Available online at www.sciencedirect.com

SCIENCE DIRECT®

Statistical Methodology 1 (2004) 105–128

Statistical Methodology

Official Journal of the International Indian Statistical Association

www.elsevier.com/locate/stamet

# The use of GARCH models in VaR estimation

Timotheos Angelidis[a,1], Alexandros Benos[a,*], Stavros Degiannakis[b,2]

[a] *Department of Banking and Financial Management, University of Piraeus, 80, Karaoli & Dimitriou Street, Piraeus GR-185 34, Greece*

[b] *Department of Statistics, Athens University of Economics and Business, 76, Patision Street, Athens GR-104 34, Greece*

Received 26 February 2004

**Abstract**

We evaluate the performance of an extensive family of ARCH models in modeling the daily Value-at-Risk (VaR) of perfectly diversified portfolios in five stock indices, using a number of distributional assumptions and sample sizes. We find, first, that leptokurtic distributions are able to produce better one-step-ahead VaR forecasts; second, the choice of sample size is important for the accuracy of the forecast, whereas the specification of the conditional mean is indifferent. Finally, the ARCH structure producing the most accurate forecasts is different for every portfolio and specific to each equity index.



*JEL classification:* C22; C52; C53; G15

*Keywords:* Value at Risk; GARCH estimation; Backtesting; Volatility forecasting; Quantile loss function

* Corresponding author. Tel.: +30 210 4142 187.

*E-mail addresses:* taggel@unipi.gr (T. Angelidis), abenos@unipi.gr (A. Benos), sdegia@aueb.gr (S. Degiannakis).

1 Tel.: +30 210 4142 187.

2 Tel.: +30 210 8203 120.

1572-3127/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2004.08.004

## 1. Introduction

Following the increase in financial uncertainty in the 1990's, resulting in famous financial disasters, (see [36,21,16] for more information), there has been intensive research from financial institutions, regulators and academics to better develop sophisticated models for market risk estimation. The most well known risk measure is Value-at-Risk (VaR), which refers to a portfolio's worst outcome that is expected to occur over a predetermined period and at a given confidence level.

VaR is an estimation of the tails of the empirical distribution. Many applications presume that asset returns are normally distributed, while it is widely documented that they exhibit skewness and excess kurtosis, resulting in an underestimation or overestimation of the true VaR. Venkataraman [55] and Zangari [58] proposed the use of a mixture of normal distributions, which is fat tailed and able to capture the extreme events compared to the "classical" approaches more easily. Billio and Pelizzon [3] introduced a multivariate switching regime model in order to calculate the VaR for ten Italian stocks and for several portfolios that have been generated by them. They tested the performance of their model using two backtesting measures and concluded that a switching regime specification is more accurate than other known methods, such as RiskMetrics™ [51] or GARCH (1, 1) under normal and Student's-$t$ distribution. Guermat and Harris [28] estimated an exponentially weighted maximum likelihood model for three equity portfolios: forecasts improved at higher confidence levels. Giot and Laurent [26] estimated the daily VaR for stock indices using a skewed Student's distribution and pointed out that it performed better than the pure symmetric one, as it reproduced the characteristics of the empirical distribution more accurately. One year later, they used the same distributional assumption for six commodities [27] and proposed to practitioners this model because it was easily implemented even in a spreadsheet, with predicted VaR numbers closer to expected ones. They stated, however, that more complex models (e.g. APARCH) performed better overall. Based on the influence of asymmetric effects in the accuracy of the VaR estimates, Brooks and Persand [12] concluded that models which do not allow for asymmetries either in the unconditional return distribution or in the volatility specification underestimate the true VaR.

Many researchers prefer to conduct simulations in order to calculate the VaR than to explicitly use a specific parametric distribution. Lambadiaris et al. [39] performed historical and Monte Carlo simulations on Greek stocks and bonds, using two different sample sizes. They concluded that the Monte Carlo method was more appropriate for the stock market, while for bonds, results were dependent on the backtesting procedure and the confidence level. Similarly, Cabedo and Moya [14] developed an ARMA historical simulation method, which improved the simple historical VaR estimation. They used eight years of daily oil prices in order to estimate their model and one year of the same prices to conduct out-of-sample testing. Notice that, in both articles, the importance of different sample sizes is pointed out. Jackson et al. [34] provided evidence that, at higher confidence levels, historical simulation based methods work better than other methods. In contrast, Hendricks [31], Vlaar [56] and Daníelsson [18] supported that an increase of the sample size tends to generate more accurate VaR estimations. Hoppe [32] also examined the issue of the sample size but argued that a smaller sample could lead to more accurate VaR estimates than a larger one. Frey and Michaud [25] argued that a small sample size captures better structural changes due to changes in trading behaviour. To sum up, the choice of an appropriate historical sample size as well as of the adequate model for forecasting volatility should be considered far from resolved.

The purpose of our paper is twofold. First, to implement several volatility models under three distributional assumptions and four historical sample sizes in order to estimate the 95% and 99% one-day VaR for five completely diversified equity index portfolios (S&P 500, NIKKEI 225, FTSE 100, CAC 40 and DAX 30). The different distributions (normal, Student's-*t* and Generalized Error Distribution) will allow the selection of a model for the return tails, while the four sample sizes (500, 1000, 1500 and 2000 observations) will reveal the importance of past data. For robustness purposes, we have used five different stock indices, to avoid results dependent on a specific financial market. Combined with three different extensions of the ARCH family (GARCH, TARCH and EGARCH) and more than 1800 one-step-ahead VaR estimate, our approach produces a total of more than 4 million estimated models! Despite this enormous set (484 models for each index), we still do not include all ARCH specifications available in the literature. We chose, on the contrary, to estimate models that are able to capture the most important characteristics of financial markets.

We then aim at evaluating the predictive accuracy of various models under a risk management framework. We employ a two-stage procedure to investigate the forecasting power of each volatility forecasting technique. In the first stage, two backtesting criteria are implemented to test the statistical accuracy of the models. In the second stage, we employ standard forecast evaluation methods in order to examine whether the differences between models (which have converged sufficiently), are statistically significant. We focus on out-of-sample evaluation criteria because we believe that a model that may be inadequate according to some in-sample evaluation criterion can still yield "better" forecasts in an out-of-sample framework than a correctly specified model.

Our study shows that the more flexible a GARCH model is, the more adequate it is in volatility forecasting, compared to parsimonious models, and that holds for all indices, all distributional assumptions and all confidence levels. Asymmetric models fare better than symmetric ones, as they capture more efficiently the characteristics of the underlying series. As for the choice of distribution, the leptokurtic ones provide better estimators since they perform better in the low probability regions that VaR tries to measure. Moreover, although the use of the entire set of available data is common practice in forecasting volatility, we find out that, at least for some cases, a restricted sample size could generate more accurate one-step-ahead VaR forecasts, since it incorporates changes in trading behaviour more efficiently.

The rest of the paper is organized as follows. Section 2 provides a description of the ARCH models, while Section 3 describes the evaluation framework for VaR estimates. Section 4 presents preliminary statistics for the dataset, explains the estimation procedure and presents the results of the empirical investigation of the estimated models for the five equity indices. Section 5 concludes.

## 2. Volatility models

Let $y_t = \ln(S_t/S_{t-1})$ denote the continuously compound rate of return from time $t-1$ to $t$, where $S_t$ is the asset price at time $t$. We assume that the time series of interest, $y_t$, is decomposed into two parts, the predictable and unpredictable component, $y_t = \mathbb{E}(y_t|I_{t-1}) + \varepsilon_t$, where $I_{t-1}$ is the information set at time $t-1$, $\mathbb{E}$ is the conditional mean operator and $\varepsilon_t$ is the unpredictable part, or innovation process. The conditional mean return is considered as a $k$-th order autoregressive process, $AR(k)$:

$$\mathbb{E}(y_t|I_{t-1}) \equiv c_0 + \sum_{i=1}^{k} c_i y_{t-i}.$$

The autoregressive process allows for the autocorrelation induced by discontinuous (or non-synchronous) trading in the stocks making up an index [53,43]. The unpredictable component, $\varepsilon_t$, can be expressed as an ARCH process in the following form:

$$\varepsilon_t = z_t \sigma_t,$$

where $z_t$ is a sequence of independently and identically distributed random variables with zero mean and unit variance. The conditional variance of $\varepsilon_t$ is $\sigma_t^2$, a time-varying, positive and measurable function of the information set at time $t-1$. Note that, even though the innovation process for the conditional mean is serially uncorrelated, it is not time independent.

Engle [22] introduced the ARCH($q$) model and expressed the conditional variance as a linear function of the past $q$ squared innovations

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2.$$

For the conditional variance to be positive, the parameters must satisfy $a_0 > 0$ and $a_i \geq 0$ for $i = 1, \ldots, q$. Bollerslev [5] proposed a generalization of the ARCH model, the GARCH($p, q$) model:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2, \tag{1}$$

where $a_0 > 0$, $a_i \geq 0$ for $i = 1, \ldots, q$, and $b_j \geq 0$ for $j = 1, \ldots, p$. If $\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$, then the process $\varepsilon_t$ is covariance stationary and its unconditional variance is equal to

$$\sigma^2 = a_0 \Bigg/ \left(1 - \sum_{i=1}^{q} a_i - \sum_{j=1}^{p} b_j\right).$$

A special case of the GARCH family is the Exponentially Weighted Moving Average (EWMA) alternative, used by the company RiskMetrics™, when they introduced their analytic VaR methodology. The volatility forecast is the weighted average of the previous period's forecast and the current squared return. Return variance is an exponentially declining process $\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1-\lambda)\varepsilon_{t-1}^2$, a special case of GARCH(1, 1) with a zero intercept and the two remaining parameters summing to one. RiskMetrics™ use $\lambda = 0.94$ for daily data and go 75 data points backwards in their estimation horizon.

The GARCH($p, q$) model successfully captures several characteristics of financial time series, such as thick tailed returns and volatility clustering, as Mandelbrot [46] noted: "... large changes tend to be followed by large changes of either sign, and small changes tend to be followed by small changes...". On the other hand, the GARCH structure presents some implementation drawbacks, since the variance depends only on the magnitude and not on the sign of $\varepsilon_t$, which is somewhat at odds with the empirical behaviour of stock market prices where a leverage effect may be present.[3] This term, introduced by Black [4], refers to the tendency for changes in stock returns to be negatively correlated with changes in returns' volatility so that volatility tends to rise in response to bad news, ($\varepsilon_t < 0$), and to fall in response to good news ($\varepsilon_t > 0$). Moreover, Brooks and Persand [12] state that a VaR model which does not account for asymmetries in volatility specification is most likely to generate inaccurate forecasts.

In order to capture the asymmetry observed in the data, a new class of models was introduced, termed the asymmetric ARCH models. The most popular model proposed to capture the asymmetric effects is Nelson's [48] exponential GARCH, or EGARCH$(p, q)$, model:

$$\ln(\sigma_t^2) = a_0 + \sum_{i=1}^{q} \left( a_i \left| \frac{\varepsilon_{t-i}}{\sigma_{t-i}} \right| + \gamma_i \frac{\varepsilon_{t-i}}{\sigma_{t-i}} \right) + \sum_{j=1}^{p} b_j \ln(\sigma_{t-j}^2). \tag{2}$$

In contrast to the GARCH model, no restrictions need to be imposed on the model estimation, since the logarithmic transformation ensures that the forecasts of the variance are non-negative. The parameters $\gamma_i$ allow for the asymmetric effect. If $\gamma_1 = 0$ then a positive surprise, $\varepsilon_t > 0$, has the same effect on volatility as a negative surprise, $\varepsilon_t < 0$. The presence of the leverage effect can be investigated by testing the hypothesis that $\gamma_1 < 0$.

The number of possible conditional volatility formulations is vast. The threshold GARCH, or TARCH$(p, q)$, model belongs to the most widely used:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \gamma_1 \varepsilon_{t-1}^2 d_{t-1} + \sum_{j=1}^{p} b_j \sigma_{t-j}^2, \tag{3}$$

where $d_t = 1$ if $\varepsilon_t < 0$, and $d_t = 0$ otherwise. It, hence, allows a response of volatility to news with different coefficients for good and bad news.

Although there is a vast number of ways to parameterize the asymmetry, these two families of models are the most widely known ones (see [24,12,13] among others).[4] Moreover, most econometric packages include routines for the estimation of conditional variance under these families and can, therefore, be used by a risk manager effortlessly.

As concerns the distribution of $z_t$, Engle [22], who introduced the ARCH process, assumed it is normal. Bollerslev [6], on the other hand, proposed the standardized $t$-distribution with $v > 2$ degrees of freedom with a density given by

$$D(z_t; v) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{\pi(v-2)}} \left(1 + \frac{z_t^2}{v-2}\right)^{-\frac{v+1}{2}}, \tag{4}$$

[3] In such a case, a stochastic volatility model may be more appropriate. There is some evidence that these models may perform better than multivariate GARCH models [17] but they are also more difficult to estimate.

[4] A wide range of ARCH models proposed in the literature has been reviewed by Bollerslev et al. [8], Bera and Higgins [2], Bollerslev et al. [9], Hamilton [29], and Degiannakis and Xekalaki [19].

where $\Gamma(v) = \int_0^\infty e^{-x} x^{v-1} dx$ is the gamma function and $v$ is the parameter which described the thickness of the distribution tails. The Student's-$t$ distribution is symmetric around zero and, for $v > 4$, the conditional kurtosis equals $3(v-2)/(v-4)$, which exceeds the normal value of three. For large values of $v$, its density converges to that of the standard normal. Nelson [48] suggested another "fat-tails" distribution, the generalized error distribution (GED) with density equal to

$$D(z_t; v) = \frac{v \exp(-0.5|z_t/\lambda|^v)}{2^{(1+1/v)}\Gamma(v^{-1})\lambda}, \qquad v > 0, \tag{5}$$

where $v$ is the tail-thickness parameter and

$$\lambda \equiv \left[\frac{\Gamma\left(\frac{1}{v}\right)}{2^{\frac{2}{v}}\Gamma\left(\frac{3}{v}\right)}\right]^{\frac{1}{2}}.$$

When $v = 2$, $z_t$ is standard normally distributed. For $v < 2$, its distribution has thicker tails than the normal one (for example, for $v = 1$, $z_t$ follows the double exponential distribution) while for $v > 2$, it has thinner tails (for $v = \infty$, $z_t$ has a uniform distribution on the interval $(-\sqrt{3}, \sqrt{3})$). Guermat and Harris [28] applied the Student's-$t$ distribution and Longerstay [44] used a mixture of normal distributions. The Gram–Charlier type distribution [42], the generalized $t$ distribution [9], the skewed $t$ distribution [40,41], the normal Poisson mixture distribution [35] and the normal lognormal mixture [33] have also been employed.

The common methodology used for ARCH estimation is Maximum Likelihood. Under the assumption of IID innovations and for $D(z_t; v)$ denoting their density function, the log-likelihood function of $\{y_t(\theta)\}$ for a sample of $T$ observations is given by

$$\mathbb{L}_T(\{y_t\}; \theta) = \sum_{t=1}^{T}\left[\ln[D(z_t(\theta); v)] - \frac{1}{2}\ln(\sigma_t^2(\theta))\right], \tag{6}$$

where $\theta$ is the vector of the parameters that have to be estimated for the conditional mean, conditional variance and density function, while $z_t(\theta) = \frac{\varepsilon_t(\theta)}{\sigma_t(\theta)}$. The likelihood estimator $\hat{\theta}$ maximizes (6). Therefore, the log-likelihood function for a sample of $T$ observations is as follows:

1. For normally distributed $z_t$'s:

$$\mathbb{L}_T(\{y_t\}; \theta) = -\frac{1}{2}\left[T\ln(2\pi) + \sum_{t=1}^{T} z_t^2 + \sum_{t=1}^{T}\ln(\sigma_t^2)\right]. \tag{7}$$

2. For standardized $t$-distributed $z_t$'s:

$$\mathbb{L}_T(\{y_t\}; \theta) = T\left[\ln\Gamma\left(\frac{v+1}{2}\right) - \ln\Gamma\left(\frac{v}{2}\right) - \frac{1}{2}\ln[\pi(v-2)]\right] - \frac{1}{2}\sum_{t=1}^{T}\left[\ln(\sigma_t^2) + (1+v)\ln\left(1 + \frac{z_t^2}{v-2}\right)\right]. \tag{8}$$

3. For GED distributed $z_t$'s:

$$\mathbb{L}_T(\{y_t\};\theta) = \sum_{t=1}^{T}\left[\ln\left(\frac{v}{\lambda}\right) - \frac{1}{2}\left|\frac{z_t}{\lambda}\right|^{v} - (1+v^{-1})\ln(2) - \ln\Gamma\left(\frac{1}{v}\right) - \frac{1}{2}\ln(\sigma_t^2)\right]. \tag{9}$$

Maximum likelihood estimates of the parameters are obtained by employing numerical maximization procedure according to the Marquardt algorithm [47]. We, instead, use the quasi-maximum likelihood estimator (QMLE) since, according to Bollerslev and Wooldridge [7], it is generally consistent, has a normal limiting distribution and provides asymptotic standard errors that are valid under non-normality.

Summing up, the one-step-ahead conditional variance forecast, $\hat{\sigma}^2_{t+1|t}$, for the GARCH$(p,q)$ model equals

$$\hat{\sigma}^2_{t+1|t} = a_0^{(t)} + \sum_{i=1}^{q} a_i^{(t)}\varepsilon^2_{t-i+1} + \sum_{j=1}^{p} b_j^{(t)}\sigma^2_{t-j+1}. \tag{10}$$

For the EGARCH$(p,q)$ model, we get instead

$$\ln\hat{\sigma}^2_{t+1|t} = a_0^{(t)} + \sum_{i=1}^{q}\left[a_i^{(t)}\left|\frac{\varepsilon_{t-i+1}}{\sigma_{t-i+1}}\right| + \gamma_i^{(t)}\left(\frac{\varepsilon_{t-i+1}}{\sigma_{t-i+1}}\right)\right] + \sum_{j=1}^{p} b_j^{(t)}\ln(\sigma^2_{t-i+1}). \tag{11}$$

Finally, the corresponding one-step-ahead conditional variance forecast in the case of the TARCH$(p,q)$ model is

$$\hat{\sigma}^2_{t+1|t} = a_0^{(t)} + \sum_{i=1}^{q}\left[a_i^{(t)}\varepsilon^2_{t-i+1}\right] + \gamma^{(t)}\varepsilon_t^2 d_t + \sum_{j=1}^{p}\left[b_j^{(t)}\sigma^2_{t-j+1}\right]. \tag{12}$$

Hence, it is straightforward to compute the one-step-ahead VaR forecasts under all distributional assumptions and for zero mean observations as

$$\mathrm{VaR}_{t+1|t} = F(\alpha)\hat{\sigma}_{t+1|t}, \tag{13}$$

with $F(\alpha)$ being the corresponding quantile (95th or 99th) of the assumed distribution, and $\hat{\sigma}_{t+1|t}$ being the forecast of the conditional standard deviation at time $t+1$ given the information at time $t$.

## 3. Evaluating the different approaches

Our objective is to test these different volatility forecasting techniques, not in an econometric laboratory but in a *risk management environment*, so we must first choose a metric for the "quality" of such forecasts. It is well known that there are many sources of error in VaR figures: sampling errors, data problems, inappropriate specification, model errors, etc. All these factors will cause our estimate often to be biased. We must also tackle the fact that a VaR value is inherently unobservable. We, therefore, have to first monitor the VaR values and then judge the models, by checking whether they are consistent with

Table 1
Unconditional coverage "no rejection" regions for a 5% test size

| Confidence level (%) | Evaluation sample size | | | |
|---|---|---|---|---|
| | 250 | 500 | 750 | 1000 |
| 5 | $7 \leq N \leq 19$ | $17 \leq N \leq 35$ | $27 \leq N \leq 49$ | $38 \leq N \leq 64$ |
| 1 | $1 \leq N \leq 6$ | $2 \leq N \leq 9$ | $3 \leq N \leq 13$ | $5 \leq N \leq 16$ |
| 0.5 | $0 \leq N \leq 4$ | $1 \leq N \leq 6$ | $1 \leq N \leq 8$ | $2 \leq N \leq 9$ |
| 0.1 | $0 \leq N \leq 1$ | $0 \leq N \leq 2$ | $0 \leq N \leq 3$ | $0 \leq N \leq 3$ |
| 0.01 | $0 \leq N \leq 0$ | $0 \leq N \leq 0$ | $0 \leq N \leq 1$ | $0 \leq N \leq 1$ |

subsequently realized returns given the confidence interval on which the VaR values were constructed in the first place.

Finally, an adequate model must not only generate statistically accurate VaR values, but it also has to be "preferred" over other equally adequate ones. Statistical adequacy will be tested based on Kupiec's [38] and Christoffersen's [15] backtesting measures, while the equivalent statistical adequacy of models will be performed via a *loss function*.

### 3.1. *Unconditional coverage*

Let $N = \sum_{t=1}^{T} I_t$ be the number of days over a $T$ period that the portfolio loss was larger than the VaR estimate, where

$$I_{t+1} = \begin{cases} 1, & \text{if } y_{t+1} < \mathrm{VaR}_{t+1|t} \\ 0, & \text{if } y_{t+1} \geq \mathrm{VaR}_{t+1|t}. \end{cases}$$

Hence, $N$ is the observed number of *exceptions* in the sample. The failure number [38] follows a binomial distribution, $N \sim B(T, p)$, and consequently the appropriate likelihood ratio statistic, under the null hypothesis that the expected exception frequency $N/T = p$, equals

$$2 \ln\left[\left(1 - \frac{N}{T}\right)^{T-N} \left(\frac{N}{T}\right) N\right] - 2 \ln[(1-p)^{T-N} p^N].$$

Asymptotically, this test is $\chi^2$-distributed with one degree of freedom. In Table 1, we present the no rejection regions of $N$ for various sample sizes and confidence levels. This test can reject a model for both high and low failures but, as stated by Kupiec [38], its power is generally poor. So we turn to a more elaborate criterion.

### 3.2. *Conditional coverage*

A more complete test was made by Christoffersen [15], who developed a likelihood ratio statistic to test the joint assumption of unconditional coverage and independence of failures. Its main advantage over the previous statistic is that it takes account of any conditionality in forecasts: if volatilities are low in some period and high in others, the forecast should respond to this clustering event. The Christoffersen procedure enables us to separate clustering effects from distributional assumption effects. His statistic is computed as

$$-2\ln[(1-p)^{T-N}p^N] + 2\ln[(1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}] \sim \chi^2(2), \tag{14}$$

where $n_{ij}$ is the number of observations with value $i$ followed by $j$, for $i, j = 0, 1$ and

$$\pi_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$$

are the corresponding probabilities. The values $i, j = 1$ denote that an exception has been made, while $i, j = 0$ indicates the opposite. If the sequence of values is independent, then the probabilities to observe or not a VaR violation in the next period must be equal, which can be written more formally as $\pi_{01} = \pi_{11} = p$. The main advantage of this test is that it can reject a VaR model that generates either too many or too few clustered violations, but needs several hundred observations in order to be accurate.

### 3.3. *Loss functions*

The statistical adequacy of VaR forecasts is obtained by the two previous procedures: if the null hypothesis cannot be rejected, we will characterize the model as an adequate one for volatility forecasting. These metrics, however, cannot conclude whether an "adequate" model is more accurate than another "adequate" one.

Lopez [45] suggested measuring the *accuracy* of VaR forecasts on the basis of the distance between observed returns and forecasted VaR values. He defined a penalty variable as follows:

$$\Psi_{t+1} = \begin{cases} 1 + (y_{t+1} - \mathrm{VaR}_{t+1|t})^2, & \text{if } y_{t+1} < \mathrm{VaR}_{t+1|t}, \\ 0, & \text{if } y_{t+1} \geq \mathrm{VaR}_{t+1|t}. \end{cases}$$

In his paper, a VaR model is penalized when an exception takes place, so it is preferred to another if it yields a lower total loss value, defined as the sum of these penalty variables: $\Psi = \sum_{t=1}^{T} \Psi_t$. This function incorporates both the cumulative number and the magnitude of exceptions. Compared to Kupiec's [38] binomial function, $\Psi$ adds up the magnitude term, so the larger the failure the higher the penalty added. The VaR forecast must neither overestimate nor underestimate the "true" VaR number as, in both cases, the financial institution allocates the wrong amount of capital.[5] In the former case, regulators charge it a higher than really needed amount of capital, worsening its performance; in the latter, the regulatory capital set aside may not be enough to cover it from market risks. Consequently, the "true" but unobservable VaR is proxied using the empirical distribution of future *realized* returns. For example, if $T$ observations are available for an out-of-sample evaluation, then their p-quantile is a proxy for the "true" VaR at the p% confidence level. The proposed loss function, named Quantile Loss (QL) function, has the

[5] According to the Basle Committee on Banking Supervision's January 1996 "*Amendment to the Capital Accord to Incorporate Market Risks*", the Value-at-Risk methodology can be used by financial institutions to calculate capital charges in respect of their interest rate, equity, foreign exchange and commodities risk.

Table 2
Descriptive statistics of the daily log returns, for the period of 9 July 1987 to 18 October 2002

| | S&P 500 | NIKKEI 225 | DAX 30 | CAC 40 | FTSE 100 |
|---|---|---|---|---|---|
| Mean | 0.00027 | −0.00026 | 0.00021 | 0.00020 | 0.00014 |
| Median | 0.00042 | −0.00011 | 0.00079 | 0.00034 | 0.00048 |
| Maximum | 0.08709 | 0.12430 | 0.07553 | 0.08225 | 0.07597 |
| Minimum | −0.22833 | −0.16135 | −0.13710 | −0.10138 | −0.13029 |
| Std. Deviation | 0.01140 | 0.01494 | 0.01455 | 0.01377 | 0.01091 |
| Skewness | −2.28112 | −0.05997 | −0.57695 | −0.32182 | −0.87756 |
| Kurtosis | 49.28424 | 10.10980 | 9.78510 | 7.39398 | 14.39124 |
| Jarque–Bera | 347,258.8 | 7936.4 | 7569.2 | 3143.9 | 21,348.6 |
| Probability | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

following form:

$$\Psi_{t+1} = \begin{cases} (y_{t+1} - \mathrm{VaR}_{t+1|t})^2, & \text{if } y_{t+1} < \mathrm{VaR}_{t+1|t}; \\ [\mathrm{Percentile}(y, 100p)_1^T - \mathrm{VaR}_{t+1|t}]^2, & \text{if } y_{t+1} \geq \mathrm{VaR}_{t+1|t}. \end{cases}$$

For every instant $t$, a model's loss increases either, (a) by the distance between the $\mathrm{VaR}_{t+1|t}$ forecast and the future 100p percentile of the empirical return distribution, if the return is larger than the VaR forecasted for the same period,[6] or (b) by the excess loss $((y_{t+1} - \mathrm{VaR}_{t+1})^2)$ term, if the realized return is smaller than the forecasted VaR (i.e., a larger loss than predicted occurred). Under this framework, a model will be preferred over its equivalent ones if it minimizes the QL function. Based on Diebold and Mariano [20] and Sarma et al. [52], we create a hypothesis test of the forecasting ability of the models. Let $z_{t+1} = \Psi^A_{t+1} - \Psi^B_{t+1}$, where $\Psi^A$ and $\Psi^B$ are the loss functions of models $A$ and $B$, respectively. A negative value of $z_{t+1}$ indicates that model $A$ is superior to model $B$. The Diebold–Mariano [20] statistic is the "$t$-statistic" for a regression of $z_{t+1}$ on a constant with heteroskedastic and autocorrelated consistent standard errors (HAC).[7]

## 4. Data and results

We generate out-of-sample VaR forecasts for five equity indices (CAC 40, DAX 30, FTSE 100, NIKKEI 225 and S&P 500), obtained from DataStream for the period of 9 July 1987 to 18 October 2002. For all indices, we compute daily log returns. A sample plot is enough to observe volatility clustering for all indices (Fig. 1). Table 2 provides summary statistics as well as the Jarque–Bera value. In all cases, the null hypothesis of normality is rejected at any level of significance, as there is evidence of significant excess kurtosis and negative skewness in all markets.

[6] Notice that this percentile remains constant throughout the out-of-sample testing period.
[7] For more details about such HAC standard errors, see White [57] and Newey and West [49].

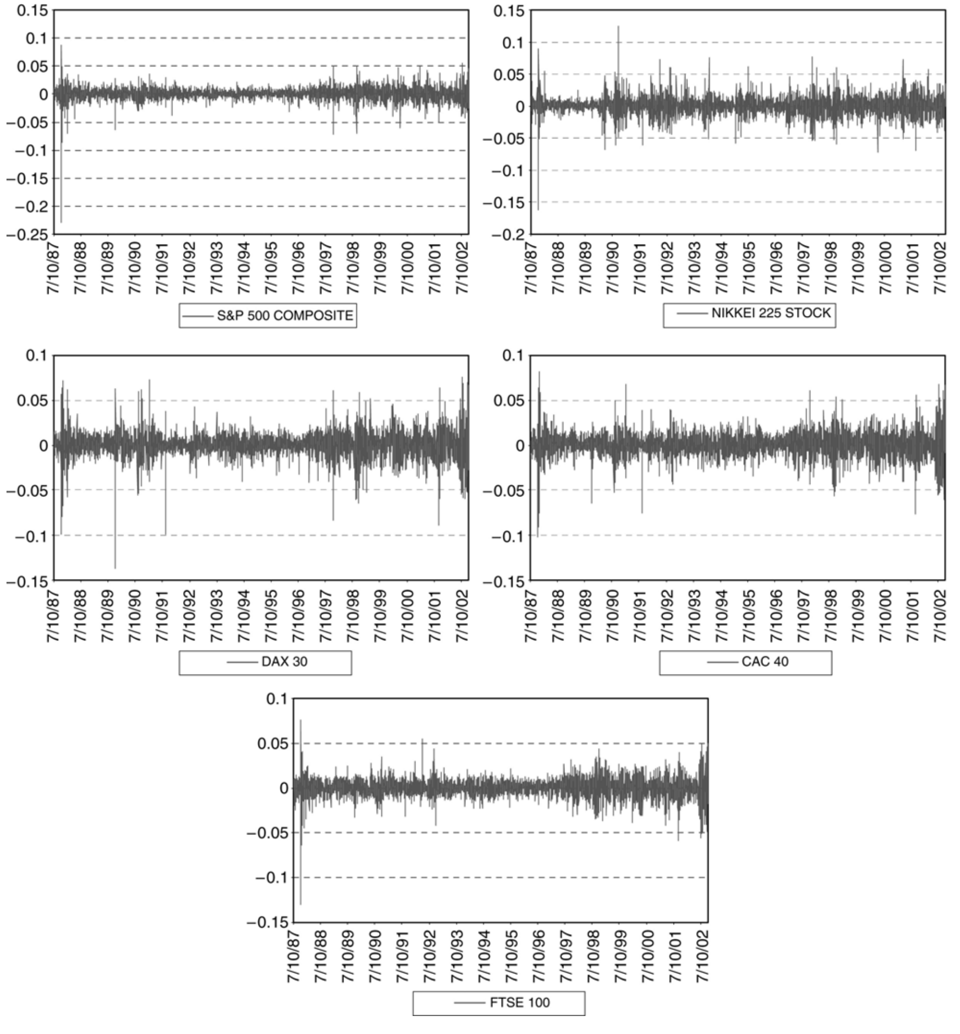


Fig. 1. Continuously Compounded Daily Returns of Equity Indices (S&P 500, NIKKEI 225, DAX 30, CAC 40 and FTSE 100) from 9 July 1987 to 18 October 2002.

We estimate the most frequently applied model to empirical studies, the AR(1) GARCH(1, 1) specification:

$$y_t = c_0 + c_1 y_{t-1} + z_t \sigma_t \tag{15}$$

$$z_t \overset{iid}{\sim} D(0, 1) \tag{16}$$

$$\sigma_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2 \tag{17}$$

assuming three different density functions for $z_t$: the normal, the Student's-$t$ and the GED ones as presented in Section 2. The models are estimated using the entire dataset available and the results are presented in Table 3. We note (a) that the conditional variance parameters are highly significant; (b) that the distribution of the $z_t$ has significantly thicker tails than the normal one, and (c) that the parameter $c_1$, which allows for the autocorrelation induced by discontinuous trading, is not significant in any case.

In our study, we model the conditional mean as a $k$-th order autoregressive process and the conditional variance as a GARCH, EGARCH or TARCH process. Using Eqs. (7)–(9) presented in Section 2, we will first apply ARCH processes to a GARCH, an EGARCH and a TARCH model, all with parameters $(p, q)$, assuming a normal distribution for innovations. We estimate them using integer values for parameters $k$, $p$, $q$ as follows: $k = 0, \dots, 4$, $p = 0, 1, 2$ and $q = 1, 2$, yielding a total of 85 models.[8] Then, we assume standardized residuals follow a thick tailed distribution (Student's-$t$ and GED) and reapply the AR(1) model to the same GARCH$(p, q)$, EGARCH$(p, q)$ and TARCH$(p, q)$ frameworks, yielding another 34 models. These latter models are able to capture volatility clustering, non-synchronous trading, leverage effects and thick tailed returns, and, moreover, converge more frequently.[9]

For all models and all equity indices, we use a rolling sample of 500, 1000, 1500 and 2000 observations with the same number of VaR forecasts for each sample size. We generate one-day VaR forecasts for both 95% and 99% confidence levels, as recommended by the Basel Committee for Banking Supervision. Model parameters were re-estimated every trading day and all tests were performed using the information available at that time. In contrast, Klaassen [37] and Hansen and Lunde [30], among others, have estimated in-sample parameters using the entire dataset and derived one-step-ahead volatility forecasts, based on these estimates. We believe estimated parameters incorporate information about trading behaviour, a variable changing through time. Therefore, their estimation should be based only on the most recently available information set.

Finally, we do not use classical in-sample model selection criteria, such as the Akaike Information one [1] or the Schwarz Bayesian Criterion [54], widely used in this literature. A good in-sample performance for a model is not a prerequisite for a good out-of-sample performance of the same model. Pagan and Schwert [50], using a collection of parametric and non-parametric models of volatility forecasting (kernels, Fourier series and two-stage OLS), found that non-parametric methods did a good job in the in-sample forecasting but parametric ones yielded superior out-of-sample forecasts. Hansen and Lunde [30] noted that "a model, which accommodates a significant in-sample relation, need not result in better out-of-sample forecasts, compared to the forecasts obtained from a more parsimonious model". Moreover, Brooks and Burke [10] argued that the "the information criteria are not a panacea in terms of generating uniformly more precise forecasts". Given these facts, we chose not to restrict our analysis to models that have been pre-selected by any standard in-sample model selection method, such as the ones cited above.

[8] Numerical maximization of the log-likelihood function for the EGARCH(2, 2) model, has failed to converge. We, therefore, excluded the five EGARCH models for these parameters from our results.

[9] If the numerical maximization of the log-likelihood function failed to converge more than four times, we excluded the models from our results. This number of failures (four) was the maximum number any model failed to converge under the normal distribution assumption. If the number of failures was less than or equal to four, VaR forecasts were computed based on previous trading day parameter estimations.

Table 3
Parameter estimates of the AR(1) GARCH(1, 1) model for five indices, using the entire dataset (9 July 1987 to 18 October 2002) and assuming three different distributions for the standardized residuals

| **Parameter** | **S&P 500** | **NIKKEI 225** | **DAX 30** | **CAC 40** | **FTSE 100** |
|---|---|---|---|---|---|
| | *Normal distribution* | | | | |
| $c_0$ | 0.000557 | 0.000549 | 0.000614 | 0.000432 | 0.000395 |
| | (0.000132) | (0.000192) | (0.000199) | (0.000197) | (0.000150) |
| $c_1$ | 0.028859 | 0.015127 | 0.037770 | 0.047177 | 0.058032 |
| | (0.017917) | (0.019737) | (0.019223) | (0.017319) | (0.017804) |
| $a_0$ | 1.64E−06 | 4.43E−06 | 6.76E−06 | 5.57E−06 | 3.22E−06 |
| | (5.53E−07) | (1.68E−06) | (3.07E−06) | (1.95E−06) | (1.01E−06) |
| $a_1$ | 0.103524 | 0.159935 | 0.136356 | 0.105451 | 0.116497 |
| | (0.037118) | (0.049919) | (0.035324) | (0.018500) | (0.026331) |
| $b_1$ | 0.890270 | 0.838772 | 0.837132 | 0.865466 | 0.859395 |
| | (0.030393) | (0.040506) | (0.033176) | (0.023878) | (0.026836) |
| Log likelihood | 12468.68 | 10913.64 | 11357.62 | 11397.68 | 12492.48 |
| | *Student's-t distribution* | | | | |
| $c_0$ | 0.000594 | 0.000275 | 0.000711 | 0.000531 | 0.000409 |
| | (0.000121) | (0.000171) | (0.000163) | (0.000178) | (0.000135) |
| $c_1$ | 0.012507 | −0.006850 | 0.017391 | 0.042970 | 0.049854 |
| | (0.016490) | (0.016965) | (0.017283) | (0.017279) | (0.017128) |
| $a_0$ | 8.77E−07 | 2.07E−06 | 2.61E−06 | 3.70E−06 | 2.21E−06 |
| | 2.23E−07 | (5.11E−07) | (6.01E−07) | (8.47E−07) | (4.60E−07) |
| $a_1$ | 0.070246 | 0.103547 | 0.103558 | 0.086050 | 0.087884 |
| | (0.008139) | (0.011004) | (0.011609) | (0.010117) | (0.010318) |
| $b_1$ | 0.925879 | 0.895057 | 0.886968 | 0.894520 | 0.891897 |
| | (0.007779) | (0.009965) | (0.011657) | (0.012412) | (0.011782) |
| $v$ | 5.60013 | 6.097681 | 7.947894 | 10.49132 | 11.07864 |
| | (0.429782) | (0.473549) | (0.547896) | (1.124427) | (0.927738) |
| Log likelihood | 12638.65 | 11076.65 | 11532.06 | 11445.73 | 12578.62 |
| | *Generalized error distribution* | | | | |
| $c_0$ | 0.000560 | 0.000306 | 0.000733 | 0.000454 | 0.000432 |
| | (0.000116) | (0.000166) | (0.000163) | (0.000178) | (0.000135) |
| $c_1$ | −0.000820 | −0.006770 | 0.017969 | 0.038205 | 0.046718 |
| | (0.015669) | (0.016497) | (0.017162) | (0.017206) | (0.017111) |
| $a_0$ | 1.11E−06 | 3.19E−06 | 4.13E−06 | 4.78E−06 | 2.67E−06 |
| | (2.60E−07) | (5.96E−07) | (8.16E−07) | (9.61E−07) | (5.60E−07) |
| $a_1$ | 0.078919 | 0.119683 | 0.118569 | 0.094767 | 0.098021 |
| | (0.007099) | (0.012156) | (0.011676) | (0.010579) | (0.011616) |
| $b_1$ | 0.915449 | 0.875434 | 0.866873 | 0.879949 | 0.879750 |
| | (0.007706) | (0.011438) | (0.013691) | (0.013724) | (0.013507) |
| $v$ | 1.250571 | 1.305231 | 1.395110 | 1.552809 | 1.545649 |
| | (0.024604) | (0.023113) | (0.021377) | (0.035018) | (0.019470) |
| Log likelihood | 12616.52 | 11040.99 | 11471.80 | 11429.26 | 12541.29 |
| Observations | 3853 | 3767 | 3835 | 3826 | 3857 |

Standard errors are presented in parentheses.

Due to the enormous size of results generated, we will proceed in their presentation as follows. First, we will study VaR forecasts estimated using the standard normal distribution assumption for all sample sizes and examine whether their performance depends on the choice of sample size. We will then go on to exhibit the contribution of alternative distributions to the VaR framework.

### *4.1. Normal distribution*

The assumption of normality produces very weak results: the vast majority of models, irrespective of the sample size chosen, understate the true one-day 95% VaR estimate since exception rates are higher than the 5% predicted level.[10] The performance of the models is heavily dependent on the stock index, with the NIKKEI 225 index faring somewhat better than the others. $P$-values for both conditional and unconditional coverage are relatively low for all indices, with European markets showing the lowest probability values.

More specifically, we observe that, in the 2000 and the 1500 sample sizes, none out of the 85 models produces an exception rate below the predicted 5%. In a very few cases, the estimated failure rate is close enough to the expected one (for example, the AR(1) GARCH(1, 1) model achieves the closest fit for the S&P 500 index). In almost all cases, either the different specification of the conditional mean or the introduction of only ARCH terms in the conditional variance does not significantly improve either the unconditional or the conditional coverage of the models. In the case of the S&P 500 index, the average $p$-value for the null hypothesis of unconditional coverage of the GARCH(1, 1) family is 56%, while the best performing family for the NIKKEI 225 and the CAC 40 indices is the EGARCH(1, 2) one, with corresponding average $p$-values equal to 79.1% and 13.6%, respectively. Similarly, for the DAX 30 and the FTSE 100 indices, we conclude that the null hypothesis of correct unconditional coverage is rejected for all the models at the 10% level of significance, while the hypothesis of conditional coverage is rejected at the 15% level.

As for the effect of the sample size, it seems that as the latter gets smaller, the results improve at all levels. This is due to the weight given to the latest observations when using a smaller sample: the *smaller* sample size captures only the *latest* market movements and the produced VaR estimate is, hence, less dependent on the long run volatility trend, in the same sense that a 200-point moving average is smoother than a 50-point one. Consequently, the VaR estimate is less often rejected at the 95% confidence level, for all models and indices. Such a behavioural trend is traced even at the 1000 observations sample, and becomes clearer at the 500 observations level, at least for three out of five indices. The only exception to this "rule of thumb" is the S&P 500, where the highest $p$-values are achieved for the largest sample size.

For the 99% VaR estimates, the exception rates are all higher than the predicted 1% for all sample sizes, with the sample size having no effect at all in improving results. As for the coverage metrics, they are very low, ranging from 45.51% conditional coverage, for the EGARCH family on the NIKKEI 225 index, down to 1.19% for the same metric with the same modeling family applied to the S&P 500 index. Such performance, at least

[10] The discussion in the Sections 4.1 and 4.2 is based on the exception rates and the $p$-values of unconditional and conditional coverage for all models, equity indices, distributional assumptions, and sample sizes that due to space limitations are available to readers upon request.

Table 4
Best performed models and the relative probability values of coverage metrics for normally distributed innovations

| Index | Unconditional coverage | | | Conditional coverage | | |
|---|---|---|---|---|---|---|
| | Size | Model | Pr. (%) | Size | Model | Pr. (%) |
| | *95% daily VaR forecasts* | | | | | |
| S&P 500 | 2000 | AR(1) GARCH(1, 1) | 72.25 | 2000 | AR(1) GARCH(1, 1) | 79.48 |
| NIKKEI 225 | 1000 | AR(0) EGARCH(1, 1) | 96.95 | 500 | AR(3) EGARCH(1, 2) | 93.14 |
| DAX 30 | 500 | AR(3) TARCH(1, 2) | 32.92 | 500 | AR(3) TARCH(1, 2) | 57.61 |
| CAC 40 | 1000 | AR(1) EGARCH(1, 2) | 54.44 | 1000 | AR(1) EGARCH(1, 2) | 56.38 |
| FTSE 100 | 500 | AR(0) EGARCH(2, 1) | 33.72 | 500 | AR(1) TARCH(2, 1) | 53.27 |
| | *99% daily VaR forecasts* | | | | | |
| S&P 500 | 1000 | AR(0) EGARCH(1, 2) | 1.40 | 1000 | AR(0) EGARCH(1, 2) | 1.19 |
| NIKKEI 225 | 2000 | AR(0) EGARCH(2, 1) | 31.87 | 2000 | AR(0) EGARCH(2, 1) | 45.51 |
| DAX 30 | 500 | AR(3) GARCH(1, 1) | 13.92 | 500 | AR(3) GARCH(1, 1) | 23.40 |
| CAC 40 | 500 | AR(2) GARCH(1, 2) | 8.70 | 500 | AR(2) GARCH(1, 2) | 15.60 |
| FTSE 100 | 1000 | AR(0) EGARCH(1, 1) | 22.56 | 1000 | AR(0) EGARCH(1, 1) | 34.59 |

for the latter index, is suggestive of a *hump* in the tail distribution of returns: this means it seriously underestimates risk at a high level of confidence whereas it produces somewhat better results at the lower confidence level, yielding respectable $p$-values for the 95% VaR metric but extremely low ones for the 99% one.

Table 4 shows the most appropriate models according to conditional and unconditional coverage, defined as the one with the highest $p$-value among all the estimated models for all sample sizes.[11] Brooks and Persand [13] also ranked models, based on the following rule: they excluded those techniques that produced an exception rate greater than the expected one. They preferred the models which generated a proportion of failures closer to the expected exception frequency. Both coverage measures suggest the same models in most of the cases. For all indices, $p$-values are deemed relatively low, with the weakest ones for the European indices: we believe that this is due to the inadequate description of such data using the normal distribution. The normal distribution performs best with the NIKKEI 225 index in both the 95% and the 99% level, since the normality assumption does not significantly increase exception rates, although, through the Jarque–Bera test, it is strictly rejected as a null hypothesis.

The main points of this thorough investigation of ARCH processes can be summarized in the following points. First of all, the GARCH term in the conditional variance plays an important role since it provides models with a longer memory and a more flexible lag structure. Second, the choice of the sample size is important in generating adequate conditional variance forecasts. To our best knowledge, there is no persistent method in the literature for choosing the appropriate sample size. In most studies, researchers make an arbitrary choice of a rolling sample. Nevertheless, Engle et al. [23] applied three different sample sizes of 300, 1000 and 5000 observations and noted “that some restrictions on the

[11] In case of identical $p$-values being achieved by more than one model, only one of these models is given.

length of forecasting sample may be profitable". Our study reaches a similar result, as different sample sizes seem to be the most appropriate for different indices.

Generally speaking, the conditional mean specification is invariant to the VaR forecast accuracy, because the adequacy of the models does not depend on the autoregressive order. However, there are some cases where the mean specification offers an increasing precision in VaR predictions. An example worth noting is the AR(2) GARCH(1, 2) model with a 500-point sample size for the CAC 40 index, which produces the most accurate forecasts for the 99% daily VaR, though the rest of the models of the GARCH(1, 2) family do not produce acceptable VaR numbers.

Lastly, the assumption of normally distributed standardized residuals provides us with adequate forecasts only at the 95% level. For each index, there is at least one family of ARCH processes, producing convenient predictions. On the other hand, normal distribution fails to produce useful estimations at the 99% level. Under the assumption that the residuals are conditionally normal distributed, their unconditional distribution has thicker tails than the normal one. However, the degree of leptokurtosis induced by the ARCH process does not often capture all of the leptokurtosis present in the data. In the next section, we introduce non-normality for the conditional distribution of the residuals in order to produce unconditional distributions with thicker tails.

### *4.2. Leptokurtic distributions*

As we have already pointed out, the normal distribution assumption does not generate accurately one-step-ahead VaR numbers. In order to model more adequately the thickness of tails, we will use two different distributional assumptions for the standardized residuals: Student's-$t$ and the Generalized Error Distribution (GED). In the previous section, we found out that the precision of forecasts did not depend on the various structures of the conditional mean. We will, therefore, apply only the AR(1) specification for the conditional mean, consistent with the non-synchronous trading effect. Models that failed to converge more than four times are totally excluded from the study.

Turning now to the distribution used for the innovations, it seems that the Student's-$t$ is a better choice overall, based on Tables 5 and 6 which present the models with the highest $p$-values for the two leptokurtic distributions. Given that the fat tails are observed in all indices, the normal distribution fails prominently if no leptokurtosis is added implicitly to a model. The GED behaves similarly to the normal for the 95% confidence level but yields better results at the 99% confidence level: given that such a distribution exhibits thicker tails than the normal, it seems that the *hump* problem is resolved. Its $p$-values for both conditional and unconditional coverage are, however, still in the 20–70% range for the S&P 500, DAX 30, CAC 40 and FTSE 100 indices. The use of the Student's-$t$ distribution improves probability values, as they range from 80% to 95%, for all volatility models and all sample sizes, for these four indices. In the case of the NIKKEI 225, it seems that the choice of Student's-$t$ distribution overcorrects for thick tails and, consequently, either the normal (at the 95% level), or the GED (at the 99% level), provide better estimates.

For the 95% VaR level, under Student's-$t$ assumption, there is evidence that GARCH or EGARCH models produce better forecasts than the corresponding TARCH model, while for the GED this is the case only for the GARCH family. By increasing the confidence

Table 5
Best performed models and the relative probability values of coverage metrics for Student's-*t* distributed innovations

| | **Unconditional coverage** | | | **Conditional coverage** | | |
|---|---|---|---|---|---|---|
| Index | Size | Model | Pr. (%) | Size | Model | Pr. (%) |
| | *95% daily VaR forecasts* | | | | | |
| S&P 500 | 1500 | AR(1) GARCH(0, 2) | 97.03 | 1000 | AR(1) EGARCH(0, 1) | 93.39 |
| NIKKEI 225 | 500 | AR(1) EGARCH(1, 2) | 6.56 | 500 | AR(1) EGARCH(1, 2) | 14.69 |
| DAX 30 | 500 | AR(1) EGARCH(0, 2) | 97.86 | 1500 | AR(1) EGARCH(1, 2) | 93.21 |
| CAC 40 | 1500 | AR(1) EGARCH(2, 1) | 97.43 | 1000 | AR(1) EGARCH(2, 1) | 93.25 |
| FTSE 100 | 1500 | AR(1) GARCH(1, 2) | 92.78 | 1500 | AR(1) TARCH(1, 1) | 93.61 |
| | *99% daily VaR forecasts* | | | | | |
| S&P 500 | 1000 | AR(1) EGARCH(2, 1) | 91.30 | 1500 | AR(1) EGARCH(0, 2) | 82.35 |
| NIKKEI 225 | 500 | AR(1) EGARCH(2, 1) | 36.26 | 500 | AR(1) EGARCH(2, 1) | 58.61 |
| DAX 30 | 1500 | AR(1) EGARCH(1, 1) | 93.43 | 1000 | AR(1) EGARCH(1, 1) | 82.56 |
| CAC 40 | 1500 | AR(1) TARCH(1, 1) | 95.11 | 1000 | AR(1) GARCH(1, 2) | 82.61 |
| FTSE 100 | 1000 | AR(1) TARCH(1, 1) | 92.04 | 1000 | AR(1) TARCH(1, 1) | 80.91 |

level of the VaR, the results become more mixed because both symmetric and asymmetric models have been selected as statistically adequate. The choice of the sample size is turning out to be one of the most important factors, as the selected model for one sample size is not statistically adequate for the other sizes, while there are extreme cases that produce totally contradictory results. For example, the best performing model for the S&P 500 index at the 95% VaR level is the AR(1) GARCH(0, 2), under the assumption of a Student's-*t* distribution and a sample size of 1500 observations. Yet, had we used the 500 observations sample size, this model would have been characterized as one of the worst. This conclusion is common to all three distributional assumptions and to both confidence levels, revealing the importance of sample size. Our findings are in line with the work of Brooks and Persand [11] who argued that there is no optimal sample size for *all* volatility forecasting techniques.

Combining leptokurtic distribution, to capture fat tails, with a low complexity volatility model, to capture volatility clustering, yields the best results for the 99% VaR level. This is due to the estimated tail-thickness parameter $(v)$, which is able to capture extreme events, a necessary condition at such confidence level. As for the 95% VaR level, the excellent performance of models using the Student's-*t* distribution was rather a surprise, since most researchers reported an overestimation of risk (see Billio and Pelizzon [3], Guermat and Harris [28]). These results are summarized in Table 7, which presents the models that have performed best according to our coverage metrics.

## 4.3. Model selection

The two backtesting measures can not compare different VaR models directly, as a greater $p$-value of a model does not indicate the superiority of that model among its competitors. Therefore, in order to evaluate the reported differences statistically, for each model that has a $p$-value for both tests greater than 10%, we compute the QL function and carry out the equality test that was described in Section 3.3. We preferred a high cut-off

Table 6
Best performed models and the relative probability values of coverage metrics for GED-distributed innovations

| Index | **Unconditional coverage** Size | Model | Pr. (%) | **Conditional coverage** Size | Model | Pr. (%) |
|---|---|---|---|---|---|---|
| | *95% daily VaR forecasts* | | | | | |
| S&P 500 | 2000 | AR(1) GARCH(1, 1) | 72.25 | 2000 | AR(1) GARCH(1, 1) | 79.48 |
| NIKKEI 225 | 500 | AR(1) GARCH(1, 2) | 61.47 | 500 | AR(1) GARCH(1, 2) | 52.22 |
| DAX 30 | 500 | AR(1) GARCH(2, 2) | 72.92 | 500 | AR(1) GARCH(2, 2) | 57.66 |
| CAC 40 | 1500 | AR(1) GARCH(2, 1) | 21.79 | 1500 | AR(1) GARCH(2, 1) | 38.84 |
| FTSE 100 | 500 | AR(1) GARCH(2, 2) | 20.47 | 500 | AR(1) GARCH(2, 2) | 38.15 |
| | *99% daily VaR forecasts* | | | | | |
| S&P 500 | 2000 | AR(1) GARCH(1, 1) | 57.23 | 2000 | AR(1) GARCH(1, 1) | 66.25 |
| NIKKEI 225 | 500 | AR(1) GARCH(1, 2) | 93.73 | 1500 | AR(1) GARCH(0, 1) | 82.87 |
| DAX 30 | 2000 | AR(1) TARCH(1, 1) | 54.33 | 2000 | AR(1) TARCH(1, 1) | 64.44 |
| CAC 40 | 2000 | AR(1) TARCH(1, 1) | 19.78 | 2000 | AR(1) TARCH(1, 1) | 26.41 |
| FTSE 100 | 1500 | AR(1) TARCH(1, 1) | 57.88 | 1500 | AR(1) TARCH(1, 1) | 66.64 |

point for the $p$-value, in order to ensure that "successful" models will not statistically over or under estimate the "true" VaR, since a high (low) VaR estimation implies that a firm must allocate more (less) capital than it is actually necessary.

Table 8 summarizes the results of the loss function approach. For each index and confidence level we present the six[12] models with the lowest QL value, while we compare the first one with the rest of them based on the procedure that is explained in Section 3.3. For example, for the S&P 500 index and the 95% VaR level, the best performed model is the AR(1) EGARCH(0, 1), which is statistically different from the other five according to the corresponding $t$-statistics. In most of the cases, the model with the lowest loss value is significantly better than the others, as the $t$-statistic indicates, while the preferred distribution is the Student's-$t$. Moreover, the simplest asymmetric volatility specifications seems to be preferred over the most complex ones, while there is no indication which sample length is the optimal. To summarize, the combination of a leptokurtic distribution and a simple asymmetric volatility model yields the best combination.

## 5. Conclusions

Following the extensive and detailed investigation of a plethora of volatility modeling techniques, briefly presented in the preceding sections, a number of comments are of order, aiming to summarize our results and give, both to the researcher and the practitioner, some fundamental guidelines with which to proceed in VaR estimation.

We have strong indications that the mean process specification plays no important role. Trying to extract autoregressive phenomena from the returns such that only the underlying volatility is left in the residuals, we experimented with a number of AR processes. Our results show that such a methodology does not add anything significant

[12] The results for all the compared models are available from the authors upon request.

Table 7
Best performed models according to probability values of coverage metrics for normal (N), Student's-$t$ (T) or GED (G) distributed innovations

| Index | **Unconditional coverage** Size | Distribution | Model | Pr. (%) | **Conditional coverage** Size | Distribution | Model | Pr. (%) |
|---|---|---|---|---|---|---|---|---|
| | *95% daily VaR forecasts* | | | | | | | |
| S&P 500 | 1500 | T | AR(1) GARCH(0, 2) | 97.03 | 1000 | T | AR(1) EGARCH(0, 1) | 93.39 |
| NIKKEI 225 | 1000 | N | AR(0) EGARCH(1, 1) | 96.95 | 500 | N | AR(3) EGARCH(1, 2) | 93.14 |
| DAX 30 | 500 | T | AR(1) EGARCH(0, 2) | 97.86 | 1500 | T | AR(1) EGARCH(1, 2) | 93.21 |
| CAC 40 | 1500 | T | AR(1) EGARCH(2, 1) | 97.43 | 1000 | T | AR(1) EGARCH(2, 1) | 93.25 |
| FTSE 100 | 1500 | T | AR(1) GARCH(1, 2) | 92.78 | 1500 | T | AR(1) TARCH(1, 1) | 93.61 |
| | *99% daily VaR forecasts* | | | | | | | |
| S&P 500 | 1000 | T | AR(1) EGARCH(2, 1) | 91.30 | 1500 | T | AR(1) EGARCH(0, 2) | 82.35 |
| NIKKEI 225 | 500 | G | AR(1) GARCH(1, 2) | 93.73 | 1500 | G | AR(1) GARCH(0, 1) | 82.87 |
| DAX 30 | 1500 | T | AR(1) EGARCH(1, 1) | 93.43 | 1000 | T | AR(1) EGARCH(1, 1) | 82.56 |
| CAC 40 | 1500 | T | AR(1) TARCH(1, 1) | 95.11 | 1000 | T | AR(1) GARCH(1, 2) | 82.61 |
| FTSE 100 | 1000 | T | AR(1) TARCH(1, 1) | 92.04 | 1000 | T | AR(1) TARCH(1, 1) | 80.91 |

Table 8
Statistical comparison between the six best performed models according to the quantile loss function

| *95% daily VaR forecasts* | | | | | *99% daily VaR forecasts* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Distr. | Loss value (%) | *t*-stat. | Model | Size | Distr. | Loss value (%) | *t*-stat. |
| S&P 500 | | | | | | | | | |
| AR(1) EGARCH(0, 1) | 2000 | T | 5.61 | – | AR(1) GARCH(0, 1) | 500 | T | 16.00 | – |
| AR(1) EGARCH(0, 1) | 1500 | T | 6.51 | −1.95[c] | AR(1) EGARCH(0, 1) | 500 | T | 16.04 | −0.08 |
| AR(1) GARCH(0, 1) | 1500 | T | 7.10 | −4.13[a] | AR(1) EGARCH(0, 1) | 1500 | T | 17.03 | −0.51 |
| AR(1) GARCH(0, 1) | 2000 | T | 7.28 | −3.76[a] | AR(1) EGARCH(0, 1) | 1000 | T | 17.24 | −1.05 |
| AR(1) EGARCH(0, 1) | 1000 | T | 7.54 | −4.83 | AR(1) GARCH(0, 1) | 1000 | T | 18.21 | −1.90[c] |
| AR(1) EGARCH(0, 1) | 500 | T | 7.54 | −4.63[a] | AR(1) EGARCH(0, 1) | 2000 | T | 18.24 | −1.39 |
| NIKKEI 225 | | | | | | | | | |
| AR(1) GARCH(0, 1) | 500 | GED | 4.85 | – | AR(1) EGARCH(0, 1) | 1500 | GED | 4.28 | – |
| AR(1) GARCH(0, 2) | 1000 | GED | 5.41 | −2.34[b] | AR(1) EGARCH(0, 1) | 1000 | GED | 5.08 | −3.66[a] |
| AR(1) TARCH(0, 2) | 1000 | GED | 5.64 | −2.82[a] | AR(1) EGARCH(0, 1) | 2000 | GED | 5.61 | −5.40[a] |
| AR(0) GARCH(0, 2) | 1000 | N | 5.72 | −1.53 | AR(1) EGARCH(0, 1) | 500 | GED | 6.71 | −4.56[a] |
| AR(1) GARCH(0, 2) | 500 | GED | 5.95 | −4.11[a] | AR(1) EGARCH(0, 2) | 1500 | GED | 7.31 | −7.75[a] |
| AR(1) EGARCH(2, 1) | 500 | GED | 6.07 | −3.61[a] | AR(1) EGARCH(0, 2) | 1000 | GED | 7.37 | −8.96[a] |
| DAX 30 | | | | | | | | | |
| AR(1) EGARCH(0, 1) | 2000 | T | 7.39 | – | AR(1) EGARCH(0, 1) | 2000 | T | 22.08 | – |
| AR(1) EGARCH(0, 1) | 500 | T | 11.86 | −8.21[a] | AR(1) EGARCH(0, 1) | 1500 | T | 27.47 | −4.37[a] |
| AR(1) GARCH(0, 1) | 1500 | T | 13.20 | −7.17[a] | AR(1) EGARCH(0, 2) | 2000 | T | 35.10 | −7.13[a] |
| AR(1) GARCH(0, 1) | 2000 | T | 14.64 | −4.91[a] | AR(1) GARCH(0, 1) | 1500 | T | 38.26 | −8.47[a] |
| AR(1) GARCH(0, 1) | 1000 | T | 15.34 | −7.74[a] | AR(1) EGARCH(0, 1) | 500 | T | 40.70 | −7.47[a] |
| AR(1) TARCH(0, 1) | 2000 | T | 15.65 | −5.71[a] | AR(1) GARCH(0, 1) | 2000 | T | 41.20 | −5.18[a] |

Table 8 *(continued)*

| *95% daily VaR forecasts* Model | Size | Distr. | Loss value (%) | $t$-stat. | *99% daily VaR forecasts* Model | Size | Distr. | Loss value (%) | $t$-stat. |
|---|---|---|---|---|---|---|---|---|---|
| CAC 40 | | | | | | | | | |
| AR(1) EGARCH(0, 1) | 2000 | T | 3.96 | – | AR(1) GARCH(0, 2) | 2000 | T | 21.66 | – |
| AR(1) EGARCH(0, 1) | 1500 | T | 4.73 | −7.54[a] | AR(1) GARCH(0, 2) | 1500 | T | 26.34 | −12.59[a] |
| AR(1) EGARCH(0, 2) | 2000 | T | 4.81 | −4.09[a] | AR(1) EGARCH(1, 2) | 2000 | T | 27.85 | −5.48[a] |
| AR(1) GARCH(0, 1) | 2000 | T | 5.10 | −3.89[a] | AR(1) EGARCH(1, 1) | 2000 | T | 27.95 | −5.86[a] |
| AR(1) GARCH(0, 1) | 1500 | T | 5.44 | −5.53[a] | AR(1) EGARCH(2, 1) | 2000 | T | 27.96 | −5.68[a] |
| AR(1) EGARCH(0, 2) | 1500 | T | 5.56 | −6.38[a] | AR(1) GARCH(1, 1) | 1500 | T | 30.37 | −5.48[a] |
| FTSE 100 | | | | | | | | | |
| AR(1) GARCH(0, 1) | 1500 | T | 6.30 | – | AR(1) EGARCH(2, 1) | 1500 | T | 23.19 | – |
| AR(1) GARCH(0, 2) | 1500 | T | 7.66 | −5.14[a] | AR(1) EGARCH(1, 2) | 1500 | T | 23.32 | −0.69 |
| AR(1) EGARCH(1, 2) | 1500 | T | 7.70 | −3.66[a] | AR(1) EGARCH(1, 1) | 2000 | T | 23.68 | −0.92 |
| AR(1) EGARCH(1, 1) | 1500 | T | 7.72 | −4.12[a] | AR(1) EGARCH(1, 2) | 2000 | T | 23.75 | −1.06 |
| AR(1) EGARCH(1, 1) | 2000 | T | 7.79 | −3.76[a] | AR(1) EGARCH(2, 1) | 2000 | T | 23.80 | −1.20 |
| AR(1) EGARCH(1, 2) | 2000 | T | 7.85 | −3.65[a] | AR(1) GARCH(1, 1) | 1500 | T | 26.19 | −1.82[c] |

The $t$-statistics are calculated according to the procedure that is explained in Section 3.3.

[a] Indicate significance at the 1% level.

[b] Indicate significance at the 5% level.

[c] Indicate significance at the 10% level.

to the VaR framework other than complexity in the estimation procedure. Moreover, using only an ARCH term (without any lagged conditional variances) yields acceptable results only when residuals are modeled under either the Student's-$t$ distribution or the GED; it is never the case for a normal distribution. Generally speaking, in the VaR framework the leptokurtic distributions and especially the Student's-$t$, are more appropriate than the normal assumption, as they generate more accurate forecasts, while there is no volatility model, which is clearly superior than the others. However, under the evaluation framework that was developed based on the proposed quantile loss function, there is strong evidence that the combination of the Student's-$t$ distribution with the simplest EGARCH models produces the most adequate VaR forecasts for the majority of the markets. Furthermore, the size of the rolling sample used in estimation turns out to be rather important: in simpler models and low confidence levels a sample size smaller than 2000 improves probability values. In more complex models, where leptokurtic distributions are used or where the confidence level chosen is high, a small sample size may lead to lack of convergence in the estimation algorithms. Finally, there is no consistent relation between the sample sizes and the optimal models, as we observe significant differences in the VaR forecasts for the same model under the four sample sizes.

## References

[1] H. Akaike, Information theory and the extension of the maximum likelihood principle, in: Proceeding of the Second International Symposium on Information Theory, Budapest, 1973, pp. 267–281.

[2] A.K. Bera, M.L. Higgins, ARCH models: properties, estimation and testing, Journal of Economic Surveys 7 (1993) 305–366.

[3] M. Billio, L. Pelizzon, Value-at-risk: a multivariate switching regime approach, Journal of Empirical Finance 7 (2000) 531–554.

[4] F. Black, Studies of stock market volatility changes, in: Proceedings of the American Statistical Association, Business and Economic Statistics Section, 1976, pp. 177–181.

[5] T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, Journal of Econometrics 31 (1986) 307–327.

[6] T. Bollerslev, A conditional heteroscedastic time series model for speculative prices and rates of return, Review of Economics and Statistics 69 (1987) 542–547.

[7] T. Bollerslev, J.M. Wooldridge, Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, Econometric Reviews 11 (1992) 143–172.

[8] T. Bollerslev, R. Chou, K.F. Kroner, ARCH modeling in finance: a review of the theory and empirical evidence, Journal of Econometrics 52 (1992) 5–59.

[9] T. Bollerslev, R.F. Engle, D. Nelson, ARCH models, in: R.F. Engle, D. McFadden (Eds.), Handbook of Econometrics, vol. 4, Elsevier, Amsterdam, 1994, pp. 2959–3038.

[10] C. Brooks, S.P. Burke, Information criteria for GARCH model selection, The European Journal of Finance 9 (2003) 557–580.

[11] C. Brooks, G. Persand, Model choice and Value-at-Risk performance, Financial Analysts Journal 58 (5) (2002) 87–98.

[12] C. Brooks, G. Persand, The effect of asymmetries on stock index return Value-at-Risk estimates, The Journal of Risk Finance (2003) 29–42.

[13] C. Brooks, G. Persand, Volatility forecasting for risk management, Journal of Forecasting 22 (2003) 1–22.

[14] D.J. Cabedo, I. Moya, Estimating oil price 'Value at Risk' using the historical simulation approach, Energy Economics 25 (2003) 239–253.

[15] P. Christoffersen, Evaluating interval forecasts, International Economic Review 39 (1998) 841–862.

[16] M. Crouhy, D. Galai, R. Mark, Risk Management, McGraw-Hill, New York, 2001.

[17] S. Danielsson, Multivariate stochastic volatility models: estimation and a comparison with VGARCH models, Department of Economics, University of Iceland, 1996 (mimeo).
[18] J. Daníelsson, The emperor has no clothes: limits to risk modeling, Journal of Banking & Finance 26 (2002) 1273–1296.
[19] S. Degiannakis, E. Xekalaki, Autoregressive conditional heteroskedasticity models: a review, Quality Technology and Quantitative Management 1 (2) (2004) 271–324.
[20] F.X. Diebold, R. Mariano, Comparing predictive accuracy, Journal of Business and Economic Statistics 13 (3) (1995) 253–263.
[21] K. Dowd, Beyond Value at Risk: The New Science of Risk Management, John Wiley & Sons, New York, 1998.
[22] R.F. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation, Econometrica 50 (1982) 987–1008.
[23] R.F. Engle, C.H. Hong, A. Kane, J. Noh, Arbitrage valuation of variance forecasts with simulated options, Advances in Futures and Options Research 6 (1993) 393–415.
[24] R.F. Engle, A.J. Patton, What good is a volatility model?, Quantitative Finance 1 (2) (2001) 237–245.
[25] R. Frey, P. Michaud, The effect of GARCH-type volatilities on prices and payoff-distributions of derivative assets—a simulation study, ETH Zurich, 1997 (working paper).
[26] P. Giot, S. Laurent, Value-at-Risk for long and short trading positions, Journal of Applied Econometrics 18 (2003) 641–664.
[27] P. Giot, S. Laurent, Market risk in commodity markets: a VaR approach, Energy Economics 25 (2003) 435–457.
[28] C. Guermat, R.D.F. Harris, Forecasting value at risk allowing for time variation in the variance and kurtosis portfolio returns, International Journal of Forecasting 18 (2002) 409–419.
[29] J.D. Hamilton, Time Series Analysis, Princeton University Press, New Jersey, 1994.
[30] P.R. Hansen, A. Lunde, A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? Brown University, 2003 (working paper).
[31] D. Hendricks, Evaluation of value-at-risk models using historical data, Federal Reserve Bank of New York, Economic Police Review 2 (1996) 39–70.
[32] R. Hoppe, VAR and the unreal world, Risk 11 (1998) 45–50.
[33] D.A. Hsieh, Modeling heteroscedasticity in daily foreign-exchange rates, Journal of Business and Economic Statistics 7 (1989) 307–317.
[34] P. Jackson, D.J. Maude, W. Perraudin, Testing Value-at-Risk approaches to capital adequacy, Bank of England Quarterly Bulletin 38 (1998) 256–266.
[35] P. Jorion, On jump processes in the foreign exchange and stock markets, Review of Financial Studies 1 (1988) 427–445.
[36] P. Jorion, Value at Risk. The New Benchmark for Controlling Derivatives Risk, McGraw-Hill, New York, 1997.
[37] F. Klaassen, Improving GARCH volatility forecasts with regime-switching GARCH, in: J.D. Hamilton, B. Raj (Eds.), Advances in Markov-Switching Models, Physica Verlag, New York, 2002, pp. 223–254.
[38] P.H. Kupiec, Techniques for verifying the accuracy of risk measurement models, The Journal of Derivatives 3 (1995) 73–84.
[39] G. Lambadiaris, L. Papadopoulou, G. Skiadopoulos, Y. Zoulis, VAR: history or simulation?, Risk, 2003.
[40] P. Lambert, S. Laurent, Modeling skewness dynamics in series of financial data, Institut de Statistique, Louvain-la-Neuve, 2000 (discussion paper).
[41] P. Lambert, S. Laurent, Modeling financial time series using GARCH-type models and a skewed Student density, Universite de Liege, 2001 (mimeo).
[42] T.K.Y. Lee, Y.K. Tse, Term structure of interest rates in the Singapore Asian dollar market, Journal of Applied Econometrics 6 (1991) 143–152.
[43] A. Lo, A.C. MacKinlay, Stock market prices do not follow random walks: evidence from a simple specification test, Review of Financial Studies 1 (1988) 41–66.
[44] J. Longerstay, An improved methodology for measuring VaR, RiskMetrics, Monitor Second Quarter, J.P. Morgan, 1996.
[45] J.A. Lopez, Methods for evaluating Value-at-Risk estimates, Federal Reserve Bank of New York, Economic Policy Review, 1998.

[46] B. Mandelbrot, The variation of certain speculative prices, Journal of Business 36 (1963) 394–419.
[47] D.W. Marquardt, An algorithm for least squares estimation of non-linear parameters, Journal of the Society for Industrial and Applied Mathematics 11 (1963) 431–441.
[48] D. Nelson, Conditional heteroskedasticity in asset returns: a new approach, Econometrica 59 (1991) 347–370.
[49] W. Newey, K. West, A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica 55 (1987) 703–708.
[50] A.R. Pagan, G.W. Schwert, Alternative models for conditional stock volatility, Journal of Econometrics 45 (1990) 267–290.
[51] RiskMetrics, Technical Document, J.P. Morgan, New York, USA, 1995.
[52] M. Sarma, S. Thomas, A. Shah, Selection of VaR models, Journal of Forecasting 22 (4) (2003) 337–358.
[53] M. Scholes, J. Williams, Estimating betas from non-synchronous data, Journal of Financial Economics 5 (1977) 309–328.
[54] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[55] S. Venkataraman, Value at risk for a mixture of normal distributions: the use of quasi-Bayesian estimation techniques, in: Economic Perspectives, Federal Reserve Bank of Chicago, 1997, pp. 2–13.
[56] P. Vlaar, Value at Risk models for Dutch bond portfolios, Journal of Banking and Finance 24 (2000) 131–154.
[57] H. White, A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity, Econometrica 48 (1980) 817–838.
[58] P. Zangari, An improved methodology for measuring VAR, RiskMetrics Monitor, Reuters/J.P. Morgan, 1996.