and the alternative ones in all DGP. This leads us to select most of the time the true distribution. 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). We do not observe any spurious powers.

The tests based on the censored likelihood score (Table 2) bring a different perspective to the analysis. Using an uncensored statistic, this test selects quite adequately the true distribution, especially when the size of the sample increases (by going from 600 observations to 1000, we double the power of the test for the two first DGP). Using the censoring statistics, we observe that in the case of medium-size samples (i.e. 600 and 1000 observations) and high censoring levels, we select more often the true distribution. Thus, the censoring procedure seems to improve the detection of differences for these sample sizes.

However, we select also more often the alternative distribution, comparing to the test with the uncensored statistic (and that holds for all censoring levels 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.

For large samples (i.e. 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. 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, at the censoring level of 10%, 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 distribution compared to the uncensored ones. The simulations highlight the need for large samples too.

For the KIMSE tests, the figures in the first column indicate the proportion of samples where the alternative has a significant lower KIMSE than the true estimated distribution. The second column indicates the proportion of samples where the true estimated

| | | (a) | (b) | | | (a) | (b) |
|---|---|---|---|---|---|---|---|
| $p_{\hat{\theta}}(0.1)$ | DGP1 | 6.7 | 8.2 | $p_{\hat{\theta}}(0.15)$ | DGP1 | 5.6 | 6.7 |
| | DGP2 | 4.7 | 12.6 | | DGP2 | 3 | 20.6 |
| | DGP3 | 6.4 | 9.4 | | DGP3 | 6.3 | 11.7 |
| $p_{\hat{\theta}}(0.2)$ | DGP1 | 5.1 | 6.2 | $p_{\hat{\theta}}(0.30)$ | DGP1 | 5.7 | 5.8 |
| | DGP2 | 2.9 | 38.3 | | DGP2 | 1.6 | 13.9 |
| | DGP3 | 6.1 | 13.5 | | DGP3 | 5.3 | 15.8 |
| Berk-Jones | DGP1 | 4.6 | 3 | | - | - | - |
| | DGP2 | 2.5 | 14 | | - | - | - |
| | DGP3 | 4 | 9.6 | | - | - | - |

Table 1. Estimated type-I errors (column (a)) and powers (column (b)) of the Berk-Jones tests and the ERL tests at four different levels (30%,20%, 15% and 10%).

| size | censoring level | DGP1 | (a) | (b) | DGP2 | (c) | (d) | DGP3 | (e) | (f) |
|---|---|---|---|---|---|---|---|---|---|---|
| 600 | KIMSE | | 14.4 | 0 | | 31.2 | 0.1 | | 21.2 | 0 |
| | 100% | | 19.7 | 0.3 | | 40.3 | 0.3 | | 11 | 0.1 |
| | 20% | | 29.5 | 21 | | 49.8 | 5.2 | | 44 | 23.1 |
| | 10% | | 13.7 | 17.7 | | 55.7 | 13 | | 41.9 | 21.9 |
| | 5% | | 19.2 | 11.6 | | 65 | 14.6 | | 47 | 24.3 |
| | 1% | | 36.9 | 8.6 | | 65.9 | 24.7 | | 53 | 25.4 |
| 1000 | 100% | | 36.7 | 0.2 | | 76.8 | 0.1 | | 18.5 | 0.3 |
| | 20% | | 38.6 | 22.5 | | 57.8 | 6.9 | | 57.8 | 23.6 |
| | 10% | | 26.2 | 17.7 | | 67.6 | 14.6 | | 55.5 | 23.4 |
| | 5% | | 29.5 | 13.1 | | 71.1 | 17.6 | | 56.3 | 23.6 |
| | 1% | | 49.1 | 8.7 | | 69.9 | 25.7 | | 54.7 | 25.1 |
| 2000 | 100% | | 70.9 | 0.1 | | 97.2 | 0 | | 29 | 0.3 |
| | 20% | | 48.8 | 23.9 | | 64.7 | 7.2 | | 60.2 | 28 |
| | 10% | | 46.2 | 21.6 | | 73 | 16 | | 60.8 | 24.9 |
| | 5% | | 51.2 | 16.9 | | 75.8 | 18.8 | | 61.7 | 23.3 |
| | 1% | | 61.3 | 7.1 | | 73 | 23.2 | | 59.9 | 27.2 |
| 3000 | 100% | | 86.8 | 0 | | 99.8 | 0 | | 37.5 | 0.1 |
| | 20% | | 51.2 | 25.9 | | 69 | 8.8 | | 66.2 | 25.4 |
| | 10% | | 55.2 | 22 | | 76.1 | 16.8 | | 66.3 | 22.9 |
| | 5% | | 57.7 | 15.4 | | 77.2 | 17.3 | | 65.3 | 21.9 |
| | 1% | | 63.3 | 8.7 | | 74.1 | 23.4 | | 60.2 | 28.9 |

Table 2. Rejection proportions of the KIMSE tests and the Kullback-Leibler divergence (KLD) tests for the three DGP, using no censoring, 20%, 10%, 5% and 1% censoring levels. Column (a), (c) and (e) indicate the proportion of samples where the true distribution has a significantly higher censored likelihood score, whereas column (b), (d) and (f) indicate the proportion of samples where the wrong distribution has a significantly higher likelihood score. For the KIMSE tests, the figures in columns (a), (c) and (e) indicate the proportions of samples where the KIMSE computed with the true model is significantly higher than with the wrong model. The figures in columns (b), (d) and (f) indicate the proportions of samples where the KIMSE computed with the wrong model is significantly higher than with the true model. For the KLD tests, we also use samples of 1000, 2000 and 3000 observations.

distribution has a significant lower KIMSE than the alternative.

## 4. Empirical illustration

In this section, we illustrate the proposed methodology on three different time series, where we test four different distributions for the innovations. Indeed, recent works emphasize the flexibility of GH subfamilies [8, 9, 33] and skewed-t distributions [22, 27], but other alternatives exist, like the sinh-arcsinh (SHASH) distributions [26]. Therefore, we will compare the NIG, the HYP (i.e. subfamilies of the GH distributions for $\lambda$ equal respectively to $-1/2$ and 1), the skewed-t [21] and the SHASH distributions. Details concerning the GH and the skewed-t alternatives can be found in Appendix B. In the next subsection, we present the important features of the SHASH distributions.

### 4.1 *The Sinh-arcsinh distribution*

This family of density functions has four parameters $(\xi, \eta, \epsilon, \delta)$, with easily interpretable meanings (location, scale, asymmetry and kurtosis) and the normal law as a central particular case (with values of the parameters being respectively 0, 1, 0 and 1). Moreover, we have analytical expression for the cdf, the quantile function and the moment generating function, that allow a quick and straightforward implementation (especially for random number generation). Besides, its flexibility allows for asymmetry and heavier as well as lighter tails than the normal law (a feature not possible with the skewed-t distributions). See Figure 1(a) and 1(b) for shapes of the SHASH with various sets of parameters. Also, these functions may overcome fitting difficulties met by GH subfamilies (see Figure 2) and avoid to deal with a flat likelihood function [4], responsible for optimization problems [32]. In their original paper, Jones and Pewsey Jones and Pewsey [26] showed that the SHASH distributions provide better fits than the normal and the Student's laws for non-mesokurtic data (the strengths of glass fibres and the ice floe snow depth). However, they did not use any financial data nor did they compare the SHASH distributions to others four-parameters families of functions. In fact, to the best of our knowledge, this distribution has never been used in Finance but only in Engineering. Therefore, we will compare its performance to the NIG, the hyperbolic (HYP) and the skewed-t distributions on financial data, using the methodology described in the previous section.

Analytically, the probability density function (pdf) of a variable X following a SHASH distribution with parameters $\xi, \eta, \epsilon, \delta$ is given by:

$$f_{\xi,\eta,\epsilon,\delta}(x) = \eta^{-1} Z_{\xi,\eta}(x)^{-1/2} \delta C_{\epsilon,\delta}((x-\xi)\eta) \exp(-S^2_{\epsilon,\delta}((x-\xi)/\eta)/2), \tag{21}$$

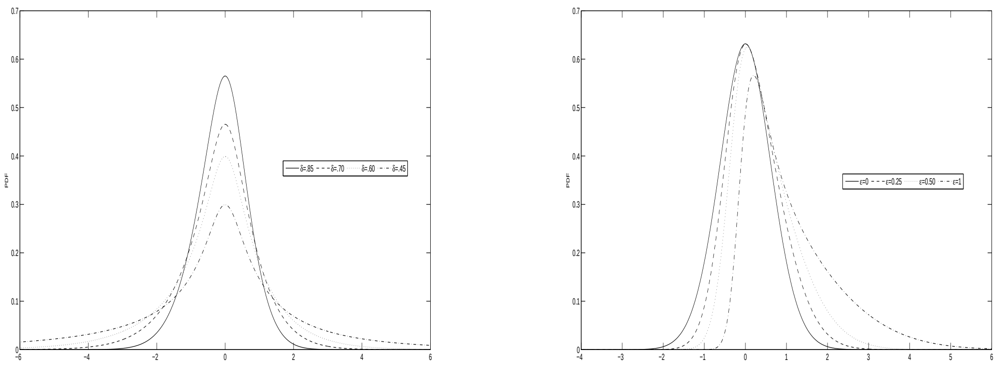
where

$$Z_{\xi,\eta}(x) = (2\pi(1 + ((x-\xi)/\eta)^2)), \tag{22}$$

$$C_{\epsilon,\delta}(x) = \cosh(\delta \sinh^{-1}(x) - \epsilon) = (1 + S^2_{\epsilon,\delta}(x))^{1/2}, \tag{23}$$

$$S_{\epsilon,\delta}(x) = \sinh(\delta \sinh^{-1}(x) - \epsilon). \tag{24}$$

Skewness increases with increasing $\epsilon$, positive skewness corresponding to $\epsilon > 0$. The kurtosis decreases with increasing $\delta$, $\delta < 1$ yielding heavier tails than the normal distribution. Equations of the two first moments can be found in Appendix B.




(a) (b)

Figure 1. Shapes of SHASH densities (a) with $\xi = 0$, $\eta = 0.6$, $\epsilon = -0.1$ and heavier tails than the Normal law ($\delta < 1$, ranging from 0.85 to 0.45) and (b) with $\xi = 0$,$\eta = 1$, $\delta = 0.95$ and positive skewness (values of $\epsilon$ ranging from 0 to 1).

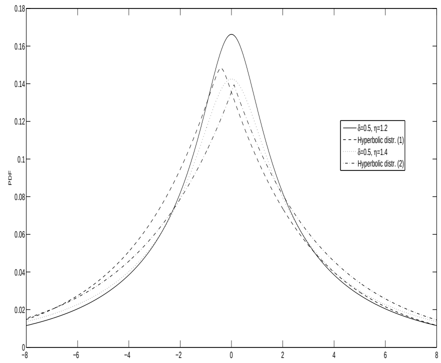



Figure 2. Solid line and dotted: SHASH functions with $\eta$ and $\delta$ parameters equal to 1.2 (resp. 1.4) and 0.5. Dashed and dashed-dotted: hyperbolic distributions with parameters obtained via MLE on samples of sinh-arcsinh distributed random variables (with the same $\eta$ and $\delta$ as above). For this kind of stochastic process, no hyperbolic distribution seems able to provide an good fit.

## 4.2 *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 3) 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.

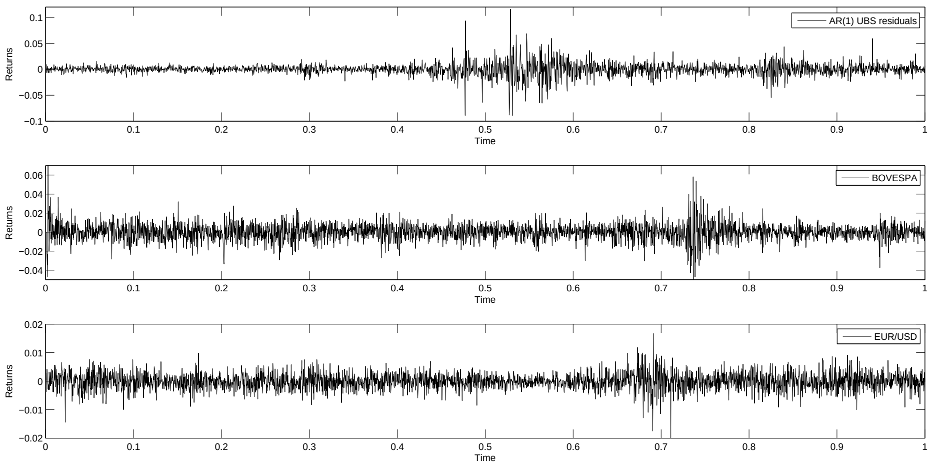



Figure 3. Daily stock returns of the AR(1) UBS residuals, the BOVESPA and the EUR/USD time series. Notice that y axis have different scales.

## 4.3 *Results*

### *4.3.1 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 4 shows the estimated conditional standard deviations with this method and Figure 5 the residuals obtained after standardization. Descriptive statistics of the residuals are presented in Table 5. 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.

### *4.3.2 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 distribution tested (Table 6). It is not very surprising, because all distributions tested

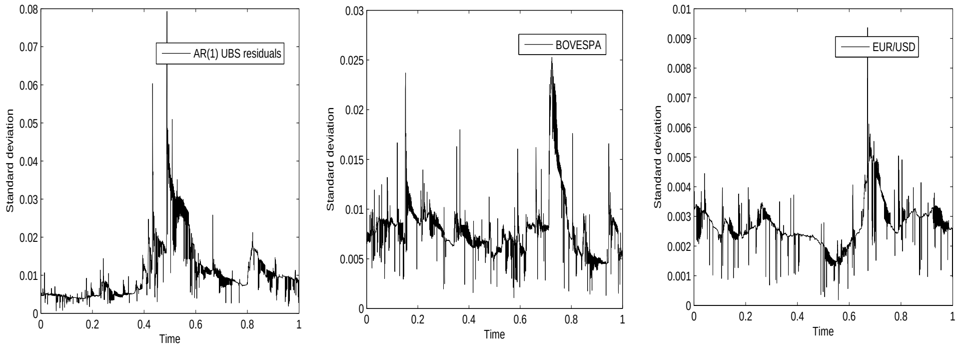



Figure 4. Conditional standard deviation estimations of the three time series using the LAVE technique with $m_0 = 5$, $\gamma = .5$ and the 200 first observations as training set.

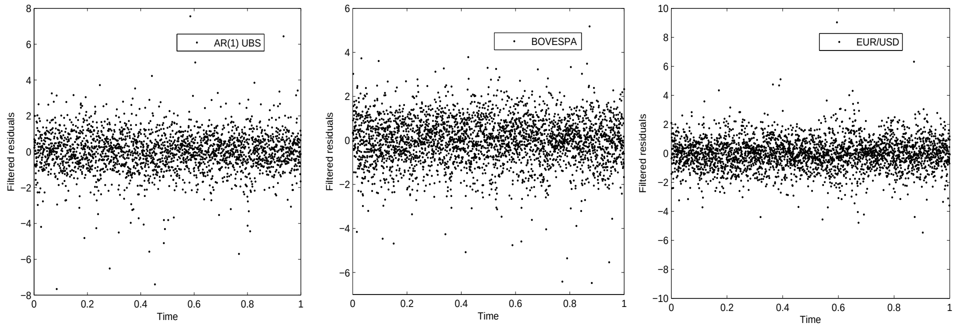



Figure 5. Scatter plots of the daily returns after standardization of the same time series (scales of the y axis are different).

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

Table 3. Descriptive statistics for the residuals after LAVE filtering.

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. Therefore, we compute the KIMSE statistic (Table 5 to 7) and we observe that:

- the skewed-t distribution has the lowest statistics for the three series,
- significant differences are detected, between the SHASH distribution and the other ones, as well as between NIG and skewed-t distributions (Table 6),
- 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 SHASH distributions appear to provide the worst fits.

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 8. Globally, SHASH distributions lead to a smaller number of rejections and provide the closest ERL to $p$ most of the time. More precisely, we identify the SHASH distribution as being the best to model the 5% quantile for all time series (with the HYP distribution in the BOVESPA time series); the NIG, the HYP and the SHASH distributions 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 9 and Table 10. If we compute the scores without censoring, results are similar to the ones deduced from the KIMSE statistics (the SHASH distributions provide the worst fits, skewed-t distributions provide the best fits). With the censoring at the 5% level, we can conclude for the UBS series that the SHASH distribution provides a significantly higher score (hyperbolic and skewed-t distributions have significantly lower scores). With the censoring at the 1% level, skewed-t distributions appear to better fit the data for the BOVESPA time series, whereas both skewed-t and NIG distributions better fit the UBS time series. No 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.

| $R_{n,f}$ | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH | 0.0063 | 0.0064 | 0.0050 |
| NIG | 0.0043 | 0.0035 | 0.0038 |
| HYP | 0.0055 | 0.0043 | 0.0036 |
| SKT | 0.0043 | 0.0030 | 0.0041 |

Table 4. B-J test statistics for the fits with SHASH, NIG, HYP and skewed-t (SKT) distributions. The 95% quantile is obtained using 1000 resamples. * indicates a rejection at the 5% test level and ** a rejection at the 1% test level. No rejection occurs.

| KIMSE | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH | 0.0237 | 0.0133 | 0.0112 |
| NIG | 0.0145 | 0.0080 | 0.0060 |
| HYP | 0.0132 | 0.0083 | 0.0075 |
| SKT | 0.0088 | 0.0056 | 0.0037 |

Table 5. Values of the KIMSE for the fits with SHASH, NIG, HYP and skewed-t distributions.

| $\Delta$ | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH - HYP | 2.9813** | 2.7072** | 1.3893 |
| SHASH - NIG | 3.6833** | 3.1434** | 2.8153** |
| SHASH - SKT | 3.7630** | 3.2454** | 2.8430** |
| HYP - NIG | -0.7755 | 0.4306 | 0.6504 |
| HYP - SKT | 2.7635** | 2.6570** | 1.4611 |
| NIG - SKT | 3.5642** | 3.1717** | 2.7049** |

Table 6. Values of the $\Delta$ statistics for the three time series. * indicates a rejection at the 5% test level and ** a rejection at the 1% level.

| $p$-value of $\Delta$ | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH - HYP | 0.0014 | 0.0034 | 0.0824 |
| SHASH - NIG | 0.0001 | 0.0008 | 0.0024 |
| SHASH - SKT | 0.0001 | 0.0006 | 0.0022 |
| HYP - NIG | 0.2190 | 0.3334 | 0.2507 |
| HYP - SKT | 0.0029 | 0.0039 | 0.0720 |
| NIG - SKT | 0.0002 | 0.0007 | 0.0034 |

Table 7. $p$-values of the $\Delta$ statistics for the three time series.

| $p_{\hat{\theta}}$ | expected coverage | | UBS | BOVESPA | EUR/USD |
|---|---|---|---|---|---|
| | 5% | SHASH | 0.0543 | 0.0590* | 0.0583 |
| | | NIG | 0.0567 | 0.0607* | 0.0632* |
| | | HYP | 0.0562* | 0.0590* | 0.0615* |
| | | SKT | 0.0609 | 0.0621* | 0.0664* |
| | 1% | SHASH | 0.0142 | 0.0101* | 0.0128 |
| | | NIG | 0.0132 | 0.0094* | 0.0142 |
| | | HYP | 0.0151* | 0.0094 | 0.0139 |
| | | SKT | 0.0161 | 0.0108* | 0.0146 |

Table 8. Values of the ERL statistic at the 5% and 1% level. * indicates a rejection at the 5% test level (two-sided test).

| $\bar{d}^{wl}$ | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH-HYP | -0.0074** | -0.0042** | -0.003* |
| 5% | 0.0050** | -0.0017* | -0.0018 |
| 1% | 0.0019** | -0.0011 | -0.0017 |
| SHASH-NIG | -0.0114** | -0.0049** | -0.0037 |
| 5% | 0.0001 | -0.0018 | -0.0024 |
| 1% | -0.0031** | -0.0014 | -0.0024 |
| SHASH-SKT | -0.0152** | -0.0057** | -0.0050 |
| 5% | 0.0041* | -0.0027 | -0.0039* |
| 1% | -0.0030** | -0.0025* | -0.0040 |
| HYP-NIG | -0.0040** | -0.0008* | -0.0007 |
| 5% | -0.0049** | -0.0002 | -0.0006 |
| 1% | -0.0050** | -0.0003 | -0.0007 |
| HYP-SKT | -0.0078** | -0.0016* | -0.002 |
| 5% | -0.0009 | -0.0009 | -0.0021 |
| 1% | -0.0049** | -0.0014** | -0.0023 |
| NIG-SKT | -0.0038** | -0.0008 | -0.0013 |
| 5% | 0.0040** | -0.0009** | -0.0015 |
| 1% | 0.0001 | -0.0012** | -0.0017 |

Table 9. $\bar{d}_{wl}$ statistic between SHASH , 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. * indicates a significant difference at the 5% test level, ** at the 1% test level.

| $\bar{d}^{wl}$ $p$-value | UBS | BOVESPA | EUR/USD |
|---|---|---|---|
| SHASH-HYP | 0.0000 | 0.0012 | 0.0388 |
| 5% | 0.0000 | 0.0325 | 0.0925 |
| 1% | 0.0000 | 0.0608 | 0.0951 |
| SHASH-NIG | 0.0000 | 0.0021 | 0.0658 |
| 5% | 0.4707 | 0.0686 | 0.1024 |
| 1% | 0.0000 | 0.0723 | 0.0989 |
| SHASH-SKT | 0.0002 | 0.0047 | 0.1008 |
| 5% | 0.04 | 0.2505 | 0.0423 |
| 1% | 0.0050 | 0.0164 | 0.0916 |
| HYP-NIG | 0.0033 | 0.0282 | 0.1831 |
| 5% | 0.0000 | 0.3470 | 0.1295 |
| 1% | 0.0000 | 0.1191 | 0.1086 |
| HYP-SKT | 0.0040 | 0.0426 | 0.1863 |
| 5% | 0.2938 | 0.0679 | 0.1123 |
| 1% | 0.0000 | 0.0018 | 0.0891 |
| NIG-SKT | 0.0097 | 0.0892 | 0.1923 |
| 5% | 0.0000 | 0.0159 | 0.1061 |
| 1% | 0.3860 | 0.0000 | 0.0813 |

Table 10. $p$-values of the standardized $\bar{d}_{wl}$ statistic using the same censoring rules as above and HAC estimators of the variance.

## 5. Conclusion

In this article, we contribute in two ways to the existing literature. First, we develop a whole methodology to compare the GOF of different density functions unconditionally on 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. 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 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 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 four different distributions (skewed-t, NIG, HYP and SHASH 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, but that NIG, HYP and SHASH distributions could be more suitable if we focus only on the left tail of the data. 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.

## Acknowledgement

This research was supported by the Fonds National de la Recherche Scientifique (Belgium) with an Aspirant (Research Fellow) grant.

## 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 nonnormal 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] 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.
[28] Lamoureux, C., Lastrapes, W., 1990. Persistence in variance, structural change, and the garch

model. Journal of Business & Economic Statistics 8, 225–234.
[29] 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.
[30] Mercurio, D., Spokoiny, V., 2004. Statistical inference for time-inhomogeneous volatility models. The Annals of Statistics 32, 577–602.
[31] Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and its Applications 9, 141142.
[32] Prause, K., 1999. The Generalized Hyperbolic model : estimation, financial derivatives, and risk measures. Ph.D. thesis.
[33] 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.
[34] Scherer, M., Rachev, S., Kim, Y., Fabozzi, F., 2012. Approximation of skewed and leptokurtic return distributions. Applied Financial Economics 22, 1305–1316.
[35] Silverman, B., 1986. Density estimation. Chapman and Hall, London.
[36] 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.
[37] Watson, G., 1964. Smooth regression analysis. Sankhya Series 26, 359–372.
[38] 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 :

| 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. ** indicates that the null hypothesis of a unit root is rejected at the 1% test level.

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 A2.

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).

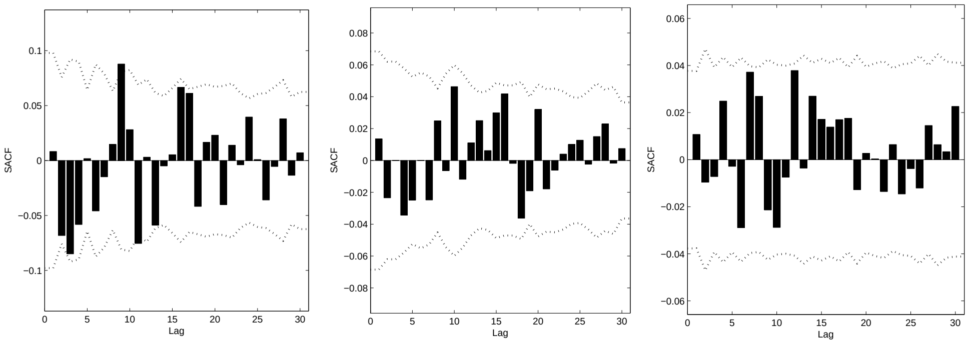



Figure A1. Plot of the ACF with robust standard errors for AR(1) UBS errors, BOVESPA and EUR/USD time series.

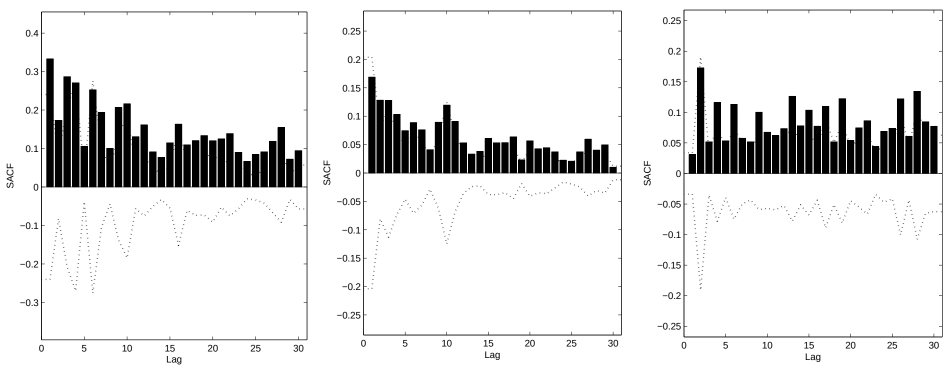



Figure A2. Plot of the ACF with robust standard errors for squared AR(1) UBS errors, BOVESPA and EUR/USD time series.

## Appendix B. Distributions references

### *Generalized Hyperbolic distribution*

The pdf of a GH function is given by [32] :

$$f_{GH}(x;\lambda,\alpha,\beta,\delta,\mu) = a(\lambda,\alpha,\beta,\delta)(\delta^2+(x-\mu)^2)^{(\lambda-\frac{1}{2})/2}G_{\lambda-\frac{1}{2}}(x), \tag{B1}$$

$$G_{\lambda-\frac{1}{2}}(x) = K_{\lambda-\frac{1}{2}}(\alpha\sqrt{\delta^2+(x-\mu)^2})\exp(\beta(x-\mu)) \tag{B2}$$

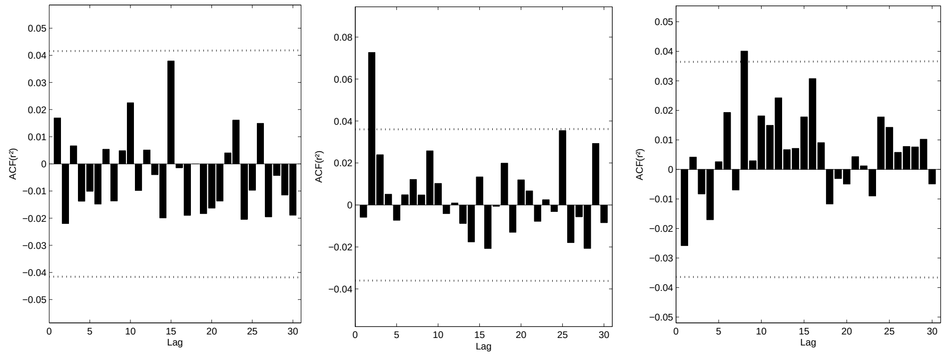



Figure A3. Sample autocorrelations for the squared estimated innovations for the five time series tested up to lag 50. If the bar is up to the dotted line, the autocorrelation at the corresponding lag is significantly different from 0 (with a level of confidence of 95%).

$$a(\lambda, \alpha, \beta, \delta) = \frac{(\alpha^2 - \beta^2)^{\lambda/2}}{\sqrt{2\pi}\alpha^{(\lambda-1/2)}\delta^{\lambda}K_{\lambda}(\delta\sqrt{\alpha^2 - \beta^2})}, \tag{B3}$$

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:

$$f_{SK}(x; \lambda, \nu) = \begin{cases} bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1-\lambda}\right)^2\right)^{-(\nu+1)/2} & x < -a/b, \\ bc\left(1 + \frac{1}{\nu-2}\left(\frac{bx+a}{1+\lambda}\right)^2\right)^{-(\nu+1)/2} & x \geq -a/b, \end{cases} \tag{B4}$$

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 :

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

### Sinh-arcsinh distribution

The two first moments are given by:

$$E(X_{\xi,\eta,\epsilon,\delta}) = \eta(\epsilon/\delta)P_{1/\delta} + \xi, \tag{B5}$$

$$E(X^2_{\xi,\eta,\epsilon,\delta}) = \frac{\eta^2}{2}\{\cosh(2\epsilon/\delta)P_{2/\delta} - 1\}, \tag{B6}$$

where

$$P_q = \frac{e^{1/4}}{(8\pi)^{1/2}}\{K_{(q+1)/2}(1/4) + K_{(q-1)/2}(1/4)\}. \tag{B7}$$

with $K$ being the Bessel function of the second kind.

### Parameters of the fitted distributions

| SHASH parameters | $\xi$ | $\eta$ | $\epsilon$ | $\delta$ |
|---|---|---|---|---|
| UBS | 0.0504 | 0.4846 | -0.0351 | 0.6583 |
| BOVESPA | 0.0967 | 0.6025 | -0.0690 | 0.7384 |
| EUR/USD | -0.0032 | 0.5250 | 0.0022 | 0.6837 |

Table B1. Estimated parameters of the SHASH distributions for the three time series.

| NIG parameters | $\alpha$ | $\beta$ | $\delta$ | $\mu$ |
|---|---|---|---|---|
| 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 | $\alpha$ | $\beta$ | $\delta$ | $\mu$ |
|---|---|---|---|---|
| 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 | $\lambda$ | $\nu$ |
|---|---|---|
| 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. $\alpha$ is the asymmetry parameters and $\nu$ the df