The backtesting of value-at-risk and expected shortfall: evidence from the Bank of Italy's foreign reserves portfolio

Marco Fruzzetti,[1] Davide Nasti[2] and Marco Orlandi[3]

Abstract

Value-at-risk (VaR) and expected shortfall (ES) are widely used by financial institutions as tools for measuring market risk. Despite its well known properties as a risk measure, a debate on ES backtestability and its use for risk models validation is still ongoing. We assess the innovative ES Acerbi-Szekely Ridge Backtesting approach by using Bank of Italy foreign reserves portfolio data. We compute the VaR and ES estimates with a well established, multivariate, conditional normal RiskMetrics model. As expected, the model passes traditional VaR backtests, showing accuracy in estimating VaR; however, ES estimates calculated using the RiskMetrics model often fail to pass the Ridge Backtest. These results show that the ES Ridge Backtesting can accurately identify mispredictions of real, thick-tailed, unconditional return distributions.

JEL classification(s): C52, G32.

Keywords: backtesting, risk measures, value-at-risk, expected shortfall.

[1] Bank of Italy, Economic Outlook and Monetary Policy Directorate, marco.fruzzetti@bancaditalia.it.

[2] Bank of Italy, Financial Risk Management Directorate, davide.nasti@bancaditalia.it.

[3] Bank of Italy, Financial Risk Management Directorate, m.orlandi@bancaditalia.it.

1. Introduction

Central banks focus on monetary policy and emergency liquidity assistance, which can involve very high financial risks. Consequently, a high risk aversion is common among central bank investment portfolio operations; large financial losses can lead to capital depletion and undermine a central bank's ability to perform its institutional functions. Although a central bank can technically operate even with negative capital, a positive net value is generally desirable because it bolsters public confidence in the institution and helps to preserve its independence. Furthermore, losses on the investment portfolio expose the bank to reputational risks. Therefore, it is essential for a central bank, when assessing the risks of its investment portfolio, to adequately take into account the interdependencies between the risks of its activities and the current and future adequacy of its assets.

To support decision-making on financial risks, a central bank must have a measurement system that is as accurate as possible. To this end, the Bank of Italy (henceforth BdI) uses specific models and seeks to verify both forecasting accuracy and the adequacy of the assumptions. Various methods can be used to validate these models, including backtesting.

Backtesting has long been a topic of interest for the Basel Committee on Banking Supervision, which has set out a risk management framework for commercial banks. And a number of central banks have also decided to adopt risk practices along the lines of these principles.

For some years, BdI has estimated the market risk of its portfolios using not only the value-at-risk (VaR) metric but also expected shortfall (ES), in line with the market risk regulations set out in the Basel Committee's Fundamental Review of the Trading Book (FRTB).

To monitor the risks of foreign exchange reserves, BdI, like other central banks, uses the RiskMetrics model to calculate a VaR. BdI is therefore interested in backtesting this risk measure. As BdI mainly focuses on the distribution tails, it has recently derived ES measures by applying a transformation to the RiskMetrics VaR and has subsequently applied a backtesting procedure to such measures.

Established backtesting techniques are available for VaR, while the literature on the ES backtesting is more fragmented. Nonetheless Acerbi and Szekely (2014), hereafter "AS", have recently proposed a well received approach that allows a validation procedure for ES models within a first-order approximation.

After a brief review of the underlying theoretical foundations, this article presents the results of a backtesting strategy for BdI's foreign reserves risk measures, with a specific focus on the AS procedure and its innovative characteristics.

2. VaR backtesting techniques

Backtesting techniques designed to assess the accuracy of VaR estimates typically focus on the **unconditional coverage** and **conditional coverage** of VaR.

The simplest test of a VaR model involves **counting the number** of days in which the actual portfolio loss is greater than the VaR forecast, which results in a VaR exception or violation.

Unconditional coverage tests aim to determine whether this number is statistically different from the number of breaches expected from a VaR estimate. For example, a VaR model with a confidence level of 95% would expect five breaches out of 100 observations. If the distance between the number of expected and realised exceptions is not (statistically) negligible, the estimates need to be improved.

In contrast, conditional coverage tests consider the exact point at which the violations occurred. Precise VaR models can react to changes in volatility and correlations to maintain **independence** of breaches. VaR modellers are interested in having a risk measure that can manage volatility clustering, as significant losses that occur in rapid succession increase the probability of default of the financial institutions compared with widely distributed individual losses. Conditional coverage tests address this problem by assessing whether the distribution of breaches over time is significantly different from the expected random one.

It is important to note that both types of test must be satisfied by an accurate VaR model. Tests that jointly examine the unconditional and conditional coverage provide an option to detect VaR measures shortfall.

Following the related literature and best practices, BdI decided to adopt the "Binomial" and the "Proportion of failures (POF)" tests for unconditional coverage tests and the "Christoffersen" and the "Time Between Failures Independence (TBFI)" for conditional coverage tests.

The sequence of successes and failures resulting from the comparison between portfolio effective results and VaR forecasts is known as a *Bernoulli trial* and the number of VaR exceptions follows a binomial distribution.

The binomial test (Haas (2001)) compares the number of VaR exceptions with the expected number, considering exceptions variability. Under the null hypothesis of having a VaR model with good predictive power, the null is rejected when the number of VaR exceptions, given a confidence level of the test (typically 95%), lies beyond the threshold value. Also, the POF test uses the binomial distribution (Kupiec (1995); Haas (2001)). In addition to the Binomial test, the POF verifies through a likelihood ratio (LR) test that the probability of exceptions is consistent with the probability p implied in the confidence level of VaR. If the data suggest that the probability of exceptions is different than p, the VaR model is rejected. According to the Neyman-Pearson Lemma, the LR test is the most powerful of its class.

The Christoffersen test (Christoffersen (1998)) examines the concept of independence between VaR violations by examining whether the probability of a violation on any given day depends on the previous day result. The relevant test statistic for verifying this kind of independence is an LR. The Christoffersen test considers only the dependence between observations on two subsequent days. However, it is possible that today's VaR violation does not depend on the violation that occurred yesterday but on the violation that occurred, for example, a week ago. The TBFI test (Haas (2001)) measures the time between exceptions. For the first VaR exception, the TBFI test is implemented as a standard time-until-first-failure (TUFF) test (Kupiec (1995); Haas (2001)); the TUFF is based on a LR and similar assumptions with respect to the POF test, measuring the number of days until the first VaR violation occurs (here the target variable is the distance, in days, between the first VaR forecast

and the first VaR violation). After calculating the LR ratio statistic for each exception, a test is obtained with a null hypothesis that assumes that each exception is independent from the others.

In order to conduct a more comprehensive analysis, we also performed two joint tests: the first is obtained combining the Christoffersen's statistic with the TBFI ("Mixed Kupiec" joint test; Haas (2001)), the second results from the sum of the Christoffersen's statistic and the POF (we named it the "Mixed Christoffersen" joint test; Nieppola (2009)).

3. ES backtesting techniques: the Acerbi-Szekely test

The literature on the backtesting of ES forecasts is quite fragmented. The ES belongs to the class of "coherent" risk measures (Acerbi and Tasche (2002)), whose concept was introduced (Artzner et al (1997)) and formalised (Artzner et al (1999)) in the late 1990s. However, only in 2011 emerged a stream of research that identified the properties of backtestable statistics and ascertained if ES belonged to this class of measures. After Gneiting's breakthrough result (Gneiting (2011)) that showed that ES is not elicitable,[4] it gradually became clear (Acerbi and Szekely (2014); Acerbi and Szekely (2017)) that ES is not backtestable either and, therefore, that an exact validation procedure cannot exist for ES models.

Nonetheless, it has been proven that VaR and ES are jointly elicitable (Fissler and Ziegel (2016).) This allows **selection procedures** based on the ranking of different risk models.[5]

Building on these concepts, Acerbi and Szekely have proposed an innovative approach, which they named *ridge backtesting*. This approach allows, within a first order approximation, a **validation procedure of an absolute type** for ES models (Acerbi and Szekely (2017); Acerbi and Szekely (2019)).

The new approach is based on the following *ridge backtesting* function:

$$Z_{ES_\alpha} = e - v - \frac{1}{\alpha}(x + v)_- \qquad (1)$$

where e, v and x are respectively the ES forecast, the VaR forecast and the observed return (X is the random variable for the portfolio returns), α is the confidence level of the ES and VaR forecasts and $(x + v)_-$ stands for $min(x + v, 0)$ [6]. The authors showed that a backtest based on Z_{ES_α} suffers from a *bias*, $B(v)$, that : a) has a prudential effect, b) is small around $v = \boldsymbol{VaR_\alpha}$[7] (Acerbi and Szekely (2017)); consequently, it is possible to use Z_{ES_α} in order to validate ES models, as long as VaR forecasts are sufficiently accurate.

4 A statistic is said to be *elicitable* if it can be obtained as the minimiser of its expected scoring function.

5 Comparative backtests can be carried out using Fissler and Ziegel's strictly consistent loss functions (containing forecasts of both risk measures, ie VaR and ES) and the Diebold-Mariano test (Diebold and Mariano (1995)).

6 The function in (1) is based on the following way of writing ES: $ES_\alpha = \min_v E\left[v + \frac{1}{\alpha}(X + v)_-\right]$ (Acerbi and Szekely (2019)).

7 Under weak regularity conditions, the sensitivity of a ridge-backtest to the auxiliary variable (in the ES case such variable is the VaR) forecast is zero at first order (Acerbi and Szekely (2017)).

Acerbi and Szekely also showed that an ES *ridge* backtest satisfies, at first order, a highly desirable property that they named *sharpness*: the expected value of Z_{ES_α} provides a measure of the average forecast error (Acerbi and and Szekely (2019)) given by the difference between: a) the average ES forecast and b) an estimator of the average true value of ES,[8] the latter quantity leads to a notion of *realised* ES which can be calculated as follows (Acerbi and Szekely (2019)):

$$\widehat{ES}_\alpha = \frac{1}{T}\sum_{t=1}^{T}[v_t + \frac{1}{\alpha}(x_t + v_t)_-] \quad (2)$$

This means that the ES *ridge* backtest depends on the amplitude of the forecast errors, in contrast to the VaR backtest (the binomial test).

Practically, the new backtesting procedure consists of a standard, one-sided hypothesis test, based on the mean backtest function:

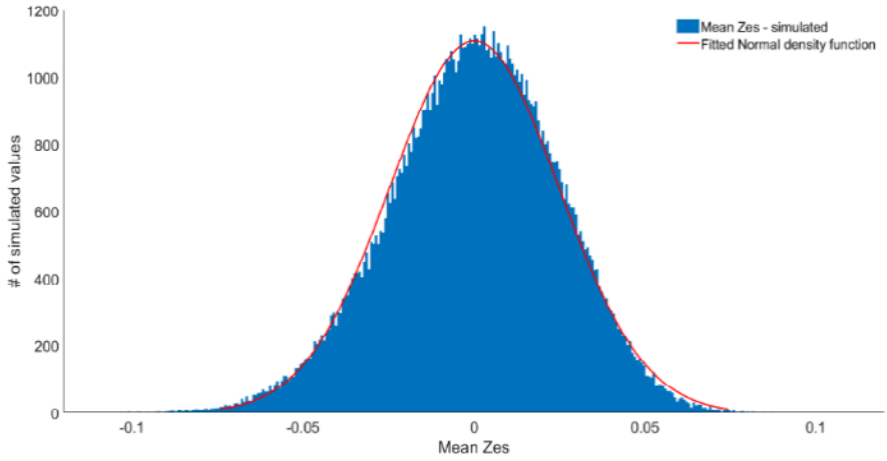
$$\bar{Z}_{ES_\alpha} = \frac{1}{T}\sum_{t=1}^{T} Z_{ES_\alpha}(e_t, v_t, X_t) \quad (3)$$

where $t = 1, \dots, T$ is a discrete sequence of time, e_t and v_t are series of ES and VaR model forecasts respectively, and $X_t \sim P_t$, being P_t the predictive distribution for X given by the model at time t. The one-sided null hypothesis ($\bar{Z}_{ES_\alpha} \geq 0$) is verified comparing the predictive distribution $P_{\bar{Z}_{ES}}$ of the test statistic and its realised value $\bar{z}_{ES}$:

$$\bar{z}_{ES} = \frac{1}{T}\sum_{t=1}^{T} Z_{ES}(e_t, v_t, x_t) \quad (4)$$

where x_t is a series of realisations of X [9]. $P_{\bar{Z}_{ES}}$ is obtained through a Monte Carlo simulation (Acerbi and Szekely (2017)); an example is shown in Graph 1.

Simulated distribution for the $\overline{Z_{ES_\alpha}}$ test statistic — Graph 1



This chart shows the $\overline{Z_{ES_\alpha}}$ test statistic distribution, obtained through a Montecarlo simulation (100,000 scenarios) on the BdI's USD portfolio (see below).

Source: Bank of Italy.

[8] Such estimate is biased for the *ridge backtesting* whereas it is correct for sharp strict backtests.

[9] So, $\bar{z}_{ES}$ is the average realisation of the backtest function over the period $t = 1, \dots, T$; considering different values of x_t, drawn from the distribution P_t, it is possible to provide the distribution of $\bar{z}_{ES}$ (see Chart 1).

One can also consider a relative and dimensionless version of the test by normalising the *ridge backtesting* function Z_{ES_α} with ES forecasts. This yields a new backtesting function denoted as $Z^{rel}_{ES_\alpha}$, which is defined as $Z^{rel}_{ES_\alpha} = \frac{Z_{ES_\alpha}}{e}$. The resulting test statistic is:

$$\bar{z}^{rel}_{ES_\alpha} = 1 - \widehat{\varphi_{ES}} \quad (5)$$

where

$$\widehat{\varphi_{ES}} = \frac{1}{T}\sum_{t=1}^{T} \frac{v_t + \frac{1}{\alpha}(x_t + v_t)_-}{e_t} \quad (6)$$

is a positively biased estimator of the average prediction ratio $\frac{ES_\alpha}{e}$ between the true and the predicted ES (Acerbi and Szekely (2019)). This estimator is conservative and provides a prudential estimate. The scaling factor that needs to be applied to ES forecasts to bring $\bar{z}^{rel}_{ES_\alpha}$ back to zero is simply the estimated ratio $\widehat{\varphi_{ES}}$.

4. Data description

To construct the reference data set for this study, we used data from 1 January 2012 to 31 March 2021. Specifically, we followed these steps:

1. We obtained a daily time series of 95% VaR measures[10] for each BdI foreign reserves portfolio;
2. We then associated each time series with actual returns in local currency.

Due to space limitations, we only present evidence relating to two of the Bank's foreign exchange reserve portfolios: those in United States dollars and in Japanese yen.

Firstly, we examined the statistical properties of the daily return series. The only notable feature that warrants consideration is the high value for the *asymmetry and kurtosis* parameters of the Japanese yen portfolio (Table 1), which may have a significant impact on our analysis.

Descriptive statistics for foreign reserves portfolios Table 1

Foreign reserves portfolio	Mean	Standard Deviation	Asymmetry	Kurtosis
$ US	0.005%	0.01%	-0.1%	5.1
¥ JP	0.001%	0.03%	-1.1%	14.7

Source: Bank of Italy.

Secondly, we had to make a decision about the sample period to use for the backtesting analysis. Regulation often requires using the latest available year of

[10] Calculated with the RiskMetrics™ methodology.

data,[11] but statistical techniques show that longer time periods can significantly increase the power of the tests.

For both risk measures, VaR and ES, we gradually increased the *backtesting period*[12] with one-day steps for VaR and 10-day steps for ES[13]. This allowed us to evaluate the outcome of the tests on the entire data set and understand its dynamics over time. To have a short-term result comparable to the "Basel" logic, we also conducted the same exercise, with a fixed backtesting period over one year of rolling data, updated on a daily basis for the VaR and every 10 days for the ES.

The RiskMetrics™ model is used in this paper to calculate risk measures. It provides a daily conditional forecast of return volatility for various risk factors. Assuming the returns' conditional distributions to be normal, the estimates of VaR and ES are at the 95% confidence level.

The daily volatility forecast is based on the assumption that only recent returns affect the volatility estimates (the EWMA decay parameter λ is 0.94). For data availability reasons, we used monthly volatility forecasts (in that case the optimum λ is 0.97), appropriately rescaled to daily frequency using the "square root of time" rule. The bias imposed by this transformation is negligible (Hendricks (1996)).

As previously mentioned, the chosen approach assumes that the risk factors returns are distributed according to a multivariate normal distribution with zero mean. This conceptual framework allows, for each confidence level, to pass from VaR to ES forecast using the following formula:

$$ES_\alpha = \frac{VaR_\alpha}{z_\alpha \cdot (1-\alpha)} \cdot \phi[\Phi^{-1}(\alpha)] \qquad (7)$$

where VaR_α e ES_α are, respectively, the VaR and ES of the portfolio at the selected confidence level α (which is 95% in the data set being examined), while z_α, ϕ e Φ are, respectively, the quantile of order α, the probability density function and the distribution function of the standard normal distribution. The ES forecasts for each foreign reserve portfolio were obtained using equation (7).[14]

5. VaR backtesting results

Table 2 shows the results of joints tests applied to the US dollar portfolio, which appear to be **contradictory**. Using the increasing time window, the Mixed Christoffersen test fails (p-value of 2%) while the Mixed Kupiec test is not rejected. **The frequency tests** suggest that the VaR model is not entirely reliable since they were **rejected, with p-values** of **3.1%** (Table 2). Further inspection reveals that the reference model tends to **overestimate** the VaR, leading to a lower number of violations than expected (Graph 2, left-hand panel). The Christoffersen independence

[11] Market risk standard set out by the Basel Committee on Banking Supervision.

[12] To be distinguished from the *lookback period*, which refers to the issue of estimating risk measures (in other words, one wonders how much history must be considered in order to produce the most accurate VaR forecast possible).

[13] The different size of the time steps is due to computational reasons.

[14] In this regard, it is interesting to put in evidence that: using RiskMetrics, for $\alpha = 0.95$ it is possible to say that $ES = 1.254 * VaR = 2.1 * standard\ deviation$. So, (conditional) gaussian models present a fundamental limit: going from basic to more sophisticated risk measures does not bring any truly new information about risk.

test and the TBFI were **not rejected** (Table 2) with regard to the **independence** of the violations.

Test results for USD portfolio
Whole period

Table 2

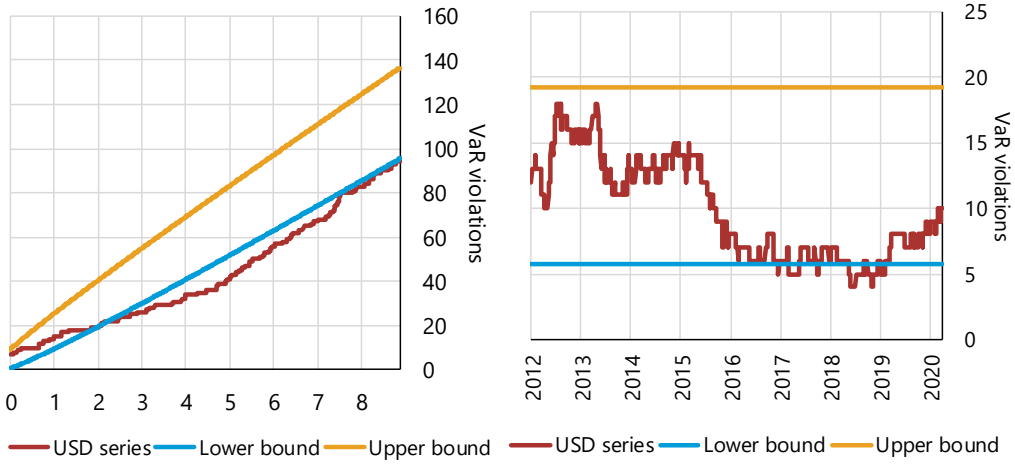
Class of test	Test	p-value / #breaches	Rejection
Joint	Mixed Kupiec	26.8%	5%
	Mixed Christoffersen	2.00%	5%
Frequency	Binomial	94	< 95.3 , > 136.5
	POF	3.1%	5%
Independence	Christoffersen	7.7%	5%
	TBFI	26.8%	5%

Source: Bank of Italy.

Lastly, the binomial test was applied to a rolling time window with a fixed length of 250 trading days (Graph 2, right-hand panel). This was to ensure compliance with the Basel framework, as the binomial test fundamental to the traffic light approach is used to validate banks' risk models.

USD portfolio binomial test results
Increasing and rolling backtesting period

Graph 2



Results of the binomial test. The amount of data used for the tests (in years, left-hand panel; equal to 250 days, right-hand panel), starting from the most recent observations, is shown on the x-axes; the number of VaR violations on the y-axes. The rejection region is represented by the area that is not included between the orange line, which indicates an underestimate of the VaR, and the blue one, which indicates an overestimate.

Source: Bank of Italy.

Regarding the Japanese yen portfolio, the results of the **joint tests**, over a large part of the increasing size windows, fall in the **rejection region** at the 5% confidence

level (Table 3, mixed Christoffersen test). However, **the frequency tests** indicate that the VaR model appears to be **well calibrated**. Both tests are passed in the presence of p-values significantly higher than the threshold (Table 3 and Graph 3, left-hand panel).

Test results for JPY portfolio
Whole period

Table 3

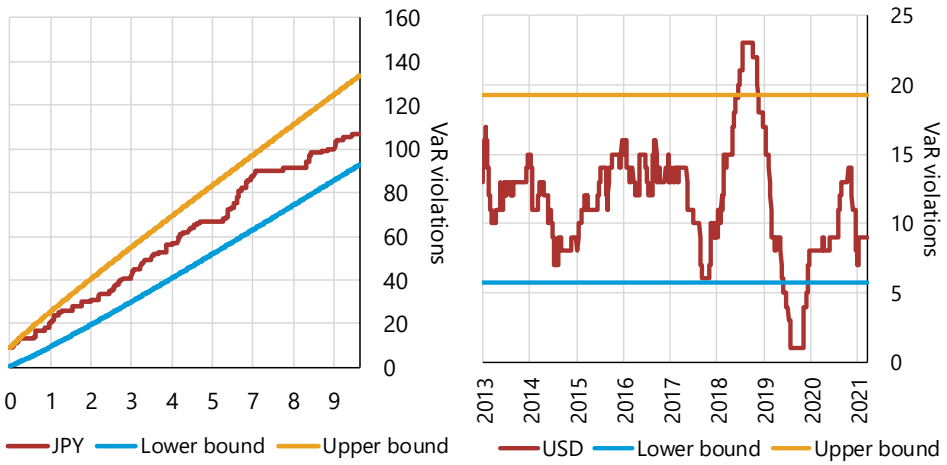
Class of test	Test	p-value / #breaches	Rejection
Joint	Mixed Kupiec	0.00%	5%
	Mixed Christoffersen	0.00%	5%
Frequency	Binomial	107	< 92.7 , > 133.4
	POF	55.30%	5%
Independence	Christoffersen	0.00%	5%
	TBFI	0.00%	5%

Source: Bank of Italy.

The results obtained with a reference time window of fixed length confirm the good performance of the model, although the series shows alternating periods of over and underestimation of risks (Graph 3, right-hand panel).

JPY portfolio binomial test results
Increasing and rolling backtesting period

Graph 3



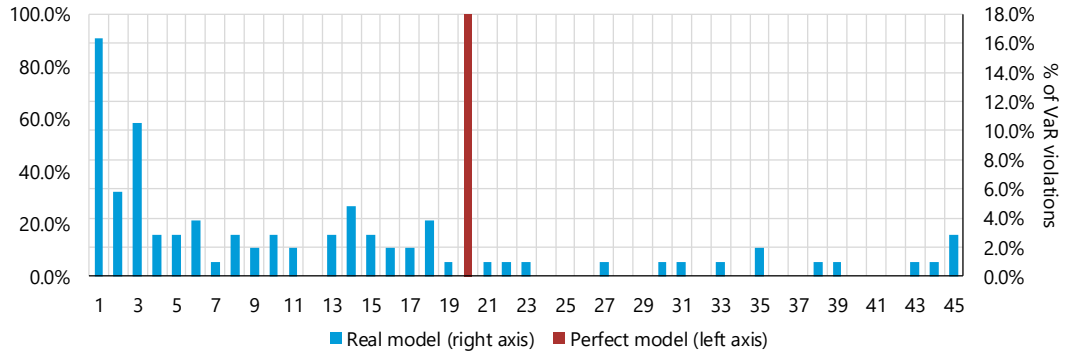
Results of the binomial test. The amount of data used for the tests (in years, left-hand panel; equal to 250 days, right-hand panel), starting from the most recent observations, is shown on the x-axes; the number of VaR violations on the y-axes. The rejection region is represented by the area that is not included between the orange line, which indicates an underestimate of the VaR, and the blue one, which indicates an overestimate.

Source: BdI.

The clustering of the violations might be the reason behind the rejection of the joint tests, as both the Christoffersen's **independence test** and the TBFI were

rejected. The Christoffersen independence test shows good results for half of the sample, where there is a 16% chance of having another violation the next day and almost a 40% chance within five days (Graph 4). This evidence highlights the limitations of a model designed to be responsive to market conditions.

TBFI test – focus on the breach distance Graph 4



Distribution of distances (in days) between VaR violations. The orange histogram represents the theoretical model of the TBFI test (left axis, all violations are spaced 1/α days apart), the blue histograms indicate the real observations (right axis).

Source: Bank of Italy.

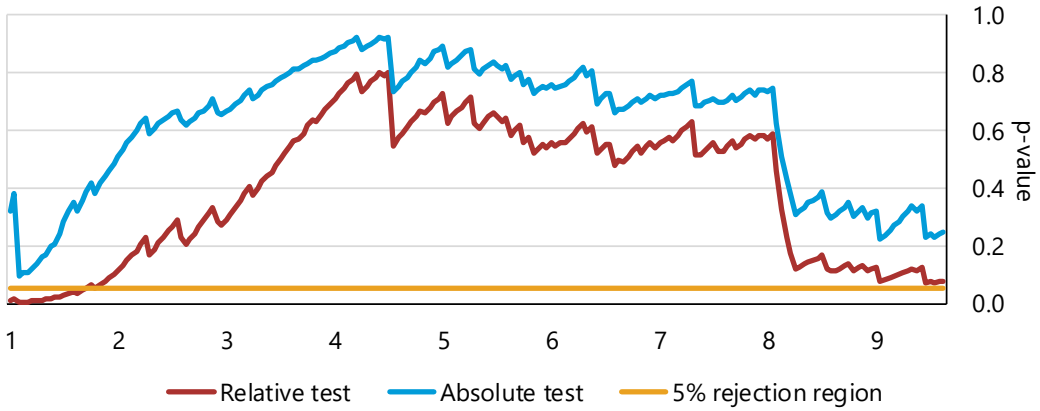
6. ES backtesting results

For the USD portfolio, **ES forecasts calculated using the** RiskMetrics™ model **show an 'average performance' that is satisfactory** when considering AS tests with increasing backtesting periods (Graph 5). The p-values for the entire data set (spanning over nine years, 01/04/2012–31/03/2021) are 0.25 and 0.08 for the absolute and relative test respectively.

Despite large fluctuations in the p-values, **absolute tests are always passed**. However, **relative tests are rejected** for backtesting periods shorter than about two years. For longer periods, the relative p-values are always greater than 0.05, but lower than the absolute ones.

USD portfolio AS test results
Increasing backtesting period

Graph 5



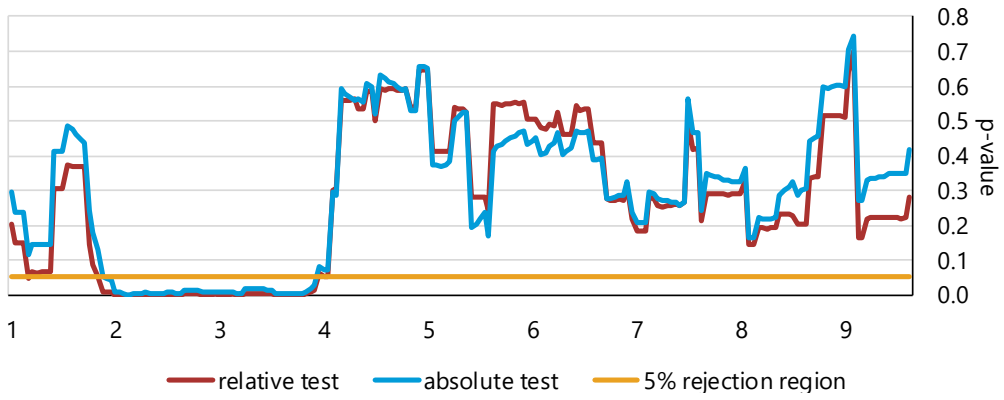
The figure shows the results of relative and absolute AS tests; the backtesting periods, indicated on the horizontal axis, are of increasing size (in years), defined starting from the most recent observation in the data set.

Source: Bank of Italy.

The AS test results over rolling windows (spanning 250 trading days) are presented in Graph 6, which sheds light on the significant drop in p-values observed in the right-hand side of Graph 5. Specifically, the rolling windows p-values approach zero in the period from the beginning of May 2013 to that of the following July.[15]

USD portfolio AS test results
Rolling backtesting period

Graph 6



The figure shows the results of relative and absolute AS tests; the backtesting periods consist of rolling observation windows of constant width (250 days), starting at different dates indicated on the horizontal axis.

Source: Bank of Italy.

[15] During this sample period, data and news began to circulate about the possible shift of the Federal Reserve towards a restrictive monetary policy stance; this led to sudden increases in Treasury yields on certain dates and consequently to significant violations of the VaR, concentrated in a short period of time.

As previously explained, the AS test is "sharp" at first order, which enables the assessment of the impact of forecast errors. Such errors can be evaluated, on average, using the concept of realised ES (2) and/or the scaling factor $\widehat{\varphi_{ES}}$ (6). However, to analyse the effect of daily forecast errors on p-values, we found that $Z_{ES_\alpha}^{rel}$ could be expressed as follows:

$$Z_{ES_\alpha}^{rel} = \frac{\alpha \cdot (e - v) - (x + v)_-}{\alpha \cdot e} \qquad (7)$$

Given that $\alpha = 0.05$, conditional on VaR violations, the term $\alpha \cdot (e - v)$ is small compared with $(x + v)_-$. Therefore, $Z_{ES_\alpha}^{rel}$ can be approximated as:

$$Z_{ES_\alpha}^{rel} \cong -\frac{(x+v)_-}{\alpha \cdot e} \quad (8)$$

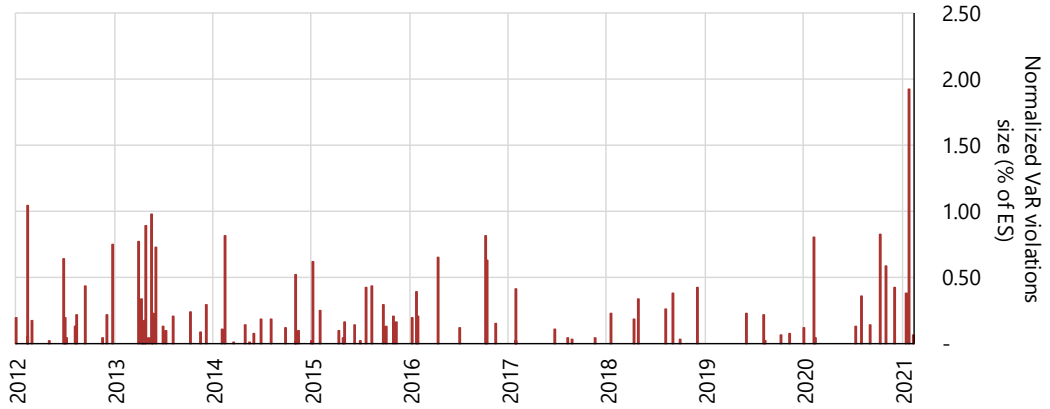
which implies:

$$-\alpha \cdot Z_{ES_\alpha}^{rel} \cong \frac{(x+v)_-}{e} \qquad (8')$$

Thus, we observed that $Z_{ES_\alpha}^{rel}$ (resp. $-\alpha \cdot Z_{ES_\alpha}^{rel}$) tends to be negative (resp. positive) conditional on VaR violations, which contributes to rejecting the null hypothesis. The normalised size of VaR forecasts violations, $\frac{(x+v)_-}{e}$ (ie the size of the VaR forecast violation as a percentage of the ES forecast), can be considered a proxy of daily relative errors on ES forecasts and used to analyse the effect of such errors on p-values. We calculated these sizes for each day in the data set.

Graph 7 illustrates the normalised violations of VaR forecasts for the USD portfolio. We observe that the low p-values reported for the 2021 period in Graph 6 are due primarily to the large violation occurred in February 2021,[16] which was seven times greater than the average.

USD portfolio normalised VaR violations — Graph 7



The figure shows the size of daily VaR violations (difference between the observed return and the VaR forecast) expressed as a percentage of the ES forecasts for the same date.

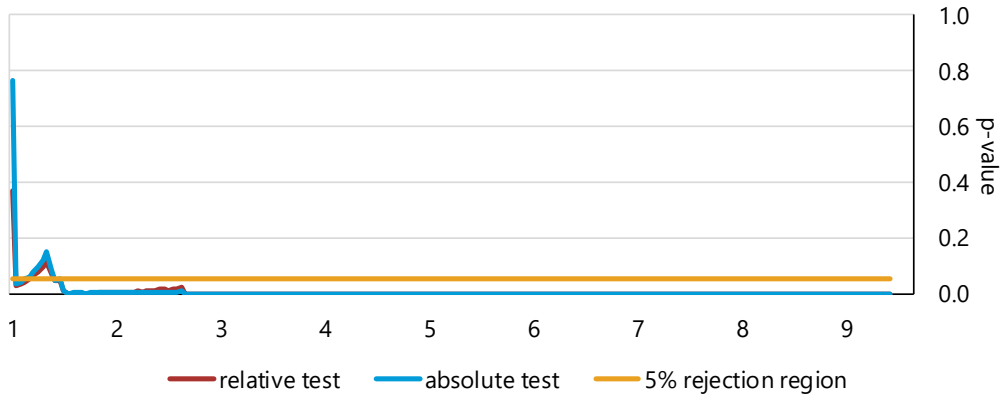
Source: Bank of Italy.

[16] This violation could be related to the 24 February FOMC meeting, whose monetary policy decision surprised markets that were expecting an interest rates increase.

Furthermore, the critical period of 2013 accounts for four out of the eight violations in the sample period that exceed a size of 0.7. The observed frequency of VaR forecasts violations in the entire data set is 0.0405, which is below the expected value of 0.05. On average, these violations have a size of 0.28.

In contrast, when considering the **JPY portfolio** and using the RiskMetrics™ model to calculate ES forecasts, the results show **unsatisfactory "average performance"** as the size of the backtesting period increases (Graph 8). For almost every backtesting period longer than one year, p-values are in fact close to zero.

JPY portfolio AS test results — Graph 8
Increasing backtesting period



The figure shows the results of the relative and absolute AS tests; the backtesting periods, indicated on the horizontal axis, are of increasing size, defined starting from the most recent observation in the data set (29/3/2021).

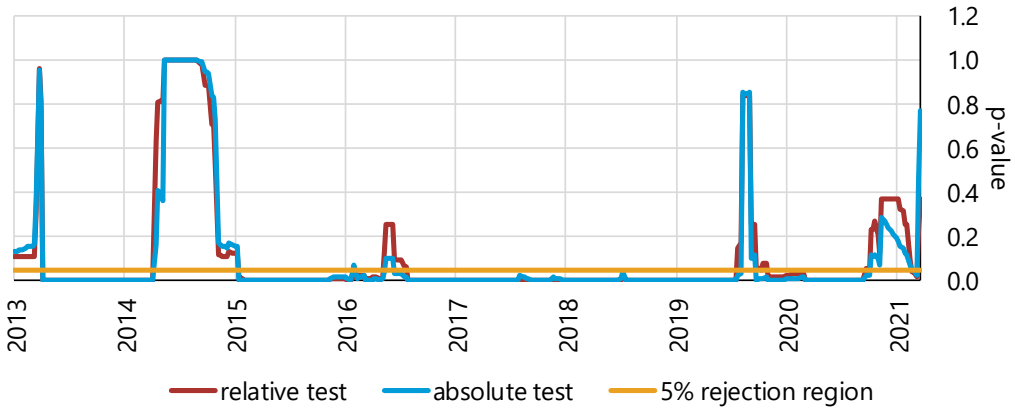
Source: Bank of Italy.

Graph 3, which presents the VaR binomial test on increasing time windows, indicates that the JPY portfolio's number of VaR violations aligns with its expected value. On average, the model correctly estimates portfolio volatility and VaR. However, the AS tests on rolling time windows are not rejected when the number of VaR violations is very close to or smaller than the lower bound of the binomial test (as shown in Graphs 3 and 9). During these periods, VaR violations are much lower than their expected value (which is 12.5 violations for 250 observations). This implies that the model overestimates the VaR as well as the ES (which is a consequence of overestimating the VaR).

In contrast, when the number of VaR violations exceeds the upper limit of the binomial test (backtesting periods starting between April and September 2015) the p-value of the AS test is very close to zero, as shown in Graph 9; even when the number of violations is close to its expected value (as depicted in Graph 3) the AS test still yields a low p-value.

JPY portfolio AS test results
Rolling backtesting period

Graph 9



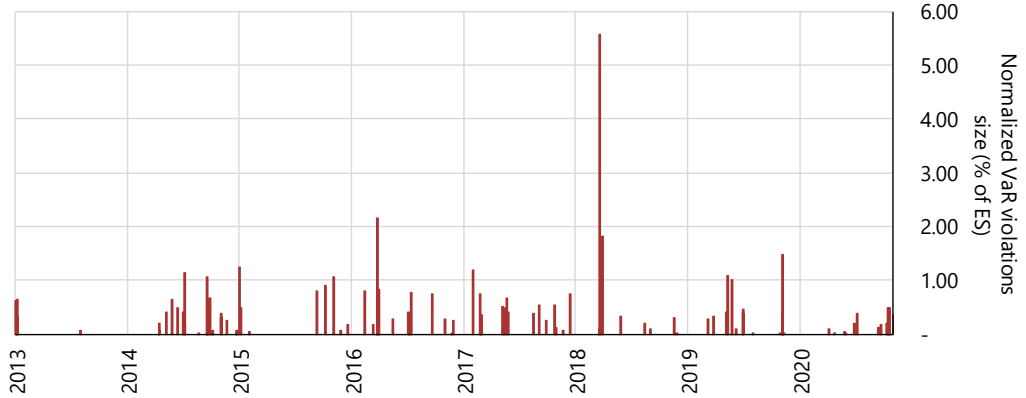
The figure shows the results of relative and absolute AS tests; the backtesting periods consist of rolling observation windows of constant width (250 days), starting at different dates indicated on the horizontal axis.

Source: Bank of Italy.

These observations, combined with the high kurtosis of the observed portfolio returns, suggest that **the AS test tends to reject the null hypothesis for the JPY portfolio**, despite the VaR forecasts being generally accurate. This is because **the model's conditional distribution fails to predict extreme returns**, resulting in a negative difference between the average forecast of the ES and the "realised" ES, which leads to the rejection of the ES model.

JPY portfolio normalised VaR violations

Graph 10



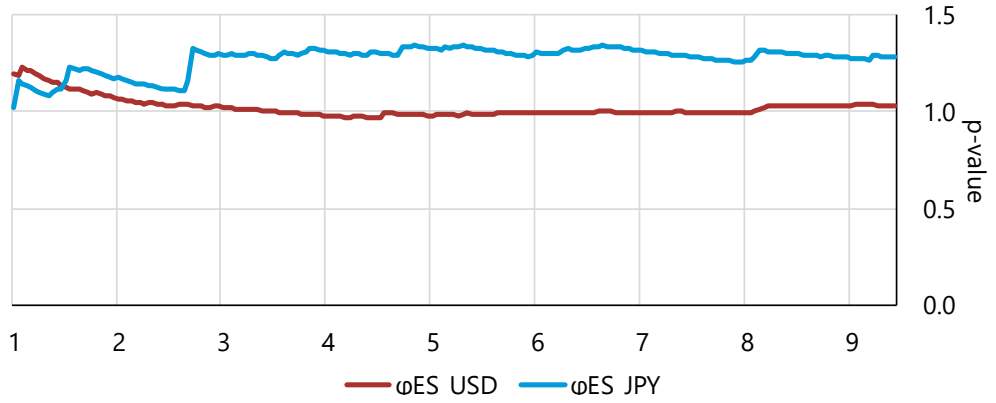
The figure shows the size of daily VaR violations (difference between the observed return and the VaR forecast) expressed as a % of the ES forecasts for the same date.

Source: Bank of Italy.

Graph 10 presents data which confirm that the frequency of VaR violations over the entire historical series is only slightly lower than expected (0.0473 versus 0.05). Additionally, for the JPY portfolio, normalised VaR violations are on average larger than for USD and show very high spikes, reaching values greater than 5.

Finally, according to Acerbi and Szekely (2019), bias effects become significant when $\widehat{\varphi_{ES}}$ deviates from 1 by more than approximately ± 60%. Graph 11 displays $\widehat{\varphi_{ES}}$ values for USD and JPY portfolios, which indicate that it stays within the threshold recommended by the authors. Moreover, the $\widehat{\varphi_{ES}}$ for USD data is very close to 1 on average, while for the JPY portfolio, it is consistently greater than 1 (with an average value of 1.27); this implies that the average ES forecasts are lower than the realised ES for this currency.

$\widehat{\varphi_{ES}}$ for USD and JPY portfolios — Graph 11



The figure shows $\widehat{\varphi_{ES}}$ values for both USD and JPY portfolios, calculated for backtesting periods of increasing size (in years).

Source: Bank of Italy.

Conclusions

This paper presents a strategy for implementing effective backtesting of the risk measures most commonly adopted for managing financial portfolios in central banks.

We used well established techniques to validate VaR forecasts while, for ES, we focused on the innovative method proposed by Acerbi and Szekely. The latter provides an absolute type validation procedure for ES models, despite the fact that this measure is not strictly backtestable. All techniques were applied to the daily time series of risk measures for two foreign reserves portfolios of the BdI.

Results show that VaR forecasts are well specified, although doubts remain about the independence of the related violations, which is a desirable feature of a sound financial risk models. Regarding ES forecasts, the presence of inadequately captured fat tails leads to the rejection of the model, which often tends to underestimate risk. Furthermore, implementing the Acerbi-Szekely test on these real data has shown its

peculiar characteristics: the Acerbi-Szekely test is a conservative framework that provides robustness to the validation process and is particularly suitable for risk-averse institutions quantifying discrepancies between prediction and actual values.

The empirical exercise has shown that the ridge backtesting could represent a powerful analysis tool for a central bank too. It is useful for a deeper assessment of the reliability of actual models, model selection when considering alternative models or, at least, for a more aware management of model risk. As regards avenues for further research, efforts should be aimed at providing solutions to overcome the problems encountered, by extending, for instance, the backtesting activity towards non-parametric models. In theoretical terms, the challenge will be defining a coherent power analysis framework for the Acerbi-Szekely test.

References

Acerbi, C and B Szekely (2014): "Backtesting expected shortfall", MSCI Inc, December.

——— (2017): "General properties of backtestable statistics", MSCI Inc, January.

——— (2019): "The minimally biased backtest for ES", *Risk*, September.

Acerbi, C and D Tasche (2002): "Expected Shortfall: A Natural Coherent Alternative to Value at Risk", *Economic Notes*, vol 31, issue 2, pp 379–88.

Artzner, P, D Heath, F Delbaen and J Eber (1997): "Thinking coherently" *Risk*, vol 10, pp 68–71.

——— (1999): "Coherent measures of risk", *Mathematical Finance*, vol 9, pp 203–28.

Christoffersen, P (1998): "Evaluating interval forecasts", *International Economic Review*, vol 39, no 4, pp 841–62.

Diebold, F and R Mariano (1995): "Comparing predictive accuracy", *Journal of Business & Economic Statistics*, vol 13, no 3, pp 253–63.

Fissler, T and J Ziegel (2016): "Higher order elicitability and Osband's principle", *The Annals of Statistics*, August, vol 44, no 4, August, pp 1680–707.

Gneiting, T "(2011): "Making and evaluating point forecasts", *Journal of the American Statistical Association*, vol 106, pp 746–62.

Haas, M (2001): "New methods in backtesting", Financial Engineering, Research Center Caesar, Bonn.

Hendricks, D (1996): "Evaluation of value-at-risk models using historical data", *Economic Policy Review*, pp 39–69.

Kupiec, P (1995): "Techniques for verifying the accuracy of risk management models", *Journal of Derivatives*, vol 3, pp 73–84.

Nieppola, O (2009): "Backtesting value-at-risk models", master's thesis, Helsinki School of Economics.