Instituto Complutense de Análisis Económico

Volatility specifications versus probability distributions in VaR forecasting

Laura Garcia-Jorcano
Department of Economic Analysis and Finance (Area of Financial Economics)
Facultad de Ciencias Jurídicas y Sociales,
Universidad de Castilla-La Mancha, Toledo, Spain

Alfonso Novales
Instituto Complutense de Análisis Económico (ICAE), and
Department of Economic Analysis, Facultad de Ciencias Económicas y Empresariales,
Universidad Complutense, Madrid, Spain

Abstract

We provide evidence suggesting that the assumption on the probability distribution for return innovations is more influential for Value at Risk (VaR) performance than the conditional volatility specification. We also show that some recently proposed asymmetric probability distributions and the APARCH and FGARCH volatility specifications beat more standard alternatives for VaR fore- casting, and they should be preferred when estimating tail risk. The flexibility of the free power parameter in conditional volatility in the APARCH and FGARCH models explains their better performance. Indeed, our estimates suggest that for a number of financial assets, the dynamics of volatility should be specified in terms of the conditional standard deviation. We draw our results on VaR forecasting performance from *i*) a variety of back testing approaches, *ii*) the Model Confidence Set approach, as well as *iii*) establishing a ranking among alternative VaR models using a *precedence* criterion that we introduce in this paper.

Keywords Value-at-risk, back testing, evaluating forecasts, Precedence, APARCH model, asymmetric distributions

Working Paper nº 1926
September, 2019

UNIVERSIDAD
COMPLUTENSE
MADRID

ISSN: **2341-2356**
WEB DE LA COLECCIÓN: http://www.ucm.es/fundamentos-analisis-economico2/documentos-de-trabajo-del-icae

Volatility specifications versus probability distributions in VaR forecasting

Laura Garcia-Jorcano [*a], Alfonso Novales [†b]

[a]Department of Economic Analysis and Finance (Area of Financial Economics), Facultad de Ciencias Jurídicas y Sociales, Universidad de Castilla-La Mancha, 45071 Toledo, Spain
[b]Instituto Complutense de Análisis Económico (ICAE), and Department of Quantitative Economics, Facultad de Ciencias Económicas y Empresariales, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain
E-mail addresses: Laura.Garcia@uclm.es (L. Garcia-Jorcano), anovales@ucm.es (A. Novales)

April 2019

Abstract

We provide evidence suggesting that the assumption on the probability distribution for return innovations is more influential for Value at Risk (VaR) performance than the conditional volatility specification. We also show that some recently proposed asymmetric probability distributions and the APARCH and FGARCH volatility specifications beat more standard alternatives for VaR forecasting, and they should be preferred when estimating tail risk. The flexibility of the free power parameter in conditional volatility in the APARCH and FGARCH models explains their better performance. Indeed, our estimates suggest that for a number of financial assets, the dynamics of volatility should be specified in terms of the conditional standard deviation. We draw our results on VaR forecasting performance from *i*) a variety of backtesting approaches, *ii*) the Model Confidence Set approach, as well as *iii*) establishing a ranking among alternative VaR models using a *precedence* criterion that we introduce in this paper.

Keywords: Value-at-risk, backtesting, evaluating forecasts, Precedence, APARCH model, asymmetric distributions

[*]https://orcid.org/0000-0002-3656-8787
[†]Corresponding author. Tel.: +34 616781602. https://orcid.org/0000-0002-2137-1789

1 Introduction

A traditional discussion in risk measurement analysis has been whether volatility models that incorporate a leverage effect, with negative innovations having a larger impact on volatility than positive innovations of the same size, lead to better Value-at-Risk (VaR) forecasts. A second modeling issue refers to whether asymmetric probability distributions for return innovations lead to an improved VaR model.[1] The goal of this paper is to examine the relative importance of the two issues for the efficiency of VaR forecasts. The question is crucial for risk managers, since there are so many potential choices for volatility model and probability distributions that it would be very convenient to establish some priorities in modeling returns for risk estimation. To that end, we have performed an extensive analysis of VaR forecasts in assets of different nature, using symmetric and asymmetric probability distributions for the innovations on volatility models with and without leverage. Even though this issue has been examined in previous work, we consider some volatility specifications and probability distributions that are still relatively new in this literature, which allows us to make some progress in modeling financial returns when forecasting Value at Risk.

We consider three general volatility specifications with leverage, GJR-GARCH, APARCH and FGARCH, with the standard symmetric GARCH model as benchmark. The FGARCH model includes as special cases many other volatility specifications, like the symmetric GARCH, GJR-GARCH and APARCH. It is, in fact, a nested family of GARCH-type models, thereby allowing for testing how simpler models fit the data. The APARCH and FGARCH models take the power on the conditional standard deviation of the innovations as a free parameter, which provides more flexibility to the dynamics of volatility, allowing for shifts and rotations in the news impact curve. The two types of asymmetry in volatility produced by shifts and rotations are distinct, and they should not be treated as substitutes for each other (Hentschel, 1995). As probability distributions for the innovations we compare the performance of the skewed Student-t distribution and skewed Generalized Error distribution as introduced in Fernandez and Steel (1998), the unbounded Johnson S_U distribution (Johnson, 1949), skewed Generalized-t distribution (Theodossiou, 1998) and Generalized Hyperbolic skew Student-t distribution (Aas and Haff, 2006), with the Normal and symmetric Student-t distributions as benchmark. An interesting feature of our work is the consideration of a variety of assets of different nature: stock market indexes, individual stocks, interest rates, commodity prices and exchange rates.

We calculate VaR forecasts following the parametric approach. An AR(1) was estimated for daily returns in all cases. The performance of VaR forecasts is examined through standard tests: the unconditional coverage test of Kupiec (1995), the independence and conditional coverage tests of Christoffersen (1998), and the Dynamic Quantile test of Engle and Manganelli (2004). VaR forecasts are also evaluated through the use of the Asymmetric Linear Tick loss function (AlTick) proposed by Giacomini and Komunjer (2005). The combination of 19 assets, 7 probability distributions, 4 volatility specifications and 4 backtests of VaR leads to an extensive set of results that need to be summarized in a search for some consistent conclusions. One of the contributions of this paper is to follow a diverse strategy to summarize test results in search of robust patterns that might suggest some preferred VaR model specifications. We proceed along several lines: *i*) comparing the number of realized and expected violations of VaR for the alternative models across

[1] Along the paper we refer to a VaR forecasting model as a combination of a probability distribution and a volatility specification for return innovations.

the set of assets, *ii*) comparing the *p*-values achieved by the different models in the VaR validation tests, *iii*) applying to the alternative models a *precedence* criterion that we introduce in this paper to rank models according to their VaR forecasting performance, *iv*) following the Model Confidence Set approach to select the most preferred models.

Our results suggest that the important assumption for VaR performance is that of the probability distribution of the innovations, with the choice of volatility model playing a secondary role. Indeed, validation tests for VaR forecasts yield very similar results for a given probability distribution as we change the volatility model. On the contrary, test results drastically change for a given volatility model when we change the assumption on the probability distribution of the innovations. In fact, the main difference arises when we move from symmetric to asymmetric probability distributions for the innovations, a result consistent with work by Lopez and Walter (2000), Angelidis and Degiannakis (2006), Gerlach et al. (2011), Dendramis et al. (2014), and Braione and Scholtes (2016). The unbounded Johnson distribution, the skew Generalized-t distribution and the skewed Generalized Error distributions precede other asymmetric distributions, like the skewed Student-t and the Generalized Hyperbolic skewed Student-t in VaR forecasting. Symmetric distributions come out as being clearly inappropriate. On the volatility side, FGARCH and APARCH volatility specifications precede other alternatives. A relevant implication is that the conditional standard deviation, rather than the conditional variance, should often be used to model the volatility dynamics of financial returns. However, these results do not seem to apply to 10-day VaR forecasting, a discrepancy that should be examined in further research.

The remainder of the paper is organized as follows: in Section 2 we present a review of the literature on the questions we analyze. In Section 3 we describe the volatility models and the probability distributions used in our analysis. In Section 4 we present preliminary statistics for our data set. In Section 5 we report the estimates of the VaR models considered. In Section 6 we provide a description of the statistical tests and the loss function used and we asses VaR performance for the different models. Finally, Section 7 concludes the paper.

2 A review of literature

Among parametric methods for VaR estimation, some authors have analyzed the improvement on VaR estimation provided by volatility models with leverage. Giot and Laurent (2003a) estimated daily VaR for stock indexes using different volatility models. They stated that more complex models like APARCH performed better than RiskMetrics or GARCH specifications (for a comparison of volatility models in VaR estimation see also El Babsiri and Zakoian, 2001). Angelidis et al. (2007) show that volatility models with leverage fare better than symmetric specifications, as they capture more efficiently the characteristics of the underlying series and provide better VaR forecasts since they perform better in the low probability regions that VaR tries to measure (see also Ane, 2006). Working with a large number of individual stocks and exchange rates, McMillan and Kambouroudis (2009) conclude that the APARCH model should be preferred for more extreme VaR forecasts, while the RiskMetrics model seems to be adequate at more moderate significance levels. In their work, RiskMetrics seems adequate in providing volatility forecasts for most Asian markets; however, the APARCH model is superior in obtaining forecasts for the G7 markets, as well as for other European markets and for the larger Asia markets.[2]

[2] An overview of the recent literature is displayed in Table A1 of the Online Appendix.

Given the widespread evidence on the skewness of the distribution of asset returns, analyzing whether the assumption of an asymmetric distribution of return innovations leads to more efficient VaR forecasts is a second methodological issue of interest. Based on the influence of leverage effects on the accuracy of VaR forecasts, Brooks and Persand (2003) concluded that models that do not allow for asymmetries either in the unconditional distribution of returns or in the volatility specification underestimate the true VaR. Giot and Laurent (2003a) used daily data for stock market indexes and individual stocks, showing that models that rely on a symmetric density for return innovations underperform with respect to skewed density models that require modeling both the left and right tails of the distribution of returns. Lee and Su (2015) estimate VaR for eight stock market indexes from Europe and Asia by a parametric GARCH approach as well as by the semi-parametric approach of Hull and White (1998). The only asymmetric distribution they consider, the skewed Generalized-t, is shown to have a better VaR forecasting performance than the Student-t, with the Normal distribution being the last in the ranking, according to the unconditional coverage test of Kupiec and two different loss functions. Corlu et al. (2016) investigate the ability of five alternative probability distributions to represent the behavior of daily equity index returns over the period 1979-2014: the skewed Student-t distribution, the generalized lambda distribution, the Johnson system of distributions, the normal inverse Gaussian distribution, and the g-and-h distribution. The explanatory power of the alternative distributions is tested using in-sample Value-at-Risk (VaR) failure rates. Their focus is on the unconditional distribution of equity returns, not on conditional distributions. They find that the generalized lambda distribution is a prominent alternative for modeling the behavior of daily equity index returns.

More recently, some papers have jointly examined the performance of both, the variance specification and the probability distribution of return innovations in VaR estimation. Gerlach et al. (2011) examine the performance of several volatility specifications: RiskMetrics, asymmetric GARCH, IGARCH, GJR-GARCH and EGARCH, under four alternative probability distributions: Gaussian, Student-t, Generalized Error Distribution and skewed Student-t in VaR forecasting at 1% and 5% significance in different time periods (pre-crisis, crisis-GFC and post-crisis) incorporating parameter uncertainty through a Bayesian approach. Results are varied and hard to summarize, but their evidence suggests a clear preference for asymmetric probability distributions for the innovations of the return process. Giot and Laurent (2003b) analyze daily returns on commodities fitting ARCH and APARCH models under a skewed Student-t probability distribution for the innovations, and using Riskmetrics as a benchmark. While the skewed Student-t APARCH model performs best in all cases, it is unclear whether the forecasting gain is enough to dominate over the computationally simpler skewed Student-t ARCH model. Bubak (2008), Tu et al. (2008), Kang and Yoon (2009) and Diamandis et al. (2011), analyze Eastern and Central European stock markets, Asian stock markets, Asian emerging markets and developed and emerging markets, respectively. Comparing a wide range of univariate conditional variance models, they show that models that incorporate an asymmetric distribution for return innovations tend to perform better than models with a symmetric distribution, in terms of both in-sample and out-of-sample (one-day-ahead) VaR forecasts. Dendramis et al. (2014) show that the VaR performance of alternative parametric models like EGARCH or the Markov regime-switching model is enhanced when combined with asymmetric probability distributions for return innovations. Tang and Shieh (2006) and Mabrouk and Saadi (2012) include Fractionally Integrated time varying GARCH models designed to capture not only volatility clustering, but also long memory in asset return volatility.

Both papers consider three probability distributions, Normal, Student-t and skew Student-t. Tang and Shieh (2006) consider FIGARCH and HYGARCH (Hyperbolic GARCH) models, showing that for the three stock index futures considered, HYGARCH models with skewed Student-t distribution perform better based on the Kupiec LR tests. Mabrouk and Saadi (2012) conclude that the skewed Student-t FIAPARCH model outperforms the alternative GARCH and HYGARCH models because it can simultaneously account for fat tails, asymmetry, volatility clustering and long memory. However, given that the VaR forecasts required by the Basel accords are short run, the inclusion of long-memory is expected not to make any fundamental difference [see for example So and Yu (2006)]. Recently, Leccadito et al. (2014) have compared the performance of a variety of volatility specifications and asymmetric distributions using multilevel VaR tests that apply independence and conditional coverage tests at different confidence levels. While the need to consider asymmetric probability distributions for return innovations seems to be well established at this point, the preference for a given volatility specification is less clear.

As in the latter group of papers, we also examined the performance of both, the variance specification and the probability distribution of return innovations in VaR estimation. We consider a complex and flexible volatility model proposed by Hentschel (1995), FGARCH, which is an omnibus model that subsumes some of the most popular GARCH models. To the best of our knowledge, there are no papers examining the performance of this model for VaR forecasting. Besides, we consider distributions that are not often considered in the literature on VaR performance, such as the skewed Generalized Error Distribution [Fernandez and Steel (1998)], Johnson S_U distribution [Johnson (1949)], skewed Generalized-t [Theodossiou (1998)] and Generalized Hyperbolic skew Student-t distribution (GHST) [Aas and Haff (2016)].

3 Volatility Models and probability distributions

Let x_t, for $t = 1,...,T$, be a time series of asset returns. It is convenient to break down the complete characterization of x_t into three components: (i) the conditional mean, μ_t (ii) the conditional variance, which contains a scale parameter that measures the dispersion of the distribution, σ_t^2 and (iii) the shape parameters, which determine the form of a conditional distribution (e.g., skewness, kurtosis) within a general family of distributions. Thus, we may write

$$x_t = \mu_t(\theta) + \varepsilon_t \qquad \mu_t(\theta) = \mathbb{E}[x_t|\mathscr{F}_{t-1}] = \mu(\theta, \mathscr{F}_{t-1}) \qquad \varepsilon_t = \sigma_t(\theta) z_t$$
$$\sigma_t^2(\theta) = \mathbb{E}[(x_t - \mu_t)^2|\mathscr{F}_{t-1}] = \sigma^2(\theta, \mathscr{F}_{t-1}) \qquad z_t \sim f(z_t|\theta)$$

The standardized innovation, $z_t = (x_t - \mu_t(\theta))/\sigma_t(\theta)$ has zero mean and a unit variance. It follows a conditional distribution f with shape parameters that capture the possible asymmetry and fat-tailedness of returns, except in the case of the Normal distribution. Vector θ contains all the parameters associated with the conditional mean and variance and the conditional distribution.

An AR(1) model for the conditional mean return is sufficient to produce serially uncorrelated innovations for all assets. We consider three general volatility models with leverage, GJR-GARCH, APARCH and FGARCH with a standard symmetric GARCH model as benchmark. As probability distributions for the innovations we compare the performance of skewed Student-t, skewed Generalized Error, unbounded Johnson S_U, skewed Generalized-t and Generalized Hyperbolic skew Student-t distributions, with the normal and symmetric Student-t distributions as benchmark.

In all models we jointly estimate by maximum likelihood the parameters in the equation for the mean return, the equation for its conditional variance and the probability distribution for the return innovations. The exception is the skewed Generalized-t distribution, for which we use a two-step estimation method because of the numerical difficulty of estimating all parameters jointly.[3]

3.1 Volatility models

The conditional variance of GARCH(1,1) model (Bollerslev, 1986) is used as a benchmark, i.e.

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where $\omega > 0$, $\alpha_1, \beta_1 \geq 0$, $\alpha_1 + \beta_1 < 1$.

The standard GARCH model captures the existence of volatility clustering but is unable to express the leverage effect, since it assumes that positive and negative error terms have the same effect on volatility. To incorporate asymmetric effects on volatility from positive and negative surprises, Glosten, Jagannathan and Runkle (1993) proposed a GJR-GARCH(1,1) model, adding the negative impact of leverage in the conditional variance equation. This model incorporates positive and negative shocks on the conditional variance asymmetrically via the use of the indicator function $I(\varepsilon_{t-i} \leq 0)$, so that the variance equation becomes,

$$\sigma_t^2 = \omega + \left[\alpha_1 \varepsilon_{t-1}^2 + \gamma_1 I(\varepsilon_{t-1} \leq 0)\varepsilon_{t-1}^2\right] + \sigma_{t-1}^2$$

The volatility effect of a unit negative shock is $\alpha_i + \gamma_i$ while the effect of a unit positive shock is α_1. A positive value of γ_1 indicates that a negative innovation generates greater volatility than a positive innovation of equal size, and on the contrary for a negative value of γ_1.

The APARCH model (Asymmetric Power ARCH model) was proposed by Ding, Granger and Engle (1993). This model can well express volatility clustering, fat tails, excess kurtosis, the leverage effect and the Taylor effect. The latter effect is named after Taylor (1986) who observed that the sample autocorrelation of absolute returns was usually larger than that of squared returns. The APARCH(1,1) is defined as,

$$\sigma_t^\delta = \omega + \alpha_1 (|\varepsilon_{t-1}| - \gamma_1 \varepsilon_{t-1})^\delta + \beta_1 (\sigma_{t-1})^\delta$$

where ω, α_1, γ_1, β_1 and δ are additional parameters to be estimated. The parameter γ_1 reflects the leverage effect $(-1 < \gamma_1 < 1)$. A positive (resp. negative) value of γ_1 means that past negative (resp. positive) shocks have a deeper impact on current conditional volatility than past positive (resp. negative) shocks. The parameter δ plays the role of a Box-Cox transformation of $\sigma_t (\delta > 0)$.

The APARCH equation is supposed to satisfy the following conditions, *i)* $\omega > 0$ (since the variance is positive), $\alpha_1 \geq 0$, $\beta_1 \geq 0$. When $\alpha_1 = 0$, $\beta_1 = 0$, then $\sigma_t^2 = \omega$, *ii)* $0 \leq \alpha_1 + \beta_1 \leq 1$. The APARCH model is a general model because it has great flexibility, having as special cases, among others, those mentioned above.

The FGARCH model (Family GARCH) of Hentschel (1995) is an omnibus model which subsumes some of the most popular GARCH models. It is similar to the APARCH model, but more

[3] In that case, we first estimated the AR(1)-GARCH conditional mean-volatility model assuming a Generalized Error distribution (GED) for the innovations, as suggested by Bali and Theodossiou (2007). The parameters of the skewed Generalized-t distribution (SGT) were estimated in a second stage using the standardized returns ($\frac{r_t - \phi_0 - \phi_1 r_{t-1}}{\sigma_t} = \frac{\varepsilon_t}{\sigma_t}$) obtained in the first step.

general, since it allows the decomposition of the residuals in the conditional variance equation to be driven by different powers for z_t and σ_t. It also allows for both shifts and rotations in the news impact curve, where the shift is the main source of asymmetry for small shocks while rotation drives the asymmetry for large shocks. The FGARCH(1,1) is defined as,

$$\sigma_t^{\lambda} = \omega + \alpha_1 \sigma_{t-1}^{\lambda} f^{\delta}(z_{t-1}) + \beta_1 (\sigma_{t-1})^{\lambda}$$

where $f^{\delta}(z_{t-1}) = (|z_{t-1} - \eta_{21}| - \eta_{11}(z_{t-1} - \eta_{21}))^{\delta}$.

Positivity of $f^{\delta}(z_{t-1})$ is guaranteed when $|\eta_{11}| \leq 1$, which ensures that neither arm of the rotated absolute value function crosses the abscissa. The parameter η_{21}, however, is unrestricted in size and sign. The magnitude and direction of a shift in the news impact curve are controlled by the parameter η_{21} while the magnitude and direction of a rotation in the news impact curve are controlled by the parameter η_{11}. Other GARCH models only permit either a shift or a rotation, but not both. Allowing for shifts in the news impact curve, the FGARCH model is more flexible than previous models, being able to capture asymmetries in volatility even in the presence of small shocks.

3.2 Probability distributions

To account for the excess skewness and kurtosis typical of financial data, the parametric volatility models presented in the previous section can be combined with skewed and leptokurtic distributions for return innovations. The skewed Student-t by Fernandez and Steel and Lambert and Laurent (2001)[4] is

$$f(z|\xi,\nu) = \frac{2}{\xi + \frac{1}{\xi}} s\{g[\xi(sz+m)|\nu] I_{(-\infty,0)}(z+m/s) + g[(sz+m)/\xi|\nu] I_{[0,\infty)}(z+m/s)\} \quad (1)$$

where $g(\cdot|\nu)$ is the symmetric (unit variance) Student-t density and ξ is the skewness parameter;[5] m and s^2 are, respectively the mean and the variance of the non-standardized skewed Student-t and are defined as,

$$E(\varepsilon|\xi) = M_1(\xi - \xi^{-1}) \equiv m$$

$$V(\varepsilon|\xi) = (M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2 \equiv s^2$$

where $M_r = 2\int_0^{\infty} s^r g(s) ds$ is the absolute moments generating function. Note that when $\xi = 1$ and $\nu = +\infty$ we get the skewness and the kurtosis of the Gaussian density. When $\xi = 1$ and $\nu > 2$ we have the skewness and the kurtosis of the (standardized) Student-t distribution.

An alternative distribution for return innovations which can capture skewness and kurtosis can be based on the Generalized Error Distribution (GED) by Nelson (1991). According to Lambert

[4] Lambert and Laurent (2001) and Giot and Laurent (2003a) have shown that for various financial daily returns, it is realistic to assume that standardized innovations $\hat{z}_t$ follows a skewed Student-t distribution.

[5] The skewness parameter $\xi > 0$ is defined such that the ratio of probability masses above and below the mean is

$$\frac{Prob(z \geq 0|\xi)}{Prob(z < 0|\xi)} = \xi^2$$

and Laurent the innovation process z_t is said to follow a (standardized) skewed Generalized error distribution, $SGED(0,1,\xi,\kappa)$, if

$$f(z|\xi,\kappa) = \frac{2}{\xi+\frac{1}{\xi}} s\{g[\xi(sz+m)|\kappa]I_{(-\infty,0)}(z+m/s) + g[(sz+m)/\xi|\kappa]I_{[0,\infty)}(z+m/s)\}$$

where $g(\cdot|\kappa)$ is the symmetric (unit variance) Generalized Error distribution, ξ is the skewness parameter, κ representing the shape parameter and $\Gamma(\cdot)$ is the gamma function. Mean (m) and standard deviation (s) are calculated in the same way as in the case of skewed Student-t distribution. As κ increases the density gets flatter and flatter while in the limit, as $\kappa \to \infty$, the distribution tends toward the uniform distribution. Special cases are the Normal when $\kappa = 2$ and the Laplace distribution when $\kappa = 1$. For $\kappa > 2$ the distribution is platykurtic and for $\kappa < 2$ it is leptokurtic.

Another alternative is the Johnson S_U distribution. It was one of the distributions derived by Johnson (1949) based on translating the Normal distribution by certain functions. Letting $Y \sim N(0,1)$, the standard Normal distribution, the random variable Z has the Johnson system of frequency curves if it is a transformation of Y by $Y = \gamma + \delta g((Z-\xi)/\lambda)$. The form of the resulting distribution depends on the choice of function g. When $g(u) = sinh^{-1}(u)$, the distribution is unbounded, called the Johnson S_U distribution. The parameters of the distribution are ξ, $\lambda > 0$, γ, $\delta > 0$.

We use a parameterization[6] of the original Johnson S_U distribution, so that the ξ and λ parameters are the mean and standard deviation of the distribution. The parameter γ determines the skewness of the distribution with $\gamma > 0$ indicating positive skewness and $\gamma < 0$ negative skewness. The parameter δ determines the kurtosis of the distribution. δ should be positive and most likely above 1.

The pdf of the Johnson's S_U, denoted here as $JSU(\xi,\lambda,\gamma,\delta)$, is defined by

$$f_Z(z) = \frac{\delta}{c\lambda}\frac{1}{\sqrt{(r^2+1)}}\frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}y^2\right]$$

where

$$y = -\gamma + \delta sinh^{-1}(r) = -\gamma + \delta \log\left[r + (r^2+1)^{1/2}\right]$$

$$r = \frac{z - (\xi + c\lambda\omega^{1/2} sinh\Omega)}{c\lambda}$$

$$c = \left\{\frac{1}{2}(\omega-1)[\omega cosh 2\Omega + 1]\right\}^{-1/2}$$

where $\omega = exp(\delta^{-2})$ and $\Omega = -\gamma/\delta$. Note that $Y \sim N(0,1)$. Here $E(Z) = \xi$ and $Var(Z) = \lambda^2$.

A very flexible distribution is the skewed Generalized-t distribution proposed by Theodossiou (1998). They developed a skewed version of the Generalized-t distribution introduced by McDonald and Newey (1988).

The skewed Generalized-t distribution has the probability density function

$$f(x|\mu,\sigma,\lambda,p,q) = \frac{p}{2v\sigma q^{1/p} B(\frac{1}{p},q)\left(\frac{|x-\mu+m|^p}{q(v\sigma)^p(\lambda sign(x-\mu+m)+1)^p}+1\right)^{\frac{1}{p}+q}}$$

[6] This parameterization is used by R rugarch package, which we use for estimating the parameters of our models.

where

$$m = \frac{2\nu\sigma\lambda q^{\frac{1}{p}} B\left(\frac{2}{p}, q - \frac{1}{p}\right)}{B\left(\frac{1}{p}, q\right)}$$

$$\nu = q^{-\frac{1}{p}} \left[(3\lambda^2 + 1) \left(\frac{B\left(\frac{3}{p}, q - \frac{2}{p}\right)}{B\left(\frac{1}{p}, q\right)} \right) - 4\lambda^2 \left(\frac{B\left(\frac{2}{p}, q - \frac{1}{p}\right)}{B\left(\frac{1}{p}, q\right)} \right)^2 \right]^{-\frac{1}{2}}$$

where $B(\cdot)$ is the beta function, and μ, σ, λ, p and q are the location, scale, skewness, peakedness and tail-thickness parameters, respectively. Note that the parameters have the following restrictions $\sigma > 0$, $-1 < \lambda < 1$, $p > 0$ and $q > 0$. The skewness parameter λ controls the rate of descent of the density around $x = 0$. The parameters p and q control the height and tails of the density, respectively. The parameter q has the degrees of freedom interpretation in case $\lambda = 0$ and $p = 2$.

More complex and novel are the distributions belonging to the generalized hyperbolic family. An special case of this family is the Generalized Hyperbolic skew Student-t distribution proposed by Aas and Haff (2006). This distribution has the important property that one tail has polynomial and the other exponential behavior. Further, it is the only subclass of the Generalize Hyperbolic family of distribution having this property. This is an alternative for modeling the empirical distribution of financial returns. It is often skewed, having one heavy and one semiheavy or more Gaussian-like tail. The skew extensions to the Student-t distribution, like that of Fernandez and Steel, have two tails behaving as polynomials. This means that they fit heavy-tailed data well, but they do not handle substantial skewness, since that requires one heavy tail and one nonheavy tail.

The probability density function of the Generalized Hyperbolic skew Student-t is given by

$$f_X(x) = \frac{2^{\frac{1-\nu}{2}} \delta^\nu |\beta|^{\frac{\nu+1}{2}} K_{\frac{\nu+1}{2}} \left(\sqrt{\beta^2(\delta^2 + (x-\mu)^2)} \right) exp(\beta(x-\mu))}{\Gamma(\frac{\nu}{2})\sqrt{\pi} \left(\sqrt{\delta^2 + (x-\mu)^2} \right)^{\frac{\nu+1}{2}}} \qquad \beta \neq 0$$

and

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\delta\Gamma(\frac{\nu}{2})} \left[1 + \frac{(x-\mu)^2}{\delta^2} \right]^{-(\nu+1)/2} \qquad \beta = 0$$

where $K_\nu(x) \sim \sqrt{\frac{\pi}{2x}} exp(-x)$ for $x \to \pm\infty$ is the modified Bessel function (Abramowitz and Stegun, 1972), μ, δ, β and ν determine the location, scale, skew and shape parameters, respectively.

When $\beta = 0$ the density $f_X(x)$ can be recognized as that of noncentral Student-t distribution with ν degrees of freedom, expectation μ and variance $\delta^2/(\nu-2)$.

4 The data

We work with daily percentage returns on five groups of assets of different nature over the sample period 01/04/2000-12/31/2015 (4173 observations). Daily returns are computed as 100 times the first difference of log prices, i.e. $100[\ln(P_{t+1}) - \ln(P_t)]\%$. The financial assets considered are: stock market indexes: IBEX 35 (€), NASDAQ 100 ($), FTSE 100 (£) and NIKKEI 225 (¥); individual stocks: IBM ($), SAN (€), AXA (€) and BP (£); interest rates: IRS 5Y (€), interest rate of GERMAN BOND 10Y (€) and interest rate of US BOND 10Y($); commodity prices CRUDE OIL BRENT ($ per barrel), NATURAL GAS ($ per Million British Thermal Units), GOLD ($ per Troy Ounce) and SILVER (Cents $ per Troy Ounce) and exchange rates EUR/USD (€), GBP/USD (£), JPY/USD (¥) and AUD/USD (Australian $). The data were extracted from Datastream. We consider a variety of diverse assets in an attempt to get broad and robust results, since time series for financial returns share well-known common stylized facts like asymmetry and high kurtosis.

Table 1 reports descriptive statistics for daily returns. All the assets have mean and median returns close to zero. Returns on interest rates are obtained as log changes in the price of implicit zero coupon bonds having the value of an interest rate as a yield. In terms of standard deviation, the sample range is higher for AUD/USD (18.7), IRS (18.0) and US BOND (17.1) and lower for JPY/USD (13.2), EUR/USD (13.4), SILVER (13.8) and the interest rate on the GERMAN BOND (13.9). The unconditional standard deviation is relatively similar for assets in the same class, except for commodities, where GAS (4.19) and OIL BRENT (2.28) are more volatile than GOLD (1.13) and SILVER (1.93). NASDAQ is more volatile than other stock market indexes and AXA is the most volatile stock. The $US exchange rate for the Australian dollar has higher standard deviation than the one for the euro, British pound or Yen. AUD/USD, SILVER, GOLD and NIKKEI have significant negative skewness, while GAS, AXA, JPY/USD and NASDAQ have high positive skewness. For all the assets considered the kurtosis is high, implying that the return distributions have much thicker tails than the Normal distribution. Kurtosis is specially large for AUD/USD, GAS, IBM and AXA while EUR/USD, while the interest rate of the GERMAN BOND and the JPY/USD exchange rate have lower kurtosis. Together with a large sample size, these values for skewness and kurtosis lead to a vary large Jarque-Bera statistic, rejecting the assumption of Normality in all cases.

5 Parameter estimates

To perform a VaR analysis we estimate four volatility models: GARCH, GJR-GARCH, APARCH and FGARCH under each of the different probability distributions assumed for the innovations: Normal, Student-t, skewed Student-t, skewed Generalized Error, Johnson S_U, skewed Generalized-t and Generalized Hyperbolic skew Student-t distributions. In Table 2 we report estimation results of the APARCH model under the Johnson S_U probability distribution for the stock market indexes and for individual stocks.[7] An AR(1) model was specified for the conditional mean return in all

[7]The estimation results for this volatility model under the different probability distributions for the stock market indexes and for individual stocks in our sample are reported in Tables A2 and A3 of the Online Appendix.

cases.[8]

The model is particularly successful in capturing the autocorrelation and heteroscedasticity exhibited by the data. The Ljung-Box Q statistic for nine lags computed on the standardized residuals does not show evidence of autocorrelation at 1% significance level. The same statistic computed with nine lags on the squared standardized residuals is not significant at 1% except for IBEX and SAN, both of them for a narrow margin. The autoregressive effect in volatility is strong, with a β_1-parameter generally above 0.90, suggesting strong memory effects. The coefficient γ_1 is positive and statistically significant for all series, indicating the existence of a leverage effect for negative returns in the conditional variance. Estimates of γ_1 are close to 1 for IBEX, NASDAQ and FTSE, suggesting that only negative shocks contribute to volatility. The skewness parameter (γ) of the Johnson S_U distribution is less than 1 for the four stock indices, suggesting the convenience of incorporating negative asymmetric features in the probability distribution in order to model innovations appropriately. Finally, the δ-parameter takes values between 0.97 and 1.54, being significantly different from 2 in most cases. This result is in line with those of Taylor (1986), Schwert (1990) and Ding et al. (1993) who indicate that there is substantially more correlation among absolute returns than among squared returns, a reflection of the 'long memory' of high-frequency financial returns. Our estimates of the APARCH model for the different asset classes (not shown in the tables) suggest that, contrary to standard practice, we should model the conditional standard deviation for stock market indexes, individual stocks and metals, the conditional variance ($\delta = 2$) for interest rates, and a value between conditional standard deviation and variance ($\delta = 1.5$) for energy commodities and exchange rates. We obtained the same parameter estimates using MatLab, R, Eviews and Gretl.

In the lower panel we present the log-likelihood values of the four volatility models (GARCH, GJR-GARCH, APARCH and FGARCH) under the JSU probability distribution. For individual stocks as well as for stock indexes the least restricted FGARCH model achieves the highest likelihood, followed by APARCH, GJR-GARCH and GARCH models. A similar result was obtained for all other assets and all the probability distributions, and Figure 1 shows mean log-likelihood values for each model specification across the set of assets. Likelihood differences are statistically significant for many assets, with the GARCH specification being rejected against GJR-GARCH, and the latter being rejected against the APARCH and FGARCH specifications. The exceptions are exchange rates and long-term interest rates, for which it is hard to discriminate among volatility specifications. On the other hand, likelihood differences between APARCH and FGARCH models are generally not statistically significant.

6 VaR Performance

We now analyze the VaR performance of our estimated models restricting our attention to the left tail of the distribution and the 1% significance level. The choice of the $\alpha = 1\%$ level is a compromise between trying to capture extreme events and trying to avoid a too low number of exceptions. Results for alternative significance levels are available from the authors upon request.

[8]Most computations were performed with the rugarch package (version 1.3-4) of R software (version 3.1.1), designed for the estimation and forecast of various univariate ARCH-type models. The exception is the estimation of models under the skewed Generalized-t and Generalized Hyperbolic skew Student-t distributions for which we used the sgt package (version 2.0) and the SkewHyperbolic package (version 0.3-2), respectively.

Considering the left tail is not a trivial choice, since results for both tails may differ significantly for asymmetric return distributions.

We estimate the one-step ahead VaR parametrically as: $VaR_{\alpha,t} = \mu_t(\theta) + \sigma_t(\theta)F^{-1}(\alpha|\theta)$, where $\mu_t(\theta)$ represents the conditional mean, $\sigma_t(\theta)$ is the conditional standard deviation and $F^{-1}(\alpha|\theta)$ denotes the corresponding quantile of the distribution of the standardized innovations z_t at a given $\alpha\%$ significance. Every day we compute 1-day ahead 1% VaR forecasts over the last five years in the sample: 2011-2015 (1260 data observations). Models were reestimated every 50 days, a choice based on several arguments: 1) estimating each day is computationally demanding because we jointly estimate the parameters for the mean return, its conditional variance, and for the probability distribution of the return innovation,[9] 2) we work with an "expanding window" that starts with a 11-year sample (2000-2010) and an out-of-sample of 5 years (2011-2015), and adding a single data to the 10-year sample does not change the estimated parameters, 3) our choice is in line with that made by different authors, like Giot and Laurent (2003a), and Diamandis et al. (2011) among others, who reestimate their models every 50 days. However, we report below some VaR backtesting results obtained reestimating the models daily.

After that, we examine the performance of VaR models through standard tests: the unconditional coverage test of Kupiec (1995), the independence and conditional coverage tests of Christoffersen (1998), the Dynamic Quantile test of Engle and Manganelli (2004), as well as by evaluating the Asymmetric Linear Tick loss function (AlTick) proposed by Giacomini and Komunjer (2005). For a comprehensive review on VaR forecasting and backtesting, see Nieto and Ruiz (2015).

The unconditional coverage test introduced by Kupiec (1995) is based on the number of violations, i.e. the number of times (T_1) that returns exceed the predicted VaR over a period of time T for a given significance level. If the VaR model is correctly specified, the failure rate ($\hat{\pi} = \frac{T_1}{T}$) should be equal to the pre-specified VaR level (α). The null hypothesis $H0 : \pi = \alpha$ is evaluated through a likelihood ratio test:

$$LR_{uc} = -2\ln\left(\frac{L(\Pi_\alpha)}{L(\widehat{\Pi})}\right) = -2\ln\left(\frac{(1-\alpha)^{T_0}\alpha^{T_1}}{(1-\hat{\pi})^{T_0}\hat{\pi}^{T_1}}\right) \quad \overset{T\to\infty}{\longrightarrow} \chi^2_1$$

where $T_0 = T - T_1$.

Two other tests by Christoffersen (1998) examine whether VaR exceedances are independent. We consider two states of nature each period: state 0 if the return does not fall below VaR: $r_t > VaR(\alpha)$, and state 1, if $r_t < VaR(\alpha)$. For the alternative hypothesis of VaR inefficiency, it is assumed that the process of violations $I_t(\alpha)$, where $I_t(\alpha) = 1$ if $r_t < VaR(\alpha)$ and $I_t(\alpha) = 0$ otherwise, can be modeled as a Markov chain with $\pi_{ij} = Pr[I_t(\alpha) = j|I_{t-1}(\alpha) = i]$. Let us then denote by T_{ij} the number of observations in state j after having been in state i in the previous period, and define $T_0 = T_{00} + T_{10}$ and $T_1 = T_{11} + T_{01}$. The two probabilities of a VaR excess (state 1), conditional on the state of the previous period π_{01} and π_{11} are estimated by $\hat{\pi}_{01} = T_{01}/(T_{00} + T_{01})$ and $\hat{\pi}_{11} = T_{11}/(T_{10} + T_{11})$. Under the null hypothesis of independence of VaR exceedances: $\pi_{01} = \pi_{11} = \pi = (T_{11} + T_{01})/T$, the likelihood function is $L(\widehat{\Pi}) = (1-\hat{\pi})^{T_0}\hat{\pi}^{T_1}$. The likelihood under the alternative hypothesis is: $L(\widehat{\Pi}_1) = (1-\hat{\pi}_{01})^{T_{00}}\hat{\pi}_{01}^{T_{01}}(1-\hat{\pi}_{11})^{T_{10}}\hat{\pi}_{11}^{T_{11}}$. The independence test of Christoffersen (1998) is a test of the hypothesis of serial independence in VaR exceedances against a first-order Markov dependence. The likelihood ratio LR_{ind} statistic is:

[9] The computation is especially laborious for the FGARCH volatility specification and the GHST probability distribution.

$LR_{ind} = -2\ln(L(\widehat{\Pi})/L(\widehat{\Pi}_1))$ with a distribution χ^2_1 . The conditional coverage test is based on the likelihood ratio statistic, $LR_{cc} = -2\ln(L(\Pi_\alpha)/L(\widehat{\Pi}_1)) = LR_{uc} + LR_{ind}$, which is asymptotically distributed χ^2_2.

While the conditional coverage test is easy to use, it is rather limited for two main reasons, *i*) the independence is tested against a very particular form of alternative dependence structure that does not take into account a dependence of order higher than one, *ii*) the use of a Markov chain only considers the influence of past violations $I_t(\alpha)$ and not the influence of any other exogenous variable. The Dynamic Quantile Test proposed by Engle and Manganelli (2004) overcomes these two drawbacks of the conditional coverage test. These authors suggest using a linear regression model that links current violations to past violations. Let us define the auxiliary variable: $Hit_t(\alpha) = I_t(\alpha) - \alpha$, so that $Hit_t(\alpha) = 1 - \alpha$ if $r_t < VaR_{t|t-1}(\alpha)$ and $Hit_t(\alpha) = -\alpha$ otherwise. The null hypothesis of this test is that the sequence of hits (Hit_t) is uncorrelated with any variable that belongs to the information set Ω_{t-1} available when the VaR was calculated and it has a mean value of zero, which implies, in particular, that the hits are not autocorrelated. The Dynamic Quantile test is a Wald test of the null hypothesis that all slopes in the regression model,

$$Hit_t(\alpha) = \delta_0 + \sum_{i=1}^{p} \delta_i Hit_{t-i} + \sum_{j=1}^{q} \delta_{p+j} X_j + \varepsilon_t$$

are zero, where X_j are explanatory variables contained in Ω_{t-1}. The test statistic has an asymptotic χ^2_{p+q+1} distribution. In our implementation of the test, we use $p = 5$ and $q = 1$ (where $X_1 = VaR(\alpha)$) as proposed by Engle and Manganelli (2004). By doing so, we are testing whether the probability of an exception depends on the level of the VaR.

To evaluate the consequences of a VaR exceedance, we use the Asymmetric Linear Tick loss function (AlTick) proposed by Giacomini and Komunjer (2005), which takes into account the magnitude of the implicit cost associated with VaR forecasting errors. Hence, it takes into consideration not only the returns that exceed the VaR, but also the opportunity cost produced by an overestimation of VaR. When there are not exceptions, the loss function penalizes for the excess capital retained:

$$L_\alpha(e_{t+1}) = \begin{cases} (\alpha - 1)e_{t+1} & \text{if } e_{t+1} < 0 \\ \alpha e_{t+1} & \text{if } e_{t+1} \geq 0 \end{cases}$$

where $e_{t+1} = r_{t+1} - VaR_{t+1}$. Giacomini and Komunjer use the asymmetric linear loss function with α equal to the significance level used to forecast VaR. The AlTick function can be seen as the implicit loss function whenever the object of interest is a forecast of a particular quantile of the conditional distribution of returns. That way, a VaR model is preferable if it has a lower average value of the loss function.

The different combinations of probability distributions and volatility specifications, applied to each of the 19 assets considered, yield a large number of VaR tests and it is hard to summarize so much information in order to achieve some clear-cut conclusion on the adequacy of each model. We will proceed in the next section along four lines: i) the frequency of rejections of a given model when applying each test to the set of assets, ii) how often the *p*-value of a given test decreases when switching between two models differing in either the probability distribution or the volatility specification, iii) selecting the preferred models by a concept of precedence among VaR models we introduce below, iv) implementing a Model Confidence Set approach to select the preferred

VaR models for each asset. This approach is based on the use of a specific loss function. The first three criteria are based on properties of the tests for validation of VaR forecasts, while the fourth criterion deals with the size of the sample returns exceeding the estimated VaR.

6.1 Frequency of violations

Violation rates for VaR close to $\alpha = 0.01$ (13 violations) are desirable. Further, under the Basel Accord, models that over-estimate risk are preferable to those that under-estimate risk levels. In our case, less than 20 violations of VaR would define the 'green zone', between 20 and 50 violations would correspond to the 'yellow zone' and the 'red zone' would be defined by more than 50 violations.[10] In fact, falling inside the green zone is not necessarily a good thing if the number of violations of VaR is too low, since then the financial institution would be taking an excessive opportunity cost of capital.

Table 3 contains a summary of backtesting results, showing for each model specification the median number of VaR violations and the number of rejections of each test across the set of 19 assets.[11] The expected number of violations (13) falls in the green zone, so a good model should be in that zone. Across the 76 VaR analysis performed (4 volatility specifications and 19 assets) models under the Normal distribution fell in the green zone 26 times out of 76 (34%), 55 times under the Student-t distribution (72%), 72 times under SKST (95%), 69 times under SGED (91%), 75 times under JSU (99%), 73 times under SGT (96%) and 70 times under GHST (92%). All the other models fell in the yellow zone. The Normal distribution falls too often in the yellow zone. The frequency of the Student-t distribution to produce a model in the green zone was not very high either. All other probability distributions lead frequently to models in the green zone. We never observed a model to fall in the red zone for any asset.

Figure 2 shows the median number of VaR violations for each combination of probability distribution and volatility specification. The Normal distribution leads to the largest median number of violations (22) across the 76 models (4 volatility specifications and 19 assets). Since the expected number of violations is 13, the Normal distribution clearly underestimates the level of risk. The GHST distribution produces the lowest median number of violations (10), with a clear overestimation of risk. All the other probability distributions have a median number of violations around 15, with a slight underestimation of risk that is more evident for the Student-t distribution. We can say that except by the Normal and GHST distributions, all other distributions perform well. Being more specific, the median frequency of violations is 1.75% for models with Normal innovations, 1.27% for Student-t innovations, 1.19% for skewed Student-t, skewed Generalized Error and skewed Generalized-t innovations, 1.11% for Johnson S_U innovations and 0.79% for Generalized Hyperbolic skew Student-t innovations. According to the frequency of violations, the

[10] In terms of Basel Accord, based on a sample of 250 observations, if the number of exceptions is less than, or equal to 4 (the green zone), the test results are consistent with an accurate model and the possibility of erroneously accepting an inaccurate model is low. At the other extreme, if there are 10 or more exceptions (the red zone), the test results are extremely unlikely to have resulted from an accurate model, and the probability of erroneously rejecting an accurate model on this basis is remote. In between these two cases we have the yellow zone, where the backtesting results could be consistent with either accurate or inaccurate models, and the supervisor should encourage a bank to present additional information about its model before taking action. We have applied to these thresholds a scale factor based on our sample size of 1260 observations.

[11] Tables A4-A7 in the Online Appendix show for each asset the number of observed violations of VaR forecasts, the statistic and p-value of each test for each combination of volatility model and probability distribution for the innovations.

unbounded Johnson S_U distribution shows the best behavior among the asymmetric probability distributions. The performance of GHST might be acceptable under some criteria, although it would lead to an excessive opportunity cost of capital.

Differences among volatility specifications are much smaller. Models with a GARCH specification fell 114 times out of 133 cases (7 probability distributions and 19 assets) in the green zone (86%), 109 times for the GJR–GARCH (82%), 108 times for APARCH (81%) and 109 times for FGARCH (82%) out of 133 VaR analysis. The median number of violations was 15, 15, 16 and 16, respectively, very similar across volatility specifications. The frequency of violations for all volatility specifications is 1.19% for GARCH and GJR-GARCH, and 1.27% for APARCH and FGARCH models. This result already suggests the need to be careful when choosing an appropriate probability distribution for return innovations. Selecting the best volatility specification is also important, but the consequences of not making the right choice do not seem to be so crucial.

It is also interesting to examine the performance by asset type. Most models tend to overestimate risk in energy commodities (OIL and GAS). The median number of violations over the set of 28 models (7 probability distributions and 4 volatility specifications) is 7 for OIL and 5 for GAS (see Figure 3). A similar result is obtained for the GBP/USD and AUD/USD exchange rates, with a median number of 10 violations in both cases, which is not the case for the two other exchange rates.[12] But the general result is that more often than not, models tend to underestimate risk in all assets, with a number of violations above the expected value of 13. Underestimation is specially evident in the non-industrial metals (GOLD and SILVER) and some Spanish stock market variables (SAN and IBEX).

Using data for NASDAQ 100, Table 4 shows that differences in backtesting statistics for $VaR_{1\%}$ when models are re-estimated every day or every 50 days are small. The largest differences arise for FGARCH volatility specifications and for the DQT test. The similarity in results reinforces our choice to re-estimate models every 50 days.

A quick glance at Table 3 already reveals that the number of rejections of the four tests is highest under the assumption of a Normal distribution for return innovations. However, comparing all the other models in terms of backtesting results is far from obvious. The next sections try to select the best performing model specifications using different approaches.

6.2 Switching between VaR models

In a comparative analysis of VaR forecasting performance, applying $n = 4$ tests to $m = 28$ alternative models representing the dynamics of $k = 19$ assets, we will generally have $m \cdot n \cdot k = 2128$ test outcomes. For 19 assets, we have a total of 216 tests performed under each probability distribution, and 378 tests under each volatility specification.[13] They produce a large amount of information, and we need to design ways to summarize that information in order to be able do draw some conclusion on the relative merits of each probability distribution and each volatility specification. This is what we do in the next sections.

We start by comparing, for each of the four VaR tests described above (Kupiec, independence, conditional coverage and Dynamic Quantile tests), the p-values of the test statistics for models that differ in either the probability distribution for the innovations or in the specification of volatility

[12] The median number of violations is also below 13 for BP, but it is so close to that target that we have to consider the difference as sampling error.

[13] Notice that the independence and the conditional coverage tests not always can be applied.

dynamics. In these tests the null hypothesis is H0: the VaR model is 'appropriate', in some sense that is specific to each test. As the probability of finding a similar sample with a more contrary evidence to H0, the *p*-value gives us a numerical indication on how favorable is our sample to H0. Hence, when comparing any two VaR forecasting models, we should prefer the one with a higher *p*-value in VaR validation tests. To summarize the results of this analysis, Table 5 displays the number of cases in which the *p*-value of the test statistic increases or decreases when we change either the probability distribution or by the specification of the volatility model. We cannot make any formal testing, but by comparing *p*-values, we are searching for patterns that might suggest that a particular model is preferred over a given alternative.

If we consider all the possible specifications sharing the same probability distribution for return innovations, we see that switching from a Normal to a Student-t distribution for return innovations increases the *p*-value of VaR tests in 160 out of a total of 216 comparisons, suggesting in those cases a more accurate VaR model.[14] Even though the test statistics are obviously subject to sampling error, that frequency of increases in *p*-value suggests that, as expected, the Student-t distribution is generally more appropriate than the Normal distribution to represent financial returns. Switching from the symmetric to the skewed Student-t distribution achieves a further increase in *p*-value in 114 comparisons, while decreasing in 75 cases. Moving from the asymmetric Student-t to the unbounded Johnson distribution achieves an increase in 91 cases while decreasing in 55 cases. Switching from the asymmetric Student-t (SKST) to other asymmetric distributions (SGT, JSU, SGED), the *p*-value increases more often than otherwise. On the contrary, if we switch from the SKST, SGED, JSU or SGT distributions to the GHST distribution, the opposite happens, with the *p*-value usually decreasing. Hence, we consider the SKST, SGED, JSU and SGT distributions to be preferable to GHST. Between these asymmetric distributions, switching to JSU or SGT leads to an increase in *p*-value in a greater number of cases.

Among volatility models, switching from the symmetric GARCH to GJR-GARCH increases the *p*-value of the statistic in 176 out of 378 comparisons. The *p*-value increases in 131 cases when switching from GJR-GARCH to APARCH, but it decreases in 167 cases. On the other hand, if we move from the APARCH to the FGARCH model, the *p*-value increases in 151 out of 378 cases, decreasing in 128 cases. Overall, the FGARCH model seems to be the preferable volatility specification. Percent differences between the number of cases in which the value of the test statistic increases or decreases when switching between volatility models are much smaller than the ones obtained when switching between two probability distributions. This suggests again that, according to the performance of the models for VaR estimation, the specification of the volatility dynamics is not as important as the choice of probability distribution for the innovation in returns.

6.3 A ranking based on backtesting results

In the previous section we have applied four tests for VaR performance: the unconditional likelihood-ratio test, the independence test, the conditional coverage test, and the dynamic quantile test, and each test has been run for a variety of models and assets. In this section we evaluate the adequacy of the different models considered for VaR forecasting by comparing the specific situations in

[14]The number of possible comparisons arises from applying all the VaR tests to all the assets. The difference between this number and the sum of increases and decreases in the *p*-value is the number of cases in which the *p*-value of the test statistic does not change.

which each model has been rejected by each test.[15]

Definition 1 *Given a confidence level between 0 and 1, we say that model M2 δ-precedes model M1 in VaR performance if i) M1 has been rejected in at least as many cases as M2, and ii) in a percentage of at least δ of the cases when M2 is rejected by a test, M1 is also rejected.*

Notice that δ does not need to be related to the confidence level at which VaR validation tests are implemented. We would expect δ to be around .90 in most practical applications. The interesting feature of this precedence criterion is that it compares any two model specifications across all the statistical tests and assets, thereby allowing us to achieve some robust results. The criterion could accommodate different weights for each test depending on the relevance we want to assign them. The precedence criterion would then use the number of rejections in each test, weighted by relevance. An interesting possibility would consist of assigning a larger weight to tests having a larger ability to discriminate among models. Weights could also be chosen as a bounded function of the size of the test rejection, either in terms of the test statistic or the *p*-value of the test.

The precedence criterion could also be used to choose among forecasting models that are required to satisfy some condition to be considered acceptable. For instance, if competing models are used over a number of periods to forecast a given variable, and there is a maximum forecast error that is acceptable, the precedence relationship would be based on the number of periods in which each model exceeds that error threshold.

Table 6 contains the information needed to establish precedence comparisons across VaR models. The upper panel corresponds to implementing the VaR validation tests at 99% confidence, while the lower panel has been obtained with test results at 95% confidence. In each panel, the upper part compares the rejections of models using probability distributions D1 (left) and D2 (right) when combined with all the volatility specifications. The lower part compares the rejections of models made up with volatility specifications M1 (left) and M2 (right) when combined with all the probability distributions. The first two columns of each panel in Table 6 show the number of cases when the two probability distributions or the two volatility specifications listed in the first column have been rejected by the data when applying the unconditional coverage tests of Kupiec. The third column displays the percentage of rejections of D2 (M2) that were also rejections of D1 (M1). We will conclude that the the probability distribution (or the volatility specification) with the lower number of rejections precedes the competitor when this percentage is below a pre-specified threshold for δ. The following three columns refer to the independence tests, and the next columns come from the conditional coverage test and the Dynamic Quantile test. The final three columns aggregate the number of rejections across tests. For instance, if we take a threshold $\delta = .90$, the independence test of Kupiec rejected 36 models made up with the Normal distribution and just 7 models with the Student-t distribution. Besides, those 7 models rejected under the Student-t distribution were also rejected under a Normal distribution. Hence, the Student-t distribution precedes the Normal distribution according to this test. The independence test rejected 7 models made up with either the Normal or the Student-t distributions. In 5 of the 7 rejections under a Student-t distribution for return innovations the model was also rejected under a Normal distribution. That

[15]However, some tests might not be feasible in some samples, which explains the empty cells for the independence test and the conditional coverage test in Tables A4 to A7 of Online Appendix.

ratio is 5/7=0.714 so that, according to the independence test, we could not conclude that models with a Student-t distribution for return innovations precede models with a Normal distribution.

The number of pairwise comparisons between probability distributions or between volatility specifications is very high because they could be made in both directions, so we show in Table 6 the more interesting ones. For instance, we do not explicitly show the comparisons between the Normal distribution and asymmetric distributions because the latter always precedes the former. Similarly, we do not show pairwise comparisons between Student-t and any asymmetric distribution other than the skewed Student-t (SKST) because the skewed Student-t tends to precede the standard Student-t, and the majority of asymmetric distributions precede in turn over the skewed Student-t distribution.[16]

Taking into account the aggregate results across the four tests we can summarize the comparisons at $\alpha = 95\%$ as in the diagram:

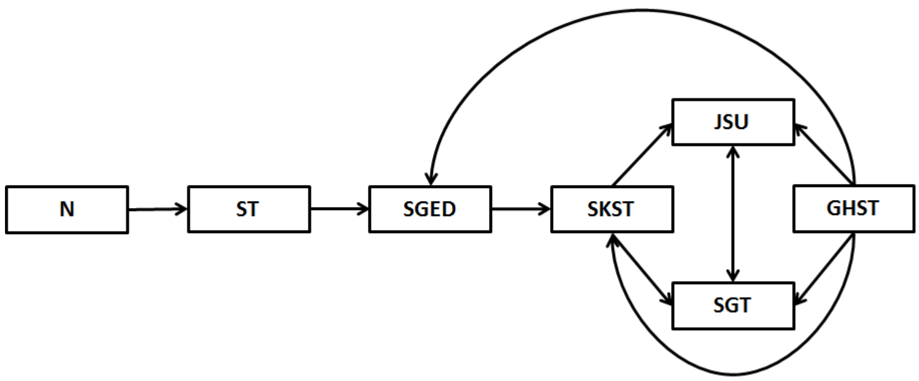

Diagram 1: Precedence relationship among probability distributions from aggregate results across the four test at $\alpha = 95\%$. Each arrow head points to a model that precedes the model where the arrow originates. A two-headed arrow indicates two models that do not precede each other.

No matter whether we take $\alpha = 99\%$ or $\alpha = 95\%$, the Normal, Student-t, SKST and SGED distributions are preceded by other alternatives, specially JSU and SGT. We observe that JSU and SGT distributions seem to precede over all others, while not being preceded by each other. According to this δ-precedence criterion the GHST distribution is judged again not to be appropriate for VaR estimation, since it is preceded by the rest of asymmetric distributions. The Normal distribution is also preceded by all other distributions.

At $\alpha = 99\%$ there is not a clear precedence ordering between volatility specifications. For $\alpha = 95\%$ the FGARCH specification seems to precede all others but, once again, differences are not as clear as when comparing probability distributions.

A preference for APARCH and FGARCH models against standard GARCH and GJR-GARCH has been a constant throughout our analysis up to this point. So, a robust conclusion is the need to incorporate a leverage effect in volatility and, possibly more important, the convenience to model standard deviations, rather than variances. The preference for asymmetric probability distributions in Table 6 is also consistent with results in Table 5 when comparing p-values of the test statistics.

[16]Even though δ-precedence is not a transitive relationship, it seems safe to focus on the models that tend to be δ-precendent.

Both analysis are based on the same information, but they use it in a very different fashion. Nothing guarantees that the conclusions on the preferred probability distributions should be the same in both analysis. On the contrary, this coincidence should be seen as a proof of the robustness of such preference.

6.4 Model Confidence Sets

The availability of several model specifications being able to adequately describe the unobserved data generating process (DGP) opens the question of selecting the 'best fitting model' according to a given optimality criterion. Recently, significant effort has been placed on developing testing procedures being able to deliver the 'best fitting' models among a set of alternatives. One of the first proposals was Diebold and Mariano (1995), but it is not applicable when the forecasts come from nested models or when the forecasts are calculated from semiparametric or non parametric methods (Giacomini and Komunjer, 2005). This has been overcome by the Reality Check (RC) approach of White (2000), the Stepwise Multiple Testing procedure of Romano and Wolf (2005), the Superior Predictive Ability (SPA) test of Hansen and Lunde (2005), the Conditional test of Giacomini and White (2006), and the Model Confidence Set (MCS) procedure developed by Hansen et al. (2011). All these approaches are relevant from an empirical point of view, especially when the set of competing alternatives is large.

We implement the Model Confidence Set (MCS) procedure developed by Hansen et al. (2011). It is a general approach to model selection that it does not require that the "true" model must be available as one of the competing models. The MCS procedure consists of a sequence of tests to construct the 'Set of Superior Models' (SSM). At each step the worst model is eliminated, until the hypothesis of Equal Predictive Ability (EPA) is not rejected for any of the models in the current SSM. At each step, each element in the SSM is characterized as having better predictive ability than models not in the set. The SSM has an interpretation similar to a confidence interval for a parameter in the sense that, with a given level of confidence, the SSM contains the best model. The EPA test statistic is evaluated under a given loss function, so that it is possible to test models on various aspects depending on the chosen loss function.[17]

Formally, the loss function $\ell_{i,t}$ associated to the i-th model $\ell_{i,t} = \ell(Y_t, \hat{Y}_{i,t})$ measures the cost associated to the difference between the observation at time t, Y_t, and $\hat{Y}_{i,t}$ the output of model i at time t. The MCS procedure starts from an initial set of models $\hat{M}^0$ of dimension m made up by all combinations of probability distributions and volatility specification considered in previous sections. Then, for a given confidence level $1-\alpha$, we obtain a smaller set, the superior set of models, SSM, $\hat{M}^*_{1-\alpha}$ of dimension $m^* \leq m$. Let us denote by d_{ij} the loss differential between models i and j, $d_{ij,t} = \ell_{i,t} - \ell_{j,t}$, $i,j = 1,...,m$, $t = 1,...,T$. The EPA hypothesis for a given set of models M can be formulated: $H_{0,M}$: $c_{ij} = 0$, for all $i,j = 1,...,m$, against the alternative: $H_{1,M}$: $c_{ij} \neq 0$, for some $i,j = 1,...,m$, where $c_{ij} = \mathbb{E}(d_{ij})$ is assumed to be finite and not time dependent. This hypothesis can be tested using the test statistic [Hansen et al. (2011)], $t_{ij} = \bar{d}_{ij} / \sqrt{\widehat{var}(\bar{d}_{ij})}$

[17]We believe that the opportunity cost of overestimating VaR is non trivial. The AlTick loss function not only penalizes underestimation but also risk overestimation, because of the excess capital retained, and therefore we prefer it over other loss functions, such as those proposed by Lopez (1998, 1999) and Sarma et al. (2003) which only penalize risk underestimation. However, it would be worthwhile to explore other loss functions that might focus on different characteristics of VaR forecasts.

for $i, j \in M$, where $\bar{d}_{ij} = n^{-1} \sum_{t=1}^{n} d_{ij,t}$ measures the relative sample loss between the i-th and j-th models, while $\widehat{var}(\bar{d}_{ij})$ is a bootstrapped estimate of $var(\bar{d}_{ij})$.

Following Hansen et al. (2011) and Radovanov and Marcikic (2014) we calculate the bootstrapped variances by a block-bootstrap procedure. To that end, we divide the time series of 1260 data observations into overlapping blocks of length p, which is usually estimated as the maximum number of significant parameters in an AR(p) process fitted to all the d_{ij} terms. Since financial returns exhibit little linear autocorrelation, an AR(1) is enough to capture the dependence structure, and we used p=1 to resample individual observations. We checked that using a block length of 2 does not change significantly the characterization of the MCS. As discussed in Hansen et al. (2011) the EPA null hypothesis maps naturally into the statistic, $T_{R,M} = \max_{i,j \in M} |t_{ij}|$. Since the asymptotic distributions of this test statistic is nonstandard, the relevant distribution under the null hypothesis was estimated using a bootstrap procedure similar to that used to estimate $var(\bar{d}_{ij})$.

Table 7 reports the frequency by which each probability distribution and each volatility specification enter into the Superior Set of Models for each asset.[18] Tests are performed at the 90% confidence level, using a block-bootstrap procedure of 10000 resamples with a block length of 1. The table shows that for some assets, like NASDAQ 100, FTSE 100, EUR/SD and JPY/USD, the SSM includes a variety of distributions and volatility specifications. That indicates that the one-step ahead 1% VaR forecasting performance of the competing combinations of probability distribution and volatility specification is relatively similar, suggesting that for these assets the use of simple models for VaR forecasting may be justified. The SGT, JSU, SGED and GHST distributions are the ones that enter most often into the MCS of the set of assets considered. Among the volatility models, FGARCH and APARCH seem to describe quite well the behavior of financial time series, although the symmetric GARCH also enters into the MCS quite often. Concerning the distribution specifications, we observe that the MCS confirms the common finding that the Normal distribution provides a poor description of the behavior of financial time series. Under the AlTick loss, the skewed Generalized-t and skewed Generalized Error distributions perform better than the Generalized Hyperbolic skew Student-t. Definitely, the Normal, Student-t and skewed Student-t distributions do not seem to be appropriate for VaR forecasting for the wide set of financial assets considered in this paper.

6.5 10-day VaR forecasting

In spite of the Basel Committee on Banking Regulation (2009) switch to require 10-day VaR estimation, there is not much work yet exploring the performance of alternative VaR models. Degiannakis and Potamia (2017) and Degiannakis et al. (2013) analyze a number of issues regarding 10-day VaR and expected shortfall forecasting. Degiannakis and Potamia (2017) conclude that the use of intra-day data does not lead to better risk estimates. They also obtain a preference for skewed distributions and a GARCH volatility specification, as well as a better forecasting performance at 97.5% than at 99% significance level. Degiannakis et al. do not find an improvement in the accuracy of risk estimation from using long memory volatility models.

A major difficulty with multi-step VaR forecasting is that the use of non-overlapping samples drastically reduces the number of VaR observations. To solve this limitation, Barone-Adesi et al.

[18] Table A8 in the Online Appendix shows the values of the AlTick loss function using for the different model specifications and assets.

(1998, 1999, 2002) propose the FHS method that extends the idea of volatility adjustment to multistep historical simulation, using overlapping data in a way that does not create blunt tails for the h-day portfolio return distribution, $h > 1$. The method consists in applying a statistical bootstrap to the standardized residuals of a parametric dynamic model of returns, to simulate log returns each day over the desired risk horizon. Typically, the model used for FHS incorporates a specification of the GARCH family for volatility dynamics. The filtering involved in FHS allows for h-day return distributions to be generated from overlapping samples, since the bootstrap allows for increasing the number of observations used for building the h-day return distribution. The advantages of the FHS approach are 1) it captures current market conditions by means of the volatility dynamics, 2) no assumptions need to be made on the distribution of the return innovations, and 3) the method allows for the computation of any risk measure at any investment horizon of interest because one can generate as many h-day returns as one likes. The drawback is that we can only apply to these multi-period returns the unconditional coverage test.

Table 8 shows the performance of the models in 10-day ahead 1% VaR forecasts for NASDAQ 100. We use an expanding window to estimate the model, starting with the 2915 observations from the 10/2/2000-12/2/2011 period. Each day we add a new observation, we estimate the models and apply filtered historical simulation (FHS) to simulate 5000 10-day future returns from which we compute VaR forecasts. The forecasting exercise extends over 1260 days, the last five years in our sample, 12/5/2011-9/30/2016, obtaining daily 10-day VaR forecasts.[19] The results we obtain for 10-day VaR forecasts can be summarized as follows: *i*) VaR violation rates are below their theoretical value of 13, indicating an overestimation of risk for all models; *ii*) models with symmetric and models with asymmetric distributions perform similarly at VaR forecasting, although GARCH volatility specifications have a better performance, with violation rates being systematically closer to their expected value;[20] *iii*) even though model specifications are not rejected, unconditional coverage test p-values are low, except for GARCH volatility.

Some of the daily changes in VaR at short horizons may be due to pure noise that gets wiped out in longer horizons. That explains that 10-day VaR forecasts are smoother than 1-day VaR forecasts and differences in estimates across models are smaller. Furthermore, under a semiparametric method for VaR forecasting as FHS, the chosen model is less relevant than under a parametric methodology.

7 Conclusions

This paper extends previous work on the forecasting performance of alternative VaR models by considering four volatility specifications: GARCH, GJR-GARCH, APARCH and FGARCH and a set of asymmetric probability distributions: skewed Student-t, skewed Generalized Error, unbounded Johnson, skewed Generalized-t and Generalized Hyperbolic skew Student-t distributions, some of them being relatively new to the financial literature. Standard symmetric distributions and GARCH models without leverage are also used as a benchmark. Our sample of daily data for assets of different nature for the January 2000-December 2015 period covers the recent financial crisis of 2007-2009.

[19]Thus, we have 10-day VaR observations that we can compare to the realized 10-day returns.

[20]This latter result is consistent with Degiannakis and Potamia (2017).

Two clear results refer to issues that have been analyzed in previous research by a number of authors: *i*) VaR models that assume asymmetric probability distributions for return innovations, like the skewed Student-t distribution, skewed Generalized Error distribution, Johnson SU distribution, and skewed Generalized-t distribution achieve better VaR performance than models with symmetric distributions, *ii*) APARCH and FGARCH models, that allow for more flexibility in modeling volatility, show a better VaR performance than more standard GARCH and GJR-GARCH volatility specifications.

Our analysis highlights other important issues. A third result is that the shape and the skew of the assumed probability distribution for innovations are even more important for the performance of a Value-at-Risk model than including a leverage effect in volatility. This corroborates results by other authors [see Lopez and Walter (2000), Angelidis and Degiannakis (2006), Gerlach et al. (2011), Dendramis et al. (2014), and Braione and Scholtes, 2016]. We provide a thorough analysis of this issue by showing that the result holds for the wide set of assets we have considered: *i*) the frequency of rejections of VaR tests in models that differ in their volatility specification is similar, while rejection frequencies among models with the same volatility specification but a different probability distribution for the innovations can differ very significantly, *ii*) changing the probability distribution in a VaR model affects the *p*-value of the statistic for VaR tests by a larger amount than changing the volatility specification, *iii*) the *precedence* criterion we have introduced in this paper establishes a clear ranking between models differing in their probability distribution, while the distinction between models that differ in their volatility specification is much less clear.

A fourth result deals with the fact that our estimates suggest that for a number of financial assets the true, unobserved volatility dynamics should not be specified in terms of the squared conditional standard deviation. Hence, models specified for the conditional variance are prone to produce biased results. Dealing with the power of the conditional standard deviation as a free parameter is an important feature of the APARCH/FGARCH volatility specifications which explains their better performance in validation tests of VaR forecasts.

According to VaR performance, switching to a Johnson SU or a skewed Generalized-t distribution tends to increase the *p*-value of VaR validation tests. In terms of the precedence criterion among VaR models we have introduced in this paper, the unbounded Johnson and skewed Generalized-t precede other asymmetric distributions like the skewed Student-t, the Generalized Hyperbolic skew Student-t and the skewed Generalized error distribution, as well as the symmetric distributions like Student-t and Normal. The skewed Generalized-t and skewed Generalized Error distributions perform better than the other distributions in terms of the Model Confidence Set procedure. According to all these analysis, FGARCH seems the preferred model to capture the volatility of financial time series, with APARCH as a close second. In summary, the combination of APARCH or FGARCH volatility with a skewed Generalized Error, skewed Generalized-t or unbounded Johnson SU distributions should be expected to provide the best VaR performance for a wide array of assets of different nature.

This evidence has been obtained in the search for broad and robust conclusions over the set of assets considered. But it could be the case that alternative VaR models provided different VaR performance for distinct asset classes. This is clearly an important issue that deserves being considered for further research.

The preference for asymmetric distributions and APARCH/FGARCH volatility disappears in 10-day VaR forecasting obtained by filtered historical simulation, with all models showing an overestimation of risk that is less obvious for GARCH volatility specifications. Following a different simulation strategy, other authors have obtained an underestimation of risk (see Degiannakis and Potamia, 2017). Since the Basel Committee on Banking Supervision (2009) requires 10-day VaR predictions, a further analysis of the different performance of alternative VaR models and simulation strategies at the different horizons remains as a central issue for further research.

Acknowledgements

The authors gratefully acknowledge financial support from the grants Ministerio de Economía y Competitividad ECO2015-67305-P, Generalitat Valenciana PrometeoII/2013/015, Programa de Ayudas a la Investigación en Macroeconomía, Economía Monetaria, Financiera y Bancaria e Historia Económica 2015-2016 from Banco de España.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aas, K., & Haff, I.H. (2006). The generalized hyperbolic skew student'st-distribution. *Journal of Financial Econometrics*, 4(2), 275-309.

Abad, P., & Benito, S. (2012). A detail comparison of value at risk estimates. *Mathematics and Computers in Simulation*, 94, 258-276.

Abad, P., Benito, S., & Lopez, C. (2015). Role of the loss function in the VaR comparison. *Journal of Risk Model Validation*, 9(1), 1-19.

Abad, P., Benito, S., Lopez, C., & Sanchez-Granero, M.A. (2016). Evaluating the performance of the skewed distributions to forecast value-at-risk in the global financial crisis. *Journal of Risk*, 18(5), 1-28.

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables (Vol. 9)*. Dover: New York.

Andersen, T., Bollerslev, T., Diebold, F., & Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica*, 71, 529-629.

Ane, T. (2006). An analysis of the flexibility of Asymmetric Power GARCH models. *Computational Statistics and Data Analysis*, 51, 1293-1311.

Angelidis, T., Benos, A., & Degiannakis, S. (2007). A robust VaR model under different time periods and weighting schemes. *Review of Quantitative Finance and Accounting*, 28, 187-201.

Angelidis, T., & Degiannakis, S. (2006). Backtesting VaR models: An Expected Shortfall approach. http://dx.doi.org/10.2139/ssrn.898473.

Azzalini, A., & Capitanio, A. (2003). An asymmetric Generalization of Gaussian and Laplace Laws. *Journal of the Royal Statistical Society*, Series B, 65, 367-389.

Bali, T.G., & Theodossiou, P. (2007). A conditional-SGT-VaR approach with alternative GARCH model. *Annals of Operations Research*, 151, 241-267.

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics*, 24, 1-14.

Barone-Adesi, G. Bourgoin, F., and Giannopoulus, K., Don't look back. *Risk*, 1998, **11**, 100-103.

Barone-Adesi, G., Giannopoulus, K. and Vosper, L., VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, 1999, **19**, 583-602.

Barone-Adesi, G., Giannopoulos, K. and Vosper, L., Backtesting derivative portfolios with filtered historical simulation (FHS). *European Financial Management*, 2002, **8**(1), 31-58.

Basel Committee on Banking Supervision (2009). Revisions to the Basel II market risk framework. Basel, Switzerland: BIS.

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

Braione, M., & Scholtes, N.K. (2016). Forecasting Value-at-Risk under different distributional Assumptions. *Econometrics*, 4(3).

Brooks, C., & Persand, G. (2003). Volatility forecasting for risk management. *Journal of Forecasting*, 22, 1-22.

Bubak, V. (2008). Value-at-Risk on Central and Eastern European stock markets: An empirical investigation using GARCH models. *Institute of Economics Studies, Charles University*, 18. https://www.econstor.eu/handle/10419/83344

Caporin, M. (2008). Evaluating Value-at-Risk measures in the presence of long memory conditional volatility. *The Journal of Risk*, 10(3), 79-110.

Christoffersen, P. (1998). Evaluating internal forecasting. *International Economic Review*, 39, 841-862.

Choi, P., & Nam, K. (2008). Asymmetric and leptokurtic distribution for heteroscedastic asset returns: the SU-normal distribution. *Journal of Empirical finance*, 15(1), 41-63.

Corlu, C.G., Meterelliyoz, M., & Tiniç, M. (2016). Empirical distributions of daily equity index returns: A comparison. *Expert System with Applications*, 54, 170-192.

Degiannakis, S., Dent, P., & Floros, C. (2013). Forecasting Value-at-Risk and Expected Shortfall using Fractionally Integrated Models of Conditional Volatility: International Evidence. *International Review of Financial Analysis*, 27, 21-33.

Degiannakis, S., & Potamia, A. (2017). Multiple-days-ahead value-at-risk and expected shortfall forecasting for stock indices, commodities and exchange rates: Inter-day versus Intra-day data. *International Review of Financial Analysis*, 49, 176-190.

Dendramis, Y., Spungin, G.E., & Tzavalis, E. (2014). Forecasting VaR models under different volatility processes and distributions of return innovations. *Journal of Forecasting*, 33, 515-531.

Diamandis, P.F., Drakos, A.A., Kouretas, G.P., & Zarangas, L. (2011). Value-at-Risk for long and short trading positions: Evidence from developed and emerging equity markets. *International Review of Financial Analysis*, 20, 165-176.

Diebold, F.X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3), 253-263.

Ding, Z., Granger, C.W.J., & Engle, R.F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1, 83-106.

El Babsiri, M., & Zakoian, J.M. (2001). Contemporaneous asymmetry in GARCH processes. *Journal of Econometrics*, 101, 257-294.

Engle R.F., & Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22, 367-381.

Ergün, A., & Jun, J. (2010). Time-varying higher-order conditional moments and forecasting intraday VaR and expected shortfall. *Quarterly Review of Economics and Finance*, 50, 264-272.

Fernandez, C., & Steel, M. (1998). On bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441), 359-371.

Gerlach, R., Chen, C.W.S., Lin, E.M.H., & Lee, W.C.W. (2011). Bayesian forecasting for financial risk management, pre and post the global financial crisis. *Journal of Forecasting*, 31(8), 661-687.

Giacomini, R., & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, 23(4), 416-431.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545-1578.

Giot, P., & Laurent, S. (2003a). Value-at-Risk for long and short trading positions. *Journal of Applied Econometrics*, 18, 641-664.

Giot, P., & Laurent, S. (2003b). Market risk in commodity markets: a VaR approach. *Energy Economics*, 25, 435-457.

Glosten, L., Jagannathan R., & Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48, 1779-1801.

Hansen, B. (1994). Autorregressive conditional density estimation. *International Economic Review*, 35, 705-730.

Hansen, P.R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)?. *Journal of Applied Econometrics*, 20(7), 873-889.

Hansen P.R., Lunde A., & Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453-497.

Hentschel, L. (1995). All in the family nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, 39, 71-104.

Hull, J., & White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of Risk*, 1(1), 5-19.

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translations. *Biometrika*, 36, 149-176.

Kang, S.H., & Yong S-M. (2009). Value-at-Risk analysis for Asian emerging markets: asymmetry and fat tails in returns innovation. *The Korean Economic Review*, 25, 387-411.

Kuester, K., Mittnik, S., & Paolella, M.S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1), 53-89.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 2, 174-184.

Lambert, P., & Laurent, S. (2001). Modelling financial time series using GARCH-type models with a skewed student distribution for the innovations. *Mimeo, Université de Liege*.

Leccadito, A., Boffelli, S., & Urga, G. (2014). Evaluating the accuracy of value-at-risk forecasts: New multilevel tests. *International Journal of Forecasting*, 30, 206-216.

Lee, C.F., & Su, J.B. (2015). Value-at-Risk estimation via a semi-parametric approach: Evidence from the stock markets. *Handbook of Financial Econometrics and Statistics*. Springer Science Business Media: New York.

Li, M. Y., & Lin, H. W. W. (2004). Estimating value-at-risk via Markov switching ARCH models - an empirical study on stock index returns. *Applied Economics Letters*, 11(11), 679-691.

Lopez, J.A. (1998). Testing your risk tests. *Financial Survey* (May-Jun), 18-20.

Lopez, J.A. (1999). Methods for evaluating Value-at-Risk estimates. *Federal Reserve Bank of San Francisco Economic Review*, 2, 3-17.

Lopez, J.A., & Walter, C.A. (2000). Evaluating covariance matrix forecasts in a Value-at-Risk framework. http://dx.doi.org/10.2139/ssrn.305279

Louzis, D.P., Xanthopoulos-Sisinis, S., & Refenes, A.P. (2014). Realized volatility models and alternative Value-at-Risk prediction strategies. *Economic Modelling*, 40, 101-116.

Mabrouk, S., & Saadi, S. (2012). Parametric Value-at-Risk analysis: Evidence from stock indexes. *The Quarterly Review of Economics and Finance*, 52, 305-321.

McDonald, J. B., & Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric theory*, 4(3), 428-457.

McMillan, D.G., & Kambourodis, D. (2009) Are RiskMetrics forecasts good enough? Evidence from 31 stock markets. *International Review of Financial Analysis*, 18, 117-124.

Mittnik, S., & Paolella, M.S. (2000). Prediction of financial downside-risk with heavy-tailed Conditional Distributions. http://dx.doi.org/10.2139/ssrn.391261

Novales, A., & Garcia-Jorcano, L. (2018). Backtesting extreme value theory models of expected shortfall. *Quantitative Finance*, 1-27. DOI: 10.1080/14697688.2018.1535182

Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347-370.

Nieto, M.R., & Ruiz, E. (2016). Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting*, 32, 475-501.

Ozun, A., Cifter, A., & Yilmazer, S. (2010). Filtered extreme-value theory for value-at-risk estimation: evidence from Turkey. *The Journal of Risk Finance*, 11(2), 164-179.

Paolella, M. (1997). Tail estimation and conditional modeling of heteroskedstic time-series. *Ph.D Thesis, Institute of Statistics and Econometrics, Christian Albrechts University of Kiel*.

Polanski, A., & Stoja, E. (2010). Incorporating higher moments into value-at-risk forecasting. *Journal of Forecasting*, 29, 523-535.

Radovanov, B., & Marcikic, A. (2014). A comparison of four different block bootstrap methods. *Croatian Operational Research Review*, 5(2), 189-202.

Romano, J.P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237-1282.

Sarma, M., Thomas, S., & Shah, A. (2003). Selection of value at risk models. *Journal of Forecasting*, 22, 337-358.

Schwert, W. (1990). Stock volatility and the crash of '87. *Review of Financial Studies*, 3, 77-102.

So, M.K.P., & Yu, P.L.H. (2006). Empirical analysis of GARCH models in value at risk estimation. *Journal of International Financial Markets, Institutions and Money*, 16, 180-197.

Tang, T.L., & Shieh, S.J. (2006). Long Memory in stock index futures markets: A value-at-risk approach. *Physica A*, 366, 437-448.

Taylor, S.J. (1986). Modelling Financial Time Series. *John Wiley and Sons, Inc*.

Theodossiou, P. (2001). Skewness and kurtosis in financial data and the pricing of options. *Working Paper. Rutgers University.*

Theodossiou, P. (1998). Financial data and skewed generalized t distribution. *Management Science*, 44, 1650-1661.

Tu, A.H., Wong, W.K., & Chang, M.C. (2008). Value-at-Risk long and short positions of Asian stock markets. *International Research Journal of Finance and Economics*, 22, 135-143.

I TABLES

I.1 STATISTICS

	Mean (bps.)	Median (bps.)	Max	Min	S.D.	Skewness	Kurtosis	J-B
IBEX	-0.47	2.89	13.48	-9.58	1.49	0.08	7.93	4234.84
NASDAQ	0.46	3.68	17.20	-11.11	1.85	0.19	9.62	7652.53
FTSE	-0.25	0	9.38	-9.26	1.21	-0.16	9.36	7042.80
NIKKEI	0.01	0	13.23	-12.11	1.50	-0.41	9.72	7979.58
IBM	0.42	0	12.26	-16.89	1.66	-0.07	11.63	12947.74
SAN	1.01	0	20.87	-15.19	2.19	0.15	9.11	6515.50
AXA	0.55	0	19.78	-20.35	2.67	0.27	10.09	8790.79
BP	-1.35	0	10.58	-14.04	1.71	-0.13	7.81	4041.28
IRS	0.55	0.48	1.92	-1.86	0.21	-0.28	8.53	5367.17
GER BOND	1.11	0.97	3.39	-2.33	0.41	-0.09	5.97	1536.83
US BOND	0.98	0.96	4.53	-5.57	0.59	-0.22	7.96	4307.77
BRENT	0.98	0	17.97	-18.72	2.28	-0.19	8.26	4831.81
GAS	0.01	0	37.81	-28.90	4.19	0.56	12.81	16946.14
GOLD	3.10	0.01	6.86	-10.16	1.13	-0.41	8.81	5991.49
SILVER	2.26	0	13.66	-12.98	1.93	-0.57	8.62	5724.23
EUR/USD	0.16	0	4.62	-3.84	0.63	0.14	5.48	1091.11
GBP/USD	-0.20	0	4.43	-3.88	0.57	-0.04	7.27	3170.80
JPY/USD	-0.41	-0.99	4.61	-3.71	0.63	0.27	6.96	2779.74
AUD/USD	0.23	1.86	6.70	-8.83	0.83	-0.82	15.13	26058.43

Table 1: Descriptive statistics for daily percent returns. Mean and median returns are in basis points. SD denotes the standard deviation, and J-B is the Jarque-Bera statistic to test for Normality. Sample: 01/04/2000-12/31/2015.

I.2 PARAMETER ESTIMATES

	JSU-APARCH							
	IBEX	NASDAQ	FTSE	NIKKEI	IBM	SAN	AXA	BP
μ	0.00753	0.03217	-0.00822	0.01139	-0.00183	-0.00943	-0.00808	-0.02175
	(0.01702)	(0.01609)	(0.01212)	(0.01853)	(0.02042)	(0.02294)	(0.03988)	(0.03988)
ϕ_1	-0.00015	-0.04282	-0.04123	-0.02483	-0.02746	-0.01694	0.02013	-0.01595
	(0.01573)	(0.01512)	(0.01549)	(0.01550)	(0.01615)	(0.01555)	(0.01613)	(0.01613)
ω	0.02018	0.01725	0.01929	0.04149	0.01594	0.02766	0.02487	0.04140
	(0.00209)	(0.00167)	(0.00212)	(0.00744)	(0.03789)	(0.00780)	(0.01239)	(0.01239)
α_1	0.06005	0.05786	0.07008	0.07674	0.06992	0.06872	0.06402	0.06569
	(0.00447)	(0.00509)	(0.00565)	(0.00886)	(0.10079)	(0.01459)	(0.01344)	(0.01344)
γ_1	1.00000	0.99999	1.00000	0.61973	0.74415	0.63103	0.78804	0.60735
	(0.00017)	(0.00021)	(0.00014)	(0.10420)	(0.63005)	(0.16579)	(0.01731)	(0.19854)
β_1	0.93389	0.93402	0.92298	0.90868	0.93797	0.93138	0.93508	0.93308
	(0.00326)	(0.00435)	(0.00334)	(0.00970)	(0.10439)	(0.01543)	(0.19854)	(0.01731)
δ	1.13403	1.26183	1.10837	1.22986	0.98521	1.13434	1.11377	1.15349
	(0.11083)	(0.11560)	(0.10984)	(0.14580)	(0.72860)	(0.15649)	(0.25054)	(0.25054)
γ skewness	-0.43680	-0.48479	-0.77439	-0.32977	-0.06871	-0.24757	-0.21555	-0.13045
	(0.10768)	(0.10903)	(0.17566)	(0.09064)	(0.06413)	(0.07955)	(0.10166)	(0.10166)
δ kurtosis	2.35587	2.25426	2.69699	2.11659	1.55408	2.08511	2.39059	2.01485
	(0.18230)	(0.18749)	(0.26567)	(0.15516)	(0.06852)	(0.14027)	(0.18108)	(0.18108)
Ljung-Box Test on Standardized Residuals Lag[5]								
statistic	3.0140	3.3590	3.3990	1.8810	1.9822	2.3797	5.9844	4.7851
p-value	0.4278	0.3466	0.3378	0.7453	0.7165	0.6009	0.0472	0.1258
Ljung-Box Test on Standardized Squared Residuals Lag[9]								
statistic	13.1100	12.3300	4.436	6.6131	0.7612	13.1766	9.9960	2.1371
p-value	0.0101	0.0153	0.5165	0.2338	0.9941	0.0097	0.0503	0.8880
Log-Likelihoods								
FGARCH	-6810.914	-7137.549	-5686.932	-6970.236	-7032.462	-8289.825	-8907.877	-7473.873
APARCH	-6816.443	-7145.028	-5703.821	-6993.738	-7033.070	-8291.966	-8916.729	-7477.180
GJRGARCH	-6829.479	-7154.870	-5713.324	-7001.968	-7054.587	-8306.280	-8926.076	-7486.198
GARCH	-6899.894	-7213.013	-5796.144	-7031.025	-7068.898	-8352.110	-8965.944	-7510.162

Table 2: Parameter estimates of the APARCH model for stock market indices and individual stocks under an unbounded Johnson probability distribution for return innovations. Estimated parameters are as indicated in the models shown in Section 3. Standard deviations are reported in parentheses. The lower panel shows the log-likelihood values of the four volatility models considered in the paper.

I.3 VaR BACKTESTING: A SUMMARY

Model	Violations	LR_{uc}	LR_{ind}	LR_{cc}	DQT
N-GARCH	21	12	4	9	13
N-GJRGARCH	22	12	3	7	11
N-APARCH	22	13	3	9	14
N-FGARCH	23	13	2	8	14
ST-GARCH	16	4	3	6	8
ST-GJRGARCH	16	6	4	6	9
ST-APARCH	17	6	3	6	10
ST-FGARCH	16	7	3	5	8
SKST-GARCH	15	1	5	5	7
SKST-GJRGARCH	14	2	4	5	6
SKST-APARCH	16	2	3	4	7
SKST-FGARCH	16	1	4	4	7
SGED-GARCH	15	1	5	5	7
SGED-GJRGARCH	14	2	5	5	6
SGED-APARCH	15	3	4	4	7
SGED-FGARCH	15	2	3	4	8
JSU-GARCH	14	1	5	5	7
JSU-GJRGARCH	14	0	4	4	5
JSU-APARCH	15	1	4	4	6
JSU-FGARCH	15	1	4	4	7
SGT-GARCH	15	1	5	5	7
SGT-GJRGARCH	14	1	4	4	6
SGT-APARCH	15	2	4	4	7
SGT-FGARCH	16	2	3	4	8
GHST-GARCH	10	3	5	5	7
GHST-GJRGARCH	8	7	6	6	8
GHST-APARCH	12	6	4	5	8
GHST-FGARCH	12	5	2	3	7

Table 3: Summary statistics for $VaR_{1\%}$ backtesting under different models. The table shows the median number of VaR violations (number of exceedances of $VaR_{1\%}$ among 1260 observations) and the number of rejections of each test at 95% confidence level, over the set of 19 assets. The tests used were: the unconditional coverage test (LR_{uc}), independence (LR_{ind}) and conditional coverage (LR_{cc}) tests and the dynamic quantile test (DQT).

I.4 BACKTESTING RESULTS AT DIFFERENT ESTIMATION FREQUENCIES

Model	Violations		LR_{uc}		LR_{ind}		LR_{cc}		DQT	
	1 day	50 days	1 day	50 days	1 day	50 days	1 day	50 days	1 day	50 days
N-GARCH	30	30	0.00	0.00	0.19	0.19	0.00	0.00	0.00	0.00
N-GJRGARCH	**25**	**24**	0.00	0.00	**0.50**	**0.46**	0.01	0.01	**0.00**	**0.01**
N-APARCH	22	22	0.02	0.02	0.39	0.39	0.04	0.04	0.02	0.02
N-FGARCH	**25**	**24**	0.00	0.00	**0.50**	**0.40**	0.01	0.01	0.01	0.01
ST-GARCH	**21**	**22**	**0.03**	**0.02**	**0.35**	**0.39**	**0.06**	**0.04**	0.01	0.01
ST-GJRGARCH	21	21	0.03	0.03	0.35	0.35	0.06	0.06	0.02	0.02
ST-APARCH	20	20	0.05	0.05	0.32	0.32	0.09	0.09	0.03	0.03
ST-FGARCH	**20**	**19**	**0.05**	**0.09**	**0.32**	**0.28**	**0.09**	**0.14**	**0.05**	**0.06**
SKST-GARCH	20	20	0.05	0.05	0.32	0.32	0.09	0.09	0.01	0.01
SKST-GJRGARCH	16	16	0.36	0.36	0.19	0.19	0.28	0.28	0.12	0.12
SKST-APARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.12	0.12
SKST-FGARCH	17	17	0.24	0.24	0.22	0.22	0.23	0.23	0.04	0.04
SGED-GARCH	16	16	0.36	0.36	0.19	0.19	0.28	0.28	0.02	0.02
SGED-GJRGARCH	14	14	0.70	0.70	0.14	0.14	0.31	0.31	0.10	0.10
SGED-APARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.12	0.12
SGED-FGARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.02	0.02
JSU-GARCH	**19**	**18**	**0.09**	**0.15**	**0.28**	**0.25**	**0.14**	**0.18**	**0.01**	**0.02**
JSU-GJRGARCH	**15**	**14**	**0.51**	**0.70**	**0.17**	**0.14**	0.31	0.31	**0.11**	**0.10**
JSU-APARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.12	0.12
JSU-FGARCH	17	17	0.24	0.24	0.22	0.22	0.23	0.23	0.04	0.04
SGT-GARCH	**19**	**18**	**0.09**	**0.15**	**0.28**	**0.25**	**0.14**	**0.18**	**0.02**	**0.03**
SGT-GJRGARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.12	0.12
SGT-APARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.12	0.12
SGT-FGARCH	15	15	0.51	0.51	0.17	0.17	0.31	0.31	0.03	0.03
GHST-GARCH	9	9	0.28	0.28	0.05	0.05	0.08	0.08	**0.00**	**0.01**
GHST-GJRGARCH	**6**	**7**	**0.04**	**0.03**	()	**0.03**	()	**0.02**	**0.01**	**0.00**
GHST-APARCH	6	6	0.04	0.04	0.02	0.02	0.01	0.01	0.00	0.00
GHST-FGARCH	**16**	**15**	**0.36**	**0.51**	**0.19**	**0.17**	**0.28**	**0.31**	**0.11**	**0.02**

Table 4: Number of violations of $VaR_{1\%}$ among the 1260 sample observations, and test p-values for $VaR_{1\%}$ under different models for NASDAQ 100, when model parameters are re-estimated daily and every 50 days. Results are shown for the unconditional coverage test (LR_{uc}), independence (LR_{ind}) and conditional coverage (LR_{cc}) tests and the dynamic quantile test (DQT). Bold figures show the differences between both estimates.

I.5 SWITCHING BETWEEN MODELS

	LR_{uc}		LR_{ind}		LR_{cc}		DQT		**TOTAL**	
Total number of statistics	**76**		**32**		**32**		**76**		**216**	
Increases/Decreases	⇑	⇓	⇑	⇓	⇑	⇓	⇑	⇓	⇑	⇓
N→ST	64	12	8	24	30	2	58	18	160	56
ST→SKST	45	11	9	20	21	7	39	37	114	75
SKST→JSU	25	6	4	15	16	4	46	30	91	55
SKST→SGT	14	15	6	12	16	2	49	27	85	56
SKST→GHST	33	37	9	19	11	16	32	44	85	116
SKST→SGED	17	16	5	15	14	6	47	29	83	66
SGED→JSU	28	13	8	9	9	8	41	35	86	65
SGED→SGT	6	11	7	2	6	3	52	24	71	40
SGED→GHST	29	38	9	16	8	17	29	47	75	118
JSU→SGT	10	30	12	6	9	9	43	33	74	78
JSU→GHST	22	43	8	17	8	18	25	51	63	129
SGT→GHST	29	29	6	16	2	20	29	47	66	112
Total number of statistics	**133**		**56**		**56**		**133**		**378**	
Increases/Decreases	⇑	⇓	⇑	⇓	⇑	⇓	⇑	⇓	⇑	⇓
GARCH→GJRGARCH	46	56	36	19	35	21	59	74	176	170
GJRGARCH→APARCH	32	50	25	16	16	26	58	75	131	167
APARCH→FGARCH	34	44	21	13	18	16	78	55	151	128

Table 5: Number of cases in which the p-value of the test statistic increases or decreases when changing the probability distribution or the volatility model for all assets. For each test, the left (right) column shows the number of cases when the p-value increases (decreases) when switching between probability distributions (upper panel) or between volatility models (lower panel). The last two columns shows the results when aggregating results for the four tests. LR_{uc} denotes the unconditional coverage test of Kupiec and LR_{ind} and LR_{cc} are the independence and the conditional coverage tests of Christoffersen, respectively. DQT denotes the Dynamic Quantile test. Rows with bold figures show the number of times that each test is applied.

I.6 PRECEDENCE BETWEEN VaR MODELS

Confidence level 99%	LR_{uc}			LR_{ind}			LR_{cc}			DQT			TOTAL		
Total number of statistics	**76**			**32**			**32**			**76**			**216**		
D1 → D2	n1	n2	p	n1	n2	p	n1	n2	p	n1	n2	p	n1	n2	p
N → ST	36	7	1	7	7	0.714	25	13	1	44	29	1	112	56	0.964
ST → SKST	7	0	1	7	6	0.833	13	7	1	29	21	1	56	34	0.971
SKST → JSU	0	0	1	6	4	1	7	4	1	21	21	0.952	34	29	0.966
SKST → SGT	0	1	0	6	5	1	7	6	1	21	21	0.952	34	33	0.939
SGED → SKST	1	0	1	6	6	0.833	7	7	0.857	22	21	1	36	34	0.941
SGED → JSU	1	0	1	6	4	1	7	4	1	22	21	1	36	29	1
SGED → SGT	1	1	1	6	5	1	7	6	1	22	21	1	36	33	1
SGT → JSU	1	0	1	5	4	1	6	4	1	21	21	1	33	29	1
GHST → SKST	9	0	1	7	6	0.667	9	7	1	24	21	0.762	49	34	0.794
GHST → SGED	9	1	1	7	6	1	9	7	1	24	22	0.727	49	36	0.833
GHST → JSU	9	0	1	7	4	1	9	4	1	24	21	0.714	49	29	0.793
GHST → SGT	9	1	1	7	5	1	9	6	1	24	21	0.714	49	33	0.818
Total number of statistics	**133**			**56**			**56**			**133**			**378**		
M1 → M2	n1	n2	p	n1	n2	p	n1	n2	p	n1	n2	p	n1	n2	p
GARCH → GJRGARCH	10	12	0.833	9	9	0.778	16	12	0.917	48	45	0.844	83	78	0.846
GJRGARCH → APARCH	12	14	0.714	9	13	0.615	12	23	0.609	45	46	0.739	78	96	0.688
APARCH → FGARCH	14	18	0.722	13	11	0.818	23	20	0.800	46	43	0.930	96	92	0.848

Confidence level 95%	LR_{uc}			LR_{ind}			LR_{cc}			DQT			TOTAL		
Total number statistics	**76**			**32**			**32**			**76**			**216**		
D1 → D2	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*
N → ST	50	23	0.826	13	13	0.769	32	23	1	52	35	1	147	94	0.926
ST → SKST	23	6	1	13	16	0.813	23	18	1	35	27	0.963	94	67	0.940
SKST → JSU	6	3	1	16	17	0.941	18	17	1	27	25	0.960	67	62	0.968
SKST → SGT	6	6	1	16	16	0.875	18	17	1	27	28	0.964	67	67	0.955
SGED → SKST	8	6	0.833	17	16	1	18	18	1	28	27	1	71	67	0.985
SGED → JSU	8	3	1	17	17	0.941	18	17	1	28	25	1	71	62	0.984
SGED → SGT	8	6	1	17	16	1	18	17	1	28	28	0.964	71	67	0.985
SGT → JSU	6	3	1	16	17	0.882	17	17	0.941	28	25	1	67	62	0.866
GHST → SKST	21	6	0.833	17	16	0.938	19	18	0.944	30	27	0.926	87	67	0.925
GHST → SGED	21	8	0.875	17	17	0.882	19	18	0.944	30	28	0.929	87	71	0.901
GHST → JSU	21	3	1	17	17	0.882	19	17	0.941	30	25	1	87	62	0.952
GHST → SGT	21	6	1	17	16	0.875	19	17	0.941	30	28	0.929	87	67	0.925
Total number statistics	**133**			**56**			**56**			**133**			**378**		
M1 → M2	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*	n1	n2	*p*
GARCH → GJRGARCH	23	30	0.700	32	30	0.867	40	37	0.865	56	51	0.863	151	148	0.831
GJRGARCH → APARCH	30	33	0.758	30	25	0.960	37	36	0.972	51	59	0.797	148	153	0.856
APARCH → FGARCH	33	31	0.968	25	22	0.955	36	32	1	59	59	0.915	153	144	0.951

Table 6: Precedence between VaR models. The upper panel shows results from tests implemented at 1% significance, while the lower panel shows results from tests at the 5% significance level. $n1$ is the number of tests in which H0 is rejected when D1 (M1) is specified as distribution (volatility model) for the return innovations of the different assets; $n2$ is the number of tests in which H0 is rejected when D2 (M2) is the probability distribution (volatility model) for the different assets and p is the proportion of times that H0 is rejected with both D2 (M2) and D1 (M1). Rows with bold figures show the total number of tests run.

I.7 MODEL CONFIDENCE SETS

	IBX	NSQ	FTS	NKE	IBM	SAN	AXA	BP	IRS	GEB
					Volatility Models					
GARCH	0	0	0	0	1	0	0	0	1	2
GJRGARCH	0	2	2	1	0	0	0	0	0	0
APARCH	2	5	5	0	0	0	0	2	0	0
FGARCH	3	4	3	2	0	1	1	0	0	3
					Probability Distributions					
N	0	0	0	0	0	0	0	1	0	0
ST	0	1	0	0	0	0	0	0	0	0
SKST	0	2	2	0	0	0	0	0	0	0
SGED	2	2	2	2	1	0	1	1	0	1
JSU	0	2	3	0	0	1	0	0	0	2
SGT	2	3	3	1	0	0	0	0	1	1
GHST	1	1	0	0	0	0	0	0	0	1
Total Number	**5**	**11**	**10**	**3**	**1**	**1**	**1**	**2**	**1**	**5**

	USB	BRE	GAS	GLD	SLV	EUR	GBP	JPY	AUD	TOT
					Volatility Models					
GARCH	6	0	0	1	4	2	0	3	0	**20**
GJRGARCH	0	0	0	0	1	2	1	0	0	**9**
APARCH	0	0	0	0	0	2	0	4	2	**22**
FGARCH	1	1	1	0	1	6	0	4	1	**32**
					Probability Distributions					
N	0	1	1	0	0	0	1	0	2	**6**
ST	1	0	0	0	0	2	0	3	1	**8**
SKST	1	0	0	0	1	1	0	0	0	**7**
SGED	1	0	0	0	0	1	0	3	0	**17**
JSU	1	0	0	0	2	1	0	0	0	**12**
SGT	1	0	0	0	1	4	0	3	0	**20**
GHST	2	0	0	1	2	3	0	2	0	**13**
Total Number	**7**	**1**	**1**	**1**	**6**	**12**	**1**	**11**	**3**	

Table 7: Number of times that each probability distribution and volatility model enter into the Superior Set of models for each asset: IBEX 35 (IBX), NASDAQ 100 (NSQ), FTSE 100 (FTS), NIKKEI 225 (NKE), IBM, SAN, AXA, BP, IRS 5Y (IRS), GERMAN BOND 10Y (GEB), US BOND 10Y (USB), CRUDE OIL BRENT (BRE), NATURAL GAS (GAS), GOLD (GLD), SILVER (SLV), EUR/USD (EUR), GBP/USD (GBP), JPY/USD (JPY) and AUD/USD (AUD). Bold figures in the last column TOT (TOTAL) of the lower panel are aggregates for each probability distribution or volatility model. Bold figures in the last row of each panel display aggregates for each asset.

I.8 10-day VaR BACKTESTING UNDER FHS

Model	Violations	LR_{uc}
N-GARCH	8	0.17
N-GJRGARCH	7	0.09
N-APARCH	7	0.09
N-FGARCH	7	0.09
ST-GARCH	9	0.29
ST-GJRGARCH	7	0.09
ST-APARCH	7	0.09
ST-FGARCH	7	0.09
SKST-GARCH	10	0.46
SKST-GJRGARCH	7	0.09
SKST-APARCH	7	0.09
SKST-FGARCH	7	0.09
SGED-GARCH	9	0.29
SGED-GJRGARCH	7	0.09
SGED-APARCH	7	0.09
SGED-FGARCH	7	0.09
JSU-GARCH	10	0.46
JSU-GJRGARCH	7	0.09
JSU-APARCH	7	0.09
JSU-FGARCH	7	0.09
SGT-GARCH	12	0.89
SGT-GJRGARCH	8	0.17
SGT-APARCH	7	0.09
SGT-FGARCH	6	0.04
GHST-GARCH	11	0.66
GHST-GJRGARCH	7	0.09
GHST-APARCH	7	0.09
GHST-FGARCH	4	0.00

Table 8: Number of violations of 10-day $VaR_{1\%}$ among 1250 sample observations and unconditional coverage test p-values for 10-day $VaR_{1\%}$ under different models for NASDAQ 100. VaR forecasts were obtained by Filtered Historical Simulation (FHS).

II FIGURES

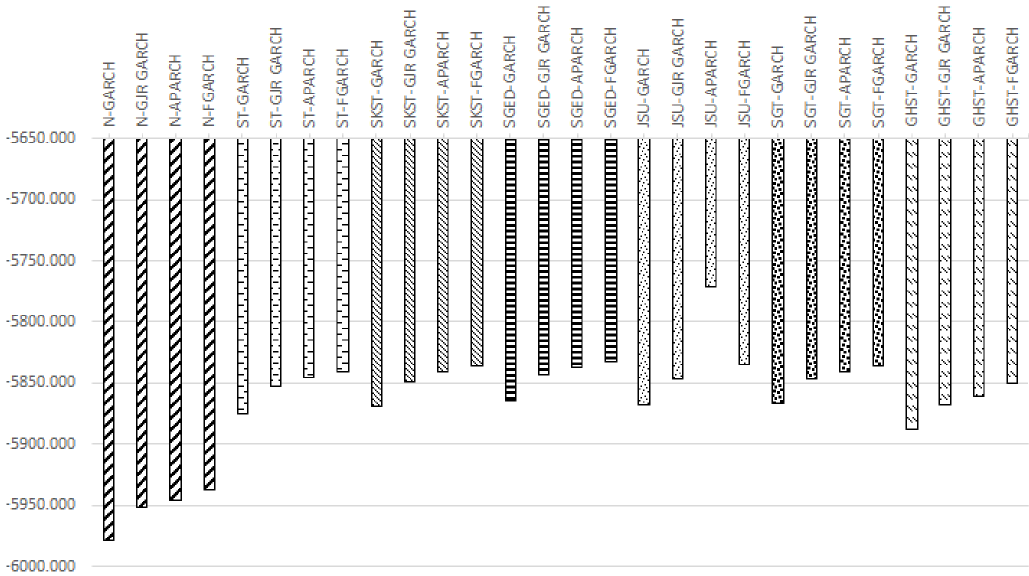

Figure 1: Mean log-likelihoods for each model over the set of 19 assets.

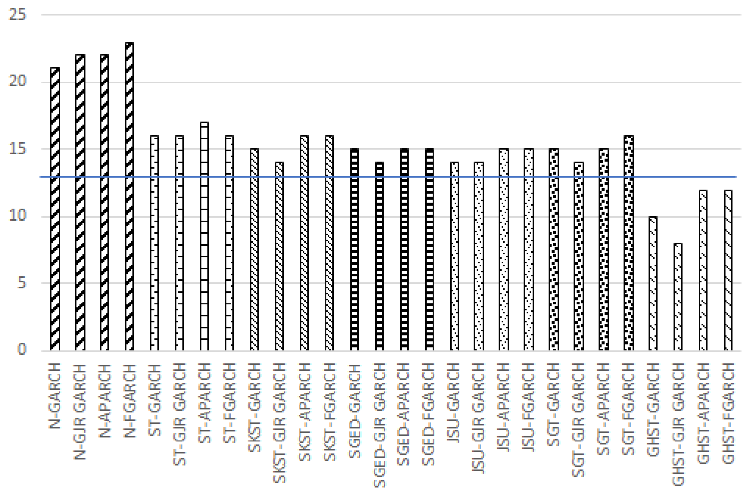

Figure 2: Median number of VaR violations for each model over the set of 19 assets.

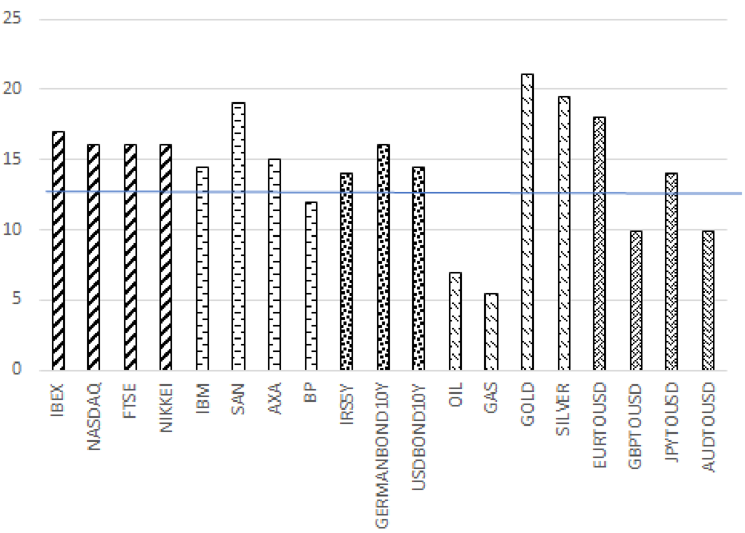

Figure 3: Median number of VaR violations for each asset over the set of 28 models.