ON THE PERFORMANCE OF BACKTESTING PROCEDURES FOR EXPECTED SHORTFALL

Pedro Fernández García

Trabajo de investigación 019/007

Master en Banca y Finanzas Cuantitativas

Director/a: Dr. Ángel León Valle
Dr. Juan Mora López

Universidad Complutense de Madrid

Universidad del País Vasco

Universidad de Valencia

Universidad de Castilla-La Mancha

www.finanzascuantitativas.com

On the performance of backtesting procedures for Expected Shortfall

Author: Pedro Fernández García

Supervisors: Angel León Valle and Juan Mora López

Master in Quantitative Finance and Banking

Universidad Complutense de Madrid
Universidad del País Vasco
Universidad de Valencia
Universidad de Castilla-La Mancha

July 2019

Abstract

The performance of various recently proposed backtesting procedure for expected shortfall is compared through several Monte Carlo experiments with rolling-window estimates, which mimic the way how these procedures are used in practice. Also there is an application in precious metals to show empirical evidence.

Keywords: Backtesting, Expected Shortfall, Value at Risk, violations, rejections

Contents

1 Introduction

In recent years, the banking legislation on capital requirements has been harder and harder to prevent financial entities bankruptcies, specially from the economical and financial crisis of 2008. For that purpose the Basel Committee on Banking Supervision established rules to calculate risk measures. The Value at Risk (VaR), which is the maximum possible loss with a certain confidence level, was used for a long time to measure risk because it has good properties. Some of those properties are monotonicity (if a portfolio has systematically lower value than another the first has less risk), translation invariance (if it is added a cash amount to the portfolio, its risk is reduced in that amount), positive homogeneity (if it is increased the size of the portfolio, its risk increase in the same proportion) or elicitability (the risk measure can be defined as the result of minimizing a expected scoring function). However, VaR does not accomplish one of the conditions a risk measure must satisfy and specifically, the subadditivity property (i.e., the portfolio risk must be lower than the sum of the individual risks of each component, what proves that diversification is beneficial). For that reason, together with the impossibility to differentiate between two portfolios with the same VaR but one with worse losses at the tail, the Expected Shortfall (ES) has become the new regulation risk measure. Even if it is not elicitable, the ES verifies the properties of a coherent risk measure and so, it does solve the VaR tail problems.

In short, financial companies started to implement programs to forecast ES in addition to VaR to make sure that their calculations were well done. Many researchers obtained a variety of ES backtests, each based on different premises. The main objective of this paper is to compare some of those ES backtests trying to obtain a kind of ranking between them for small samples.

This paper is structured in different sections to reach that objective. In Section 2 we validate the econometric programs and introduce the models and the risk measures that we are going to use in the paper. Section 3 presents the different ES backtests that we will compare in this study. Section 4 includes the results of Monte Carlo experiments, both in size and in power. Section 5 provides an empirical application of backtesting ES with precious metal prices. Section 6 concludes. Finally, Sections 7 and 8 are the references and the annex, respectively.

2 Model for returns, risk measures and quality validation

In this study, we calculate the VaR and ES risk measures under the parametric method. To do it, we estimate the parameters for the mean model, the variance model and the innovations distribution. To fit the parameters, we will use the Maximum Likelihood method by using Matlab. To ensure that the econometric programs created to estimate the parameters are enough powerful with a reduced sample, we simulate alternative asset return series.

2.1 Model for returns and risk measures

We model the asset return dynamics as in Acereda, León and Mora (2019). Specifically, the conditional variance is driven by the NGARCH(1,1) model. It is an extension of the popular GARCH (1,1) model that incorporates a new parameter (c) in order to fit the leverage effect according to the empirical evidence. With respect to the mean, we will use an AR(1) model like most of the financial papers do. Regarding the innovations, we will use three different distributions: the standard Normal, the Student t and the Skewed t by Hansen (1994). The Normal distribution is symmetric and with excess kurtosis of zero (i.e., kurtosis is three), the Student t is symmetric and exhibits fatter tails than the Normal (i.e., kurtosis higher than three) and the Skewed t which nests the Student t since it allows to capture skewness. In short, the asset return model is given by

$$\begin{aligned} \text{Returns} &: R_t = \mu_t + \sigma_t Z_t \\ \text{AR(1)} &: \mu_t = \phi_0 + \phi_1 R_{t-1} \\ \text{NGARCH(1,1)} &: \sigma_t^2 = b_0 + b_1\sigma_{t-1}^2 + b_2(R_{t-1} - \mu_{t-1} - c\sigma_{t-1})^2 \end{aligned}$$

$$\begin{aligned} \text{Innovations} : Z_t \sim \quad & N(m_Z = 0, v_Z = 1) \text{ or} \\ & t_k(m_Z = 0, v_Z = 1) \text{ or} \\ & skt_{k,s}(m_Z = 0, v_Z = 1) \end{aligned}$$

We define R_t as the asset return at time t, μ_t is the conditional mean process and σ_t^2 is the conditional variance process and Z_t is a random variable (innovation) such that $Z_t \sim iid$ (independent and identically distributed). Note that Z_t follows one of the three distributions presented previously. All these innovations are standardized to have mean (m_Z) of zero and variance (v_Z) of one. To standardize, we obtain $Z_t = (Y_t - m_Y) \ / \sqrt{v_Y}$ where Y_t is the random variable following the typical distribution.

The AR(1) parameter ϕ_1 should be in the interval $(-1, 1)$ for mean stationarity. Respecting the NGARCH(1,1), the parameters b_0, b_1, b_2 must be positive to ensure a non-negative variance and the inequality of $b_2(1 + c^2) + b_1 < 1$ must hold for variance stationarity. Finally, some restrictions on the parameters implied in the density of Z_t (i.e, the k for the case of the Student t and (k, s) for the Skewed t) will be commented later.

Let $f_Y(\cdot)$, $VaR_{Y,\alpha}(\cdot)$ and $ES_{Y,\alpha}(\cdot)$ be the probability density function (pdf), the α Value at Risk of Y_t and the α Expected Shortfall of Y_t. Note that in this paper we define $VaR_\alpha(X) = inf\{x \in \mathbb{R} : P(X < x) \geq \alpha\}$ and $ES_\alpha(X) = E(X \mid X \leq VaR_\alpha(X))$,i.e., negative numbers, while other papers the risk measures are defined with positive values. Also note that $Y_t = m_Y + \frac{\sqrt{v_Y}}{\sigma_t}(R_t - \mu_t)$; as a consequence, the pdf, the α Value at Risk and the α Expected Shortfall of R_t conditional on past information are:

$$\begin{aligned} f_t(r \quad &| \quad \theta) = \frac{v_Y^{1/2}}{\sigma_t} f_Y(m_Y + \frac{v_Y^{1/2}}{\sigma_t}(R_t - \mu_t)), \\ VaR_{t,\alpha}(r \quad &| \quad \theta) = \mu_t + \frac{\sigma_t}{v_Y^{1/2}}(VaR_Y(\alpha) - m_Y), \text{ and} \\ ES_{t,\alpha}(r \quad &| \quad \theta) = \mu_t + \frac{\sigma_t}{v_Y^{1/2}}(ES_Y(\alpha) - m_Y) \end{aligned}$$

There are some equivalent specifications of a skewed Student-t distribution. Here we use the parameterization of Hansen (1994) but it can be changed to the Zhu and Galbraith (2010) parameterization changing the parameter s by the parameter $s' = \frac{1-s}{2}$. A random variable Y is said to have a skewed Student-t distribution with parameters s and k if:

$$Y = \begin{cases} (s-1)|T| & \text{if } U \leq \frac{1-s}{2} \\ (s+1)|T| & \text{if } U > \frac{1-s}{2} \end{cases}$$

where T is a Student t random variable with k degrees of freedom, U is a uniform $(0,1)$ random variable, T and U are independent and s is a parameter in $(-1,1)$.Then, the cdf and pdf of Y are:

$$F_Y(y) = \begin{cases} (1-s)\, F_k\left(\frac{y}{(1-s)}\right) & \text{if } y \leq 0 \\ (1+s)\, F_k\left(\frac{y}{1+s}\right) - s & \text{if } y > 0 \end{cases}$$

$$f_Y(y) = \begin{cases} f_k\left(\frac{y}{1-s}\right) & \text{if } y \leq 0 \\ f_k\left(\frac{y}{1+s}\right) & \text{if } y > 0 \end{cases}$$

where $F_k(\cdot)$ and $f_k(\cdot)$ denote the cdf and the pdf of a Student-t distribution with k degrees of freedom. It follows from here that if $k > 1$ then

$$m_Y \equiv E(Y) = \frac{2s\sqrt{k}\Gamma(\frac{k-1}{2})}{\sqrt{\pi}\Gamma(\frac{k}{2})},$$

and if $k > 2$ then

$$v_Y = var(Y) = \frac{4k((1-\frac{1-s}{2})^3 + \left(\frac{1-s}{2}\right)^3)}{k-2} - m_Y^2$$

The α VaR of Y is:

$$VaR_{Y,\alpha} = \begin{cases} (1-s)\, F_k^{-1}\left(\frac{\alpha}{1-s}\right) & \text{if } \alpha \leq \frac{1-s}{2} \\ (-1-s) F_k^{-1}\left(\frac{1-\alpha}{(1+s)}\right) & \text{if } \alpha > \frac{1-s}{2} \end{cases}$$

Finally, the α Expected Shortfall of Y can be found using its pdf:

$$ES_{Y,\alpha} = \begin{cases} -\frac{1}{\alpha}\frac{k}{k-1}\left(1-s\right)^2 f_k\left[F_k^{-1}\left(\frac{\alpha}{1-s}\right)\right]\left(1+\frac{F_k^{-1}(\frac{\alpha}{1-s})^2}{k}\right) & \text{if } \alpha \leq \frac{1-s}{2} \\ -\frac{4}{\alpha}\frac{k}{k-1}\left[(1-\frac{1-s}{2})^2 f_k\left[F_k^{-1}\left(\frac{1-\alpha}{1+s}\right)\right](1+\frac{F_k^{-1}(\frac{1-\alpha}{1+s})^2}{k}) - s f_k(0)\right] & \text{if } \alpha > \frac{1-s}{2} \end{cases}$$

The parameters of the model can be estimated by Maximum Likelihood (ML). When assuming for Y_t a skewed Student-t distribution, the log-likelihood is

$$\log L_R(\theta) = \sum\nolimits_{t=1}^{T} \frac{1}{2}\log v_Y - \frac{1}{2}\log\sigma_t^2 + \log f_Y(m_Y + \frac{v_Y^{1/2}}{\sigma_t}(R_t - \mu_t)),$$

which is maximized subject to the restrictions specified above. When assuming for Y_t a Student-t distribution or a normal distribution, the same procedure applies, with obvious changes. With the formulas described above for a skewed t, we can obtain the formulas of a Student t using $s = 0$, and for the normal distribution using $s = 0;\ k = \infty$.

2.2 Quality validation of ML estimates

To test if the econometric programs are good asymptotically with small samples to do the research, we perform a small Monte Carlo experiment with values of the parameters similar to those that are typically encountered in applications and with the distributions described before, generating series with similar descriptive statistics to most financial returns series. In our estimations we use the sample mean as μ_0, the sample standard deviation as σ_0, and μ_0 as R_0. If the estimation parameters are close to the true ones, then we conclude that the programs are good. To determine when the parameters are close enough, we are going to make a hypothesis contrast, taking as null hypothesis that each estimated coefficient is equal to the true parameter. We take $\widetilde{\kappa} = 0.05$ as nominal size or probability to reject null hypothesis, being this true. If the p-value of the contrast is greater than $\widetilde{\kappa}$ then the estimation approximates the true parameter.

We have done 10000 samples of 5000 observations each and, if the estimations are good enough, we should have that true parameters and the average of estimated coefficients are similar, doing that only 5% of times we reject the null hypothesis (because we select $\widetilde{\kappa} = 0.05$). Also, the standard deviation of the estimated coefficients and the average of estimated standard errors should be similar too. The results are in Tables 1 (Normal innovations), 2 (Student's t innovations) and 3 (Hansen's skewed t innovations).

Table 1: Monte Carlo experiment, AR-NGARCH with Normal distribution, results with 10000 samples of size 5000

	true coefficients	average of coefficients	std. deviation of coefficients	average of std. errors	prob. rejection H_0 ($\widetilde{\kappa} = 0.05$)
ϕ_0	0.01	0.0093437	0.014351	0.014438	0.0487
ϕ_1	0.05	0.049871	0.014365	0.014525	0.0471
b_0	0.035	0.035686	0.0049537	0.0050055	0.0469
b_1	0.85	0.84837	0.011327	0.011373	0.0492
b_2	0.07	0.069498	0.0076648	0.0075375	0.0599
c	0.90	0.91812	0.10597	0.10346	0.0507

Table 2: Monte Carlo experiment, AR-NGARCH with Student t distribution, results with 10000 samples of size 5000

	true coefficients	average of coefficients	std. deviation of coefficients	average of std. errors	prob. rejection H_0 ($\widetilde{\kappa} = 0.05$)
ϕ_0	0.01	0.010239	0.014415	0.014483	0.0464
ϕ_1	0.05	0.049739	0.014324	0.014438	0.0475
b_0	0.035	0.036105	0.026808	0.0055215	0.0505
b_1	0.85	0.84829	0.015939	0.0128	0.0504
b_2	0.07	0.069591	0.0092235	0.0080083	0.0635
c	0.90	0.91623	0.11358	0.11054	0.0507
k	10	10.239	1.4248	1.3857	0.0457

Table 3: Monte Carlo experiment, AR-NGARCH with Skewed Student t distribution, results with 10000 samples of size 5000

	true coefficients	average of coefficients	std. deviation of coefficients	average of std. errors	prob. rejection H_0 ($\widetilde{\kappa} = 0.05$)
ϕ_0	0.01	0.010512	0.014682	0.014685	0.0512
ϕ_1	0.05	0.049561	0.014463	0.014535	0.0484
b_0	0.035	0.035879	0.012998	0.0052708	0.0551
b_1	0.85	0.84823	0.015517	0.012614	0.0552
b_2	0.07	0.069572	0.0083875	0.0080268	0.0586
c	0.90	0.91703	0.11795	0.11427	0.0485
k	10	10.232	1.4276	1.3859	0.0475
s	-0.06	-0.0602	0.020298	0.020294	0.0498

As we can see, the results are as we expected. The true coefficients and the average of coefficients are very close and the same happens with the standard deviation of coefficients and the average of standard errors. The rejection rates are similar to 5%, with some differences because of the randomness, although there are some parameters like b_2 that overrejects a little. Therefore, we are going to use these programs to do our research.

3 Backtesting studies

In this section we are going to describe the different backtests of Expected Shortfall that we will use in the Monte Carlo simulation. The selected backtests are the proposed by Kratz, Lok & McNeil (2018), the Unconditional test by Du & Escanciano (2017), the first two tests by Acerbi & Székely (2014) and the tail test by Berkowitz (2001).

3.1 Kratz, Lok & McNeil (2018)

First of all, we are going to backtest the Expected Shortfall using Kratz, Lok and McNeil's multinomial tests. Kratz et al. (2018) shows in their paper that it can be approached the Expected Shortfall indirectly, with a multilevel VaR, i.e., as an average of different VaR levels, and then can be used this property, together with the binomial distribution of the VaR exceptions, to do a multinomial backtest.

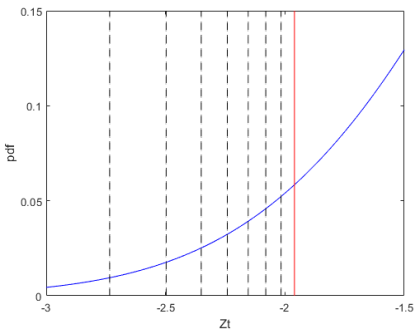
For that purpose, we must choose how many VaR levels, N, are we going to use (in our case $N = 8$), and calculate the VaR values parametrically. Each VaR coverage level is defined by:

$$\begin{aligned} \alpha_j &= \alpha - \frac{j-1}{N}\alpha; \quad j = 1, ..., N; \quad N \in \mathbb{N} \\ \alpha_0 &= 1 \text{ and } \alpha_{N+1} = 0 \end{aligned}$$

[1]

We consider $\alpha = 2.5\%$, which in this paper equals the 97.5% level that Basel Committee on Banking Supervision said. If the true returns (R_t) violates in a temporal moment j estimated VaR_t levels, then we denote $X_t = j$. As a visual example it can be seen in Graphic 1 the N VaR levels for a normal distribution.

Graphic 1: Different VaR levels for a normal distribution to make the multinomial tests.



Blue line is the probability density function, red line denotes VaR_α and each dashed line is a VaR_{α_j}.

[1] The α_j values have been adapted to reflect the coverage level. Kratz et al. uses the confidence level as Basel, which equivalates $\alpha_{coverage} = (1 - \alpha_{confidence})$

We define then, the observed cell counts as the number of times that the variable X_t takes a determined value j by:

$$O_j = \sum_{t=1}^{n} 1_{[X_t=j]}, \quad j = 0, 1, ..., N$$

where $1_{[X_t=j]}$ is the indicator function that is worth one if X_t takes the value j and zero if not. Under the unconditional coverage hypothesis

$$[P(X_t \leq j) = 1 - \alpha_{j+1} \text{ for all } t]$$

and the independence hypothesis

$$[X_t \text{ is independent of } X_s \text{ for } s \neq t]$$

the random vector $(O_0, ..., O_N)$ should follow the multinomial distribution

$$(O_0, ..., O_N) \sim MN(n, (\alpha_0 - \alpha_1, ..., \alpha_N - \alpha_{N+1}))$$

where n is the number of trials or experiments that we realize.

With this idea, if the parameters estimated are $1 = \theta_0 < \theta_1 < .. < \theta_N < \theta_{N+1} = 0$, and considering the model $(O_0, ..., O_N) \sim MN(n, (\theta_0 - \theta_1, ..., \theta_N - \theta_{N+1}))$, the null and alternative hypothesis to test are:

$$\begin{aligned} H_0 &: \theta_j = \alpha_j \text{ for } j = 1, .., N \\ H_1 &: \theta_j \neq \alpha_j \text{ in other case} \end{aligned}$$

With all of this tools, we can calculate the next 3 statistic test that Kratz et al. proposed:

Pearson chi-squared test:

$$S_N = \sum_{j=0}^{N} \frac{\left(O_j - n\left(\alpha_j - \alpha_{j+1}\right)\right)^2}{n\left(\alpha_j - \alpha_{j+1}\right)} \underset{H_0}{\overset{d}{\sim}} \chi^2_N$$

Nass test:

$$\tilde{c} = \frac{2E(S_N)}{var(S_N)}, \qquad \nu = \tilde{c}E(S_N)$$
$$E(S_N) = N, \quad var(S_N) = 2N - \frac{N^2+4N+1}{n} + \frac{1}{n}\sum_{j=0}^{N} \frac{1}{\alpha_j - \alpha_{j+1}}$$

$$\tilde{c}S_N \underset{H_0}{\overset{d}{\sim}} \chi^2_\nu,$$

Likelihood Ratio Test (LRT):

$$\tilde{S}_N = 2 \sum_{j:0\leq j\leq N, O_j \neq 0} O_j \log\left(\frac{O_j}{n(\alpha_j - \alpha_{j+1})}\right) \underset{H_0}{\overset{d}{\sim}} \chi^2_N$$

Under the null hypothesis our estimated model of returns satisfy the unconditional coverage hypothesis and the independence hypothesis, what means that our Expected Shortfall is well calculated. However, if the statistic is greater than the critical value ($\chi^2_{N,(1-\widetilde{\kappa})}$ for Pearson and LRT or $\chi^2_{v,(1-\widetilde{\kappa})}$ for Nass) we must reject the null hypothesis and our ES estimation will not be good enough.

3.2 Du & Escanciano (2017)

The Du and Escanciano's Expected Shortfall backtests are very similar to the usual VaR backtests of Kupiec (1995) and Christoffersen (1998) because these backtests are based on cumulative violations. These cumulative violations are the integral of the violations over the coverage level in the left tail. For this study, we are going to focus on calculate their Unconditional ES backtest.

To do this backtest we need to define the hit or α-violation function at time t as:

$$h_t(\alpha) = 1_{(R_t \leq VaR_{t,\alpha})}$$

where $1_{(.)}$ is the indicator function that, in this case, shows when there is a violation of $VaR_{t,\alpha}$.

This hit function is the same as the X_t in Kratz backtest, but with $N = 1$. Starting from the idea that the violations follows a Bernoulli distribution with mean α, and centered violations are a martingale difference sequence, VaR backtests were made. Du & Escanciano wanted to find a similar property to apply it to ES backtests and, if Expected Shortfall is the integral of VaR, then they could use the integral of $h_t(\alpha)$, the cumulative violation process $H_t(\alpha)$, to make a backtest.

By Fubini's theorem, they discovered that $H_t(\alpha)$ has mean $\frac{\alpha}{2}$ and therefore $H_t(\alpha) - \frac{\alpha}{2}$ follows a martingale difference sequence too. We define u_t as the innovations cdf , i.e, $u_t = F(Z_t)$. Then:

$$\begin{aligned} H_t(\alpha) &= \frac{1}{\alpha}\int_0^{\alpha} h_t(u)du = \frac{1}{\alpha}\int_0^{\alpha} 1_{(u_t \leq u)}du \\ &= \frac{1}{\alpha}(\alpha - u_t)\, 1_{(u_t \leq \alpha)} \end{aligned}$$

Like violations (h_t), cumulative violations (H_t) are distribution-free, since u_t is an i.i.d. uniform [0,1] variable. The advantage of cumulative violations are that contain information about all the tail distribution, while violations only provide punctual information.

The unconditional backtest for ES is an analogue of Kupiec's VaR backtest (1995) and uses a t-test for the hypothesis:

$$\begin{aligned} H_0 &: E[H_t(\alpha)] = \frac{\alpha}{2} \\ H_1 &: E[H_t(\alpha)] \neq \frac{\alpha}{2} \end{aligned}$$

To calculate the statistic test we need the next result: $E\left[H_t^2(\alpha)\right] = \frac{\alpha}{3}$, and hence, $var[H_t(\alpha)] = \alpha\left(\frac{1}{3} - \frac{\alpha}{4}\right)$. Then, the Unconditional test is defined:

$$U_{ES} = \frac{\sqrt{n}\left(\overline{H(\alpha)} - \frac{\alpha}{2}\right)}{\sqrt{\alpha\left(\frac{1}{3} - \frac{\alpha}{4}\right)}} \underset{H_0}{\overset{d}{\sim}} N(0,1)$$

where n is the number of tested observations and $\overline{H(\alpha)}$ is the sample mean of estimated cumulative violations $\left(\widehat{H_t(\alpha)}\right)$.

$$\overline{H(\alpha)} = \frac{1}{n}\sum_{t=1}^{n}\widehat{H_t}(\alpha)$$

So we reject the null hypothesis, and hence the ES forecast, if $U_{ES} < N(0,1)_{\frac{\tilde{\kappa}}{2}}$ or $U_{ES} > N(0,1)_{\frac{1-\tilde{\kappa}}{2}}$, because this is a two tail contrast, and not reject if $N(0,1)_{\frac{\tilde{\kappa}}{2}} < U_{ES} < N(0,1)_{\frac{1-\tilde{\kappa}}{2}}$.

3.3 Acerbi & Székely (2014)

Acerbi and Székely (2014) propose in their paper 3 different statistical backtests of Expected Shortfall. In this paper we are going to use the first two statistics, Z_1 & Z_2 which are centered in estimate the tail distribution. Both tests are very similar, rejecting only if there are evidence of risk underestimation, but they have slightly different assumptions and different null and alternative hypothesis. They need the assumption of continuity of the cumulative distribution function (cdf) and probability density function (pdf) of returns.

The first test is inspired by the definition of Expected Shortfall and from that it can be derived:

$$E\left[\frac{R_t}{ES_{t,\alpha}} - 1 \mid R_t < VaR_{t,\alpha}\right] = 0$$

In this test, it's supposed that the sample observations are independent and $VaR_{t,\alpha}$ has been tested already and it has not been rejected. Then the first test of Acerbi et al. concentrates in testing the magnitude of the realized exceptions against the predicted by the model chosen. Their first test statistic is defined:

$$Z_1(R) = \sum_{t=1}^{n}\frac{R_t h_t}{N_T ES_{\alpha,t}} - 1$$

where h_t is the hit function which scores 1 if there are violation and 0 if not

$$h_t(\alpha) = 1_{(R_t < VaR_{\alpha,t})}$$

and N_T denotes the total number of violations of $VaR_{t,\alpha}$

$$N_T = \sum_{t=1}^{n} h_t$$

The null and alternative hypothesis that are contrasted in this test are:

$$\begin{aligned} H_0 &: P_t^{\alpha} = F_t^{\alpha} \ \ \forall t \\ H_1 &: ES_{t,\alpha}^{F} \leq ES_{t,\alpha}^{P}, \text{ for all } t \text{ and } < \text{ for some } t \\ &\quad VaR_{t,\alpha}^{F} = VaR_{t,\alpha}^{P}, \text{ for all } t \end{aligned}$$

where P represents the predicted distribution model,conditional to previous information, used to forecast VaR and ES, while F represents the real conditional distribution model. Then P_t^{α} is the estimated tail distribution and F_t^{α} is the

real tail distribution for $R < VaR_{t,\alpha}$. This test suppose VaR is well calculated and rejects only if there are underestimation of risk. Under these conditions:

$$\begin{aligned} E_{H_0}\left[Z_1 \mid N_T > 0\right] &= 0 \\ E_{H_1}\left[Z_1 \mid N_T > 0\right] &> 0 \end{aligned}$$

If the risk model is correct, then $Z_1 \approx 0$ but if the statistic is positive enough, $Z_1 > 0$, there will be evidence to reject null hypothesis and the ES will be miscalculated.

On the other hand, the second test serves to contrast VaR & ES at the same time. Based on the fact that $ES_{t,\alpha} = E\left[\frac{R_t h_t}{\alpha}\right]$, the statistic Z_2 is defined:

$$Z_2(R) = \sum_{t=1}^{n} \frac{R_t h_t}{n\alpha ES_{\alpha,t}} - 1$$

The null and alternative hypothesis in this test are

$$\begin{aligned} H_0 &: P_t^{\alpha} = F_t^{\alpha} \ \ \forall t \\ H_1 &: ES_{t,\alpha}^{F} \leq ES_{t,\alpha}^{P}, \text{ for all } t \text{ and } < \text{ for some } t \\ &\quad\ VaR_{t,\alpha}^{F} \leq VaR_{t,\alpha}^{P}, \text{ for all } t \end{aligned}$$

This test does not need the independence assumption, and evaluates magnitude and frequency of violations at the same time. If, like test 1, the statistic is positive enough, we will reject null hypothesis and ES model should be corrected and probably VaR too.

The main problem of these tests are the absence of known asymptotic distribution. Unlike the other backtests presented in this paper, Acerbi & Székely's backtests need to simulate the statistic distribution to achieve critical values and reject or not reject null hypothesis. First of all, a large number (M) of return series are simulated following the estimated model. Afterwards, for each return series, statistics Z_1 and Z_2 are calculated. Finally, critical values are calculated as the $(1-\widetilde{\kappa})$ quantile of Z_1 and the $(1-\widetilde{\kappa})$ quantile of Z_2, so that a $\widetilde{\kappa}$ proportion of statistics were rejected.

$$\begin{aligned} &1)\ R_t^i \sim P_t \ \ \forall t, \forall i = 1, ..., M \\ &2)\ Z_j^i = Z_j(R_t^i) \ \ j = 1, 2 \\ &3)\ Z_{j,\widetilde{\kappa}}^{crit} = inf\{z_j \in \mathbb{R} : P(Z_j < z_j) \geq 1 - \widetilde{\kappa}\} \end{aligned}$$

Now with the critical values, it is possible to make the contrast. In our case, we are going to do an approximation of the critical values for the study that will be commented later.

3.4 Berkowitz (2001)

Berkowitz (2001) proposed in his paper two tests for density forecast, specially for cases with small samples. One of the tests is centered in all the distribution while the other, which will be selected for this paper, is centered only in the tail distribution, ignoring the rest. Berkowitz tests are based on the idea that, if the distribution has been well estimated, the distribution of the variable (u_t) obtained from the estimated cumulative distribution function of the innovations should be an i.i.d. uniform distribution $U(0,1)$. It is difficult to devise parametric tests when the null hypothesis is that a variable follows an uniform distribution $U(0,1)$. For that reason Berkowitz suggested to create the backtest transforming the variable using the inverse of the standard normal distribution function of u_t.

Thanks to the next two propositions, Berkowitz created his likelihood ratio tests. The first proposition says that if a series $u_t = \int_{-\infty}^{R_t} f(v)dv$ is distributed as an i.i.d. $U(0,1)$ then

$$z_t = \Phi^{-1}\left[\int_{-\infty}^{R_t} f(v)dv\right] \text{ is an i.i.d. } N(0,1)$$

and the second one indicates the next result

$$\log\left[\frac{f(R_t)}{\hat{f}(R_t)}\right] = \log\left[\frac{g(z_t)}{\phi(z_t)}\right]$$

where $\Phi^{-1}(\cdot)$ is the inverse normal distribution, $\phi(\cdot)$ is the normal pdf, $g(z_t)$ is the pdf of z_t and $f(R_t)$ & $\hat{f}(R_t)$ are the real and estimated pdf of R_t respectively. With all of this and focusing on tail losses, Berkowitz proposed a Likelihood Ratio Test based on a censored likelihood which compare the forecasted tail density with the observed one.

To calculate the statistic, first of all we have to obtained the u_t series with our estimation models. Then, we have to transform that possible uniform u_t series to a possible normal series $z_t = \Phi^{-1}[u_t]$. To restrict to the α lower tail (in our case the coverage level of 2.5%), we have to calculate the cutoff point $VaR_\alpha^{Berk} = \Phi^{-1}[\alpha]$ and define a new variable:

$$z_t^* = \begin{cases} VaR_\alpha^{Berk} & \text{if } z_t \geq VaR_\alpha^{Berk} \\ z_t & \text{if } z_t < VaR_\alpha^{Berk} \end{cases}$$

With this new variable that, if we have estimated the returns correctly, should follow the tail distribution of a standard normal, we can calculate the log-likelihood function for joint estimation of μ^{Berk} (the mean parameter) and σ^{Berk} (the standard deviation parameter)

$$\begin{aligned} L\left(\mu^{Berk}, \sigma^{Berk} \mid z^*\right) &= \sum_{z^* < VaR_\alpha^{Berk}} \log\left[\frac{1}{\sigma^{Berk}}\phi\left(\frac{z^* - \mu^{Berk}}{\sigma^{Berk}}\right)\right] \\ &+ \sum_{z^* = VaR_\alpha^{Berk}} \log\left[1 - \Phi\left(\frac{VaR_\alpha^{Berk} - \mu^{Berk}}{\sigma^{Berk}}\right)\right] \end{aligned}$$

Therefore, the backtest can be based on the likelihood of a censored normal. The LR test by Berkowitz requires as hypothesis that:

$$\begin{aligned} H_0 &: \mu^{Berk} = 0 \; and \; \sigma^{Berk} = 1 \\ H_1 &: \mu^{Berk} \neq 0 \; and/or \; \sigma^{Berk} \neq 1 \end{aligned}$$

Then Berkowitz shows that we can evaluate a restricted likelihood $L(0,1)$ and compare it to the likelihood of the parameters that maximizes it, $L\left(\hat{\mu}^{Berk}, \hat{\sigma}^{Berk}\right)$. Berkowitz tail backtest statistic is:

$$LR_{tail}^{Berk} = -2\left[L(0,1) - L\left(\hat{\mu}^{Berk}, \hat{\sigma}^{Berk}\right)\right] \underset{H_0}{\overset{d}{\sim}} \chi_2^2$$

This test rejects the Expected Shortfall if $LR_{tail}^{Berk} > \chi^2_{2,(1-\widetilde{\kappa})}$. That means that there is some mismatch in the first 2 moments either because there are very large losses relative to the forecast or because there are too small.

4 Monte Carlo experiments results

After explain each of the backtests of the Expected Shortfall, in this section we are going to compare them. For this purpose we are going to perform Monte Carlo experiments, creating different matrix of returns starting from the same seed. To do this, we simulate 1000 samples of $n + T + 200$ random numbers between 0 and 1 and we obtain the necessary innovations (using the same seed). Applying the AR-NGARCH model (or the AR-GARCH, according to each case) and discarding the 200 first observations to not depend on the initial point (leaving samples of $n + T$ observations), we obtain our simulated profits and losses. With our returns, we estimate the parameters of the model with a rolling window of $T = 2500$ observations and we are going to estimate new parameters each 10 observations (2 market weeks in practice) to reduce estimation times and to be more applicable for banks. That is, we have taken the $T = 2500$ first observations to estimate the parameters and $VaR_{t,\alpha}$ & $ES_{t,\alpha}$ for the period 2501:2510 and calculated the tools described before (as $u_t, h_t, ...$). With observations 11:2510 we have repeated the operation for the period 2511:2520 and so on until the last observation. With this method we obtain $n = 250$ *or* 500 predictions of mean, variance, $VaR_{t,\alpha}$ and $ES_{t,\alpha}$ and we compare them with the true returns for each sample.

With all that data, we can calculate each of the backtests for each of the 1000 samples. Considering a coverage level of $\widetilde{\kappa} = 0.05$, we calculate for each statistic its p-value. If the p-value is greater than $\widetilde{\kappa} = 0.05$, then we have not got evidence to reject null hypothesis and the $ES_{t,\alpha}$ will be well calculated. We will reach to the same result if we use critical values instead of p-values to compare the statistic value and to reject or not reject.

In the case of Acerbi & Székely's tests, to calculate the p-values or critical values, we will use an approximation. We are going to simulate $M = 10000$ samples of returns following the estimated distribution (from the null hypothesis, with their same sizes both in sample and out of sample), with the parameter values $(\phi_0, \phi_1, b_0, b_1, b_2, c) = (0.01, 0.05, 0.035, 0.85, 0.07, 0.90)$ [if the null hypothesis is a GARCH model then $c = 0$]. In this way, the critical values from each sample in the size and the power are the same, so the computational cost is reduced drastically and the results can be compared better. After creating the returns, we can continue with steps 2) and 3), described in Acerbi & Székely's backtests, to estimate the distribution and critical values. There is another approximation in Annex 1 that shows similar results than this one.

In the next subsections we are going to see the size and the power of each ES backtest. We will see what proportion of times (of the thousand samples) we reject null hypothesis H_0 ($\frac{n^o \ of \ times \ that \ it's \ rejected \ H_0}{1000}$).

4.1 Size

In this subsection we will see how the backtests behave if the assumption of mean model, variance model and innovations distributions are estimated with the adequate method, i.e., the real returns are predicted well.

We are going to use AR-NGARCH with the parameters $(\phi_0, \phi_1, b_0, b_1, b_2, c) = (0.01, 0.05, 0.035, 0.85, 0.07, 0.90)$, that are reasonable similar to real financial series, to simulate Profits & Losses. These returns are considered the real ones in

these experiments. Then we estimate the $VaR_{t,\alpha}$ & $ES_{t,\alpha}$ with the Maximum Likelihood estimations of the parameters, following the same model, and we will see how many times we reject or not the forecasted Expected Shortfall with each backtest. This value is the size of the contrast.

We will do this with innovations following the standard normal distribution, the Student t with $k = 10$ degrees of freedom and the Hansen's Skewed t with $k = 10$ and $s = -0.06$.The results are in Table 4:

Table 4: Size backtest experiments, AR-NGARCH with different distributions
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Prob. reject H0	n	Pearson	Nass	LRT	Uncond.	Z_2	Z_1	Berk.
Normal	250	0.090	0.050	0.031	0.046	0.044	0.040	0.056
	500	0.071	0.053	0.069	0.052	0.060	0.050	0.062
Student t	250	0.091	0.049	0.028	0.054	0.054	0.052	0.053
$(k = 10)$	500	0.074	0.058	0.068	0.059	0.062	0.056	0.066
Skewed Student t	250	0.114	0.069	0.037	0.052	0.061	0.055	0.054
$(k = 10;\ s = -0.06)$	500	0.081	0.064	0.074	0.048	0.088	0.037	0.064

Sizes for AR(1)-NGARCH(1,1) returns with different innovations. 'Pearson', 'Nass' & 'LRT' refers to Kratz et al. backtests, 'Uncond.' is Du & Escanciano unconditional test, 'Z_1'& 'Z_2' are the first and second Acerbi & Székely's backtests and 'Berk.' refers to the tail LRT by Berkowitz

As we can see, the results are, as it's expected, very close to 5%. That means that the size contrasts are well performed. Also we can observe that there are some differences between each backtest and between the same backtest but with different out of sample sizes (n). If we compare each backtest made with $n = 250$ and $n = 500$ we can see that the size value is more accurate with an out of sample greater in the Pearson test, as it's said in Kratz et al. paper, but in Z_2 and Berkowitz test, the size values increase above the 5%. The LRT by Kratz et al. has small size values which increases with bigger n values. With the rest of backtests there are no significative variations. Respect the size comparisons between backtests seems that Pearson test is the worst and the correction implemented in Nass test works very well. The rest of backtests performs fairly well, but the most stable backtest around 5% seems to be the Unconditional test by Du and Escanciano.

4.2 Power

The power of a contrast is the probability to reject null hypothesis when the alternative hypothesis is correct. In this subsection we are going to see how good are each backtest to reject false models, although the real and estimated models were very similar. Regarding the contrast powers we are going to realize different tables in which we will change some parameters of skewness or kurtosis one by one in the alternative hypothesis.

For the first tables of powers we will compare how well performs each test if we change the innovations distribution. We will estimate the innovations with

normal distribution but the real generator process was made with H_1 model, innovations with Student t distribution, which has more kurtosis the lower the degrees of freedom (k) are. We can see the results in tables 5a & 5b.

$H_0 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Student t innovations with k degrees of freedom

Table 5: Power backtest experiments, changing Student t degrees of freedom
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$							
k	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
100	0.091	0.051	0.027	0.055	0.048	0.047	0.059
20	0.106	0.071	0.039	0.087	0.073	0.124	0.108
10	0.153	0.098	0.061	0.134	0.093	0.218	0.197
9	0.165	0.103	0.063	0.137	0.104	0.244	0.211
8	0.164	0.116	0.073	0.145	0.115	0.286	0.247
7	0.187	0.124	0.083	0.159	0.123	0.335	0.288
6	0.199	0.140	0.090	0.173	0.134	0.407	0.343
5	0.233	0.163	0.115	0.183	0.148	0.511	0.436

Panel b: $n = 500$							
k	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
100	0.069	0.056	0.078	0.059	0.067	0.063	0.069
20	0.110	0.088	0.086	0.088	0.092	0.154	0.131
10	0.145	0.117	0.124	0.160	0.134	0.356	0.298
9	0.167	0.133	0.131	0.173	0.145	0.405	0.350
8	0.200	0.161	0.148	0.189	0.164	0.469	0.400
7	0.223	0.181	0.167	0.202	0.176	0.530	0.487
6	0.257	0.219	0.202	0.224	0.205	0.624	0.559
5	0.296	0.260	0.258	0.246	0.231	0.711	0.650

The case where $k \to \infty$, is the normal distribution case, so with $k = 100$ we should have results very similar to normal size cases. This happens, as we can see, both for $n = 250$ and for $n = 500$. Also we observe that almost in all cases the sizes increase as the degrees of freedom descend. That is a good signal because it means that if there are more kurtosis, i.e., there are more large losses, the probability to reject the normal case increase. Most of the backtests do not reach the 30% size level. Only the first test by Acerbi & Székely, which suppose that the VaR backtest was satisfactory, and the tail Likelihood Ratio Test by Berkowitz have large powers, what means that are more sensible to tail mismatches.

If we compare the results between powers with an out of sample size of $n = 250$ and $n = 500$, we can see that, in general, the backtests performs better with $n = 500$ (larger n).

After comparing changes on degrees of freedom, we are going to keep them constant and change the Hansen's skewed parameter (s) in Skewed Student t

distribution. The alternative hypothesis will generate skewed student t innovations while we estimate it with normal distribution, i.e., we will not capture the innovations skewness or kurtosis. We can see the results for different s values in the interval $(-1, 1)$ in tables 6a & 6b.

$H_0 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Skewed Student t innovations with skewed parameter s and 100 degrees of freedom

Table 6: Power backtest experiments, changing Skewed Student t skewness parameter
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$							
s	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
−0.80	0.733	0.638	0.495	0.729	0.656	0.662	0.743
−0.60	0.730	0.663	0.499	0.750	0.714	0.563	0.719
−0.40	0.608	0.516	0.346	0.617	0.571	0.339	0.544
−0.20	0.334	0.251	0.155	0.334	0.299	0.137	0.238
0.00	0.091	0.051	0.027	0.055	0.048	0.047	0.059
0.20	0.021	0.011	0.010	0.108	0.001	0.018	0.243
0.40	0.004	0.002	0.002	0.418	0.000	0.010	0.634
0.60	0.000	0.000	0.000	0.868	0.000	0.000	0.957
0.80	0.000	0.000	0.000	1.000	0.000	0.000	1.000

Panel b: $n = 500$							
s	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
−0.80	0.897	0.880	0.806	0.927	0.914	0.922	0.947
−0.60	0.901	0.885	0.802	0.941	0.938	0.856	0.946
−0.40	0.783	0.748	0.613	0.863	0.856	0.606	0.814
−0.20	0.446	0.394	0.270	0.512	0.526	0.250	0.428
0.00	0.069	0.056	0.078	0.059	0.067	0.063	0.069
0.20	0.008	0.005	0.261	0.337	0.000	0.017	0.398
0.40	0.002	0.001	0.752	0.888	0.000	0.010	0.907
0.60	0.000	0.000	0.994	1.000	0.000	0.002	1.000
0.80	0.000	0.000	1.000	1.000	0.000	0.000	1.000

The results are so diverse. As we can see, when $s = 0.00$ we have the Student t distribution with $k = 100$, i.e., the same cases as tables 5a and 5b, very close to normal cases. On the other hand, if we decrease s parameter to negative values, then we have negative skewness and with normal distribution estimations, there are more and more violations of $VaR_{t,\alpha}$ & $ES_{t,\alpha}$. That is reflected in fast increases in powers as s goes down. However, we can observe very different reactions to positive values of s. Acerby and Székely's tests tend to reduce power values to zero, due to their hypothesis contrast that only rejects if there are an undervaluation of risk measures. Since $s > 0$ creates positive skewness and less rejections (because risk measures are overestimated), then it's so hard to reject the backtest. As regards the Unconditional test by Du & Escanciano and the tail Likelihood Ratio test by Berkowitz, the results are pretty good. These backtests tend to power values of 100% quickly with $n = 500$ and, with $n = 250$, they increase more slowly but even faster than with $s < 0$. But

with the tests by Kratz et al. (Pearson, Nass and LRT), the results are more problematic. With $n = 250$ the three backtests tend to zero. This happens because with so few violations, very large samples are required to generate rejections. We can see it in Annex 2. However, with $n = 500$, the LRT is the only one of those tests that begins to show good results.

These experiments encourage financial institutions to choose Du & Escanciano and/or Berkowitz backtests, and Basel Committee on Banking Supervision will prefer Acerbi & Székely tests because the overestimation of risk is not relevant to them.

Finally, the last power experiment will be changing only the variance model, varying the asymmetric NGARCH parameter. We will create returns following an AR-NGARCH model with normal innovations, but the estimations will be created following an AR-GARCH model instead. We can see the results for different c values in the interval $[-1, 1]$ in tables 7a & 7b. Remember that for the returns to be stationary the parameters must accomplish in this case:

$$0.07(1 + c^2) + 0.85 < 1 \rightarrow 0.07c^2 < 0.08 \rightarrow c^2 < 1.143$$

$H_0 : R_t \sim AR(1)\text{-}GARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with asymmetric parameter c and Normal innovations

Table 7: Power backtest experiments, changing asymmetric NGARCH parameter c
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$							
c	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
−1.00	0.073	0.042	0.026	0.038	0.033	0.109	0.062
−0.80	0.066	0.036	0.019	0.032	0.035	0.089	0.051
−0.60	0.077	0.041	0.018	0.036	0.042	0.069	0.047
−0.40	0.082	0.045	0.027	0.040	0.047	0.058	0.047
−0.20	0.084	0.047	0.026	0.040	0.055	0.065	0.060
0	0.098	0.061	0.027	0.048	0.057	0.056	0.055
0.20	0.102	0.058	0.028	0.058	0.067	0.056	0.064
0.40	0.110	0.074	0.037	0.067	0.081	0.057	0.068
0.60	0.119	0.081	0.042	0.081	0.091	0.056	0.064
0.80	0.127	0.084	0.043	0.097	0.121	0.073	0.071
1.00	0.162	0.099	0.049	0.119	0.135	0.096	0.102

Panel b: $n = 500$							
c	Pearson	Nass	LRT	Uncond	Z_2	Z_1	Berk
-1.00	0.045	0.032	0.077	0.039	0.025	0.124	0.070
-0.80	0.045	0.033	0.069	0.040	0.034	0.090	0.052
-0.60	0.052	0.042	0.074	0.040	0.041	0.074	0.062
-0.40	0.068	0.050	0.075	0.044	0.040	0.055	0.056
-0.20	0.061	0.041	0.073	0.048	0.051	0.057	0.059
0.00	0.063	0.046	0.062	0.051	0.062	0.055	0.065
0.20	0.076	0.060	0.071	0.058	0.074	0.058	0.059
0.40	0.087	0.063	0.071	0.069	0.082	0.057	0.075
0.60	0.099	0.081	0.077	0.082	0.105	0.056	0.081
0.80	0.123	0.091	0.091	0.109	0.132	0.077	0.092
1.00	0.141	0.115	0.104	0.156	0.173	0.111	0.133

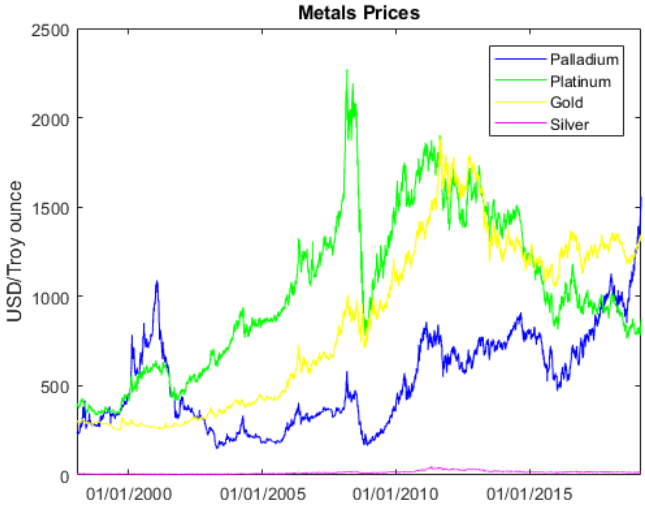
At first sight we can say that there are no big differences between powers changing c parameter. The case where $c = 0.00$ is the size, close to $\widetilde{\kappa} = 0.05$, but the maximum power value, with $c = 1.00$, does not reach the 20% neither with sample $n = 250$ nor $n = 500$. This result shows us that it is more important the skewness from the innovations distribution than the asymmetry from the variance model. Comparing the results between samples, we see that for all backtests except Pearson, the results are more powerful with larger n.The Pearson test is different because it reacts in a peculiar way to changes to sample size. For that reason, it is verified that the modified test (Nass) performs better.

We can see that it becomes hard to detect wrong variance models. With positive c values the power tests increase a little, but with negative ones most of the tests are not powerful given the previous sample sizes. It is verified that Z_1 performs well, while both Berkowitz and LRT perform slightly. Some experiments have been implemented with larger sample sizes and we have seen that the results improve a little with mainly large absolute values of c.

5 Empirical work

In this part of the paper we are going to apply the different backtests studied before on real returns. In our case, we are going to use data from different precious metal prices provided by Datastream, particularly we will use 5500 data from palladium, platinum, gold and silver from 29/01/1998 to 28/02/2019. All metal prices are in United States Dollars per Troy ounce. Gold and silver prices are from Handy & Harman source (from New York), palladium prices are from the London Metal Exchange and platinum prices are from Thomson Reuters source (from New York too). To get a first impression about the data we plot in Graphic 2 the evolution of these metal prices.

Graphic 2: Evolution of precious metal prices



Starting from the prices (P_t), we calculated the logarithmic returns (in percentage) as follows and obtained the next descriptive statistics (Table 8).

$$R_t = 100 * \ln\left(\frac{P_t}{P_{t-1}}\right)$$

Table 8: Descriptive statistics of precious metal returns

	Palladium	Platinum	Gold	Silver
Mean	0.0342	0.0146	0.0268	0.0173
Median	0	0	0	0
Maximum	15.8406	9.5846	7.0060	13.6648
Minimum	−17.8590	−12.4017	−9.5962	−12.9970
Std. Deviation	2.1043	1.3847	1.0647	1.8050
Skewness	−0.2270	−0.4822	−0.2025	−0.5263
Kurtosis	9.2915	8.5634	9.5150	8.9632

As we can see, all of these metal returns have a positive but close to zero mean and they have negative skewness and large kurtosis, which means that the returns follows a model with asymmetric parameters and/or have skewed innovations. The volatility clusters that we can see in Graphic 3 (which is ahead) show us that GARCH models can fit well. For all that reasons, we suspect that the metal returns can follow an AR(1)-NGARCH(1,1) with Hansen's skewed t innovations distribution and an AR(1)-GARCH(1,1) with normal innovations will be slightly good but insufficient. To make comparisons we are going to estimate both models.

As we did with the Monte Carlo experiments, we are going to use an in sample rolling window of $T = 2500$ observations to estimate the model parameters, and each 10 observations we move the window to reestimate them. The ES coverage level will be $\alpha = 2.5\%$ and we will calculate the backtests for all the out of sample ($n = 3000$), for approximately each two years ($n = 500$, see Annex 3) and for each year ($n = 250$).

First we are going to estimate with the $AR(1)$-$NGARCH(1,1)$-$skt_{k,s}$ model. The average of each estimated parameter for each metal returns are in Table 9:

Table 9: Mean of estimated parameters for AR-NGARCH-skt model

	Palladium	Platinum	Gold	Silver
$\widehat{\widehat{\phi_0}}$	0.0256	0.0451	0.0551	0.0513
$\widehat{\widehat{\phi_1}}$	0.0373	0.0052	−0.0229	−0.0245
$\widehat{\widehat{b_0}}$	0.0762	0.0230	0.0101	0.0251
$\widehat{\widehat{b_1}}$	0.8892	0.9247	0.9420	0.9504
$\widehat{\widehat{b_2}}$	0.1069	0.0649	0.0403	0.0371
$\widehat{\widehat{c}}$	0.0484	−0.1028	−0.5010	−0.4597
$\widehat{\widehat{k}}$	4.2474	5.7485	4.7570	4.0649
$\widehat{\widehat{s}}$	−0.0212	−0.0639	−0.0193	−0.0657

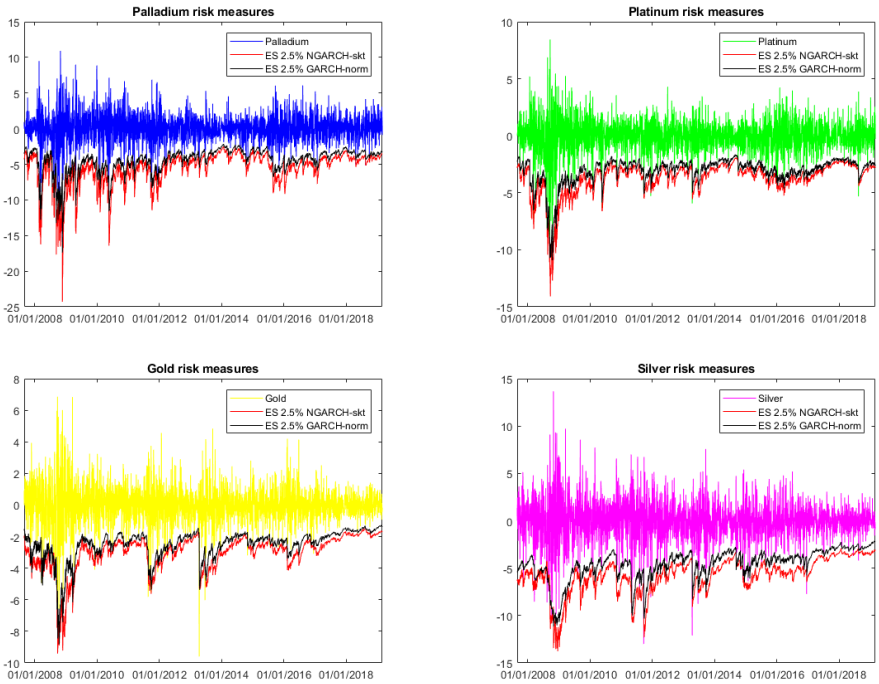
After that, we estimate with the $AR(1)$-$GARCH(1,1)$-*normal* model. The average of each estimated parameter are in Table 10:

Table 10: Mean of estimated parameters for AR-GARCH-norm. model

	Palladium	Platinum	Gold	Silver
$\widehat{\widehat{\phi_0}}$	0.0427	0.0371	0.0398	0.0245
$\widehat{\widehat{\phi_1}}$	0.0425	0.0429	−0.0016	0.0136
$\widehat{\widehat{b_0}}$	0.0723	0.0325	0.0195	0.0427
$\widehat{\widehat{b_1}}$	0.9101	0.9139	0.9262	0.9435
$\widehat{\widehat{b_2}}$	0.0789	0.0709	0.0594	0.0489

With the estimations, we calculate the $VaR_{t,2.5\%}$ and $ES_{t,2.5\%}$ for each temporal moment and for each model. We can see the resulting picture in Graphic 3:

Graphic 3: Returns and loss NGARCH risk measures of metal prices



The risk measures adapt quite well to the returns, at first sight, even with the sudden changes at crisis period due to high volatility, but most times, the Expected Shortfall with asymmetric parameters (with NGARCH and skewed t) is more conservative than symmetric ES. Due to this, using all the sample, $n = 3000$, the quantity of VaR violations at 1.25% ($V\,(0.0125)$) and cumulative violations at 2.5% ($CV\,(0.025)$) are less for the first case than for the second. It can be viewed in Table 11:

Table 11: Descriptive analysis of loss violations between models. Sample size $n = 3000$

Models	$AR(1)$-$NGARCH(1,1)$-*skt*			$AR(1)$-$GARCH(1,1)$-*norm*		
	$V(0.0125)$	$CV(0.025)$	$n * 0.0125$	$V(0.0125)$	$CV(0.025)$	$n * 0.0125$
Palladium	41	39.3956	37.5	59	56.3796	37.5
Platinum	35	37.7645	37.5	58	56.0311	37.5
Gold	45	41.5695	37.5	62	57.7059	37.5
Silver	38	38.8728	37.5	58	56.7262	37.5

To analyze better if we can reject the model hypothesis for the Expected Shortfall risk measures, on Table 12 there are the p-values of the backtests. P-values lower than 5% detect rejection of ES estimation and lower than 0.01% detect strong rejections (transposed to Basel traffic light, first case would be

equivalent to yellow light, which means that there would be penalties, and second case would be equivalent to red light, which means that there would be penalties and intervention too)

Table 12: P-values of loss backtesting contrasts for n=3000 (full sample size)

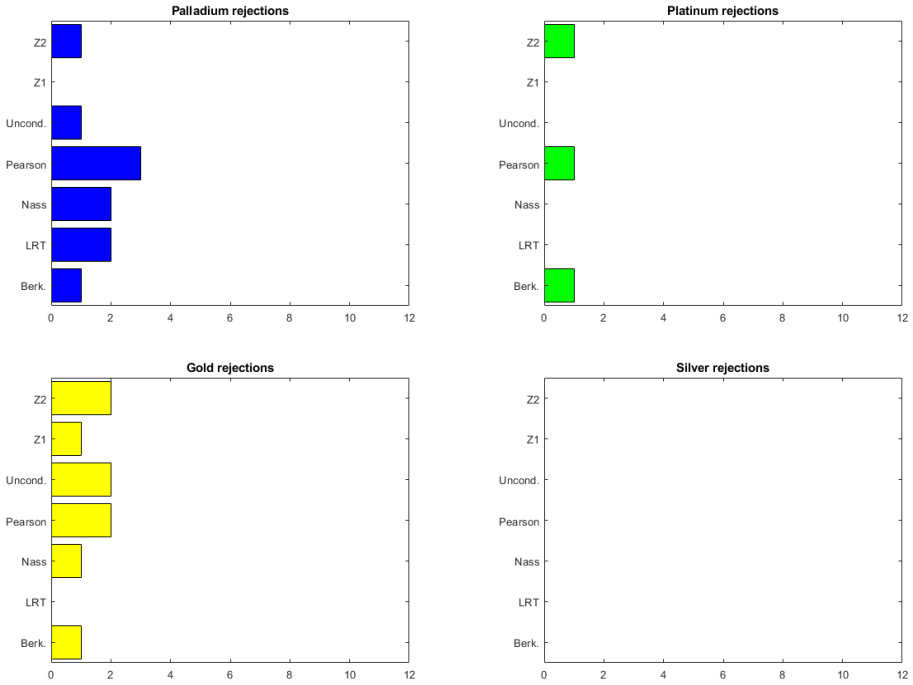
		Palladium	Platinum	Gold	Silver
AR NGARCH skt	Berk	0.7629	0.4208	0.6981	0.8712
	Uncond	0.7019	0.9574	0.4113	0.7816
	Z1	0.9826	0.9988	0.7986	0.8598
	Z2	0.5261	0.4053	0.1592	0.6009
	Pearson	0.4612	0.8935	0.2043	0.9962
	Nass	0.4588	0.8860	0.2078	0.9953
	LRT	0.4803	0.9208	0.3229	0.9956
AR GARCH normal	Berk	0.0000	0.0000	0.0000	0.0000
	Uncond	0.0001	0.0002	0.0000	0.0001
	Z1	0.0000	0.0000	0.0000	0.0000
	Z2	0.0008	0.0008	0.0004	0.0006
	Pearson	0.0000	0.0005	0.0000	0.0000
	Nass	0.0000	0.0006	0.0000	0.0000
	LRT	0.0000	0.0077	0.0000	0.0000

As expected, with asymmetric model there are no rejection in any case but with the symmetric model always is rejected the null hypothesis and, in many cases, with very small p-values (equivalent to red light).

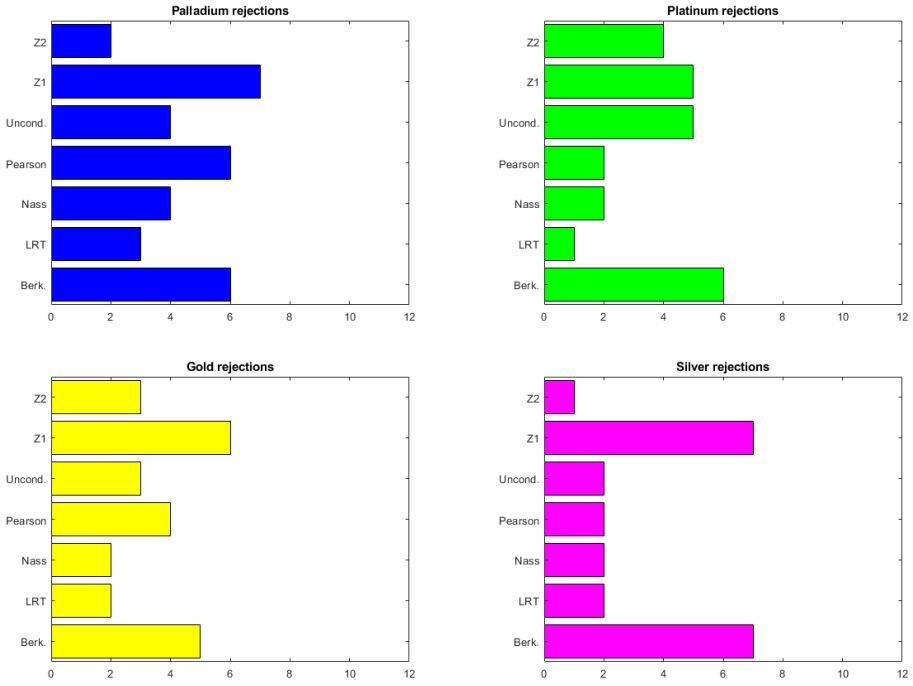
With this big sample the results are evident, but in real life, the backtests have to be realized each year or each two years. Now we are going to show the results for $n = 250$. The calculations for $n = 500$ will be shown in Annex 3.

As we divide the sample into parts of 250, we have 12 estimations of each backtest (one for each "year"). In Graphics 4 and 5 are the times that each backtest is rejected at $\widetilde{\kappa} = 0.05$ for each precious metal series. Results without bar indicate that there are no rejections and ES works well for all the sample. Besides, if a plot have long bars means that that backtest rejects H_0 most times and the forecast is wrong.

Graphic 4: Times the loss asymmetric ES is rejected at 5% in each backtest



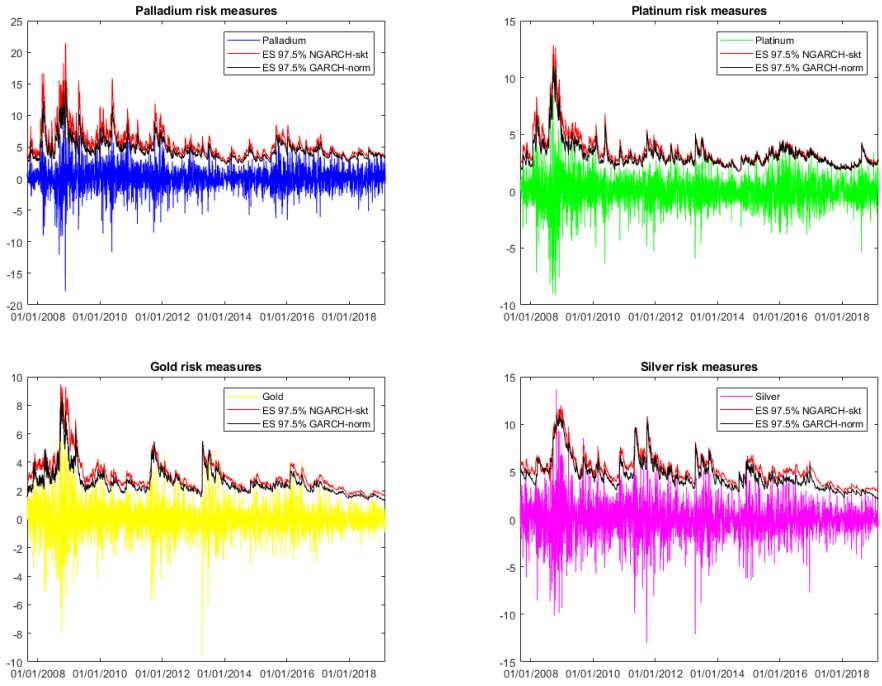
Graphic 5: Times the loss symmetric ES is rejected at 5% in each backtest



In Graphic 4 we can see the results for AR-NGARCH-skt model, where there are few rejections at 5% coverage level, even having no rejections for any silver test. However, seeing Graphic 5, where there are the results for AR-GARCH-normal model, the conclusions are very different. Rejection times go up to 6 or 7 for Berkowitz tests and Acerbi and Székely Z_1 tests. Also, the rest of backtests rejections increase too, so this model is not good enough both for large and small samples, while the asymmetric model is.

We can do the same study with the short position, i.e., if you sell these precious metals instead of buying them. The corresponding risk measure is $ES_{t,97.5\%}$ and their evolutions are plotted in Graphic 6:

Graphic 6: Returns and profit NGARCH risk measures of metal prices



As well as we did with the long position, we make an analysis of profit violations and calculate the p-values for each backtest with all the sample. The results are in Table 13 and Table 14 respectively:

Table 13: Descriptive analysis of profit violations between models. Sample size $n = 3000$

Models	$AR(1)$-$NGARCH(1,1)$-*skt*			$AR(1)$-$GARCH(1,1)$-*norm*		
	$V(0.0125)$	$CV(0.025)$	$n * 0.0125$	$V(0.0125)$	$CV(0.025)$	$n * 0.0125$
Palladium	35	36.3771	37.5	53	47.8667	37.5
Platinum	34	32.8336	37.5	42	39.4417	37.5
Gold	23	30.3320	37.5	48	45.8412	37.5
Silver	39	36.4702	37.5	44	43.1375	37.5

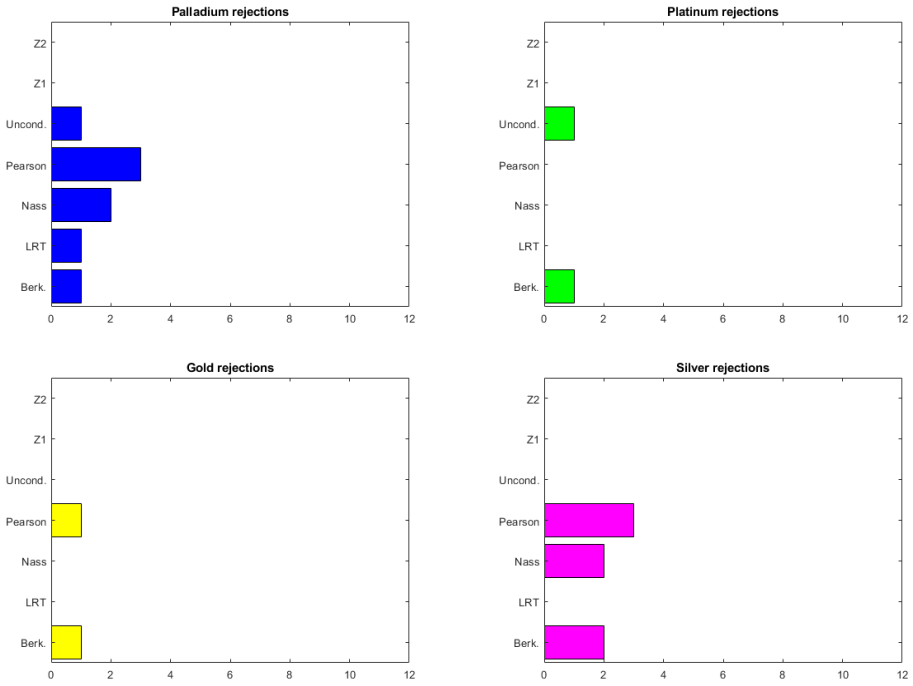
Table 14: P-values of profit backtesting contrasts for n=3000 (full sample size)

		Palladium	Platinum	Gold	Silver
AR NGARCH skt	Berk	0.098	0.369	0.020	0.070
	Uncond	0.821	0.346	0.148	0.835
	Z1	0.992	0.971	0.997	0.993
	Z2	0.713	0.841	0.894	0.722
	Pearson	0.101	0.384	0.305	0.101
	Nass	0.105	0.384	0.306	0.105
	LRT	0.086	0.440	0.241	0.088
AR GARCH normal	Berk	0.000	0.000	0.017	0.001
	Uncond	0.036	0.695	0.092	0.255
	Z1	0.000	0.000	0.026	0.000
	Z2	0.123	0.522	0.054	0.483
	Pearson	0.072	0.052	0.147	0.055
	Nass	0.076	0.055	0.151	0.059
	LRT	0.096	0.027	0.190	0.062

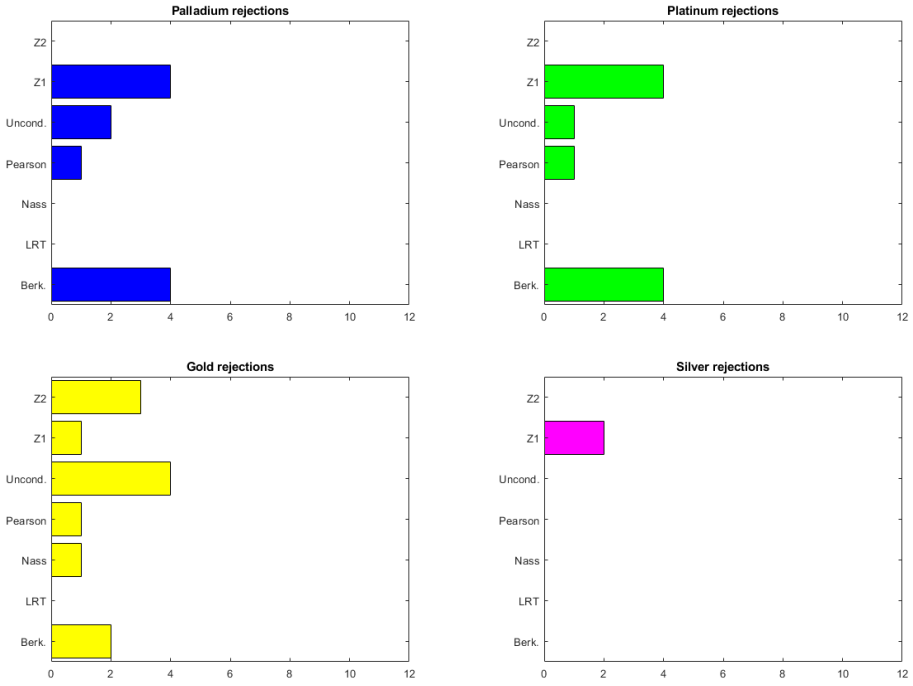
We can see that the results vary respect the other tail. With the asymmetric model is rejected the Gold returns with the test by Berkowitz. Furthermore, the rest of tests are not as powerful than in long position. With the symmetric model the opposite happens, increasing the p-values and reaching to not reject in some cases, although there are more rejections than with asymmetric model.

Doing the same with samples of $n = 250$ the results are the next:

Graphic 7: Times the profit asymmetric ES is rejected at 5% in each backtest



Graphic 8: Times the profit symmetric ES is rejected at 5% in each backtest



The graphics show that, in the profit tail, the number of rejections increase a few, but not with silver returns. With all of this seems that both models forecast quite well for Silver (only in short position). Also seems that centering in tail distributions is a better idea that trying to forecast all the distribution. The study of the performance of backtesting procedures for ES with extreme value theory models will be considered for subsequent papers.

6 Conclusions

Risk management has been changing rapidly and financial institutions have been adapted to it. The Expected Shortfall is the future risk measure and good estimations are needed to maintain the banking system without collapses and bankruptcies. To ensure this, ES backtests are required, but in recent years have appeared many different papers with alternative backtesting approaches based in diverse properties. In this paper we have compared some of them.

We conclude that, if there is uncertainty about which type of misspecification have been produced, the tests by Berkowitz and Du & Escanciano usually perform better. Acerby and Székely's tests are relevant too because they bring good results to supervisors focusing only on risk underestimation (they do not take into account risk overestimation). We have also seen that Pearson test needs Nass correction with small samples to perform well.

Furthermore, we have corroborated that models with a GARCH variance process and normal innovation distribution are a good starting point. Nevertheless, they are insufficient to predict the evolution of financial series because most of them have positive excess kurtosis and non zero skewness (principally negative) which have not been detected by the model. The inclusion of NGARCH models and, especially, the Hansen's skewed t innovation distribution help to fit the returns and, therefore, the risk measures, being more difficult to have higher loss levels than expected.

In future research, the backtesting analysis can be improved with larger sample sizes, reestimating the rolling window parameters each day instead of every 2 weeks or adding another tests such as the Righi and Ceretta approach (2013) or the Graham and Pál approach (2014). Moreover, the Extreme Value Theory can be applied to focus on both tails of the distribution instead of modeling the whole distribution. In this way, it is possible to improve the results since it has been shown in Section 5 that the behavior in both tails can be different.

In short, neither the absence of elicitability property nor a small sample size can inhibit the Expected Shortfall backtest since there are many tests performing well which are useful to select between models.

7 References

References

[1] Zhu, D. and Galbraith, J.W. (2009) "A Generalized Asymmetric Student-t Distribution with Application to Financial Econometrics" *Journal of Econometrics,* Vol.157, No. 2, 297-305

[2] Hansen, B. E. (1994) "Autoregressive conditional density estimation" *International Economic Review*, 35, 705-730.

[3] Kratz, M., Lok, Y.H. and McNeil, A.J. (2018) "Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall" *Journal of Banking and Finance*, 88, 393-407

[4] Du, Z. and Escanciano, J.C. (2017) "Backtesting Expected Shortfall: Accounting for Tail Risk" *Management Science,* Vol. 63, No. 4, 940-958

[5] Wehn, C.S. (2018) "Back to backtesting: integrated backtesting for value-at-risk and expected shortfall in practice" *Journal of Risk Model Validation* 12(4), 17-39

[6] Acereda, B., León, A.M. and Mora, J. (2019) "Estimating the Expected Shortfall of Cryptocurrencies: An Evaluation Based on Backtesting" *Finance Research Letters,* Forthcoming

[7] Novales, A. and Garcia-Jorcano, L. (2019) "Backtesting extreme value theory models of expected shortfall" *Quantitative Finance*, Vol. 19, No. 5, 799–825

[8] Acerbi C. and Székely B. (2014) "Backtesting Expected Shortfall." *Risk Magazine,* 27, 76-81

[9] Berkowitz J. (2001), "Testing Density Forecasts, With Applications to Risk Management" *Journal of Business & Economic Statistics,*Vol. 19, No. 4, 465-474

[10] Engle, R.F. and Ng, V.K., (1993) "Measuring and testing the impact of news on volatility" *The journal of finance*, 48(5), 1749-1778.

8 Annex

8.1 Annex 1: Alternative Acerbi & Székely approximation

In this annex we are going to make another approximation to Acerbi & Székely's tests. As we said in Section 4, we made an approximation in the critical values to save computational time due to the test nature. We did it also to have the same critical values for all the hypothesis and because the variation is small. In this part we created the critical values following Acerbi and Székely's steps, but changing the estimated distribution, which has different parameters each 10 observations to the average of estimated distributions, i.e., we calculated the null hypothesis parameters for each series (25 or 50 parameter sets for each series) and made the average of parameters for all the series

$$\overline{\widehat{\theta}} = \frac{1}{1000} \sum_{r=1}^{1000} \frac{1}{n/10} \sum_{i=1}^{n/10} \widehat{\theta}_{r,i}$$

Using this new method to calculate the critical values, the power for the first and second test are indicated in the following tables as Z_1^m & Z_2^m. Also, there are the powers obtained with the previous method to compare results. Each table uses the same hypothesis as in Section 4.

$H_0 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Student t innovations with k degrees of freedom

Table 5 extension:
Power backtest experiment, changing Student t degrees of freedom
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$				
k	Z_2	Z_1	Z_2^m	Z_1^m
100	0.048	0.047	0.048	0.047
20	0.073	0.124	0.072	0.123
10	0.093	0.218	0.093	0.218
9	0.104	0.244	0.104	0.241
8	0.115	0.286	0.113	0.286
7	0.123	0.335	0.123	0.335
6	0.134	0.407	0.134	0.407
5	0.148	0.511	0.148	0.509

Panel b: $n = 500$				
k	Z_2	Z_1	Z_2^m	Z_1^m
100	0.067	0.063	0.067	0.063
20	0.092	0.154	0.092	0.154
10	0.134	0.356	0.134	0.358
9	0.145	0.405	0.145	0.405
8	0.164	0.469	0.164	0.469
7	0.176	0.530	0.176	0.531
6	0.205	0.624	0.205	0.624
5	0.231	0.711	0.231	0.709

$H_0 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with Skewed Student t innovations with skewed parameter s and 100 degrees of freedom

Table 6 extension:
Power backtest experiment, changing Skewed Student t skewness parameter
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$				
s	Z_2	Z_1	Z_2^m	Z_1^m
-0.80	0.656	0.662	0.691	0.654
-0.60	0.714	0.563	0.743	0.550
-0.40	0.571	0.339	0.620	0.328
-0.20	0.299	0.137	0.321	0.133
0.00	0.048	0.047	0.048	0.047
0.20	0.001	0.018	0.001	0.017
0.40	0.000	0.010	0.000	0.010
0.60	0.000	0.000	0.000	0.000
0.80	0.000	0.000	0.000	0.000

Panel b: $n = 500$				
s	Z_2	Z_1	Z_2^m	Z_1^m
-0.80	0.914	0.922	0.918	0.912
-0.60	0.938	0.856	0.943	0.835
-0.40	0.856	0.606	0.864	0.558
-0.20	0.526	0.250	0.530	0.235
0.00	0.067	0.063	0.067	0.063
0.20	0.000	0.017	0.000	0.018
0.40	0.000	0.010	0.000	0.011
0.60	0.000	0.002	0.000	0.003
0.80	0.000	0.000	0.000	0.000

$H_0 : R_t \sim AR(1)\text{-}GARCH(1,1)$ with Normal innovations

$H_1 : R_t \sim AR(1)\text{-}NGARCH(1,1)$ with asymmetric parameter c and Normal innovations

Table 7 extension:
Power backtest experiment, changing asymmetric NGARCH parameter c
$T = 2500 \quad \alpha = 0.025 \quad N = 8 \quad \widetilde{\kappa} = 0.05$

Panel a: $n = 250$				
c	Z_2	Z_1	Z_2^m	Z_1^m
−1.00	0.033	0.109	0.026	0.103
−0.80	0.035	0.089	0.033	0.089
−0.60	0.042	0.069	0.042	0.070
−0.40	0.047	0.058	0.047	0.059
−0.20	0.055	0.065	0.055	0.066
0	0.057	0.056	0.057	0.056
0.20	0.067	0.056	0.067	0.055
0.40	0.081	0.057	0.081	0.057
0.60	0.091	0.056	0.091	0.055
0.80	0.121	0.073	0.118	0.068
1.00	0.135	0.096	0.128	0.084

Panel b: $n = 500$				
c	Z_2	Z_1	Z_2^m	Z_1^m
−1.00	0.025	0.124	0.022	0.117
−0.80	0.034	0.090	0.034	0.093
−0.60	0.041	0.074	0.041	0.077
−0.40	0.040	0.055	0.040	0.057
−0.20	0.051	0.057	0.051	0.057
0	0.062	0.055	0.062	0.055
0.20	0.074	0.058	0.074	0.058
0.40	0.082	0.057	0.082	0.057
0.60	0.105	0.056	0.104	0.056
0.80	0.132	0.077	0.130	0.077
1.00	0.173	0.111	0.169	0.097

As we can see, these results are very similar to the results calculated previously, with small differences. So both approximations can be useful.

8.2 Annex 2: Examples of insufficient size to obtain good power results

In this examples, we will show why the size samples of $n = 250, 500$ could be insufficient to reject models with positive skewness. Let us assume that the distribution is so asymmetric that there are no violations to any VaR level (in this case we use $N = 8$ VaR levels and $\widetilde{\kappa} = 0.05$); then, the results for Pearson, Nass and LRT backtests of ES are:

For $n = 250$

$$\begin{aligned}\hat{S}_{Pearson,250} &= \frac{(250 - 250(1-0.025))^2}{250(1-0.025)} + 8 * \frac{(0 - 250 * 0.003125)^2}{250 * 0.003125} = \\ &= 6.410 < \chi^2_{8,(1-0.05)} = 15.507\end{aligned}$$

$$\begin{aligned}\tilde{c}_{250} &= \frac{2*8}{2*8 - \frac{8^2+4*8+1}{250} + \frac{1}{250} * (\frac{1}{0.975} + 8 * \frac{1}{0.003125})} = 0.619 \\ \nu_{250} &= 0.619 * 8 = 4.950 \\ \hat{S}_{Nass,250} &= 0.619 * 6.410 = \\ &= 3.967 < \chi^2_{4.950,(1-0.05)} = 10.994\end{aligned}$$

$$\begin{aligned}\hat{S}_{LRT,250} &= 2 * \left[250 \log\left(\frac{250}{250(1-0.025)}\right)\right] = \\ &= 12.659 < \chi^2_{8,(1-0.05)} = 15.507\end{aligned}$$

Therefore, in this case H_0 is not rejected with any of these three statistics.

For $n = 500$

$$\begin{aligned}\hat{S}_{Pearson,500} &= \sum_{j=0}^{N} \frac{(500 - 500(1-0.025))^2}{500(1-0.025)} + 8 * \frac{(0 - 500 * 0.003125)^2}{500 * 0.003125} = \\ &= 12.821 < \chi^2_{8,(1-0.05)} = 15.507\end{aligned}$$

$$\begin{aligned}\tilde{c}_{500} &= \frac{2*8}{2*8 - \frac{8^2+4*8+1}{500} + \frac{1}{500} * (\frac{1}{0.975} + 8 * \frac{1}{0.003125})} = 0.765 \\ \nu_{500} &= 0.765 * 8 = 6.116 \\ \hat{S}_{Nass,500} &= 0.765 * 12.821 = \\ &= 9.801 < \chi^2_{6.116,(1-0.05)} = 12.765\end{aligned}$$

$$\begin{aligned}\hat{S}_{LRT,500} &= 2 * \left[500 \log\left(\frac{500}{500(1-0.025)}\right)\right] = \\ &= 25.318 > \chi^2_{8,(1-0.05)} = 15.507\end{aligned}$$

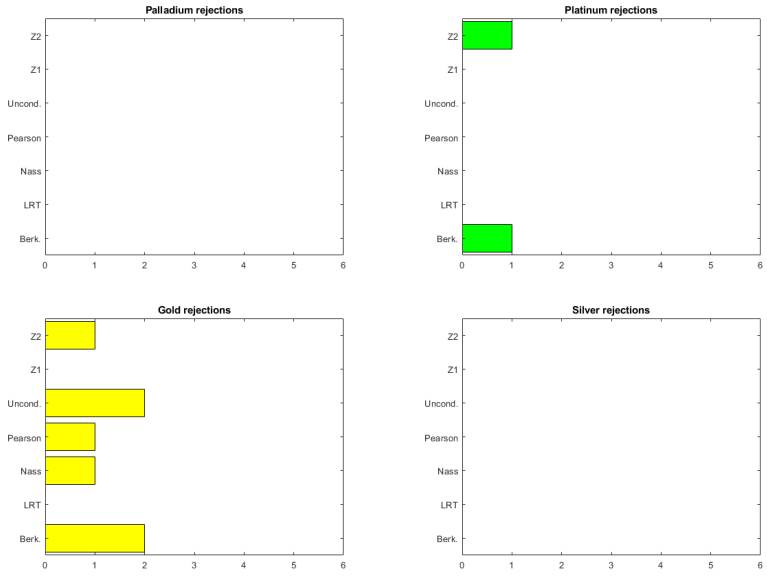
In this case H_0 is not rejected by Pearson or Nass but the LRT detects the risk overestimation and rejects H_0.

As we saw before in Table 6b, the LRT with 500 of sample size already detects bad estimations. If we change the sample size n (as we have done) or the contrast coverage level ($\widetilde{\kappa}$), then the results would change.

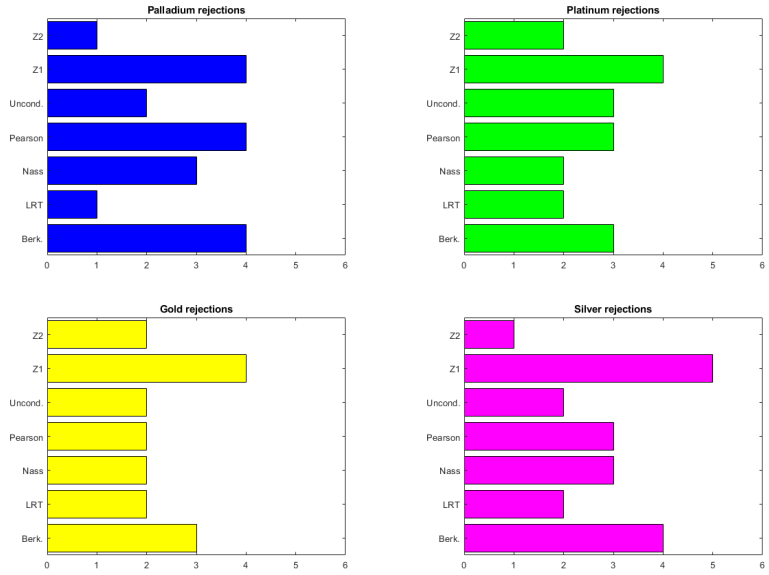
8.3 Annex 3: Results of metal returns study with $n = 500$

In this annex we show another results of backtesting Expected Shortfall on the precious metal returns. In the main work we estimate the tests with 250 and 3000 sample sizes, but, as we did with the size and power experiments, we are going to estimate it with $n = 500$ too (2 years). We expect that the results have greater rejection rates than with 250 for the symmetric ES and less rejection rates for the asymmetric, showing preference to this last model.

Graphic 9: Times the loss asymmetric ES is rejected at 5% in each backtest

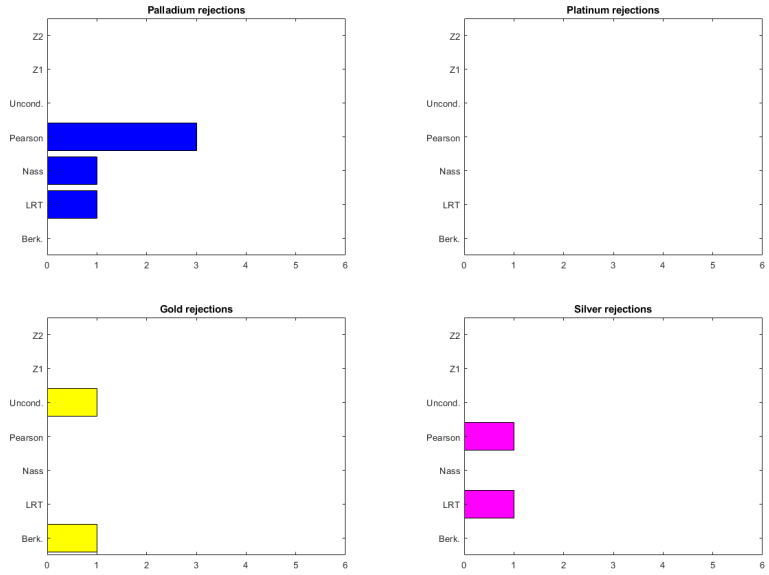


Graphic 10: Times the loss symmetric ES is rejected at 5% in each backtest

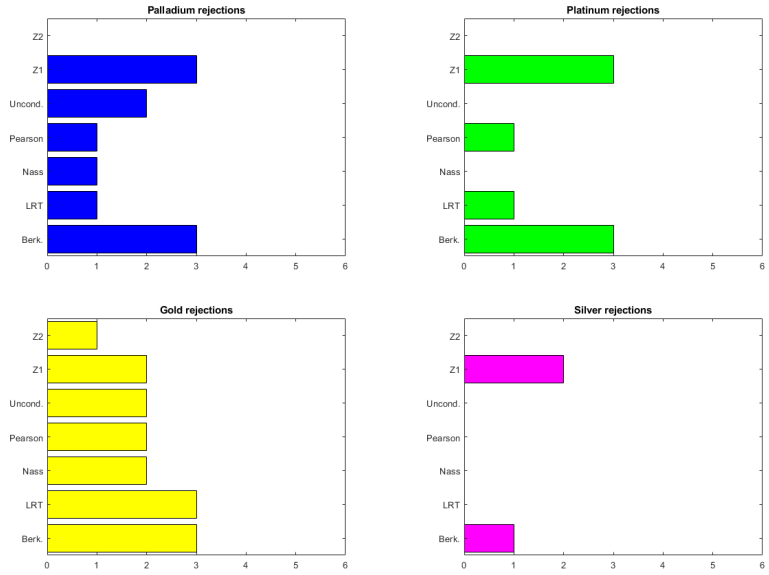


Looking to Graphics 9 and 10 we can corroborate our speculations. As with the other samples, to include asymmetry is a good idea and the few times that with AR-NGARCH-skt the risk measure is rejected, the p-values are not lower enough to do the banking supervisors an intervention ($0.01\% <$ rejection p-values $< 5\%$).For short position the results with $n = 500$ are in Graphics 11 and 12:

Graphic 11: Times the profit asymmetric ES is rejected at 5% in each backtest



Graphic 12: Times the profit symmetric ES is rejected at 5% in each backtest



Similar results as long position, except for silver returns that with a GARCH-normal model there are few violations, but they are in the LR test by Berkowitz and in Z1.