Instituto Complutense de Análisis Económico

Backtesting Extreme Value Theory models of expected shortfall

Alfonso Novales
Instituto Complutense de Análisis Económico (ICAE), and
Department of Economic Analysis, Facultad de Ciencias Económicas y Empresariales,
Universidad Complutense, 28223 Madrid, Spain

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

Abstract

We use stock market data to analyze the quality of alternative models and procedures for fore- casting expected shortfall (ES) at different significance levels. We compute ES forecasts from conditional models applied to the full distribution of returns as well as from models that focus on tail events using extreme value theory (EVT). We also apply the semiparametric filtered historical simulation (FHS) approach to ES forecasting to obtain 10-day ES forecasts. At the 10-day hori- zon we also combine FHS with EVT. The performance of the different models is assessed using six different ES backtests recently proposed in the literature. Our results suggest that conditional EVT-based models produce more accurate 1-day and 10-day ES forecasts than do non-EVT based models. Under either approach, asymmetric probability distributions for return innovations tend to produce better forecasts. Incorporating EVT in parametric or semiparametric approaches also improves ES forecasting performance. These qualitative results are also valid for the recent crisis period, even though all models then underestimate the level of risk. FHS narrows the range of numerical forecasts obtained from alternative models, thereby reducing model risk. Combining EVT and FHS seems to be best approach for obtaining accurate ES forecasts

Keywords extreme value theory; skewed distributions; expected shortfall; backtesting; filtered historical simulation

Working Paper nº 1924
September, 2019

UNIVERSIDAD COMPLUTENSE MADRID

ISSN: **2341-2356**

Backtesting Extreme Value Theory models of expected shortfall

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

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

August 2018

Abstract

We use stock market data to analyze the quality of alternative models and procedures for forecasting expected shortfall (ES) at different significance levels. We compute ES forecasts from conditional models applied to the full distribution of returns as well as from models that focus on tail events using extreme value theory (EVT). We also apply the semiparametric filtered historical simulation (FHS) approach to ES forecasting to obtain 10-day ES forecasts. At the 10-day horizon we also combine FHS with EVT. The performance of the different models is assessed using six different ES backtests recently proposed in the literature. Our results suggest that conditional EVT-based models produce more accurate 1-day and 10-day ES forecasts than do non-EVT based models. Under either approach, asymmetric probability distributions for return innovations tend to produce better forecasts. Incorporating EVT in parametric or semiparametric approaches also improves ES forecasting performance. These qualitative results are also valid for the recent crisis period, even though all models then underestimate the level of risk. FHS narrows the range of numerical forecasts obtained from alternative models, thereby reducing model risk. Combining EVT and FHS seems to be best approach for obtaining accurate ES forecasts.

Keywords: extreme value theory; skewed distributions; expected shortfall; backtesting; filtered historical simulation

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

1 Introduction

The Basel Committee on Banking Supervision has recently chosen expected shortfall (ES) as the market risk measure to be used for banking regulation purposes, replacing value at risk (VaR). The change is motivated by the superior properties of ES as a measure of risk, since it is based on information taken from the full tail of the distribution of returns. However, in spite of its advantages as a measure of risk, ES is still less used than VaR. The main drawback with the use of ES for risk regulation is the unavailability of simple tools for the evaluation of ES forecasts. The reason is that backtesting ES is much harder than backtesting VaR, which is usually done by comparing whether the observed percentage of outcomes covered by the risk measure is consistent with the intended level of coverage. That difficulty led the Basel Committee to reconsider requiring the backtesting of ES, and its consultative document Basel Committee (2016) proposed calculating risk and capital using ES, but conducting backtesting only on VaR. However, it is important that the capital reserves indicated by the VaR calculation can be tested, and the adequacy of the level of reserves should be subject to a valid statistical test. The current emphasis of the Basel Committee on ES makes clear that ES backtesting certainly will be on the future agenda for capital requirements at financial institutions. The goal of this paper is precisely to advance in the application of alternative approaches to ES backtesting under different modeling choices and comparing the performance of the different out-of-sample ES forecasts.

There is not much work evaluating and comparing the performance of ES forecasting models. Taylor (2007) proposes exponentially weighted quantile regression (EWQR) to estimate VaR and ES. He considered 1-day forecasting of conditional quantiles and their associated ES at different significance levels. ES estimates are evaluated employing an approach similar to that of McNeil and Frey (2000) to conclude that the best performance for ES estimation was achieved by the EWQR. Alexander and Sheedy (2008) develop a two-stage methodology for conducting stress tests in which an initial shock event is linked to the probability of its occurrence. Working with three pairs of major currencies they found their results compared favorably with the traditional historical scenario stress testing approach. Jalal and Rockinger (2008) use a circular block bootstrap to take adequately into account the possible dependencies among exceedances. Applying the two-step procedure of McNeil and Frey (2000), they found that ES forecasts captured actual shortfalls satisfactorily. Ergün and Jun (2010) show that the autoregressive conditional density model of Hansen (1994) with a time-varying conditional skewness parameter seems to provide better ES forecasts, beating forecasts based on other GARCH-based models as well as those based on the EVT approach. Kourouma et al. (2011) introduce a validation test for ES that they employ to compare unconditional and conditional ES forecasting models at 1-day, 5-day and 10-day horizons using as the conditional model a GJR-GARCH with normal distributed return innovations. Wong et al. (2012) compare conditional models with GARCH and APARCH volatility specifications and normal, Student-t, and skew Student-t distributions with an EVT model under normality using the saddlepoint backtest proposed by Wong (2008, 2010). Gerlach and Chen (2014) extend the standard daily return-based conditional autorregresive expectile model class to incorporate the intraday range as an explanatory variable, in several model specifications. The 1-day ES forecasts are assessed with the ESRate criterion, defined as the proportion of observations for which the actual return is greater than the predicted ES level. Righi and Ceretta (2015) evaluate unconditional, conditional, and quantile/expectile regression-based models for ES forecasting using the ES backtest proposed by McNeil and Frey (2000) as well as the Righi and Ceretta (2013) test,

which is based on the truncated distribution of returns beyond VaR. Clift, Costanzino, and Curran (2016) apply the three approaches to ES backtesting recently proposed by Wong (2008), Acerbi and Szekely (2014), and Costanzino and Curran (2015). For ES forecasting they only consider a constant volatility model and a GARCH volatility specification under normality. These papers provide evidence of the gains that can be achieved by using asymmetric probability distributions and EVT for ES forecasting.[1]

Risk analysis has rarely been implemented beyond a 1-day horizon when forecasting the ES of financial assets, even though risk horizons longer than one day are particularly important for risk liquidity management, for long term strategic asset allocation as well as to compute capital requirements. Moreover, the Basel Committee obliges banks to compute their risk levels at a 10-day horizon. The difficulty in doing this is obtaining enough homogeneous data on 10-day returns over non-overlapping periods. This explains the extended use of the scaling law, whose use is also proposed in the Basel Committee supervision documents, even though it is known that it may lead to severe biases in many realistic situations. We get around this limitation by using filtered historical simulation (FHS) to obtain time series for 10-day returns. From them, we estimate 10-day VaR and ES by applying the same methodologies as for 1-day ES forecasting.

We want to make some progress in the comparison of different relevant approaches to VaR and ES forecasting. Using stock market data, we take into account volatility clustering and leverage effects in return volatility by using the APARCH model [Ding, Granger, and Engle (1993)] under different probability distributions for the standardized innovations: Gaussian, Student-t, skewed Student-t [Fernandez and Steel(1998)], skewed generalized error [Fernandez and Steel (1998)], and Johnson S_U [Johnson (1949)]. Some existing methodologies for ES validation [McNeil and Frey (2000), Berkowitz (2001), Kerkhof and Melenberg (2004), and Wong (2008)] have been shown to be subject to a variety of limitations, so we apply some recently proposed approaches to ES backtesting that overcome such limitations: the tests of Righi and Ceretta (2013), the first two tests of Acerbi and Szekely (2014), the test of Graham and Pál (2014), the quantile-space unconditional coverage test of Costanzino and Curran (2015), and the conditional test of Du and Escanciano (2016). We provide a detailed description of these ES backtests in Section 4.2.

We use standard parametric methods for the full distribution of returns and also methods that focus on the tail of this distribution. We generate 1-day VaR and ES forecasts following two approaches. On the one hand, we use parametric expressions that are well-known in the risk literature. Alternatively, we employ the semiparametric FHS approach that combines Monte Carlo simulation and bootstrapping techniques. In both cases we apply these methods by themselves but we also combine them with an extreme value theory (EVT) approach for modeling the tail of the distribution of returns. To apply a parametric version of EVT we follow the approach by McNeil and Frey (2000), whereas under the semiparametric approach we implement EVT following the proposal of Daníelsson and de Vries (2000) that we describe below. For 10-day forecasts, we use the FHS approach by itself, but also combine it with EVT modeling of the tail of the return distribution.

Some alternatives have been recently proposed to estimate multi-period value at risk and expected shortfall in a GARCH framework with leverage and asymmetric probability distributions.

[1] In other studies VaR is the primary measure of interest, with ES left as a secondary consideration. Examples are Zhou (2012), Degiannakis, Floros, and Dent (2013), and Tolikas (2014), where not much focus is placed on ES forecasting patterns.

So and Wong (2012) propose analytical estimates with a better performance than alternatives like RiskMetrics or the application of the square root rule for the variance. Degiannakis et al. (2014) show the interest of considering integrated variance models to estimate multi-period risk measures. Lönnbark (2016) has proposed a Monte Carlo simulation approach to estimate multiple period value at risk and expected shortfall that compares favorably with analytical approximation alternatives to the distribution of multi-period returns. Comparing the performance of such alternative approaches with the one we follow remains as an interesting question for future research.

We contribute to the literature in four ways:

1. First, we evaluate the improvement that can be achieved by incorporating FHS in standard parametric forecasting. We compare the performance of the parametric approach to ES forecasting with the semiparametric filtered historical simulation approach. This is especially important for VaR and ES forecasting at horizons more distant than 1 day. In both analyses we compare the performance of VaR and ES estimates obtained under normality or Student-t assumptions with those obtained under some asymmetric probability distributions for return innovations that are relatively new to the financial literature.

2. Second, in the parametric and semiparametric analysis we apply the five different approaches for ES backtesting described above.[2]

3. Third, we estimate and test VaR and ES forecasting models at a 10-day horizon, an analysis that has seldom been considered in the financial literature.

4. Fourth, we examine the accuracy of risk models for VaR and ES forecasting at 1-day and 10-day horizons during the pre-crisis and crisis periods for different significance levels.

To the best of our knowledge, this is the first time that a systematic test of ES forecasting models has been made that considers a variety of probability distributions and two alternatives to the standard parametric approach, such as EVT and the semiparametric FHS, with a detailed attention to 1-day and 10-day VaR and ES forecasts as well as to the results pre-crisis and during the crisis.

The remainder of the paper is organized as follows. In Section 2 we describe the standard risk measures and their mathematical properties. In Section 3 we show the data and estimation models. We describe the parametric and semiparametric procedures we have followed to estimate the different models and generate VaR and ES forecasts. In Section 4 we review the VaR and ES backtesting approaches we follow in this paper. In Section 5 we report the results of the out-of-sample 1-day ES forecasting exercise in terms of the tests available for ES performance. In Section 6 we describe a robustness analysis examining two issues: the different results obtained for pre-crisis and crisis periods separately, and an assessment of 10-day ES forecasting. Section 7 describes our view on the use of VaR and ES for capital adequacy, and Section 7 concludes the paper.

[2] An overview of ES backtesting procedures used in recent literature can be seen in Table A1 in the online appendix.

2 Background: Standard Risk Measures and their properties

Value at risk (VaR) is a simple risk indicator that measures what loss will be exceeded only a small percentage of times in the next k trading days ($100\alpha\%$). We define VaR as a quantile of the profit/loss distribution for a given horizon and a given shortfall probability, reporting VaR as a negative number. Thus, given the log-return $r_{t,t+k}$ of a portfolio between t and $t+k$, VaR at a level α is defined by $Pr(r_{t,t+k} < VaR^{\alpha}_{t+k}) = \alpha$. For simplicity, we assume that we are predicting the VaR at some level α for 1-day returns, $r_{t,t+1} = \mu_{t+1} + \sigma_{t+1} z_{t+1}$, where μ_{t+1} is the conditional mean return in period $t+1$, σ^2_{t+1} is the conditional variance, and z_{t+1} represents the white noise time series of return innovations, which will follow a given probability distribution F. From the VaR^{α}_{t+1} definition we have, $Pr(z_{t+1} < (VaR^{\alpha}_{t+1} - \mu_{t+1})/\sigma_{t+1}) = \alpha$, which amounts to, $F((VaR^{\alpha}_{t+1} - \mu_{t+1})/\sigma_{t+1}) = \alpha$, or

$$VaR^{\alpha}_{t+1} = \mu_{t+1} + \sigma_{t+1} F^{-1}(\alpha). \tag{1}$$

Given the drawbacks of VaR as a risk measure, it is convenient to compute the ES, which accounts for the magnitudes of large losses as well as the probabilities that they occur. The ES is defined from VaR as $ES^{\alpha}_{t+k} = \mathbb{E}_{t+k}[r_{t,t+k} | r_{t,t+k} < VaR^{\alpha}_{t+k}]$ and tells us the expected value of the loss k days ahead, conditional on it being worse than the VaR. As in the case of VaR, we define ES in terms of low quantile values of profit/loss distribution, leading to a negative ES estimate for low α values. For the 1-day ES we have, $ES^{\alpha}_{t+1} = \mathbb{E}_{t+1}[r_{t,t+1} | r_{t,t+1} < VaR^{\alpha}_{t+1}] = \mu_{t+1} + \sigma_{t+1}\mathbb{E}_{t+1}[z_{t+1} | z_{t+1} < (VaR^{\alpha}_{t+1} - \mu_{t+1})/\sigma_{t+1}]$. Finally, using (1) we get

$$ES^{\alpha}_{t+1} = \mu_{t+1} + \sigma_{t+1}\mathbb{E}_{t+1}[z_{t+1} | z_{t+1} < F^{-1}(\alpha)]. \tag{2}$$

If we assume the existence of an absolutely continuous cdf F, ES is defined as

$$\mathbb{E}_{t+1}[z_{t+1} | z_{t+1} < F^{-1}(\alpha)] = \frac{1}{\alpha}\int_0^{\alpha} F^{-1}(s)ds = \frac{1}{\alpha}\int_{-\infty}^{F^{-1}(\alpha)} r f(r) dr.$$

The subadditivity property fails to hold for VaR in general, so VaR is not a coherent measure.[3] Indeed, examples [see e.g. Embrechts et al. (2009)] can be given where VaR is superadditive, i.e. $VaR^{\alpha}(\sum_{i=1}^{n} Y_i) < \sum_{i=1}^{n} VaR^{\alpha}(Y_i)$. Whether or not VaR is subadditive depends on the properties of the joint loss distributions. The lack of subadditivity contradicts the notion that there should be a diversification benefit associated with merging portfolios. As a consequence, a decentralization of risk management using VaR is difficult since we cannot be sure that by aggregating VaR numbers for different portfolios or business units we will obtain a bound for the overall risk of the enterprise. Subadditivity of VaR requires strong assumptions like a joint elliptical distribution among returns or an Archimedean survival dependence structure, which are often inconsistent with the properties of actual data [Embrechts et al. (2002, 2009)]. Using majorization theory, Ibragimov and Walden (2007), Ibragimov (2009) demonstrated that the VaR measure is subadditive for the infinite variance stable distributions provided the mean return is finite, and Garcia et al. (2007) extended the result to general Pareto distributions. Daníelsson et al. (2005), Ibragimov (2005) and

[3] Artzner et al. (1999) state four axioms which any risk measure used for effective risk regulation and management should satisfy: i) homogeneity, ii) subadditivity, iii) monotonicity, and iv) translation invariance. Such risk measures are said to be coherent.

Garcia et al. (2007) also discuss cases of VaR subadditivity for distributions with Pareto type tails when the variance is finite. Daníelsson et al. (2013) identify sufficient conditions for VaR to be subadditive in the relevant tail region for fat-tailed and dependent distributions.[4]

An additional limitation is that VaR at the level α gives no information about the severity of tail losses which occur with a probability less than α, in contrast to ES at the same confidence level. When looking at aggregated risks $\sum_{i=1}^{n} Y_i$, it is well known [Acerbi and Tasche (2002)] that ES is a coherent risk measure. In particular, in contrast to VaR, ES is generally subadditive. A potential deficiency of ES when compared with the VaR approach to risk measurement refers to forecasting and backtesting ES. Gneiting (2011) showed that ES is not elicitable. He proved that the existence of convex level sets is a necessary condition for the elicitability of a risk measure and disproved their existence for ES. That means that it is not possible to find a scoring function $s(x,y)$ such that the ES forecast x of the true ES y can be obtained as the x that minimizes $s(x,y)$ (see Gneiting (2011) and Emmer et al. (2015) among others). Many authors have interpreted Gneiting's findings as evidence that it is not possible to backtest ES at all [see, for instance, Carver (2013)], in spite of the fact that successful attempts of backtesting ES had been made before 2011.[5] The paper by Gneiting changed the focus of the discussion from how ES could be backtested to the question of whether it was even possible to do so. Not everybody has interpreted Gneiting's findings as evidence that ES is not backtestable. Following Gneiting's findings, Emmer et al. (2013), showed that ES is conditionally elicitable for continuous distributions with finite means. In the same paper, these authors also made a careful comparison of different measures and their mathematical properties. They concluded that ES is the most appropriate risk measure even though it is not elicitable. A similar discussion of the implications of different risk measures and their effect on regulation can be found in Chen (2014).

That point was settled recently by Fissler, Ziegel, and Gneiting (2015) and by Acerbi and Szekely (2014), who demonstrated that lack of elicitability is not an impediment to backtesting ES. ES cannot be backtested through any scoring function but there is no reason why this could not be done using another method that does not exploit the property of elicitability.

3 Data and Estimation Models

We work with daily percentage returns on assets over the sample period 10/2/2000 - 9/30/2016 (4175 sample observations). Daily returns are computed as 100 times the difference of log prices, i.e. $100[\ln(P_{t+1}) - \ln(P_t)]\%$. The financial assets considered are International Business Machines [IBM] (\$), Banco Santander [SAN] (€), AXA [AXA] (€), and BP [BP] (£). The data were extracted from Datastream. Table 1 reports descriptive statistics for the daily percentage return series. All of them have a mean close to zero. Median returns are zero. SAN has the largest total range $(max - min)$ and BP has the smallest range. The unconditional standard deviation (S.D.) is around 2, with AXA having the highest and IBM the lowest S.D. All assets have negative skewness, except AXA. For all assets considered, the kurtosis is high, implying that the distributions of these returns have tails much thicker than does the normal distribution. Accordingly, the Jarque-Bera statistic (J-B) is statistically significant, rejecting the assumption of normality in all cases.

[4] Specifically, they show that VaR is subadditive in the relevant tail region when asset returns exhibit multivariate regular variation, for both independent and cross sectionally dependent returns, provided the mean is finite.

[5] For example, Kerkhof and Melenberg (2004) found methods that performed better than comparable VaR backtests.

Along the paper we work with conditional models for forecasting VaR and ES. We will consider an APARCH specification for volatility [Ding, Granger, and Engle (1993)], which is not too restrictive since it includes as special cases some of the most standard conditional volatility models. Garcia-Jorcano and Novales (2017) show that APARCH volatility fits the data for a variety of assets better than alternative models nested in APARCH. The success in capturing the heteroscedasticity exhibited by the data may be due to the increased flexibility of the APARCH model in dealing with the power on the conditional standard deviation as a free parameter. These authors also present evidence suggesting that estimates of risk measures are much more sensitive to the choice of probability distribution than to the choice of volatility model. An AR(1) model was considered for the conditional mean return, which was enough to produce serially uncorrelated innovations.

For a given return series $r_1, ..., r_T$, we estimate the AR(1)-APARCH(1,1) model

$$r_t = \phi_0 + \phi_1 r_{t-1} + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t, \qquad t = 1, 2, ...T,$$

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

where ω, α_i, γ_i, β_j, and δ are parameters of the volatility model to be estimated. The parameter γ_i reflects the leverage effect $(-1 < \gamma_i < 1)$. A positive (resp. negative) value of γ_i 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 $\sigma_t (\delta > 0)$.

We jointly estimate by maximum likelihood the parameters in the mean return equation, the equation for the conditional standard deviation, and the probability distribution for return innovations. As distributions, we alternatively consider the Gaussian, Student-t, skewed Student-t, skewed generalized error, and Johnson S_U distributions and work with the residuals of the model.[6] In addition, through the usual diagnostics performed on the standardized residuals and their squared values, we assessed that returns are properly filtered.[7]

3.1 A parametric approach to VaR and ES estimation

Under a conditional volatility model like the one above, the risk measures become

$$VaR_t^\alpha = \mu_t + \sigma_t F^{-1}(\alpha),$$

$$ES_t^\alpha = \mu_t + \sigma_t \left(\frac{1}{\alpha} \int_0^\alpha F^{-1}(s) ds \right),$$

[6] We provide a description of asymmetric probability distributions in Appendices I.1 - I.3. All computations were performed with the R software (version 3.1.1) package *rugarch* (version 1.3-4) designed for the estimation and forecast of various univariate ARCH-type models.

[7] An alternative leptokurtic and asymmetric distribution that has been considered in this context is the skewed generalized-t (SGT) distribution proposed by Theodossiou (1998). The SGT distribution has the attractive feature of encompassing most of the distributions that are usually assumed for standardized returns, such as the Gaussian, generalized error distribution (GED), Student-t and skewed Student-t distributions. Recently, Ergen (2015) has considered the skewed-t distribution proposed by Azzalini and Capitanio (2003) and Aas and Haff (2006) propose the use of the generalized hyperbolic skew Student-t distribution for unconditional and conditional VaR forecasting.

$$SD_t^{\alpha} = \left[\sigma_t^2 \frac{1}{\alpha} \int_0^{\alpha} \left(F^{-1}(s) - \left(\frac{1}{\alpha} \int_0^{\alpha} F^{-1}(s) ds \right) \right)^2 ds \right]^{1/2}.$$

The SD measure in the last expression is the dispersion around the expected value truncated by the VaR. It will play an important role in the ES backtesting approach of Righi and Ceretta that we describe below.

Another possibility is to estimate the conditional quantile using the EVT approach. EVT is concerned with the distribution of the smallest order statistics and it considers only the tail of the distribution of returns without making any specific assumption concerning the center of the distribution [Rocco (2014)]. For more details, see Longin (2005). Although EVT is interesting for risk modeling, the stylized facts make the i.i.d. assumption inappropriate for most financial data. To address this issue, Daníelsson and de Vries (2000) and McNeil and Frey (2000) suggest applying the EVT analysis to the filtered standardized residuals z_t from a previously estimated model, as proposed by Diebold, Schuermann, and Stroughair (2000). This is possible because under a correct specification of the conditional mean and variance, the filtered residuals will be approximately i.i.d., an assumption of EVT modeling.

The tail index parameter in EVT can be estimated nonparametrically without assuming any particular model for the tail with the Hill estimator [Hill (1975)] and Pickands estimator [Pickands (1975)]. Tail parameters in EVT can also be estimated from two parametric approaches based on classical methods such as maximum likelihood. In one, block maxima (BM), the sample is split into m subsamples of n observations. The maximum values of each subsample, when properly normalized, converge to a generalized extreme value (GEV) distribution [see for example Longin (2005) and Diebold, Schuermann, and Stroughair (2000)]. In the paper we use an alternative EVT parametric approach, peaks-over-threshold (POT), which is based on the generalized Pareto distribution (GPD). The GPD distribution can be seen as the limiting tail distribution for a wide variety of commonly studied continuous distributions. Under this method, any observations that exceed a given high threshold, u, are modeled separately from non-extreme observations. McNeil and Frey (2000) show that the EVT method based on the GPD yields quantile estimates that are more stable than those obtained using the Hill estimator. The weakness of this approach is the lack of objective information for the choice of threshold, which affects the numerical values and the properties of the implied quantile estimates.

POT is the typical approach used in finance for parametric EVT estimation. It essentially consists in fitting a GPD to the innovations obtained from filtering returns using an estimated conditional volatility model. Under the i.i.d. assumption, we consider the distribution function of excesses $Y = u - Z$ over a high, fixed threshold u, $F_u(y) = P(Y = u - Z \leq y | Z < u) = [F(u) - F(u - y)]/[F(u)]$, $y \geq 0$.[8] Pickands (1975) shows that the generalized Pareto distribution (GPD) arises naturally as the limit distribution of the scaled excesses of identical and independently distributed (i.i.d.) random variables over high thresholds. We say that excesses from a given threshold follow a generalized Pareto distribution $Y = u - Z \sim GPD(\xi, \beta)$ if

$$F_u(y) \approx GPD_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta} \right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\frac{y}{\beta} \right), & \xi = 0. \end{cases}$$

[8] Note that we focus on the lower tail of the data, and we have adapted all the formulations accordingly.

$GPD_{\xi,\beta}(y)$ has support $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\beta/\xi$ if $\xi < 0$, where $\beta > 0$ is a scale parameter and ξ is the tail shape parameter, which is crucial because it governs the tail behavior of $GPD_{\xi,\beta}(y)$. The case $\xi > 0$ corresponds with heavy-tailed distributions whose tails decay like power functions, such as the Pareto, Student-t, Cauchy, Burr, log-gamma and Fréchet distributions. For example, in this case, the tail index parameter equal to $1/\xi$ corresponds to the degrees of freedom of the Student-t distribution. The case $\xi = 0$ corresponds with distributions such as the normal, exponential, gamma, and lognormal distributions, whose tails essentially decay exponentially. The final group of distributions are short-tailed distributions ($\xi < 0$) with a finite right endpoint, such as the uniform and beta distributions.[9]

Consider now the following equality for points $z < u$ in the left tail of F:

$$F(z) = F(u) - F_u(u-z)F(u) = F(u)(1 - F_u(u-z)).$$

If we estimate the first term on the right-hand side of the equation using the proportion of tail data T_u/T, and if we estimate the second term by approximating the excess distribution with a generalized Pareto distribution fitted by maximum likelihood, we get the tail estimator

$$\widehat{F}(z) = \frac{T_u}{T}\left(1 + \hat{\xi}\frac{u-z}{\hat{\beta}}\right)^{-1/\hat{\xi}}.$$

It is very important to note that the distribution F of the conditional model and the distribution $GPD_{\xi,\beta}$ for the excesses over threshold, $\{y\}$, are not linked. Thus, it is possible to use any conditional model to filter the data before applying EVT to z. In our analysis we assume a variety of asymmetric distributions for F that give rise to different conditional EVT estimates. Under the EVT approach, the risk measures obtained are

$$VaR_t^\alpha = \mu_t + \sigma_t F_z^{-1}(\alpha) = \mu_t + \sigma_t\left(u + \frac{\beta}{\xi}\left[1 - \left(\frac{\alpha}{T_u/T}\right)^{-\xi}\right]\right),$$

$$ES_t^\alpha = \mu_t + \sigma_t\left(\frac{1}{\alpha}\int_0^\alpha F_z^{-1}(s)ds\right) = \mu_t + \sigma_t\left(\frac{VaR_t^\alpha}{1-\xi} - \left(\frac{\beta + \xi u}{1-\xi}\right)\right),$$

$$SD_t^\alpha = \left[\sigma_t^2\frac{1}{\alpha}\int_0^\alpha\left(F_z^{-1}(s) - \left(\frac{1}{\alpha}\int_0^\alpha F_z^{-1}(s)ds\right)\right)^2 ds\right]^{1/2}.$$

To summarize, McNeil and Frey proceed as follows. In the first step, they filter the dependence in the time series of returns by computing the residuals of a GARCH-type model, which should be i.i.d. if the GARCH-type model correctly fits the data. In the second step, they model the extreme behavior of the residuals using the tail approach explained above. Finally, in order to produce a VaR forecast of original returns, they trace back the steps by first producing the α-quantile forecast

[9]The implied assumption is that the tail of the underlying distribution begins at the threshold u. From our sample of T data a random number of observations, T_u, will exceed this threshold. If we assume that the T_u excesses over the threshold are i.i.d. with an exact GPD distribution, Smith (1987) has shown that the maximum likelihood estimates $\hat{\xi} = \hat{\xi}_N$ and $\hat{\beta} = \hat{\beta}_N$ of the GPD parameters ξ and β are consistent and asymptotically normal as $T_u \to \infty$, provided $\xi > -1/2$. Under the weaker assumption that the excesses are i.i.d. from a $F_u(y)$ which is only approximately GPD he also obtains asymptotic normality results for ξ and β.

for the GARCH-type filtered residuals and transforming the α-quantile forecast for the original returns using the conditional forecast at the required horizon.

It is worth emphasizing that the GARCH-EVT approach incorporates the two ingredients required for an accurate evaluation of the conditional VaR, i.e. a model for the dynamics of the first and second return moments, and an appropriate model for the conditional distribution of returns. An obvious improvement of this approach as compared to the unconditional EVT is that it incorporates in VaR forecasting changes in expected return and volatility. For instance, if we assume a change in volatility over the recent period, the GARCH-EVT is able to incorporate this new feature in its VaR evaluation, whereas the unconditional EVT would remain stuck at the average level of volatility over the estimation sample.

Chan and Gray (2006) describe the conditional EVT and its application to the forecasting of VaR of daily electricity prices. McNeil and Frey (2000) propose filtering returns by estimating a GARCH model, then applying EVT to the tails of the empirical distribution of innovations while bootstrapping from the central part of the distribution. They verify that the generalized Pareto distribution of EVT results in better estimates for ES than does unconditional EVT, suggesting that the ability to capture changes in volatility is crucial for VaR computation. Jalal and Rockinger (2008) show that this procedure appears to perform a remarkable job when combined with a well-chosen threshold estimation, such as that in Gonzalo and Olmo (2004). Kourouma et al. (2011) conclude that the conditional EVT model is more accurate and reliable for VaR forecasting, according to the rate of violations as well as by application of the Wald, Kupiec, and Christoffersen tests. They also consider EVT a better model for ES forecasting according to an ES test they introduce, based on the average difference between realized returns and the predicted ES.

To implement EVT in practice the analyst needs to choose a threshold that will define the tail of the distribution of returns. The selection of a threshold involves a trade-off between bias and variance. The lower the threshold, the greater the estimation bias, while the higher the threshold, the greater the variance of the estimators, and therefore the greater the degree of uncertainty about the true parameter values. The literature about threshold selection is still scarce for practical cases in which the condition of i.i.d. observations is not reasonable. For instance, Chavez-Demoulin and McGill (2012) determine that a quantile between 0.08 and 0.05 would be an appropriate threshold for a set of high-frequency data from a stock market, using the sample mean excess plot (SMEP) introduced by Davison and Smith (1990). Herrera (2013) performs a sensitivity analysis of VaR estimates for a set of stock market indices, concluding that a quantile between 0.10 and 0.08 could be justified.

To select the threshold we follow the approach used by Chavez-Demoulin and McGill (2012): if the tail of a given variable, defined by the values to the left of the u_0 threshold, follows a GPD, then the excess distribution over higher thresholds u, $u > u_0$, remains a GPD with the same ξ parameter but with a scaling (β_u) that grows linearly with the threshold u. Provided $\xi < 1$, the mean excess function is given by $e(u) = \mathbb{E}[z-u|z>u] = \frac{\beta_u}{1-\xi} = \frac{\beta_{u_0}+\xi u}{1-\xi}$. Hence, the generalized Pareto distribution should also be suitable for the excesses over any thresholds $u > u_0$, subject to the appropriate change of the scale parameter to β_u. The sample mean of the excesses of the threshold u, $e_{T_u}(u) = \frac{\sum_{i=1}^{T_u}(z_i-u)\mathbb{I}_{z_i>u}}{T_u}$, where T_u is the number of observations that exceed u, provides an empirical estimate of $\mathbb{E}[z-u|z>u]$. Therefore, we should expect the sample mean excess to change linearly with u on the range of values of the threshold u for which the GPD model is appropriate. Figure A1 in the online appendix is a sample mean excess plot (SMEP)

for the four stocks for quantile thresholds from 0.80 to 0.99.[10] The horizontal axis is labeled by standardized innovations, not probabilities. We work with the distribution of losses, which is why the x-axis shows positive values relative to quantiles. The variance of the excesses increases with the threshold, leading to wider confidence intervals. According to Figure A1, a good compromise for the choice of the threshold u seems to be between the 0.85 and 0.97 quantiles, corresponding in the GPD to standardized losses between 1% and 2%, depending on the asset. Gray rectangles in Figure A1 indicate pieces of the the SMEP that are approximately straight lines. Visual inspection suggests that if the mean excess plot becomes linear then we might select as our threshold u_0 a value around 1%. To simplify the discussion we decided to use the 0.90 quantile for the four assets, which amounts to using the 0.10 quantile for standardized innovations.

3.2 Parameter estimates

Tables A2a - A2e in the online appendix display maximum likelihood estimates for the EVT models for the four stocks under the five probability distributions, for a given threshold u. In all cases, we use 10% of the data to determine the threshold, for the reasons given in the previous paragraph. The autoregressive effect in volatility specification is strong, with β_1 around 0.93, suggesting strong memory effects. The estimated γ_1 coefficient is positive and statistically significant at 10% in most cases,[11] indicating the existence of a leverage effect for negative returns in conditional volatility. It is also important that the skewness parameter for the SKST and SGED distributions is less than 1 and for the Johnson S_U distribution it is less than 0 for the four stocks, suggesting the utility of incorporating negative asymmetry to model innovations appropriately.[12] The shape parameter is low, implying high kurtosis. The parameter δ takes values between 1.04 and 1.22, and differs significantly from 2. This result suggests that more attention should be paid to modeling the dynamics of the conditional standard deviation rather than the conditional variance, as has been pointed out for a variety of assets by Garcia-Jorcano and Novales (2017).

When applying EVT, we generate the residuals obtained by filtering stock returns using the estimated AR(1)-APARCH(1,1) model, and estimate the parameters of the generalized Pareto distribution from the standardized residuals. For all asset returns, the estimated tail index ξ of the generalized Pareto distribution is positive. The left tail of the GPD distribution is fat and the probability of occurrence of extreme losses is higher than predicted using the normal distribution. The estimated tail indices of IBM and SAN are higher than those of AXA and BP, reflecting the thicker left tails of their return distributions.[13]

As an illustration, we now examine the estimation results for IBM in more detail. The maximum likelihood estimates of the generalized Pareto distribution parameters for IBM are $(\hat{\xi}, \hat{\beta}) = (0.39, 0.51)$, with standard errors of 0.12 and 0.07, respectively. Figure A2 in the online appendix shows a well-defined likelihood profile for this asset with a maximum log-likelihood of -91.877 reached for $\hat{\xi} = 0.39$. Thus, the model we have fitted has very heavy tails with finite variance. We consider the tail of the IBM return distribution as defined by the threshold $u = 1.0533$. That leaves us with 126 exceedances (10% of 1260 data points). Figure 1 shows the fitted GPD

[10] Confidence bands are constructed applying the delta method, assuming that the sample mean follows a normal distribution

[11] Except for IBM under fat-tailed distributions.

[12] Although the Johnson S_U skew parameter is not significant at 5% for IBM and at 10% for BP.

[13] In the estimation of EVT models, we use the R packages *ismev* (version 1.41) and *evir* (version 1.7-3).

model for the excess return distribution, $F_u(y)$, where $y = z - u$, superimposed on points plotted at empirical estimates of excess probabilities for each loss (126 losses).[14] Note the close correspondence between the empirical estimates and the GPD curve. Under the EVT approach the filtered residuals from all models considered fit the GPD curve very similarly, especially when the filtered residuals come from asymmetric distributions. Figure 2 shows the estimated tail probabilities. The points in the graph show the empirical tail probabilities for the 126 threshold exceedances. The smooth curve running through the points is the tail estimator, defined for the right tail by

$$1 - \widehat{F}(z) = \frac{T_u}{T}\left(1 + \hat{\xi}\frac{z-u}{\hat{\beta}}\right)^{-1/\hat{\xi}}.$$

3.3 Estimating risk by a semiparametric approach: filtered historical simulation

The standard historical approach is often limited to the 1-day horizon because of the lack of enough historical data to use non-overlapping h-day returns. Using overlapping h-day returns would distort the tail behavior of the return distributions leading to significant error in VaR and ES forecasts at extreme quantiles. A way out of this difficulty is to estimate innovation quantiles non-parametrically using bootstrapping, which does not need to assume any particular probability distribution [Ruiz and Pascual (2002)]. Bootstrap procedures have the advantage that they allow for the construction of confidence intervals for VaR estimates. Pascual, Ruiz, and Romo (2006) propose a bootstrap procedure that allows for the incorporation of parameter uncertainty. Barone-Adesi (1998), Giannopoulus (1999), and Vosper (2002) propose the filtered historical simulation (FHS) method that extends the idea of volatility adjustment to multi-step historical simulation, using overlapping data in a way that does not create blunt tails for the h-day portfolio return distribution. 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.

FHS is in fact a hybrid method combining some attractive features of both historical and Monte Carlo VaR models. The advantages of 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.

Suppose that at a time s, we want to simulate returns for the next h days. We select $\{z^*_{s+1}, z^*_{s+2}, ..., z^*_{s+h}\}$ at random with replacement (statistical bootstrap) from the set of standardized innovations from our model $\{\hat{z}_1, \hat{z}_2, ..., \hat{z}_s\}$ after filtering out using the APARCH and AR models. We use the APARCH model to simulate future returns for dates $t = s+1, s+2, ..., s+h$:

$$\sigma^*_t = (\hat{\omega} + \hat{\alpha}_1(|\varepsilon^*_{t-1}| - \hat{\gamma}_1\varepsilon^*_{t-1})^{\hat{\delta}} + \hat{\beta}_1(\sigma^*_{t-1})^{\hat{\delta}})^{1/\hat{\delta}}, \tag{3}$$

$$\varepsilon^*_t = z^*_t\sigma^*_t, \tag{4}$$

[14] Figures 1 and 2 show the right tail, considering losses as positive numbers.

$$r_t^* = \hat{\phi}_0 + \hat{\phi}_1 r_{t-1}^* + \varepsilon_t^*. \tag{5}$$

The algorithm contains the following steps:

1. Select $\{z_{s+1}^*, z_{s+2}^*, ..., z_{s+h}^*\}$, drawn randomly with replacement from $\{\hat{z}_1, \hat{z}_2, ..., \hat{z}_s\}$.
2. Take as initial values the estimates from the previous iteration: $\sigma_s^* = \hat{\sigma}_s$, $\varepsilon_s^* = \hat{\varepsilon}_s$, $r_t^* = r_t$.
3. For $t = s+1, s+2, ..., s+h$:
   - Plug σ_{t-1}^* and ε_{t-1}^* into equation (3) to get σ_t^*.
   - Plug z_t^* (from step 1) and σ_t^* into equation (4) to get ε_t^*.
   - Plug r_{t-1}^* and ε_t^* into equation (5) to get r_t^*.
   - The simulated log return over h days ($r_{s,s+h}^*$) is the sum $r_{s,s+1}^* + r_{s+1,s+2}^* + ... + r_{s+h-1,s+h}^*$.
4. Repeating this procedure N times yields N simulated h-day returns, $r_{i,s,s+h}^*, i = 1, 2, ..., N$.

Once we have these trajectories, we compute h-day VaR and ES forecasts by

$$VaR_{s+h}^{\alpha} = Percentile\left\{r_{i,s,s+h}^*, i = 1, ..., N; 100\alpha\right\}, \quad i = 1, 2, ..., N,$$

$$ES_{s+h}^{\alpha} = (N\alpha)^{-1} \sum_{i=1}^{N} (r_{i,s,s+h}^* \mathbb{1}_{\{r_{i,s,s+h}^* < VaR_{s+h}^{\alpha}\}}),$$

where $\mathbb{1}$ is the indicator function equal to 1 if the h-day return $r_{i,s,s+h}^*$ is lower than VaR and equal to 0 otherwise. Thus, the ES is just the mean of the simulated returns below VaR. Finally,

$$SD_{s+h}^{\alpha} = \left\{ (N\alpha)^{-1} \sum_{i=1}^{N} [(r_{i,s,s+h}^* \mathbb{1}_{\{r_{i,s,s+h}^* < VaR_{s+h}^{\alpha}\}}) - ES_{s+h}^{\alpha}] \right\}^{1/2},$$

and, thus SD is just the standard deviation around the ES, considering only the values below VaR.

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, estimate the models and apply the algorithm to generate $N = 5000$ h-day return simulations from which we compute forecasts for VaR and ES. The forecasting exercise extends over 1260 days, the last five years in our sample, 12/5/2011-9/30/2016, obtaining daily h-day forecasts of the VaR, ES, and SD risk measures.

To combine FHS with EVT we generate $N = 5000$ simulations for the h-day returns using a combination of bootstrapping in-sample residuals from the fitted models (i.e. FHS) and GPD simulation. We apply the following algorithm, which was also proposed independently by Daníelsson and de Vries (2000):

1. Use bootstrapping to randomly sample from the standardized innovations for each future period and for each of the N trajectories.
2. If a selected innovation z^* is below the threshold (u), we draw a realization y from the previously estimated GPD($\hat{\xi}, \hat{\beta}$). The value y is taken as the excess below the threshold u, i.e. the numerical value of the innovation to be used in simulation will be: $z^* = u - y$.

3. Otherwise, return the standardized innovations themselves.

4. Finally, we trace back from simulated standardized innovations to recover the returns and we end up with N sequences of hypothetical daily returns for day $s+1$ through day $s+h$. From these, we calculate the hypothetical h-day returns as $r^*_{s,s+h} = \sum_{j=1}^{h} r_{i,s+j}$ for $i = 1, 2, ..., N$, and we can calculate the h-day VaR, h-day ES, and h-day SD as described above.

5. We repeat this procedure for $s+1, s+2, s+3, ..., s+1259$, to cover the out-of-sample period.

4 Alternative approaches to backtesting value at risk and expected shortfall

In spite of having better properties as a measure of risk, ES is still less used than VaR, essentially because backtesting ES is much harder than backtesting VaR. Recently, some ES backtesting procedures have been developed, such as the residual approach introduced by McNeil and Frey (2000), the censored Gaussian approach proposed by Berkowitz (2001), and the functional delta approach of Kerkhof and Melenberg (2004). However, these approaches have some drawbacks. They rely on asymptotic test statistics that might be inaccurate when the sample size is small, and this could penalize financial institutions because of an incorrect forecasting of ES. Further, these tests compute the required p-value based on the full sample size rather than conditioning on the number of exceptions. The saddlepoint techniques introduced by Wong (2008) are accurate and yield reasonable test power even for a small sample size. They allow for detecting the deficiency of a risk model based on just one or two exceptions before any more data is observed. Nonetheless, they still have the limitations of relying on a Gaussian distribution and using a full distribution conditional standard deviation as the dispersion measure.

Some tests have recently been proposed to backtest ES that overcome these limitations. Emmer, Kratz, and Tasche (2015) proposed a new ES backtest based on a simple linear approximation. The ES forecast is obtained as the average of quantiles at different VaR levels, and it is considered acceptable if all the VaR forecasts pass the Kupiec test. The test by Righi and Ceretta (2013) verifies whether the average observed deviation from ES is zero, using the distribution of returns truncated to the left of VaR. Later, Acerbi and Szekely (2014) introduced three model-free, non-parametric backtesting methodologies for ES showing them to be more powerful than the Basel VaR test. Their tests are straightforward to apply but require simulation analysis (as does the Righi and Ceretta test) to compute critical values and p-values. Graham and Pál (2014) proposed a tractable and intuitive extension of the Lugannani-Rice approach of Wong (2008). Costanzino and Curran (2015) developed a methodology that can be used to backtest any spectral risk measure, such as ES, exploiting the fact that ES is an average of a continuum of VaR levels. They introduced an unconditional ES backtest similar to the unconditional VaR backtest of Kupiec, to test whether the average cumulative violation is equal to $\alpha/2$. Later, Du and Escanciano (2016) proposed backtesting for ES based on cumulative violations, which is the natural analogue of the commonly used conditional backtest for VaR, extending the results obtained by Costanzino and Curran (2015). The tests by Costanzino and Curran and Du and Escanciano can be thought of as the continuous limit of the Emmer, Kratz, and Tasche (2015) idea in that they are joint tests of a continuum of VaR levels. These are the tests we apply in the following sections for validating ES forecasts.

The following sections review the VaR and ES backtesting approaches we use in the paper.

4.1 VaR backtesting

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 α quantile used in the estimation of VaR. The null hypothesis $H0 : \pi = \alpha$ is evaluated using the likelihood ratio test

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

where $T_0 = T - T_1$.

Two other tests by Christoffersen (1998) examine whether VaR exceedances are independent. We consider two states of nature for 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 $\pi_{ij} = Pr[I_t(\alpha) = j | I_{t-1}(\alpha) = i]$. We 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 the 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 statistic is $LR_{ind} = -2\ln(L(\widehat{\Pi})/L(\widehat{\Pi}_1))$ with a χ^2_1 distribution. The second test is a conditional coverage test, based on the likelihood ratio statistic, $LR_{cc} = -2\ln(L(\Pi_\alpha)/L(\widehat{\Pi}_1)) = LR_{uc} + LR_{ind}$, which is asymptotically χ^2_2 distributed.

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 dependence of order higher than one, and *ii*) the use of a Markov chain takes into account only 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. We define the auxiliary variable $Hit_t(\alpha) = I_t(\alpha) - \alpha$, so that $Hit_t(\alpha) = 1 - \alpha$ if $r_t < VaR_t^\alpha$ and $Hit_t(\alpha) = -\alpha$ otherwise. The null hypothesis for 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

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

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 $X_{1t} = VaR_t^{\alpha}$) 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.

4.2 Backtesting ES

4.2.1 The Righi and Ceretta approach

The ES backtest approach of Righi and Ceretta (2013) extends and improves those previously introduced in the literature in three main ways. First, they use the dispersion of the truncated distribution by the estimated VaR upper limit, instead of the whole probability function. They refer to this dispersion as the shortfall deviation (SD). Second, they do not limit the approach to the Gaussian case. They permit other probability distribution functions and even an empirical distribution, making this approach more flexible. Finally, their approach allows testing separately whether each VaR violation differs significantly from the ES, and this facilitates using a faster model for error verification, which is extremely useful since prompt action is often required in order to avert extreme financial losses due to market risk.

The SD is the square root of the truncated variance for some quantile conditional on the probability α. We obtain the 1-day SD as $SD_{t+k}^{\alpha} = (VAR_{t+k}[r_{t,t+k}|r_{t,t+k} < VaR_{t+k}^{\alpha}])^{1/2}$, and since $r_{t,t+1} = \mu_{t+1} + \sigma_{t+1}z_{t+1}$, by standardization we get $SD_{t+1}^{\alpha} = (\sigma_{t+1}^2 VAR_{t+1}[z_{t+1}|z_{t+1} < F^{-1}(\alpha)])^{1/2}$.

The SD is a better estimate than the whole sample standard deviation because, when extreme negative returns occur it is the risk in the left tail that concerns risk managers and financial institutions. Furthermore, to quantify precisely how far a loss was from its expected value, one needs to use some dispersion measure intrinsic to this expectation rather than one linked with the absolute distribution expectation.

Righi and Ceretta propose to backtest if the day k violation is significantly worse from that expected for certain α VaR quantile, $BT_{t+k} = (r_{t,t+k} - ES_{t+k}^{\alpha})/SD_{t+k}^{\alpha}$, where how far the occurred loss is from its expected value is computed in units of the dispersion measure. This test has as null hypothesis $H_0 : BT_{t+k} = 0$ against the alternative that $H_1 : BT_{t+k} < 0$. We use the test in Righi and Ceretta (2015) and we focus on performing a single test all the out-of-sample observations, i.e. $H_0 : \mathbb{E}[BT_t] = 0$ against $H_1 : \mathbb{E}[BT_t] < 0$. This is in contrast to Righi and Ceretta (2013) who test for each day of the out-of-sample period.

As $r_{t,t+k} = \mu_{t+k} + \sigma_{t+k}z_{t+k}$, we can write the expression for the test statistic in a form ready to use with sample return data,

$$BT_{t+k} = \frac{z_{t+k} - E_{t+k}[z_{t+k}|z_{t+k} < F^{-1}(\alpha)]}{(VAR_{t+k}[z_{t+k}|z_{t+k} < F^{-1}(\alpha)])^{1/2}},$$

where z_{t+k} denotes the standardized innovations once the μ_t and σ_t models have been estimated under some specific assumptions.

This is a one-tailed test with the alternative hypothesis that the observed loss is worse than the expected one. To robustly obtain the statistical probability linked with the calculated value of BT_{t+k}, i.e. with no need to rely on any assumption about the distribution of this ratio, Monte Carlo simulations are needed (see Righi and Ceretta (2013)).

4.2.2 The Acerbi and Szekely approaches

We use the first two of the three tests introduced in Acerbi and Szekely (2014), each under slightly different assumptions, and with somewhat different null and alternative hypotheses. The intuition underlying the design of the test statistics Z_i, $i \in 1,2$, is the same in both tests. These tests are non-parametric and free from distributional assumptions. They depend neither on the form nor on the parameters of the parent distribution, although they need the assumption of continuity of the distribution function and the probability density function of returns, together with the independence of the sample observations.[15] Acerbi and Szekely (2014) propose an algorithm based on Monte Carlo simulations to estimate the critical values and the p-values of the test statistics. We use it here, as well as for the Righi and Ceretta tests, simulating 10000 processes of length 1000.

The two test statistics are as follows:[16]

Statistic Z_1: Testing ES after VaR.

$$Z_1 = \frac{1}{N_T} \sum_{t=1}^{T} \frac{I_t r_t}{ES_t^{\alpha}} - 1,$$

where $N_T = \sum_{t=1}^{T} I_t > 0$ with $I_t = \mathbb{1}_{\{r_t < VaR_t^{\alpha}\}}$ being the indicator of VaR breaches and T being the length of the out-of-sample period. The null hypothesis is $H_0 : P_t^{\alpha} = F_t^{\alpha} \quad \forall t$ where F_t^{α} is the tail of cumulative distribution of forecasts at time t when $r_t < VaR_t^{\alpha,F}$ and P_t^{α} represents the tail of the unknown distribution from which the realized events, r_t, are drawn. The VaR and expected shortfall under the theoretical and the empirical distributions are denoted by $VaR_t^{\alpha,P}$, $ES_t^{\alpha,P}$, $VaR_t^{\alpha,F}$, and $ES_t^{\alpha,F}$. The alternative hypothesis is

$$\begin{aligned} H_1 : \quad & ES_t^{\alpha,P} \leq ES_t^{\alpha,F} \qquad \forall t \text{ and } < \text{for some } t, \\ & VaR_t^{\alpha,P} = VaR_t^{\alpha,F} \quad \forall t. \end{aligned}$$

We see that the predicted VaR^{α} is still correct given H_1, in line with the idea that this test is subordinate to a preliminary VaR test. This test is in fact completely insensitive to an excessive number of exceptions as it is an average taken over exceptions themselves.

Under these conditions $\mathbb{E}_{H_0}[Z_1|N_T > 0] = 0$ and $\mathbb{E}_{H_1}[Z_1|N_T > 0] > 0$. Hence, the realized value Z_1 is expected to be zero, and its positivity signals a problem.

Statistic Z_2: Testing ES directly.

$$Z_2 = \frac{1}{T\alpha} \sum_{t=1}^{T} \frac{I_t r_t}{ES_t^{\alpha}} - 1,$$

provided that $N_T > 0$. H_0 is as in the previous test and the alternative hypothesis is

$$\begin{aligned} H_1 : \quad & ES_t^{\alpha,P} \leq ES_t^{\alpha,F} \qquad \forall t \text{ and } < \text{for some } t, \\ & VaR_t^{\alpha,P} \leq VaR_t^{\alpha,F} \quad \forall t. \end{aligned}$$

[15] Continuity and strict monotonicity allow for expected shortfall to be expressed as the expected value of returns below the value at risk.

[16] We have adapted the test statistics to apply to negative values of VaR_{α} and ES_{α}. Acerbi and Szekely (2014) define them for positive ES values.

We note that $\mathbb{E}_{H_0}[N_T] = T\alpha$. We have again $\mathbb{E}_{H_0}[Z_2] = 0$ and $\mathbb{E}_{H_1}[Z_2] > 0$.

Unlike the Z_1 statistic, the sum of the VaR breach event returns is now divided by the expected value. The Z_2 statistic will tend to reject a large number of VaR breach events of small magnitude. This leads to the difference in H_1 between the two statistics. Rejecting the H_0 of Z_2 includes rejecting $VaR_t^{\alpha,F}$ as being correctly specified.

Under the null hypothesis the number of theoretical VaR breaches is $\mathbb{E}_{H_0}[N_T] = T\alpha$. The relationship between the two test statistics of Acerbi and Szekely is $Z_2 = (1+Z_1)N_T/T\alpha - 1$. This shows that while Z_1, being just an average taken over excesses, is insensitive to an excessive number of exceptions, Z_2 depends on that number through the ratio $N_T/T\alpha$. This is why, when the number of violations exceeds the theoretical level, p-values for the Z_2-test are lower than for the Z_1 test. Therefore, an ES model will pass the Z_2 test when not only the magnitude but also the frequency of the excesses is statistically equal to the expected one.[17]

4.2.3 The Graham and Pál Approach

The goal of the Graham and Pál (2014) test is to quantify how extreme each VaR violation is in relation to its forecast distribution. The approach can be intuitively described as an extension of the "hit" time series concept, wherein each of the "1" values (when VaR violation occurs) is modified so as to measure the distance between a violation and its corresponding VaR threshold. We obtain a time series with negative values consisting of the differences between each percentile smaller than the VaR threshold percentile and the VaR threshold percentile itself. If there is no VaR violation on a specific observation date, then a value of zero is still recorded. We expect this series to be uniformly distributed within the tail region if the distribution of the series of forecasts accurately represents the portfolio/asset's *P&L*.

The central risk concept employed in this backtest is that of tail risk, as defined by Wong (2010). The tail risk at significance level α (TR_α) is related to VaR and ES in each period by

$$TR_\alpha = \int_{-\infty}^{q(\alpha)} (r - q(\alpha)) f(r) dr = \alpha(ES_\alpha - VaR_\alpha),$$

where $VaR_\alpha = q(\alpha) = F^{-1}(\alpha)$. The tail risk will always be negative, and we can consider $\alpha^{-1}TR$ as the difference between the ES and the VaR.

Given a sample of T returns $r_1, r_2, ..., r_T$, the sample unbiased estimator for the tail risk at confidence level $(1-\alpha)$ can be calculated by

$$\widehat{TR}_\alpha = \bar{X} = \frac{1}{T}\sum_{t=1}^{T} (r_t - q_t(\alpha)) \mathbb{1}_{\{r_t < q_t(\alpha)\}} = \frac{1}{T}\sum_{t=1}^{T} X_t,$$

where $X_t = (r_t - q_t(\alpha)) \mathbb{1}_{\{r_t < q_t(\alpha)\}}$. We note that the range of X is $\text{range}(X) = (-\infty, 0]$.

To proceed, we simply transform the realized losses and forecast distributions through the probability integral transform (PIT) to the exponential context to ensure that our sample estimator

[17] Acerbi and Szekely (2014) show that the Z_2 test is more powerful than the Z_1 test when the null and alternative hypothesis differ in volatility, while Z_1 is more powerful than Z_2 in the case of different tail indices.

for the average tail risk is created through identically distributed sample values.[18] The exact transformation is simple. We only need to locate each VaR violation as a percentile value within its forecast distribution. For a return r_t exceeding the VaR the random variable X is redefined as

$$X_t = (\ln F_t(r_t) - \ln F_t(q_t(\alpha)))\mathbb{1}_{\{r_t<q_t(\alpha)\}} = (\ln p_t - \ln \alpha)\mathbb{1}_{\{F_t(r_t)<\alpha\}} = (\ln p_t - \ln \alpha)\mathbb{1}_{\{\ln p_t<\ln \alpha\}},$$

where $p_t = F_t(r_t)$, and the transformations within the indicator function are possible by monotonicity.

If the forecast CDFs $F_t(\cdot)$ are correct estimates of the real and unobservable *P&L* distributions, then the series p_t is distributed uniformly $\mathbb{U}(0,1)$. Moreover, if the sequence of forecast CDFs is correctly conditionally calibrated, then the corresponding p_t sequence is i.i.d. $\mathbb{U}(0,1)$.[19] Now we need to determine the exact distribution of $\{X_t\}$. Knowing its CDF and PDF allows us to determine its moments and cumulants, and both of these are used in determining its theoretical average value and in calculating the sample value of the test statistic using the small-sample asymptotic technique described next [details can be found in Graham and Pál (2014)].

Under the assumption that the sequence of forecast *P&L* distributions is correctly conditionally calibrated, we know that $\{r_t\} := \{\ln p_t\}$ is i.i.d. $Exp(-\infty, 0)$ with PDF $\phi_E(\cdot)$ and CDF $\Phi_E(\cdot)$ both equal to e^r for $-\infty < r < 0$. The moment-generating and cumulant-generating functions of X can be used in the Lugannani-Rice formula to calculate the tail probability of exceeding the sample mean $\bar{X}$ by a saddlepoint technique. These functions are

$$\left.\begin{aligned} M(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{0} e^{tx} dF_X(x) = \frac{\alpha}{t+1} + 1 - \alpha, \\ K(t) = \ln M(t). \end{aligned}\right\}$$

Omitting the remaining details,[20] to approximate the tail of the cumulative distribution function of the sample mean of the X-variable defined above we proceed as follows: first, we analytically solve the saddle-point equation $K'(s) = \frac{M'(s)}{M(s)} = -\frac{\alpha}{(s+1)[s(1-\alpha)+1]} = \bar{x} \quad \text{for} \quad \bar{x} < 0$, where s is the unique solution in the interval $(-1, \infty)$. Second, we define, for readability, $\eta = s\sqrt{TK''(s)}$ and $\varsigma = sgn(s)\sqrt{2T(s\bar{x} - K(s))}$, where $sgn(s)$ denotes the sign of s. Finally, according to Lugannani and Rice, we have that the tail probability of exceeding the sample mean $\bar{x} \neq \mu_X$ is given by

$$\mathbb{P}[\bar{X} > \bar{x}] = 1 - \Phi(\varsigma) + \phi(\varsigma)\left(\frac{1}{\eta} - \frac{1}{\varsigma} + \mathcal{O}(T^{-3/2})\right),$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ represent the standard normal CDF and PDF, respectively.

We now formulate the hypothesis test explicitly. Although the tail loss $TR_\alpha = \alpha(ES_\alpha - VaR_\alpha)$ could be used as the test statistic, Graham and Pál follow Wong (2010) and take as the test statistic the standardized variable

$$z = \alpha^{-1} TR_\alpha = ES_\alpha - VaR_\alpha.$$

[18] Rosenblatt (1952), Crnkovic and Drachman (1996), Diebold et al. (1998), and Berkowitz (2001) are often credited with introducing PIT into the financial risk management backtesting literature. Graham and Pál apply a further transformation to the exponential context because it allows to solve for the saddle point analytically. This solution is, moreover, well defined over the complete interval of interest for tail losses.

[19] Another necessary condition for the series to be i.i.d. is that the *P&L* time horizons do not overlap; otherwise, serial interdependencies may occur within the data.

[20] For more details, Lugannani and Rice (1980), Daniels (1987), and Wong (2010).

Therefore, under the exponential tail distribution null hypothesis, we have $VaR^0 = \ln\alpha$ and $ES^0 = \frac{1}{\alpha}\int_{-\infty}^{\ln\alpha} r\phi_E(r)dr = \frac{1}{\alpha}\int_{-\infty}^{\ln\alpha} re^r dr = \frac{1}{\alpha}(re^r - e^r)|_{-\infty}^{\ln\alpha} = \ln\alpha - 1$, so that $TR^0 = \alpha(ES^0 - VaR^0) = -\alpha$. For instance, with $\alpha = 0.01$, we have

$$z_0 = \alpha^{-1}TR^0 = ES^0 - VaR^0 = -1, \qquad \text{i.e.} \quad TR^0 = -0.01.$$

Accordingly, a one-tailed regulatory backtest to check whether the risk model provides sufficient risk coverage may be formulated in terms of z, with the null and alternative hypotheses defined by

$$H_0: \quad z = z_0, \qquad \text{i.e.} \quad TR_\alpha = TR^0,$$
$$H_1: \quad z < z_0, \qquad \text{i.e.} \quad TR_\alpha < TR^0.$$

The p-value of the hypothesis test can obtained using the Lugannani-Rice formula above as

$$p - value = \mathbb{P}[\bar{X} \leq \bar{x}] = 1 - \mathbb{P}[\bar{X} > \bar{x}].$$

The null hypothesis is rejected if the realized value of the sample statistic $\widehat{TR}_\alpha$ is significantly lower than the theoretical level of tail risk TR^0. If we obtain $TR_\alpha > TR^0$, we will say that the risk model captures tail risk sufficiently, or that it provides sufficient risk coverage, although risk may then be overestimated. When that happens, the logarithmic difference between the probability of an excess and the significance level for VaR_t^α follows a distribution with thicker tails than the exponential distribution. Alternatively, when the forecast CDF is a correct estimate of the real and unobservable $P\&L$ distribution, such probability differences follow an exponential distribution.[21]

4.2.4 The Costanzino and Curran and Du and Escanciano approaches

It is well known that for each coverage level, violations should be unpredictable if the risk model is appropriate, i.e. they should be a martingale difference sequence (mds). Indeed, rather than just one mds, violations form a class of mds indexed by the coverage level. The cumulative violation process accumulates all violations in its left tail, just like the ES accumulates the VaR in its left tail. Du and Escanciano (2016) suggest a Box-Pierce test to check for the mds property. Their Box-Pierce test is the analogue for ES of the conditional backtest proposed by Christoffersen (1998) and Berkowitz, Christoffersen and Pelletier (2011) for VaR.

This approach is developed from Costanzino and Curran (2015). It is based on the idea that ES is an average of a continuum of VaR levels and it can be thought of as the continuous limit of the Emmer, Kratz, and Tasche idea in that it is a joint test of a continuum of VaR levels. Unlike the test proposed by Du and Escanciano, the test proposed by Costanzino and Curran does not test independence, but it is the first proposed coverage test for spectral risk measures. It essentially amounts to a joint test of a continuum of weighted VaR quantiles to give a single decision at a fixed confidence level. The key of the method is to show that the spectral measure failure rate is asymptotically normal under the null hypothesis and therefore admits a formal Z-test.

The cumulative violation process is defined by

$$H_t(\alpha) = \frac{1}{\alpha}\int_0^\alpha h_t(u)du,$$

[21] That amounts to return violations, in probability terms, following a uniform (0,1) distribution.

where $h_t(u) = \mathbb{1}_{(r_t \leq VaR_t(u))}$ is the u-violation or hit at time t. Since $h_t(u)$ has mean u, by the Fubini Theorem $H_t(\alpha)$ has mean $1/\alpha \int_0^\alpha u du = \alpha/2$. Moreover, again by the Fubini Theorem, the mds property of the class $\{h_t(\alpha) - \alpha : \alpha \in [0,1]\}_{t=1}^{\infty}$ is preserved by integration, which means that $\{H_t(\alpha) - \alpha/2\}_{t=1}^{\infty}$ is also mds.

For computational purposes, it is convenient to define $u_t = F(r_t, \Omega_{t-1})$ where $F(\cdot, \Omega_{t-1})$ denotes the conditional cumulative distribution function of r_t given Ω_{t-1}. Using the fact that $h_t(u) = \mathbb{1}_{(r_t \leq VaR_t(u))} = \mathbb{1}_{(u_t \leq u)}$, we obtain,[22]

$$H_t(\alpha) = \frac{1}{\alpha} \int_0^\alpha \mathbb{1}_{(u_t \leq u)} du = \frac{1}{\alpha}(\alpha - u_t)\mathbb{1}_{(u_t \leq \alpha)}.$$

As for violations, cumulative violations are distribution-free, since $\{u_t\}_{t=1}^{\infty}$ comprises a sample of i.i.d. $\mathbb{U}(0,1)$ variables. Working with violations avoids approximations, as the previous integral can be computed exactly. Unlike violations, cumulative violations contain information on the tail risk. When violations are zero, cumulative violations are also zero, but when a violation occurs, the cumulative violation measures how far the actual value of r_t is from its quantile.

The variables $\{u_t\}_{t=1}^{\infty}$ necessary to construct $\{H_t(\alpha)\}_{t=1}^{\infty}$ are generally unknown, since the distribution of the data F is unknown. In practice, researchers and risk managers specify a parametric conditional distribution $F(\cdot, \Omega_{t-1}, \theta_0)$, where θ_0 is some unknown parameter in $\Theta \subset \mathbb{R}^p$, and proceed to estimate θ_0 before producing VaR and ES forecasts. With the parametric model, we can define the "generalized errors", $u_t(\theta_0) = F(r_t, \Omega_{t-1}, \theta_0)$ and the associated cumulative violations, $H_t(\alpha, \theta_0) = \frac{1}{\alpha}(\alpha - u_t(\theta_0))\mathbb{1}_{(u_t \leq \alpha)}$.

Very much like for VaRs, the arguments above provide a theoretical justification for backtesting ES by checking whether $\{H_t(\alpha, \theta_0) - \alpha/2\}_{t=1}^{\infty}$ has zero mean (unconditional ES backtest) and it is serially uncorrelated (conditional ES backtest). The unconditional backtest for ES is a standard t-test for the null hypothesis

$$H_{0u} = \mathbb{E}[H_t(\alpha, \theta_0)] = \alpha/2.$$

Note that a simple calculations show that $\mathbb{E}[H_t^2(\alpha, \theta_0)] = \alpha/3$, and hence, $Var(H_t(\alpha)) = \alpha(1/3 - \alpha/4)$. Therefore, a simple t-test statistic is

$$U_{ES} = \frac{\sqrt{T}(\bar{H}(\alpha) - \alpha/2)}{\sqrt{\alpha(1/3 - \alpha/4)}} \xrightarrow{d} N(0,1),$$

where T is the size of the out-of-sample period which is used to evaluate (backtest) the ES model and $\bar{H}(\alpha)$ denotes the sample mean of $\{\hat{H}_t(\alpha)\}_{t=1}^{T}$. The U_{ES} statistic has a standard normal limit distribution when the estimation period is much larger than the evaluation period.

Next, the conditional backtest has the null hypothesis

$$H_{0c} : \mathbb{E}[H_t(\alpha, \theta_0) - \alpha/2 | \Omega_{t-1}] = 0,$$

which is the analogue of the null hypothesis of the conditional backtest for VaR.[23] Define the lag-j

[22] $H_t(\alpha) = \frac{1}{\alpha}\int_0^\alpha \mathbb{1}_{(u_t \leq u)} du = \frac{1}{\alpha}\mathbb{1}_{(u_t \leq \alpha)} \int_0^\alpha \mathbb{1}_{(u_t \leq u)} du = \frac{1}{\alpha}\mathbb{1}_{(u_t \leq \alpha)} \int_{u_t}^\alpha 1 du = \frac{1}{\alpha}\mathbb{1}_{(u_t \leq \alpha)}(\alpha - u_t)$.

[23] Note that we use the conditional mean restriction in the definition of autocorrelations. As a result, tests based on γ_{Tj} are expected to have power against deviations from H_{0c}, where $H_t(\alpha)$ are uncorrelated but have mean different from $\alpha/2$.

autocovariance and autocorrelation of $\{H_t(\alpha)\}_{t=1}^{T}$ for $j \geq 0$ by

$$\gamma_{Tj} = \frac{1}{T-j} \sum_{t=j+1}^{T} (H_t(\alpha) - \alpha/2)(H_{t-j}(\alpha) - \alpha/2) \quad \text{and} \quad \rho_{nj} = \frac{\gamma_{Tj}}{\gamma_{T0}}.$$

Simple conditional tests can be constructed using $\hat{\rho}_{nj}$, for example the Box-Pierce test statistic

$$C_{ES}(m) = n \sum_{j=1}^{m} \hat{\rho}_{nj}^2 \xrightarrow{d} \chi_m^2.$$

The C_{ES} statistic has a chi-square distribution with m degrees of freedom when the estimation period is much larger than the evaluation period.

5 Evaluating 1-day ES forecasts

5.1 ES forecasts under the parametric approach

In this section we show the results from VaR and ES forecasts following a standard time-varying parametric approach. We restrict our attention to the left tail of the distribution and the 1%, 2.5% and 5% significance levels, and we compute recursive VaR and ES forecasts from an expanding window. First, each model is estimated using 2915 daily observations from the 10/2/2000-12/2/2011 sample period. After that, we increase the initial sample by one data point each day until the end of 2016, to compute 1-day ahead VaR and ES forecasts over five years: 2012-2016 (1260 data observations). Over this forecasting period, models are estimated every 50 days, a choice intended to reduce the computational cost while avoiding frequent parameter variation that might be due to pure noise.

We follow other authors like Giot and Laurent (2003a, 2003b), McMillan and Speight (2004) and McMillan and Kambouroudis (2009), who use an expanding window. They usually start the estimation process by excluding from the complete sample for which data is available, generally 5 to 10 years, the observations chosen as out of the sample period, over which one-step (day) forecasts are obtained. However, Alexander and Sheedy (2008) conclude that the estimation window is an important source of model risk. Considering a range of possible estimation windows (250, 500, 1000, and 2000 days) their results show that large windows should be preferred to smaller estimation windows for VaR risk estimation, especially in conditional models. Righi and Ceretta (2015) also use estimation windows of different sizes to forecast VaR and ES with a variety of unconditional and conditional models. They conclude that the larger the window, the more conservative risk predictions tend to be. Conditional models exhibit more homogeneity than unconditional models concerning these estimation windows because conditional models rely on parametric filtering and not only on the empirical data, which is sensitive to the bandwidth used in the estimation. With a starting window of 2,915 observations that is increased every day, we seem to be on the safe side, according to the papers mentioned. However, the sensitivity of VaR and ES estimates to the size of the estimation window is clearly a question that deserves further exploration.

Table A3 in the online appendix displays descriptive statistics for returns for the in-sample (10/2/2000-12/2/2011) and out-of-sample (12/5/2011-9/30/2016) periods. Skewness is negative,

except for SAN and AXA in the in-sample period. Kurtosis is higher than 3 for the four stocks in both periods. We are thus confronted with fat tail distributions and the Jarque-Bera statistic clearly rejects the null hypothesis of a normal distribution. VaR and ES forecasts based on the assumption of a normal distribution of returns are therefore inappropriate, so we forecast both risk measures, not only using the information provided by the full distribution but also using the information from extreme events, as explained in subsection 3.1. Applying EVT to these leptokurtic distributions seems justified as it should allow for a better estimation of extreme variations in financial returns.

Figure 3 shows IBM daily percentage returns (1260 data) together with out-of-sample $VaR_{1\%}$ and $VaR_{5\%}$ forecasts from an AR(1) model for returns with a JSU-APARCH(1,1) model for return innovations. Such forecasts are compared with those obtained by applying extreme value theory (EVT), fitting a GPD density to the tail of the distribution. The differences in VaR calculated with the two models are small for the 5% quantile but they become more important for the 1% quantile. VaR forecasts under EVT indicate higher losses than are indicated by the VaR estimated without the use of EVT. Figure 4 shows $ES_{1\%}$ and $ES_{5\%}$ forecasts obtained with EVT and without EVT. We can see that the forecast of average losses exceeding VaR under the GPD distribution in the EVT approach is greater than that obtained from a JSU distribution in the non-EVT approach, especially for the more extreme quantiles.[24]

Assuming that $\xi < 1$, the ratio of the two risk measures predicted under the EVT approach, behaves for small values of the quantile probability α as,

$$\lim_{\alpha \to 0} \frac{ES_\alpha}{VaR_\alpha} = \begin{cases} (1-\xi)^{-1}, & \xi \geq 0, \\ 1, & \xi < 0. \end{cases}$$

It is essentially determined by the shape parameter ξ of the GPD distribution when we go far enough out into the tail. Figure A3 in the online appendix shows the evolution of the ratio for the model AR(1)-JSU-APARCH(1,1) for IBM, with the estimated parameter $\xi = 0.392$. When $\alpha \to 0$, this ratio tends to $(1-\xi)^{-1} = 1.644$.

We now examine our forecasts for the complete out-of-sample period (5 years, 1260 data). Since we generate time series of VaR and ES forecasts, we just summarize the results for the 5-year period. Tables 2 - 5 show the average of out-of-sample 1-day VaR and ES forecasts ($\overline{VaR}$, $\overline{ES}$), and the violation ratio (Viol) of the underlying VaR and the backtesting results for the different models. Our discussion here focuses on the general patterns that appear in these forecasting results. As is expected for leptokurtic distributions, average VaR forecasts are larger in absolute value at extreme significance levels when using EVT, the opposite being the case for $\alpha = 0.05$. The average ES forecasts from conditional EVT-based models can be seen to be "more negative" than forecasts from conditional models not based on EVT. As shown in Figure 4, differences on ES forecasts at the 1% significance level are larger than those at the 5% significance level.

It seems desirable that a good ES model have a violation ratio close to the theoretical one. Indeed, as we have already seen, some validation tests for ES are based on this comparison. Conditional EVT-based models tend to yield a violation ratio very close to the theoretical one. Departures from the theoretical violation ratio are larger for models not using EVT, especially when assuming the normal and Student-t distributions for return innovations. In general, the violations

[24] Figures 3 and 4 show only the negative returns so as to maintain a clear perspective on the different VaR and ES estimates.

ratio suggests that conditional EVT-based models forecast the VaR quantile correctly, corroborating Kuester, Mittnik, and Paolella (2006), who attest the superiority of this approach. Furthermore, we will show below that EVT-based models not only yield an accurate violation ratio but they also perform well at ES backtesting. On the other hand, conditional ES models not based on EVT have a violation ratio higher than expected, although they improve under heavy-tailed distributions, corroborating Mabrouk and Saadi (2012). They forecast ES worse than EVT-based models.

To save space we just show in the tables p-values for the different tests, omitting the numerical values of the test statistics. The Acerbi and Szekely tests and the Graham and Pál tests yield significant evidence against models not based on EVT. For these three tests we observe large differences in p-values between conditional models based on EVT and non-EVT based conditional models in favor of the former, which seem to produce better risk forecasts. This is even clearer at lower significance levels, revealing the fact that without close attention to extreme returns it is hard to capture tail risk with precision. Furthermore, at the 1% significance level, p-values for the Acerbi and Szekely and Graham and Pál tests for the conditional models not based on EVT theory are very close to 0, with positive realized values for Z_1 and Z_2 (not shown in the tables). Hence, these tests reject H_0 because of significant evidence of risk underestimation. The rejection is still more apparent assuming normally distributed return innovations. In models without EVT, the Graham and Pál test discriminates against the normal and Student-t distribution for almost all significance levels for the four stocks. This is also often the case with the unconditional coverage test, LR_{uc}. On the other hand, in EVT-based models the tests do not discriminate among the results for VaR and ES validation obtained under the alternative probability distributions. It seems that when EVT is applied, the choice of probability distribution for non-extreme returns is not a critical issue. The two-sided tests of Costanzino and Curran and Du and Escanciano do not yield evidence against models that do not incorporate EVT. With the only exception of SAN at 1% significance, they do not even reject non-EVT based models. This suggests that the problem with such models is not the clustering of VaR exceedances but, rather, their size.

We indicate in boldface the p-values of the Righi and Ceretta, Acerbi and Szekely, and Graham and Pál tests when the test statistics have the sign opposite to the one supposed in the alternative hypothesis. This situation arises for EVT-based models for the four stocks, which means that we are overestimating risk. Being one-sided tests, the null hypothesis cannot be rejected in these settings. We review the structure of the tests. The Righi and Ceretta test considers $H_0 : \mathbb{E}(BT_t) = 0$ against $H_1 : \mathbb{E}[BT_t] < 0$, but with some models we obtain $\mathbb{E}[BT_t] > 0$, reflecting that most excesses fall between VaR and ES, not beyond ES, especially under the EVT approach. The first test by Acerbi and Szekely specifies $H_0 : \mathbb{E}[Z_1] = 0$ against $H_1 : \mathbb{E}[Z_1] > 0$ and the second one, $H_0 : \mathbb{E}[Z_2] = 0$ against $H_1 : \mathbb{E}[Z_2] > 0$. However, with some models, especially models based on EVT, we obtain $\mathbb{E}[Z_1] < 0$ and $\mathbb{E}[Z_2] < 0$. In the first test, this means that the average of realized excesses is lower in absolute value than the predicted ES. In the second test, it may indicate that both the average excess and the number of excesses are lower than expected. The one-sided Graham and Pál test considers $H_0 : TR_\alpha = TR^0$ against $H_1 : TR_\alpha < TR^0$ where TR^0 is equal to $-\alpha$ under the exponential assumption. However, the test statistic obtained with EVT-based models has the wrong sign for AXA, indicating that the actual, unobserved tail risk is lower than what these models detect. There is also a suggestion of risk overestimation for BP under normality at 2.5% and 5% significance.

In short, bold figures in the tables signal frequent overestimation of risk for EVT-based ES

models that is not detected by one-sided tests. In these cases the number of violations does not differ much from the theoretical value, reflecting good VaR forecasts. However, the sign of the test statistic is contrary to that in the null hypothesis, showing an overestimation of ES that implies the level of capital required is too high. On the other hand, we have checked that the absolute value of the statistic is generally very small, suggesting that the estimation error may be statistically acceptable and the excess in the cost of capital is generally small. The possible overvaluation of risk can be seen in Figure 8, that shows the tail probability distributions estimated for IBM. The results for other assets are similar. Colored lines show the estimated tail probabilities and the rectangles display observed relative frequencies. Estimated parameters for each distribution are shown in parenthesis in the footnote to the figure. We observe that most probability distributions other than GPD tend to undervalue the weight of extreme returns. Such undervaluation is especially obvious for the normal distribution. On the contrary, the GPD is well suited to capturing tail risk appropriately, and it avoids underestimating extreme risks, although at the price of slight overvaluation of the risk of medium range losses.

The tests by Costanzino and Curran and Du and Escanciano are two-tailed and hence both risk undervaluation and overvaluation can lead to a rejection of the null hypothesis. These tests are based on the cumulative violation process. Unlike violations, cumulative violations H_t contain information on tail risk and, therefore, they provide a more complete description of the risk involved in a given distribution of returns. Their main advantage is that the distribution of the test statistic is available for finite out-of-sample sizes, which leads to size and power properties better than for other tests. The *p*-values for these tests shown in Tables 2 - 5 generally lead to not rejecting the null hypothesis. The lack of rejection is an indication that the overestimation of risk by EVT-based models is not very important. Evidence against the ES models considered arises from the unconditional Costanzino and Curran test, which often rejects non-EVT based models under normal and Student-t distributions, especially in the more extreme 1% and 2.5% tails. The two conditional tests do not discriminate among models or among probability distributions. The only exception is the rejection of non-EVT based models for SAN by the $C_{ES}(5)$ test because of the autocorrelation of cumulative violations over the first five lags at the 1% significance level. To help understand these results, Figures 6 and 7 show the cumulative violations $\{\hat{H}_t(0.05)\}$, $\{\hat{H}_t(0.025)\}$, and $\{\hat{H}_t(0.01)\}$ of IBM and SAN in the out-of-sample period for the JSU-APARCH and JSU-EVT-APARCH models. We do not observe large values of $\{\hat{H}_t(\alpha)\}$, but we observe some clusters of cumulative violations, which suggest deviations from the martingale difference sequence hypothesis that would be implied by an appropriate ES forecast. In fact, we reject the null hypothesis of the unconditional test, $\mathbb{E}[\hat{H}_t(\alpha)] \neq \alpha/2$, for the JSU-APARCH, at the 1% significance level for IBM, but not for SAN. To complete the picture, Figures 8 and 9 show non-significant autocorrelations in the hit sequence over the first twelve lags for IBM returns, e.g. $Cov(H_t(\alpha), H_{t-j}(\alpha)) = 0$. On the contrary, there are some significant autocorrelations for SAN with both approaches, especially at 1% significance. In fact, with other distributions, at the 1% significance level, the *p*-values of the $C_{ES}(5)$ statistic are also equal to 0 for this stock. The number of extreme losses and the average losses are not very large but they are highly correlated.

Table 6 reports the expected value of violations ($n\alpha$), the number of violations ($V(\alpha) = \sum_{t=1}^{N} \hat{h}_t(\alpha)$), and the cumulative violations $\left(CV(\alpha) = \sum_{t=1}^{N} \hat{H}_t(\alpha)\right)$ for all assets in the out-of-sample period with the JSU-APARCH and JSU-EVT-APARCH models. The number of observed violations is closer to the theoretical level under the EVT approach, except for SAN and BP at

the 5% and 1% significance levels, respectively. This is additional evidence that VaR forecasts are more accurate using the EVT approach. The mean absolute deviation in the number of violations with respect to the theoretical level in Table 6 is 4.8 for the non-EVT models and 2.4 for the EVT-based models. Furthermore, at the more extreme 2.5% and 1% confidence levels, the number of violations by non-EVT models is above the theoretical level, suggesting an undervaluation of risk. An even more relevant result concerns cumulative violations since they exploit information on tail risk. Table 6 shows that in 11 of the 12 comparisons cumulative violations, measured by $CV(\alpha)$, show larger losses for non-EVT based models. Mean cumulative violations are 19 and 17 under non-EVT based and EVT-based ES forecasting models, respectively.

Generally speaking, we have obtained that, for our sample of stocks, conditional EVT-based models produce better VaR forecasts and yield the best results for ES forecasts according to different ES backtests. In many cases, we obtain p-values close to 1 with EVT-based models. The success of EVT models for ES forecasting corroborates findings of Marinelli et al. (2007), Jalal and Rockinger (2008), and Wong et al. (2012). However, we must bear in mind that the Righi and Ceretta, Acerbi and Szekely, and Graham and Pál tests are one-sided by nature and are focused on risk undervaluation. Consequently, in these tests risk overestimation does not lead to a rejection of the null hypothesis, and this seems to occur frequently in ES forecasting with EVT-based models.

As a preliminary test to check the robustness with respect to the choice of threshold of our findings on the overestimation of risk, we also used the 0.08 and 0.15 thresholds for the conditional model under a JSU distribution for IBM and AXA. The evidence for the overestimation of risk by EVT models did not change significantly. This result is in line with Kourouma et al. (2011), who observe a strong overestimation of ES based on the conditional EVT model, although they only use their own ES backtest, which we described in the introduction.

5.2 ES forecasts under filtered historical simulation

As an alternative, we evaluate the performance of 1-day out-of-sample ES forecasts from semi-parametric FHS using the test of Righi and Ceretta and the two tests of Acerbi and Szekely because they are suitable for non-parametric VaR and ES forecasts. We observed in Tables 2-5 an underestimation of risk under non-EVT based models. EVT-based models increase the numerical risk estimates, although at the cost of frequent slight overestimation of risk. Tables A4 - A7 in the online appendix show average VaR ES forecasts ($\overline{VaR}$, $\overline{ES}$), the violation ratio of the underlying VaR, and backtesting results. Comparing with Tables 2 - 5 leads to several observations suggesting that the overestimation of risk by EVT-based models can be corrected by the use of FHS: *i*) conditional EVT-based models do not always yield "more negative" average ES values than conditional models not based on EVT; *ii*) unlike Tables 2 - 5, EVT based models do not yield a lower violation rate than non-EVT based models; *iii*) using FHS, average VaR and ES forecasts are closer to those obtained under the parametric approach for non EVT-based models than for EVT-based models; and *iv*) VaR violation rates are better than those obtained under the parametric approach. These four observations suggest that application of FHS avoids the overestimation of risk to a great extent. As a matter of fact, *v*) the overvaluation of risk as signaled by a test statistic having a sign opposite to H_1 in the one-tailed tests is much less frequent than under the parametric approach in Tables 2 - 5. Two additional relevant results are: *vi*) Models that do not incorporate EVT seem again unsuitable in terms of ES forecasts, being rejected often by the Acerbi and Szekely Z_1 and Z_2 tests for $ES_{1\%}$ and $ES_{2.5\%}$. Less discrimination is obtained at 5% significance level. For instance,

at this level, all models display good ES performance for BP at 10% significance, although the Z_2 test suggests that ES is possibly overvalued. *vii*) Average ES values over the out-of-sample period (5 years, 1260 data) are now more similar among models than under the parametric approach in Tables 2 - 5. This observation is important because it amounts to a reduction in model risk, i.e. a reduction in the uncertainty that arises about the true value of VaR and ES due to the availability of forecasts coming from a variety of alternative models.

The conclusions obtained when applying ES backtests under the parametric and FHS approaches are similar, which is reassuring. Differences between conditional models based on EVT and those not based on EVT are more evident under the parametric approach, because the power and flexibility of conditional volatility models is diluted by historical simulation. The dilution depends on the number of realizations generated for FHS estimation.

6 Robustness analysis

In this section we report on the results obtained from two different tests for robustness. First, we split the sample into pre-crisis and post-crisis subperiods to analyze the performance of ES forecasts in stable and stressed times. Second, we consider a 10-day risk horizon for ES forecasts.

6.1 Pre-crisis and crisis periods

The pre-crisis period is defined so as to have the same number of observations as the crisis period (1239 data points). For the pre-crisis period we used the sample 10/2/2000-6/30/2005 to compute 1-day forecasts over 7/1/2005-6/29/2007. For the crisis period, we used 10/1/2002-6/29/2007 as the in-sample period and 7/2/2007-6/29/2009 as the out-sample period. To save space, we just summarize the conclusions, but detailed results for both periods are available from the authors upon request. We observe that models not based on EVT with asymmetric distributions and all EVT-based models are preferred because they have more flexibility for capturing the risk in both pre-crisis and crisis periods.

We summarize the results: *i*) asymmetric distributions perform slightly better in ES forecasting than symmetric distributions (with or without EVT) in both periods (pre-crisis and crisis); *ii*) during the pre-crisis period the ratio of violations is close to the expected ratio (α) for most models, but the performance according to this criterion is much better for EVT-based models; *iii*) both classes of models undervalue risk during the crisis systematically, exhibiting numbers of violations above the theoretical one, suggesting that these models do not fully adapt to the occurrence of tail events; *iv*) cumulative violations display significant autocorrelation during the crisis period, especially when forecasting $ES_{1\%}$ and $ES_{2.5\%}$; *v*) in general, *p*-values obtained in all tests during the pre-crisis period are higher than those obtained in the crisis period, suggesting that the utility of the models for ES forecasting in the crisis period is more questionable; and *v*) EVT-based models are preferred in terms of ES backtesting in both periods.

6.2 10-day ES forecasting

It is well known that the variance of a Gaussian variable follows a simple scaling law. Indeed, the Basel Committee, in its 1996 Amendment (Basel II), states that it will accept a simple $\sqrt{h}$ scaling of 1-day VaR for deriving the 10-day VaR required in calculating market risk and the related risk capital, and the Basel Committee proposed in 2016 (Basel III) to use the square root of the time scaling rule to calculate ES for risk horizons longer than one-day. However, the stylized facts on financial market volatility and research findings have repeatedly shown that the scaling rule is inappropriate, most likely because financial returns do not follow an i.i.d. stable distribution. Furthermore, this type of scaling for volatility adjustment is incompatible with a mean reversion volatility model because it assumes that volatility remains constant or fluctuates around a local mean over the risk horizon and does not revert to mean at all. Obviously, under the scaling rule, the longer the risk horizon, the higher the error in VaR forecasting.

The standard historical approach is usually limited to the 1-day horizon because we simply do not have enough relevant historical data to use non-overlapping h-day returns when h is 10 or more. On the other hand, using overlapping h-day returns would distort the tail behavior of return distributions, leading to significant error in VaR and ES forecasts at extreme quantiles. We use filtered historical simulation, which allows us to generate a 10-day return distribution from overlapping samples by increasing the number of observations used through a bootstrapping procedure. The drawback is that we can only apply to these multi-period returns the unconditional coverage test, the Righi and Ceretta test, and the two tests by Acerbi and Szekely. Tables A8 - A11 in the online appendix show the average values of VaR and ES forecasts, the violation ratios of the underlying VaR, and backtesting results for the distinct models for 10-day VaR and ES forecasting. Since the out-of-sample period comprises 1260 observations, we have 1250 10-day ES observations that we can compare to the realized 10-day returns.

The results we obtain for 10-day ES prediction can be summarized as follows: *i*) VaR violation rates tend to be below their theoretical values; *ii*) the unconditional coverage test often rejects VaR forecasts, although the rejection only applies to non-EVT based models; *iii*) the filtered historical simulation yields similar 1-day VaR and ES forecasts from the different models, significantly reducing model risk and suggesting the convenience of using this semiparametric approach; *iv*) among conditional models not based on EVT, models with symmetric and models with asymmetric distributions perform similarly at VaR and ES forecasting, and the same observation applies to EVT-based conditional models; *v*) unfortunately, ES backtests again indicate risk overestimation for EVT and non-EVT based models.

7 Value at risk and expected shortfall as indicators for capital adequacy

The January 2016 Basel Committee decision to use expected shortfall as the criterion to evaluate capital adequacy brought up a natural discussion on the preference for value at risk versus expected shortfall as a more appropriate risk measure. Most research on this issue suggests better theoretical properties for expected shortfall (see, among others, Yamai and Yoshiba, 2005). There is also evidence on its superior performance in quiet times, especially at low significance levels, although the evidence seems to indicate that both risk measures show failures in stressed market times.

It is by now clear that the lack of elicitability of ES does not preclude backtesting, as we have seen in this paper and in many others cited in the list of references. Our analysis of expected shortfall backtesting suggests a general overestimation of risk. That may be safe from the regulators' point of view, but it increases costs for financial institutions. In fact, the results we have shown suggest that both value at risk and expected shortfall tend to overestimate risk in tranquil times while having a serious difficulty in estimating the right level of risk in stressed times. To avoid that bias, the large shocks at the start of a crisis starts ¿está bien puesto aquí starts? should feed into the volatility model for market returns to have an effect on value at risk and expected shortfall estimates, since parameter estimates and estimates for the probability distribution of returns change slowly with new information. It may be in that specific situation when we need a good volatility model that can make volatility forecasts adjust quickly to large market shocks.

Analyzing hypotheses of this type in different historical periods and markets seems as an interesting research area. A good estimate of expected shortfall will usually require a good value at risk estimate, but model validation should increasingly rely on backtesting expected shortfall, more so than on backtesting value at risk, since expected shortfall is the basis to determine the level of regulatory capital. Maybe the point is that neither measure should be used by itself. Both provide specific but complementary information, even though a full picture of the risk faced by a particular portfolio investment will usually need a more complete analysis.

An even more important issue refers to the level of precision that can be attained in their estimation, especially in the case of expected shortfall, for which a larger amount of data is needed. As in any statistical estimation, achieving a given level of confidence would require accepting some range of numerical values for value at risk and expected shortfall, and explicitly acknowledging this would have significant consequences for risk regulation and management. More research is needed to find methods to establish such confidence intervals, given the substantive implications that the level of regulatory capital has for financial institutions as well as the need to guarantee the safety of any investment portfolio during stressed times.

8 Conclusions

In spite of the substantial theoretical evidence documenting the superiority of expected shortfall (ES) over VaR as a measure of risk, financial institutions and regulators have only recently embraced ES as an alternative to VaR for financial risk management. One of the major obstacles in this transition has been the unavailability of simple tools for the evaluation of ES forecasts. While the Basel rules for VaR tests are based on counting the number of exceptions, assessing the adequacy of a ES model requires the consideration of the size of tail losses beyond the VaR boundary. In recent years various different approaches have been proposed in the literature for ES backtesting, but, to the best of our knowledge, this paper makes the first extensive comparison of a variety of alternative ES backtesting procedures.

We use the daily market closing prices for IBM, Santander, AXA, and BP from 10/2/2000 to 9/30/2016, and we consider some flexible families of asymmetric distributions for asset returns that include more standard probability distributions as special cases. The normal and Student-t distributions are used as benchmarks. We use an APARCH volatility specification for all assets, because it has the flexibility to deal with the power in the conditional deviation variable as a free parameter, and it includes a number of well-known models as special cases. Once we estimate the

dynamics of returns and the parameters of the probability distribution for return innovations, we forecast returns and volatility applying the standard parametric approach as well as the semiparametric filtered historical simulation (FHS) approach to forecast VaR and ES. Finally, we analyze the performance of 1-day and 10-day ES forecasts obtained from both methods under the different probability distributions.

As the true temporal dependency of financial returns is a complex issue, the standard approach to risk management can be improved by considering a two-step procedure proposed by McNeil and Frey (2000) that applies extreme value theory (EVT). First, return data is filtered with an estimated conditional model for sample returns and their volatility under a given probability distribution and, second, a generalized Pareto distribution for return innovations is estimated after filtering to remove autocorrelation and GARCH effects. If the estimated model is well specified, standardized innovations will have an i.i.d. structure. This two-step procedure leads to a significant improvement in performance, since VaR and ES forecasts then incorporate changes in expected returns and volatility over time. As in the standard approach, we then forecast VaR and ES at different significance levels at 1-day and 10-day horizons and compare the results with those obtained under the standard parametric approach.

In standard conditional models fitted to the full distribution of return innovations, we observe that asymmetric distributions play an important role in capturing tail risk. This is because some stylized facts of financial returns such as volatility clusters, heavy tails, and asymmetry are reflected suitably by these asymmetric distributions. When we apply EVT to return innovations by modeling the tail with a GPD we obtain good ES forecasts regardless of the probability distribution used for returns. It seems that considering the return innovations in the tail of the distribution is more important than discriminating among probability distributions when forecasting ES. Moreover, each combination of APARCH volatility and probability distribution under the EVT approach dominates the similar specification under the standard approach fitted to the full distribution. Conditional EVT models turn out to be more accurate and reliable than standard conditional models not based on EVT both for forecasting VaR as well as for predicting losses beyond VaR.

Except for the Costanzino and Curran and Du and Escanciano tests, which are two-tailed tests, the ES tests we consider focus on a possible undervaluation of risk. We have pointed out that in some cases backtesting does not reject the model specification because the sample evidence is contrary to both the null hypothesis and the alternative hypothesis. In other words, some ES models are not rejected in spite of the fact that they overvalue risk, although the rejection fails by a small amount in most cases. When using ES to build an institution's reserves to cover potential losses in times of crisis, risk undervaluation may be fatal, but overvaluation will lead to inefficient use of capital. This is a relevant consideration that should be taken into account for ES model validation.

We have also shown that using FHS can be very useful. First, qualitative results under FHS are very similar to those obtained under the parametric approach, which is reassuring. EVT-based models dominate non-EVT based models for forecasting both VaR and ES, and asymmetric probability distributions yield more accurate ES forecasts. Second, ES forecasts are much more similar for different probability distributions, and also between forecasts from EVT-based models and non-EVT based models. This implies considerable reduction in model risk, i.e. the uncertainty in ES forecasting because of having alternative model specifications. Given the extreme importance

of these forecasts for capital requirements at financial institutions, reducing model risk is a central issue in tail risk estimation. Furthermore, the evidence on overestimation of risk essentially disappears at the 1-day horizon when we use FHS, but not when forecasting at 10-day horizons.

Other than showing a clear preference for an EVT approach as well as rejecting symmetric probability distributions for modeling return innovations, none of the tests we have considered discriminates much among alternative probability distributions. It seems that once we use EVT, the choice of probability distribution for the non-extreme observations does not seriously condition VaR and ES estimates. However, the recommendation to use FHS under an EVT specification for VaR and ES forecasting and the possibility of exploiting its potential for risk estimation at longer horizons are clear conclusions of this research.

Acknowledgements

The authors gratefully acknowledge financial support from the grants ECO2015-67305-P, PrometeoII /2013/015, Programa de Ayudas a la Investigación from Banco de España, and Programa de Financiación de Universidad Complutense de Madrid - Santander Universidades.

References

Aas, K. and Haff, I.H., The Generalized Hyperbolic Skew Student's t-Distribution. *Journal of Financial Econometrics*, 2006, **4**(2), 275-309.

Alexander, C. and Sheedy, E, Developing a stress testing framework based on market risk models. *Journal of Banking and Finance*, 2008, **32**, 2220-2236.

Artzner, P., Delbaen, F., Eber, J.M. and Heath, M., Coherent measures of risks. *Mathematical Finance*, 1999, **9**, 203-228.

Acerbi, C. and Szekely, B., Backtesting Expected Shortfall. *Publication of MSCI*. https://www.msci.com/www/research-paper/research-insight-backtesting/0128184734, 2014.

Acerbi, C. and Tasche, D., On the coherence of Expected Shortfall. *Journal of Banking and Finance*, 2002, **26**, 1487-1503.

Azzalini, A. and Capitanio, A., Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2003, **65**(2), 367-389.

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, Standards: Minimum Capital requirements for market risk. *Bank for International Settlements*, 2016.

Berkowitz, J., Testing density forecasts, with applications to risk management. *Review of Financial Studies*, 2001, **14**, 371-405.

Berkowitz, J., Christoffersen, P.F. and Pelletier, D., Evaluating Value-at-Risk models with desk-level data. *Management Science*, 2011, **57**(12), 2213-2227.

Carver, L., Mooted VaR substitute cannot be backtested, says top quant. *Risk, March*, 2013, **8**.

Chan, K.F., and Gray, P., Using extreme value theory to measure value-at-risk for daily electricity spot prices. *International Journal of Forecasting*, 2006, **22**(2), 283-300.

Chavez-Demoulin, V. and McGill, J., High-frequency financial data modeling using Hawkes processes. *Journal of Banking and Finance*, 2012, **36**, 3415-3426.

Chen, J.M., Measuring market risk under the basel accords: VaR, stressed VaR, and Expected Shortfall. *Aestimatio, The IEB International Journal of Finance*, 2014, **8**, 184-201.

Christoffersen, P.F., Evaluating interval forecasts. *International Economic Review*, 1998, **39**, 841-862

Clift, S.S., Costanzino, N. and Curran, M., Empirical Performance of Backtesting Methods for Expected Shortfall. http://dx.doi.org/10.2139/ssrn.2618345, 2016.

Cont, R., Deguest, R. and Scandolo, G., Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 2010, **16**(6), 593-291.

Costanzino, N. and Curran, M., Backtesting General Spectral Risk Measures with application to Expected Shortfall. http://dx.doi.org/10.2139/ssrn.2514403, 2015.

Crnkovic, C. and Drachman, J., Quality control. *Risk*, 1996, **9**, 139-143.

Daniels, H.E., Tail Probability Approximations. *International Statistic Review*, 1987, **55**, 37-48.

Daníelsson, J., Jorgensen, B.N., Mandira, S., Samorodnitsky, G. and de Vries, C.G., Subadditivity re-examined: the case for Value-at-Risk. *www.riskresearch.org*, 2005.

Daníelsson, J., Jorgensen, B.N., Samorodnitsky, G., Sarma, M. and de Vries, C.G., Fat tails, VaR and subadditivity. *Journal of Econometrics*, 2013, **172**(2), 283-291.

Daníelsson, J. and de Vries, C.G., Value-at-Risk and extreme returns. *Annales d'Economie et de Statistique*, 2000, 239-270.

Davison, A. and Smith, R., Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B*, 1990, 393-442.

Degiannakis, S., Dent, P. and Floros, C., A Monte Carlo Simulation Approach to Forecasting Multi-period Value-at-Risk and Expected Shortfall Using the FIGARCH-skT Specification. *The Manchester School*, 2014, **82**(1), 71-102.

Degiannakis, S., Floros, C. and Dent, P., Forecasting value-at-risk and expected shortfall using fractionally integrated models of conditional volatility: International evidence. *International Review of Financial Analysis*, 2013, **27**, 21-33.

Diebold, F.X., Gunther, T.A. and Tay, A.S., Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 1998, **39**(4), 863-883.

Diebold, F.X., Schuermann, T. and Stroughair, J.D., Pitfalls and opportunities in the use of extreme value theory in risk management. *The Journal of Risk Finance*, 2000, **50**, 264-272.

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

Du, Z. and Escanciano, J.C., Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 2016, **63**(4), 940-958.

Embrechts, P, Lambrigger, D. and Wüthrich, M., Multivariate extremes and the aggregation of dependence risks: examples and counter-examples. *Extremes*, 2009, **12**(2), 107-127.

Embrechts, P., McNeil, A. and Straumann, D., Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 2002, **1**, 176-223.

Embrechts, P., Nešlehová, J. and Wüthrich, M.V., Additivity properties for Value-at-Risk under Archimedean dependence and heavy-tailedness. *Insurance: Mathematics and Economics*, 2009, **44**(2), 164-169.

Emmer, S., Kratz, M. and Tasche, D., What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk*, 2015, **18**(2).

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

Ergen, I., Two-step methods in VaR prediction and the importance of fat tails. *Quantitative Finance*, 2015, **15**(6), 1013-1030.

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

Fernandez, C. and Steel, M., On Bayesian Modelling of Fat Tails and Skewness. *Journal of the American Statistical Association*, 1998, **93**(441), 359-371.

Fissler, T., Ziegel, J.F. and Gneiting, T., Expected Shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*, 2015.

Garcia, R., Renault, É. and Tsafack, G., Proper conditioning for coherent VaR in portfolio management. *Management Science*, 2007, **53**(3), 483-494.

Garcia-Jorcano, L. and Novales, A., Volatility specifications versus probability distributions in VaR estimation. *Manuscript*, 2017.

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

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

Gerlach, R. and Chen, C. W., Bayesian expected shortfall forecasting incorporating the intraday range, *Journal of Financial Econometrics*, 2014, **14**(1), 128-158.

Gneiting T., Making and evaluation point forecasts. *Journal of the American Statistical Association*, 2011, **106**, 746-762.

Gonzalo, J., and Olmo, J., Which extreme values are really extreme?. *Journal of Financial Econometrics*, 2004, **2**(3), 349-369.

Graham, A. and Pál, J., Backtesting value-at-risk tail losses on a dynamic portfolio. *Journal of Risk Model Validation*, 2014, **8**(2), 59-96.

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

Herrera, R., Energy risk management through self-exciting marked point process. *Energy Economics*, 2013, **38**, 64-76.

Hill, B.M., A simple general approach to inference about the tail of a distribution. *The Annals of Statistic*, 1975, **3**(5), 1163-1174.

Ibragimov, R., New Majorization Theory In Economics And Martingale Convergence Results In Econometrics. *Ph.D. Thesis, Yale University*, 2005.

Ibragimov, R., Portfolio diversification and value at risk under thicktailedness. *Quantitative Finance*, 2009, **9**, 565-580.

Ibragimov, R. and Walden, J., The limits of diversification when losses may be large. *Journal of banking and finance*, 2007, **31**(8), 2551-2569.

Jalal, A. and Rockinger, M., Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. *Journal of Empirical Finance*, 2008, **15**, 868-877.

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

Kerkhof, J. and Melenberg, B., Backtesting for risk-based regulatory capital. *Journal of Banking and Finance*, 2004, **28**, 1845-1865.

Kourouma, L., Dupre, D., Sanfilippo, G. and Taramasco, O., Extreme value at risk and expected shortfall during financial crisis. http://dx.doi.org/10.2139/ssrn.1744091, 2011.

Kuester, K., Mittnik, S. and Paolella, M.S., Value-at-Risk prediction: a comparison of alternative strategies. *Journal of Financial Econometrics*, 2006, **4**, 53-89.

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

Lambert, P. and Laurent, S., Modelling Financial Time Series using GARCH-type models with a skewed student distribution for the innovations. *Mimeo, Université de Liege*, 2001.

Lönnbark, C., Approximation methods for multiple period Value at Risk and Expected Shortfall prediction. *Quantitative Finance*, 2016, **16**(6), 947-968.

Longin, F., The choice of the distribution of asset returns: how extreme value theory can help? *Journal of Banking and Finance*, 2005, **29**, 1017-1035.

Lugannani, R. and Rice, S.O., Saddlepoint approximation for the distribution of the sum of independent random variables. *Advanced Applied Probability*, 1980, **12**, 475-490.

Mabrouk, S. and Saadi, S., Parametric value-at-risk analysis: evidence from stock indices. *The Quartely Review of Economics and Fiannce*, 2012, **52**, 305-321.

Marinelli, C., D'Addona, S. and Rachev, S., A comparison of some univariate models for value-at-risk and expected shortfall. *International Journal of Theoretical and Applied Finance*, 2007, **10**, 1043-1075.

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

McMillan, D.G. and Speight, A.E., Daily volatility forecasts: Reassessing the performance of GARCH models. *Journal of Forecasting*, 2004, **23**(6), 449-460.

McNeil, A. and Frey, R., Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 2000, **7**, 271-300.

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

Pascual, L., Romo, J. and Ruiz, E., Bootstrap prediction for returns and volatilities in garch models. *Computational Statistics and Data Analysis*, 2006, **50**(9), 2293-2312.

Pickands, J., Statistical inference using extreme order statistics. *The Annals of Statistics*, 1975, **3**, 119-131.

Righi, M.B., and Ceretta, P.S., A comparison of Expected Shortfall estimation models. *Journal of Economics and Business*, 2015, **78**, 14-47.

Righi, M.B. and Ceretta, P.S., Individual and flexible Expected Shortfall backtesting. *Journal of Risk Model Validation*, 2013, **7**(3), 3-20.

Rocco, M., Extreme value theory in finance: a survey. *Journal of Economic Surveys*, 2014, **28**(1), 82-108.

Rosenblatt, M., Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 1952, **23**, 470-472.

Ruiz, E. and Pascual, L., Bootstrapping financial time series. *Journal of Economic Surveys*, 2002, **16**(3), 271-300.

Smith, R., Estimating tails of probability distributions. *The Annals of Statistics*, 1987, **15**, 1174-1207.

So, M.K. and Wong, C.M., Estimation of multiple period expected shortfall and median shortfall for risk management. *Quantitative Finance*, 2012, **12**(5), 739-754.

Taylor, J.W., Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics*, 2007, **6**(3), 382-406.

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

Tolikas, K., Unexpected tails in risk measurement: Some international evidence. *Journal of Banking and Finance*, 2014, **40**, 476-493.

Wong, W.K., Fan, G. and Zeng, Y., Capturing tail risks beyond VaR. *Review of Pacific Basin Financial Markets and Polices*, 2012, **15**(03), 1250015.

Wong, W.K., Backtesting value-at-risk based on tail losses. *Journal of Empirical Finance*, 2010, **17**(3), 526-538.

Wong, W.K., Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking and Finance*, 2008, **3**, 1404-1415.

Yamai, Y. and Yoshiba, T., 2005. Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking and Finance*, **29**(4), 997-1015.

Zhou, J., Extreme risk measures for REITs: A comparison among alternative methods. *Applied Financial Economics*, 2012, **22**, 113-126.

I APPENDICES

I.1 Skewed Student-t distribution

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 distribution of Fernandez and Steel and Lambert and Laurent (2001)[25] is

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

where $g(\cdot|\nu)$ is the symmetric (unit variance) Student-t density and ξ is the skewness parameter;[26] m and s^2 are, respectively the mean and the variance of the non-standardized skewed Student-t and are defined by

$$\mathbb{E}(\varepsilon|\xi) = M_1(\xi - \xi^{-1}) \equiv m,$$
$$\mathbb{V}(\varepsilon|\xi) = (M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2 \equiv s^2,$$

where $M_r = 2\int_0^\infty s^r g(s)ds$ is the absolute moment 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.

I.2 Skewed generalized error distribution

An alternative distribution for return innovations which can capture skewness and kurtosis can be based on the generalized error distribution (GED) of Nelson (1991). According to Lambert 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 represents the shape parameter, and $\Gamma(\cdot)$ is the gamma function. The mean (m) and standard deviation (s) are calculated in the same way as for the 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 distribution, when $\kappa = 2$, and the Laplace distribution, when $\kappa = 1$. For $\kappa > 2$ the distribution is platykurtic and for $\kappa < 2$ it is leptokurtic.

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

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

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

I.3 Johnson S_U distribution

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 of the form $Y = \gamma + \delta g((Z-\xi)/\lambda)$. The form of the resulting distribution depends on the choice of function g. When $g(u) = sinh^{-1}(u)$, the distribution is unbounded, and is called the Johnson S_U distribution. The parameters of the distribution are ξ, $\lambda > 0$, γ, and $\delta > 0$.

We use a parametrization[27] of the original Johnson S_U distribution, so that the parameters ξ and λ are the mean and the 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. The parameter δ must be positive and is usually greater than 1.

The pdf of the Johnson S_U, denoted here by $JSU(\xi, \lambda, \gamma, \delta)$, 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 $\mathbb{E}(Z) = \xi$ and $\mathbb{V}(Z) = \lambda^2$.

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

II TABLES

II.1 DESCRIPTIVE STATISTICS

	Mean (bps.)	Median (bps.)	Max	Min	S.D.	Skewness	Kurtosis	J-B
IBM	0.83	0	11.35	-16.89	1.58	-0.22	12.33	15194.87
SAN	1.56	0	20.88	-22.17	2.26	-0.07	10.50	9793.17
AXA	1.47	0	19.78	-20.35	2.69	0.19	10.24	9155.81
BP	-0.69	0	10.58	-14.04	1.69	-0.19	8.01	4390.88

Table 1: Descriptive statistics for daily percent returns. Sample: 10/2/2000 - 9/30/2016 (4175 daily observations). Mean and median returns in basis points. S.D. is the standard deviation. J-B is the Jarque-Bera test statistic.

II.2 BACKTESTING 1-DAY VaR AND ES

IBM									**1% significance level**					
	$\overline{VaR}$	$\overline{ES}$	**Viol**	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-2.83	-3.25	0.014	0.15	-	-	0.32	0.00	0.01	0.01	0.00	0.00	0.65	0.97
ST	-3.15	-4.20	0.010	0.91	-	-	0.95	0.07	0.00	0.00	0.00	0.01	0.72	0.99
SKST	-3.19	-4.27	0.010	0.91	-	-	0.95	0.07	0.00	0.00	0.00	0.02	0.73	0.99
SGED	-3.20	-3.92	0.010	0.91	-	-	0.95	0.02	0.00	0.00	0.00	0.01	0.72	0.99
JSU	-3.23	-4.21	0.010	0.91	-	-	0.95	0.06	0.00	0.00	0.00	0.02	0.73	0.99
N-EVT	-3.52	-5.93	0.010	0.91	-	-	0.96	**0.38**	**0.96**	**1.00**	0.30	0.26	0.76	0.99
ST-EVT	-3.59	-6.06	0.010	0.91	-	-	0.94	**0.35**	**0.92**	**0.99**	0.30	0.26	0.77	0.99
SKST-EVT	-3.58	-6.05	0.010	0.91	-	-	0.94	**0.35**	**0.93**	**1.00**	0.30	0.26	0.77	0.99
SGED-EVT	-3.56	-5.92	0.010	0.91	-	-	0.95	**0.34**	**0.89**	**0.98**	0.29	0.25	0.77	0.99
JSU-EVT	-3.58	-6.05	0.010	0.91	-	-	0.94	**0.34**	**0.90**	**1.00**	0.30	0.26	0.76	0.99
IBM									**2.5% significance level**					
	$\overline{VaR}$	$\overline{ES}$	**Viol**	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-2.38	-2.85	0.022	0.52	0.15	0.28	0.34	0.01	0.01	0.03	0.00	0.13	0.79	0.84
ST	-2.39	-3.30	0.024	0.79	0.19	0.41	0.31	0.13	0.03	0.05	0.01	0.29	0.86	0.89
SKST	-2.42	-3.35	0.021	0.41	0.58	0.61	0.24	0.12	0.01	0.08	0.03	0.39	1.00	0.91
SGED	-2.52	-3.25	0.018	0.11	0.42	0.20	0.30	0.05	0.01	0.32	0.00	0.48	0.63	0.91
JSU	-2.46	-3.35	0.019	0.16	0.46	0.28	0.28	0.10	0.01	0.32	0.02	0.46	0.81	0.92
N-EVT	-2.37	-4.04	0.022	0.52	0.15	0.28	0.34	**0.36**	**0.79**	**0.95**	0.49	0.35	0.70	0.91
ST-EVT	-2.41	-4.12	0.022	0.52	0.62	0.72	0.35	0.31	0.68	0.94	0.47	0.38	1.00	0.93
SKST-EVT	-2.41	-4.12	0.024	0.79	0.19	0.41	0.31	**0.34**	**0.73**	**0.96**	0.47	0.38	0.99	0.93
SGED-EVT	-2.40	-4.06	0.022	0.52	0.62	0.72	0.61	0.32	0.66	0.92	0.48	0.35	0.69	0.91
JSU-EVT	-2.40	-4.12	0.023	0.65	0.66	0.82	0.54	0.32	0.68	0.99	0.47	0.38	1.00	0.93
IBM									**5% significance level**					
	$\overline{VaR}$	$\overline{ES}$	**Viol**	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-2.00	-2.51	0.037	0.03	0.35	0.06	0.35	0.05	0.04	0.49	0.00	0.17	0.37	0.55
ST	-1.86	-2.70	0.044	0.29	0.30	0.33	0.54	0.18	0.13	0.40	0.09	0.33	0.11	0.41
SKST	-1.88	-2.73	0.044	0.29	0.30	0.33	0.54	0.20	0.11	0.39	0.16	0.22	0.11	0.40
SGED	-1.98	-2.74	0.037	0.03	0.35	0.06	0.24	0.15	0.07	0.75	0.02	0.06	0.36	0.58
JSU	-1.91	-2.75	0.040	0.11	0.20	0.12	0.32	0.18	0.11	0.67	0.15	0.14	0.13	0.42
N-EVT	-1.74	-3.00	0.053	0.61	0.19	0.38	0.36	**0.48**	**0.69**	**0.90**	0.52	0.41	0.41	0.67
ST-EVT	-1.76	-3.05	0.049	0.90	0.53	0.81	0.66	**0.40**	**0.73**	**0.97**	0.54	0.37	0.13	0.46
SKST-EVT	-1.76	-3.05	0.050	1.00	0.29	0.57	0.56	**0.41**	**0.76**	**0.96**	0.54	0.37	0.13	0.46
SGED-EVT	-1.75	-3.01	0.053	0.61	0.41	0.62	0.56	**0.44**	**0.69**	**0.96**	0.54	0.36	0.23	0.59
JSU-EVT	-1.75	-3.05	0.051	0.90	0.31	0.60	0.55	**0.41**	**0.63**	**0.93**	0.54	0.37	0.13	0.45

Table 2: Mean VaR forecasts ($\overline{VaR}$), mean ES forecasts ($\overline{ES}$), violation ratio (Viol), and backtesting results (p-values) for VaR and ES forecasts for IBM. LR_{uc} is the unconditional coverage test of Kupiec (1995), LR_{ind} and LR_{cc} are the independence and conditional coverage tests of Christoffersen (1998), DQT is the dynamic quantile test of Engle and Manganelli (2004), BT_T is the test of Righi and Ceretta (2015), Z_1 and Z_2 are the tests of Acerbi and Szekely (2014), TR is the test of Graham and Pál (2014), and U_{ES}, $C_{ES}(1)$, and $C_{ES}(5)$ are the unconditional and the conditional ($lags = 1$ and $lags = 5$) tests of Costanzino and Curran (2015) and Du and Escanciano (2016). A p-value in bold indicates that the statistic obtained in this test has a sign opposite to that specified for the alternative hypothesis. Hyphens indicate that the independence and conditional coverage tests cannot be applied.

SAN	1% significance level													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-4.64	-5.32	0.021	0.00	-	-	0.00	0.05	0.01	0.00	0.00	0.00	0.58	0.06
ST	-4.98	-6.14	0.015	0.09	-	-	0.01	0.14	0.01	0.00	0.01	0.06	0.70	0.00
SKST	-5.17	-6.40	0.014	0.15	-	-	0.01	0.16	0.01	0.00	0.03	0.28	0.77	0.00
SGED	-5.17	-6.17	0.014	0.15	-	-	0.01	0.11	0.01	0.01	0.00	0.22	0.77	0.00
JSU	-5.24	-6.47	0.013	0.24	-	-	0.02	0.17	0.01	0.00	0.02	0.35	0.79	0.00
N-EVT	-5.67	-7.76	0.009	0.64	-	-	0.01	**0.96**	**0.99**	**1.00**	0.52	0.24	0.84	0.09
ST-EVT	-5.73	-7.81	0.009	0.64	-	-	0.02	**0.96**	**0.99**	**1.00**	0.52	0.22	0.85	0.07
SKST-EVT	-5.73	-7.77	0.009	0.64	-	-	0.02	**0.96**	**0.97**	**0.99**	0.52	0.22	0.85	0.06
SGED-EVT	-5.72	-7.72	0.009	0.64	-	-	0.02	**0.97**	**1.00**	**1.00**	0.52	0.22	0.85	0.08
JSU-EVT	-5.73	-7.77	0.009	0.64	-	-	0.02	**0.96**	**0.99**	**0.99**	0.52	0.22	0.85	0.07
SAN	**2.5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-3.91	-4.67	0.033	0.10	0.69	0.24	0.00	0.08	0.02	0.02	0.00	0.00	0.44	0.52
ST	-3.97	-5.10	0.033	0.10	0.54	0.22	0.00	0.20	0.06	0.01	0.01	0.04	0.43	0.35
SKST	-4.11	-5.29	0.028	0.53	0.81	0.80	0.03	0.19	0.09	0.07	0.04	0.20	0.44	0.24
SGED	-4.17	-5.23	0.027	0.66	0.80	0.88	0.02	0.17	0.04	0.03	0.00	0.27	0.46	0.20
JSU	-4.14	-5.35	0.027	0.66	0.80	0.88	0.02	0.20	0.03	0.03	0.05	0.28	0.45	0.20
N-EVT	-4.19	-5.80	0.028	0.53	0.81	0.80	0.03	**0.96**	**0.96**	**0.99**	0.53	0.43	0.48	0.17
ST-EVT	-4.23	-5.84	0.026	0.79	-	-	0.01	**0.94**	**0.95**	**0.99**	0.53	0.42	0.48	0.14
SKST-EVT	-4.22	-5.81	0.026	0.79	-	-	0.01	**0.94**	**0.93**	**0.95**	0.54	0.41	0.48	0.14
SGED-EVT	-4.22	-5.79	0.026	0.79	-	-	0.01	**0.95**	**0.94**	**0.97**	0.54	0.41	0.48	0.15
JSU-EVT	-4.22	-5.82	0.026	0.79	-	-	0.01	**0.94**	**0.96**	**0.99**	0.54	0.41	0.48	0.14
SAN	**5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-3.29	-4.12	0.052	0.70	0.34	0.59	0.08	0.14	0.06	0.06	0.00	0.07	0.70	0.52
ST	-3.21	-4.32	0.056	0.31	0.54	0.50	0.03	0.23	0.15	0.06	0.01	0.09	0.99	0.48
SKST	-3.30	-4.48	0.050	1.00	0.43	0.73	0.13	0.23	0.06	0.05	0.09	0.34	0.82	0.52
SGED	-3.36	-4.48	0.050	1.00	0.43	0.73	0.13	0.24	0.12	0.16	0.01	0.47	0.60	0.52
JSU	-3.31	-4.52	0.050	1.00	0.43	0.73	0.13	0.24	0.12	0.13	0.11	0.43	0.77	0.53
N-EVT	-3.26	-4.59	0.053	0.61	0.32	0.53	0.08	**0.93**	**0.95**	**0.97**	0.50	0.49	0.72	0.60
ST-EVT	-3.29	-4.61	0.052	0.70	0.34	0.59	0.08	**0.91**	**0.92**	**0.97**	0.51	0.50	0.82	0.53
SKST-EVT	-3.28	-4.59	0.052	0.70	0.34	0.59	0.08	**0.92**	**0.90**	**0.94**	0.51	0.50	0.86	0.55
SGED-EVT	-3.27	-4.57	0.052	0.70	0.34	0.59	0.08	**0.92**	**0.98**	**0.99**	0.51	0.49	0.77	0.57
JSU-EVT	-3.28	-4.59	0.052	0.70	0.34	0.59	0.08	**0.91**	**0.93**	**0.96**	0.51	0.50	0.84	0.55

Table 3: Mean VaR forecasts ($\overline{VaR}$), mean ES forecasts ($\overline{ES}$), violation ratio (Viol), and backtesting results (p-values) for VaR and ES forecasts for SAN. LR_{uc} is the unconditional coverage test of Kupiec (1995), LR_{ind} and LR_{cc} are the independence and conditional coverage tests of Christoffersen (1998), DQT is the dynamic quantile test of Engle and Manganelli (2004), BT_T is the test of Righi and Ceretta (2015), Z_1 and Z_2 are the tests of Acerbi and Szekely (2014), TR is the test of Graham and Pál (2014), and U_{ES}, $C_{ES}(1)$, and $C_{ES}(5)$ are the unconditional and the conditional ($lags = 1$ and $lags = 5$) tests of Costanzino and Curran (2015) and Du and Escanciano (2016). A p-value in bold indicates that the statistic obtained in this test has a sign opposite to that specified for the alternative hypothesis. Hyphens indicate that the independence and conditional coverage tests cannot be applied.

AXA	1% significance level													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-4.40	-5.04	0.021	0.00	0.11	0.00	0.00	0.13	0.03	0.00	0.00	0.00	0.68	0.84
ST	-4.65	-5.62	0.017	0.03	-	-	0.01	0.25	0.06	0.01	0.03	0.08	0.70	0.99
SKST	-4.75	-5.77	0.013	0.36	-	-	0.13	0.19	0.01	0.00	0.09	0.25	0.75	0.99
SGED	-4.76	-5.61	0.014	0.15	-	-	0.06	0.18	0.00	0.00	0.00	0.21	0.76	0.99
JSU	-4.79	-5.81	0.012	0.51	-	-	0.66	0.20	0.02	0.01	0.09	0.32	0.76	0.99
N-EVT	-5.43	-6.40	0.009	0.64	-	-	0.95	**1.00**	**1.00**	**1.00**	**0.66**	0.12	0.84	1.00
ST-EVT	-5.45	-6.46	0.009	0.64	-	-	0.95	**1.00**	**1.00**	**1.00**	**0.66**	0.13	0.84	1.00
SKST-EVT	-5.45	-6.45	0.009	0.64	-	-	0.95	**1.00**	**0.99**	**0.99**	**0.66**	0.13	0.84	1.00
SGED-EVT	-5.43	-6.43	0.009	0.64	-	-	0.95	**1.00**	**0.99**	**1.00**	**0.66**	0.12	0.84	1.00
JSU-EVT	-5.44	-6.45	0.009	0.64	-	-	0.95	**1.00**	**1.00**	**1.00**	**0.66**	0.13	0.84	1.00
AXA	**2.5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-3.71	-4.42	0.034	0.05	0.61	0.13	0.13	0.15	0.05	0.03	0.00	0.00	0.41	0.55
ST	-3.75	-4.73	0.033	0.07	0.58	0.17	0.16	0.24	0.08	0.03	0.01	0.01	0.33	0.55
SKST	-3.83	-4.85	0.032	0.14	0.50	0.27	0.28	0.26	0.09	0.02	0.06	0.06	0.29	0.58
SGED	-3.89	-4.80	0.032	0.14	0.50	0.27	0.29	0.26	0.04	0.99	0.01	0.11	0.35	0.64
JSU	-3.86	-4.88	0.032	0.14	0.50	0.27	0.28	0.28	0.11	0.00	0.07	0.10	0.29	0.60
N-EVT	-4.09	-4.99	0.027	0.66	0.30	0.53	0.49	**1.00**	**0.99**	**1.00**	**0.70**	0.27	0.57	0.82
ST-EVT	-4.11	-5.03	0.027	0.66	0.30	0.53	0.54	**1.00**	**1.00**	**1.00**	**0.68**	0.31	0.31	0.73
SKST-EVT	-4.10	-5.02	0.027	0.66	0.30	0.53	0.54	**1.00**	**0.97**	**0.99**	**0.69**	0.30	0.29	0.72
SGED-EVT	-4.09	-5.01	0.026	0.79	0.27	0.52	0.56	**1.00**	**0.97**	**1.00**	**0.69**	0.29	0.37	0.75
JSU-EVT	-4.10	-5.02	0.027	0.66	0.30	0.53	0.54	**1.00**	**0.95**	**1.00**	**0.69**	0.30	0.30	0.72
AXA	**5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-3.12	-3.91	0.056	0.37	0.73	0.63	0.72	0.20	0.13	0.07	0.00	0.01	0.66	0.52
ST	-3.06	-4.05	0.057	0.25	0.73	0.49	0.74	0.26	0.24	0.10	0.01	0.02	0.70	0.59
SKST	-3.12	-4.14	0.056	0.31	0.73	0.56	0.76	0.29	0.12	0.07	0.07	0.13	0.61	0.56
SGED	-3.17	-4.15	0.052	0.70	0.66	0.84	0.70	0.27	0.08	0.04	0.02	0.24	0.55	0.55
JSU	-3.13	-4.16	0.056	0.37	0.73	0.63	0.82	0.29	0.17	0.08	0.09	0.17	0.59	0.56
N-EVT	-3.11	-3.96	0.056	0.37	0.73	0.63	0.72	**1.00**	**0.96**	**1.00**	**0.57**	0.45	0.58	0.51
ST-EVT	-3.14	-4.01	0.055	0.44	0.72	0.70	0.95	**1.00**	**0.95**	**0.98**	**0.56**	0.44	0.55	0.55
SKST-EVT	-3.13	-3.99	0.056	0.37	0.73	0.63	0.82	**1.00**	**0.95**	**0.98**	**0.56**	0.45	0.55	0.55
SGED-EVT	-3.12	-3.98	0.056	0.37	0.73	0.63	0.72	**1.00**	**0.96**	**0.99**	**0.56**	0.45	0.55	0.52
JSU-EVT	-3.13	-3.99	0.056	0.37	0.73	0.63	0.82	**1.00**	**0.95**	**0.95**	**0.57**	0.45	0.55	0.55

Table 4: Mean VaR forecasts ($\overline{VaR}$), mean ES forecasts ($\overline{ES}$), violation ratio (Viol), and backtesting results (p-values) for VaR and ES forecasts for AXA. LR_{uc} is the unconditional coverage test of Kupiec (1995), LR_{ind} and LR_{cc} are the independence and conditional coverage tests of Christoffersen (1998), DQT is the dynamic quantile test of Engle and Manganelli (2004), BT_T is the test of Righi and Ceretta (2015), Z_1 and Z_2 are the tests of Acerbi and Szekely (2014), TR is the test of Graham and Pál (2014), and U_{ES}, $C_{ES}(1)$, and $C_{ES}(5)$ are the unconditional and the conditional ($lags = 1$ and $lags = 5$) tests of Costanzino and Curran (2015) and Du and Escanciano (2016). A p-value in bold indicates that the statistic obtained in this test has a sign opposite to that specified for the alternative hypothesis. Hyphens indicate that the independence and conditional coverage tests cannot be applied.

BP	1% significance level													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-3.23	-3.70	0.031	0.15	-	-	0.34	0.06	0.03	0.02	0.00	0.00	0.69	0.98
ST	-3.46	-4.28	0.012	0.51	-	-	0.83	0.20	0.01	0.00	0.06	0.12	0.74	0.99
SKST	-3.53	-4.37	0.012	0.51	-	-	0.83	0.22	0.00	0.01	0.11	0.20	0.76	0.99
SGED	-3.54	-4.21	0.011	0.70	-	-	0.94	0.16	0.02	0.04	0.02	0.19	0.78	1.00
JSU	-3.56	-4.38	0.012	0.51	-	-	0.83	0.24	0.02	0.01	0.10	0.24	0.77	0.99
N-EVT	-3.75	-4.58	0.007	0.28	-	-	0.99	**0.64**	**0.99**	**1.00**	0.39	0.45	0.82	1.00
ST-EVT	-3.81	-4.70	0.007	0.28	-	-	0.99	**0.55**	**0.98**	**0.99**	0.41	0.47	0.81	1.00
SKST-EVT	-3.80	-4.70	0.007	0.28	-	-	0.99	**0.56**	**0.99**	**1.00**	0.41	0.47	0.81	1.00
SGED-EVT	-3.77	-4.65	0.007	0.28	-	-	0.99	**0.58**	**0.98**	**1.00**	0.41	0.45	0.82	1.00
JSU-EVT	-3.80	-4.69	0.007	0.28	-	-	0.99	**0.56**	**0.99**	**1.00**	0.41	0.46	0.81	1.00
BP	**2.5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-2.72	-3.24	0.031	0.19	0.75	0.40	0.38	0.18	0.07	0.03	0.00	0.05	1.00	0.89
ST	-2.76	-3.55	0.029	0.33	0.80	0.61	0.61	0.29	0.13	0.04	0.10	0.23	0.83	0.92
SKST	-2.81	-3.62	0.028	0.53	0.81	0.80	0.76	0.30	0.18	0.06	0.20	0.42	0.78	0.94
SGED	-2.86	-3.58	0.025	0.93	0.73	0.94	0.93	0.26	0.07	0.25	0.09	0.44	0.77	0.94
JSU	-2.83	-3.64	0.028	0.53	0.81	0.80	0.76	0.31	0.20	0.11	0.22	0.50	0.77	0.94
N-EVT	-2.85	-3.63	0.024	0.79	0.69	0.89	0.93	**0.90**	**0.97**	**0.99**	**0.59**	0.25	0.88	0.94
ST-EVT	-2.89	-3.70	0.025	0.93	0.76	0.95	0.86	**0.80**	**0.96**	**0.99**	0.58	0.27	0.69	0.95
SKST-EVT	-2.88	-3.70	0.026	0.79	0.79	0.93	0.81	**0.81**	**0.95**	**1.00**	0.58	0.28	0.70	0.94
SGED-EVT	-2.86	-3.67	0.025	0.93	0.76	0.95	0.91	**0.84**	**0.98**	**0.99**	0.58	0.27	0.80	0.95
JSU-EVT	-2.88	-3.70	0.025	0.93	0.76	0.95	0.87	**0.81**	**0.99**	**1.00**	0.58	0.28	0.71	0.95
BP	**5% significance level**													
	$\overline{VaR}$	$\overline{ES}$	Viol	LR_{uc}	LR_{ind}	LR_{cc}	DQT	BT_T	Z_1	Z_2	TR	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$
N	-2.28	-2.86	0.044	0.29	0.27	0.31	0.62	0.17	0.11	0.49	0.00	0.24	0.69	0.66
ST	-2.23	-3.01	0.047	0.60	0.19	0.37	0.52	0.26	0.12	0.39	0.13	0.28	0.79	0.60
SKST	-2.27	-3.06	0.044	0.36	0.25	0.34	0.61	0.26	0.11	**0.63**	0.24	0.43	0.79	0.64
SGED	-2.31	-3.07	0.043	0.23	0.29	0.28	0.63	0.27	0.16	**0.81**	0.18	0.40	0.75	0.68
JSU	-2.27	-3.08	0.044	0.36	0.25	0.34	0.61	0.27	0.13	**0.74**	0.27	0.49	0.79	0.65
N-EVT	-2.20	-2.94	0.049	0.90	0.46	0.75	0.54	**0.81**	**0.90**	**0.96**	**0.55**	0.43	0.74	0.64
ST-EVT	-2.23	-3.00	0.048	0.79	0.49	0.76	0.79	**0.70**	**0.90**	**0.97**	0.53	0.44	0.82	0.63
SKST-EVT	-2.22	-2.99	0.048	0.70	0.18	0.37	0.70	**0.71**	**0.90**	**0.99**	0.52	0.45	0.81	0.63
SGED-EVT	-2.21	-2.97	0.048	0.79	0.49	0.76	0.57	**0.73**	**0.88**	**0.97**	0.53	0.45	0.79	0.64
JSU-EVT	-2.22	-2.99	0.047	0.60	0.19	0.37	0.52	**0.70**	**0.93**	**0.98**	0.53	0.45	0.81	0.63

Table 5: Mean VaR forecasts ($\overline{VaR}$), mean ES forecasts ($\overline{ES}$), violation ratio (Viol), and backtesting results (p-values) for VaR and ES forecasts for BP. LR_{uc} is the unconditional coverage test of Kupiec (1995), LR_{ind} and LR_{cc} are the independence and conditional coverage tests of Christoffersen (1998), DQT is the dynamic quantile test of Engle and Manganelli (2004), BT_T is the test of Righi and Ceretta (2015), Z_1 and Z_2 are the tests of Acerbi and Szekely (2014), TR is the test of Graham and Pál (2014), and U_{ES}, $C_{ES}(1)$, and $C_{ES}(5)$ are the unconditional and the conditional ($lags = 1$ and $lags = 5$) tests of Costanzino and Curran (2015) and Du and Escanciano (2016). A p-value in bold indicates that the statistic obtained in this test has a sign opposite to that specified for the alternative hypothesis. Hyphens indicate that the independence and conditional coverage tests cannot be applied.

II.3 DESCRIPTIVE ANALYSIS OF VIOLATIONS

	IBM		**SAN**		**AXA**		**BP**	
	JSU	**JSU-EVT**	**JSU**	**JSU-EVT**	**JSU**	**JSU-EVT**	**JSU**	**JSU-EVT**
$n\alpha$	12.6	12.6	12.6	12.6	12.6	12.6	12.6	12.6
V(0.01)	13	13	17	11	15	11	15	9
CV(0.01)	10.41	7.63	7.06	4.73	7.28	3.96	7.74	6.11
$n\alpha$	31.5	31.5	31.5	31.5	31.5	31.5	31.5	31.5
V(0.025)	24	29	34	33	40	34	35	32
CV(0.025)	16.05	14.72	17.63	15.04	19.78	14.04	15.73	13.83
$n\alpha$	63	63	63	63	63	63	63	63
V(0.05)	51	64	63	66	70	70	56	59
CV(0.05)	26.56	30.03	32.32	31.48	35.77	32.02	31.65	30.93

Table 6: Descriptive analysis of violations under the AR(1)-APARCH(1,1)-JSU and EVT-AR(1)-APARCH(1,1)-JSU models. V and CV denote the number of violations and cumulative violations, respectively.

III FIGURES

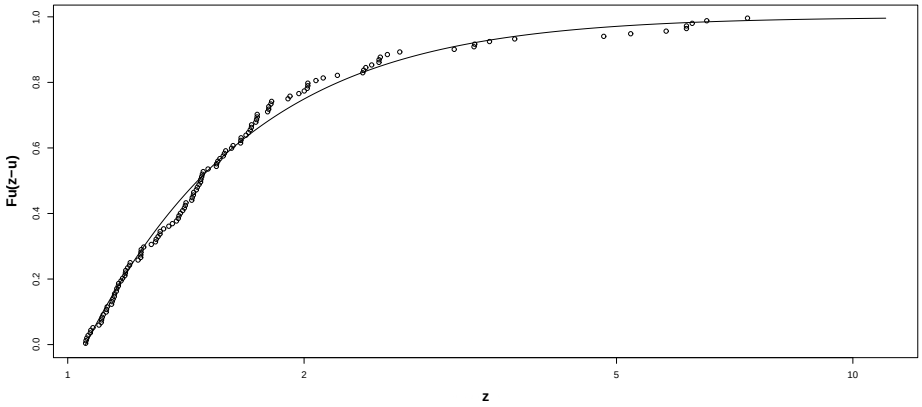

Figure 1: Empirical distribution of threshold excesses for IBM filtered residuals under the EVT-AR(1)-APARCH(1,1)-JSU model versus the fitted GPD.

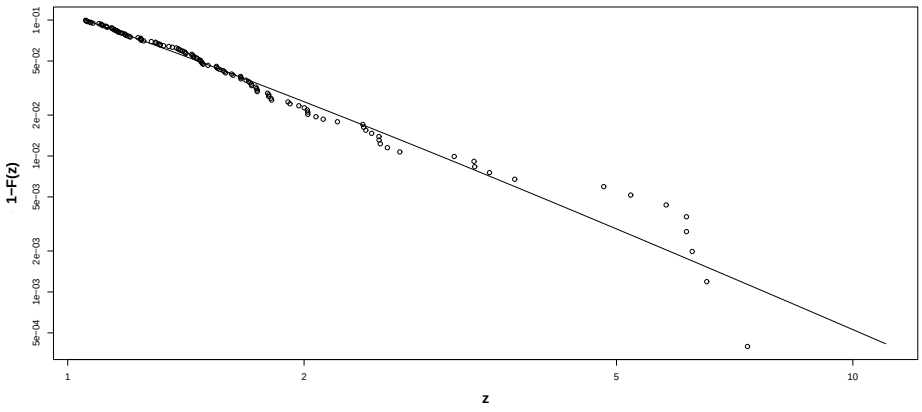

Figure 2: The smooth curve through the points shows the estimated tail of filtered residuals for IBM under the AR(1)-APARCH(1,1)-JSU model using the tail estimator. Points are plotted at empirical tail probabilities calculated from the empirical distribution function.

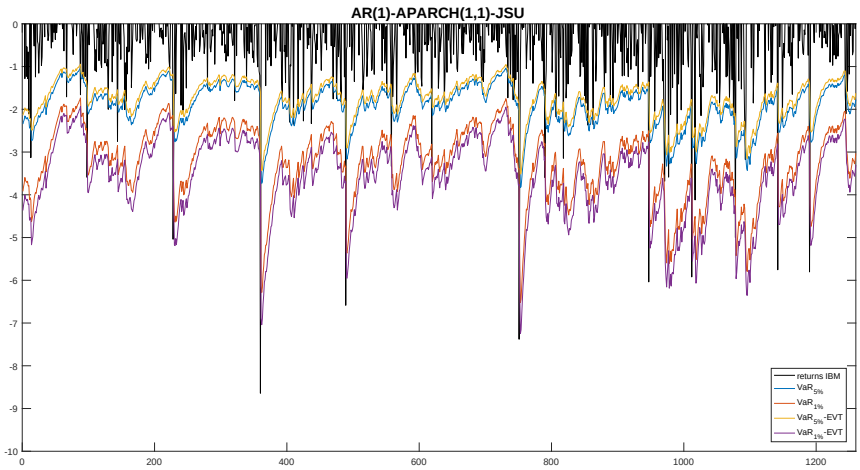

Figure 3: IBM daily percent returns and $VaR_{1\%}$ and $VaR_{5\%}$ forecasts with the full sample as well as using only extreme values. We only show the negative returns so as to maintain a clear perspective on the different VaR estimates.

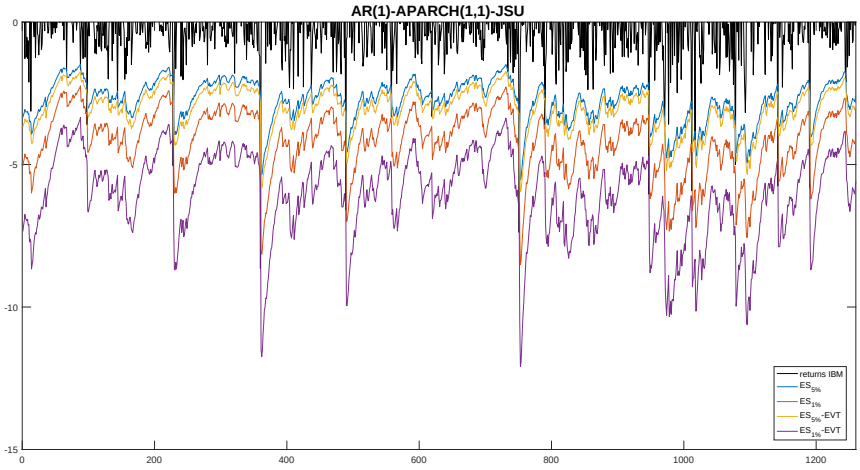

Figure 4: IBM daily percent returns and $ES_{1\%}$ and $ES_{5\%}$ forecasts with the full sample as well as using only extreme values. We only show the negative returns so as to maintain a clear perspective on the different ES estimates.

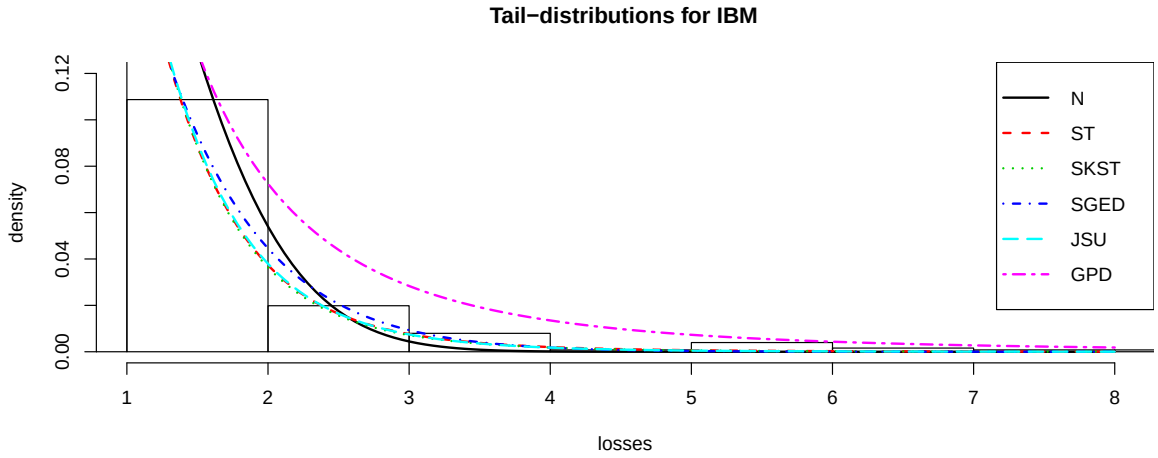

Figure 5: Estimated tail-distributions for IBM. N is the normal, ST is the Student-t (4.67), SKST is the skewed Student-t (0.97, 4.69), SGED is the skewed generalized error (0.99, 1.15), and JSU is the Johnson SU (-0.092, 1.53) distribution. Numerical estimates for the parameters are enclosed in brackets.

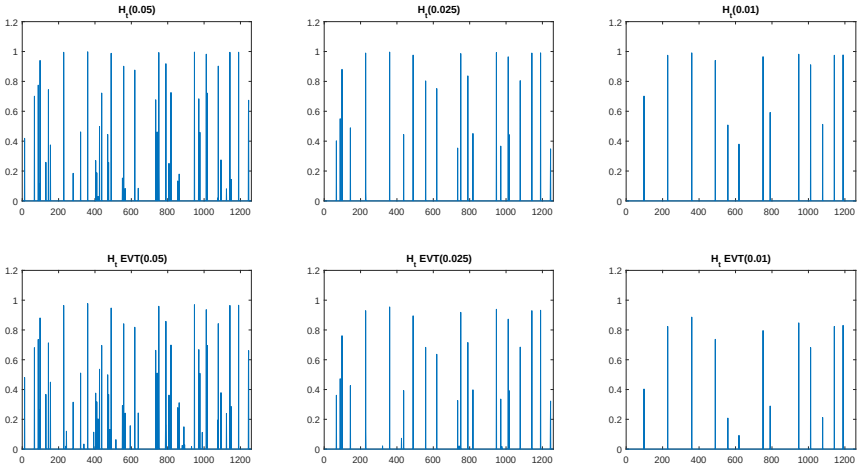

Figure 6: Cumulative hits (violations) of IBM under the JSU-APARCH and JSU-EVT-APARCH models for different α.

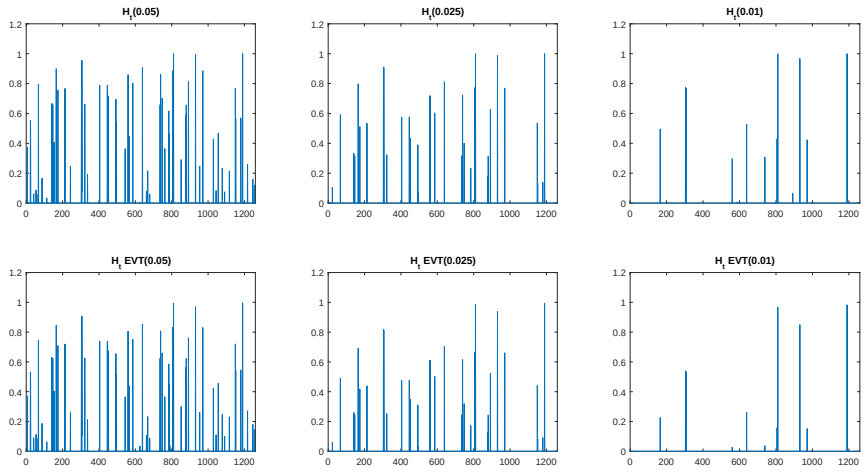

Figure 7: Cumulative hits (violations) of SAN under the JSU-APARCH and JSU-EVT-APARCH models for different α.

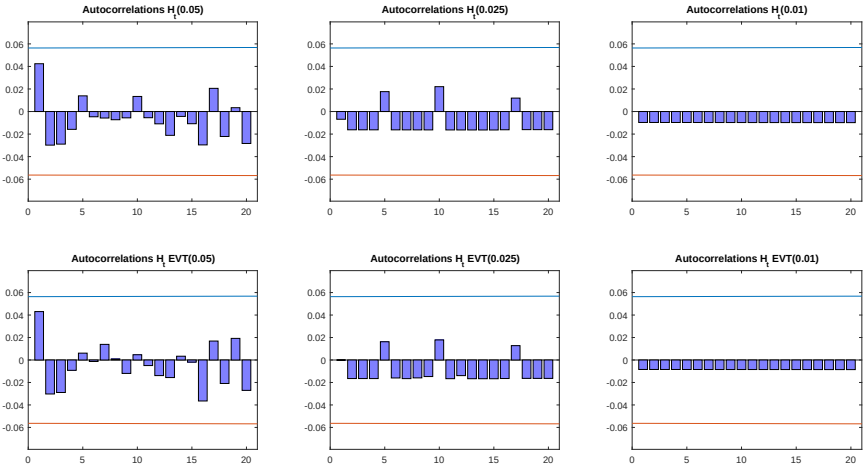

Figure 8: Sample autocorrelations of cumulative hits (violations) of IBM under the JSU-APARCH and JSU-EVT-APARCH models for different values of α.

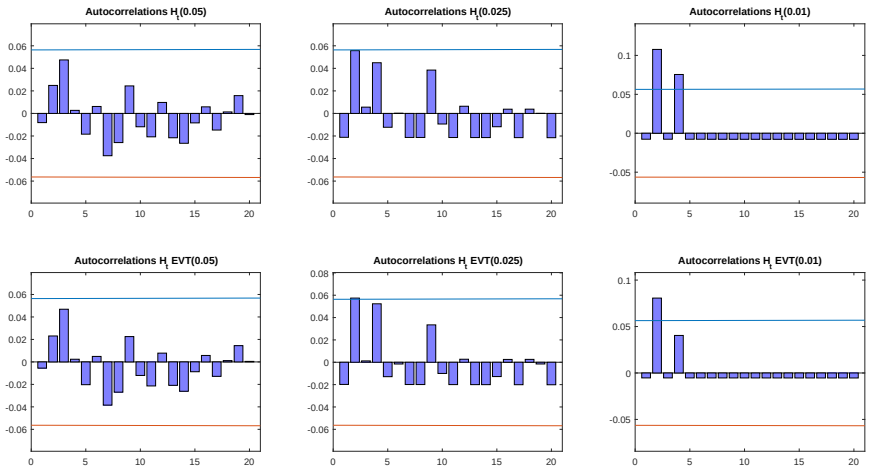

Figure 9: Sample autocorrelations of cumulative hits (violations) of SAN under the JSU-APARCH and JSU-EVT-APARCH models for different values of α.