INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Backtesting Expected Shortfall: Comparative study and impact analysis on capital requirements

João Miguel Garcias Catarino

Master in Finance

Supervisor:
PhD, Joaquim Paulo Viegas Ferreira de Carvalho, Assistant Professor, ISCTE-IUL

October, 2023

iscte

BUSINESS
SCHOOL

Department of Finance

Backtesting Expected Shortfall: Comparative study and impact analysis on capital requirements

João Miguel Garcias Catarino

Master in Finance

Supervisor:
PhD, Joaquim Paulo Viegas Ferreira de Carvalho, Assistant Professor, ISCTE-IUL

October, 2023

Acknowledgments

Firstly, I extend my gratitude to my supervisor Paulo Viegas de Carvalho for his support, guidance, and mentorship throughout the course of my research.

I am also profoundly thankful to all of my family, friends, and my Soraia for their unending encouragement and understanding. Their support has been a cornerstone during this academic journey.

Resumo

O documento Fundamental Review of the Trading Book trouxe algumas mudanças nas normas de Basileia para o cálculo dos requisitos de capital, particularmente com a substituição do Value at Risk pelo Expected Shortfall. Essa mudança gerou alguma discussão, uma vez que não existe uma resposta definitiva para a questão de como fazer backtesting ao último. Esta dissertação avalia os resultados de diferentes métodos de backtesting para testar o rigor das estimativas de Value at Risk e de Expected Shortfall de uma carteira de índices internacionais. O objetivo é avaliar se é possível incluir um requisito direto de backtesting ao Expected Shortfall em futuras alterações regulamentares, bem como na validação interna de modelos de risco. Para fazer essa avaliação, as diferentes metodologias são apresentadas de forma que sejam comparáveis com a abordagem de semáforo do FRTB, em termos de resultados dos testes e o seu impacto nos requisitos de capital ao longo dos últimos vinte anos. Os resultados mostram que, embora seja possível fazer backtesting ao Expected Shortfall, incluí-lo como requisito nos quadros de Basileia exigiria mudanças adicionais na regulamentação para evitar um impacto considerável e possivelmente instável nos requisitos de capital dos bancos.

Palavras Chave: Expected Shortfall, Value at Risk, Backtesting, Requisitos de Capital, FRTB
Classificação JEL: G28, G32

Abstract

The Fundamental Review of the Trading Book brought some changes to the Basel standards for the calculation of capital requirements, particularly with the replacement of Value at Risk with Expected Shortfall. This change came with some discussion, as there is no definitive answer to the question of how to backtest the latter. This dissertation applies existing backtesting methods to test the accuracy of both Value at Risk and Expected Shortfall estimates of an international equity index portfolio. The aim is to assess if it is possible to include Expected Shortfall backtesting in future regulation changes, as well as in internal risk model validation. To do this assessment, different methodologies will be presented in a way that allows comparisons against the FRTB's traffic-light-approach, in terms of testing results and their impact on the capital requirements over the past twenty years. The results show that even though backtesting Expected Shortfall is possible, including it in the Basel framework would require further changes in the regulation to prevent considerable and possibly unstable shifts in the banks' capital requirements.

Keywords: Expected Shortfall, Value at Risk, Backtesting, Capital Requirements, FRTB
JEL Classification: G28, G32

Contents

List of Tables

List of Figures

Glossary

A - Amber Zone

AR - Autoregressive

ARMA - Autoregressive Moving Averages

BCBS - Basel Committee on Banking Supervision

C&C - Constanzino and Curran's test

c.d.f. - Cumulative Distribution function

d.f. - Degrees of Freedom

ES - Expected Shortfall

EUR - Euro

EWMA - Exponentially Weighted Moving Averages

FRTB - Fundamental Review of the Trading Book

G - Green Zone

GARCH - Generalized Autoregressive Conditional Heteroskedasticity

GJR - Glosten, Jagannathan, and Runkle

HKD - Hong Kong Dollar

HS - Historical Simulation

i.i.d. - Independent and Identically Distributed

IMA - Internal Models Approach

JPY - Japanese Yen

LRT - Likelihood Ratio Test

LR - Likelihood Ratio

M&P - Moldenhauer and Pitera's test

MN - Multinomial Distribution

P&L - Profit and Loss

p.d.f. - Probability Density Function

R - Red Zone

RM - RiskMetrics

USD - United States Dollar

VaR - Value at Risk

Z1 - Acerbi and Szekey's first test

Z2 - Acerbi and Szekey's second test

Z3 - Acerbi and Szekey's third test

ZR - Acerbi and Szekey's Ridge test

1 Introduction

For the past years, Value at Risk (VaR) has been the standard method of quantifying market risk, recommended by international regulators to determine minimum capital requirements that banks must hold.

However, VaR proved to be too insensitive to extreme events, as it fails to consider the severity of potential losses beyond a certain threshold, leading to insufficient liquidity levels if those events occur (and they have). Essentially, VaR aims to predict the worst expected outcome within a certain confidence level. However, confidence is not certainty, and VaR does not provide any information on what happens if the losses are even worse than the worst expectation. For that reason, the latest international regulatory framework (Fundamental Review of the Trading Book, published in 2016, revised in 2019) by the Basel Committee on Banking Supervision (BCBS) points to the replacement of VaR with Expected Shortfall (ES) as a more prudent way of determining capital adequacy. Expected Shortfall provides the expected loss in the presence of extreme events, that is, those when VaR breaches occur. ES, however, has some shortcomings when compared to VaR, with the most relevant of them regarding model backtesting.

To assess the accuracy of risk models and their parameters, they must be tested against the realized returns of the assets in question. Additionally, capital surcharges are applied depending on backtesting results. While for VaR that is straightforward, consisting essentially of counting the number of losses that surpass it, it is not the case for ES. To overcome this problem, the FRTB recommends a backtesting procedure that is still based on the VaR estimate instead of ES. This approach stirred some debate in the field, as the measure being used to set the capital amounts is not the one that is being tested, which means that the procedure is not precisely testing the coverage of the capital requirements. Although VaR is easier to backtest than ES, this does not mean that there is no way of backtesting the latter, as several methods have been developed.

This dissertation compares some of these methods in terms of admissibility, computational burden, and model dependence. The tests included in the analysis are two of the three multinomial tests proposed by Kratz et al. (2018), four of the five tests developed by Acerbi and Szekely (2014; 2017), the cumulative violations test (Constanzino and Curran, 2015; Du and Escanciano, 2016) and the secured position test of Moldenhauer and Pitera (2019). The goal of this study is not exactly to propose changes in the regulation, but rather to better understand just how feasible backtesting ES directly can be in the context of both internal and external validation (regulation), how it could be performed, and the impact it can have on capital requirements.

The study reveals that the presented Expected Shortfall backtesting methods lead to different outcomes when it comes to model rejections, but are generally less prudent than the FRTB's traffic-light approach. Regarding the capital requirements, the simplest way of using Expected Shortfall is by calculating the capital surcharge from the ratio of the average value of VaR breaches to the average ES estimate, leading to a simple, model-free formula. This is based on the so-called *Ridge Backtest* (Acerbi and Szekely, 2017) and results show that this approach is likely to lead to considerably higher capital requirements. To the best of my knowledge, this idea and conclusion are new to the existing literature. Using other methods would require parametric assumptions to determine the capital add-ons, and their impact would therefore be dependent on those assumptions.

The debate between VaR and ES and the search for the most adequate risk measures have long been present in the literature, more so since the latest regulation's shift towards ES. With this study, I hope to bring a new, different judgment into the discussion and contribute to banks' and regulators' pursuit of prudent practices regarding market risk.

To do that, I start by comparing VaR and ES in common ground in Section 3, exploring each measure's attractive properties and disadvantages, paying special attention to the backtesting problem, with an analysis of FRTB's approach, other VaR backtesting methods, and finally the proposed procedures for ES backtesting. Across sections 4 and 5, both VaR and ES will be computed with different specifications for a hypothetical portfolio, representative of banks' assets. The backtesting methods will then be applied to each of the measures to assess and compare their model rejection frequencies and possible effects on the capital requirements. Conclusions are presented in section 6.

2 Literature Review

The 2007-2008 financial crisis exposed the Basel II standards, in practice at the time, as inadequate and insufficient when it came to market risk capital requirements. To overcome the weaknesses of the framework, BCBS published the Fundamental Review of the Trading Book to be implemented into the new minimum capital requirements for market risk as part of Basel III. One of the proposed changes for banks using internal models, the so-called Internal Models Approach (IMA), is the use of Expected Shortfall at a 2.5% significance level ($\mathrm{ES}_{2.5\%}$), replacion Value at Risk at a 1% significance level ($\mathrm{VaR}_{1\%}$). By doing this, banks account for tail risk and guarantee more conservative capital levels. To qualify for the IMA, banks must (among other prerequisites) conduct a regular backtesting procedure, based on the 1-day $\mathrm{VaR}_{1\%}$ estimates of the prior 12 months. This includes a traffic-light test, depending on the number of VaR breaches recorded.

For a given significance level α, VaR_α is derived from the α-quantile of the P&L distribution. This can be estimated in different manners, namely by assigning a parametric distribution to the returns, using past returns to obtain the quantile, or via Quantile Regression, among other methodologies. Expected Shortfall (ES) is computed after VaR, as it consists of the expected value of the losses that surpass that point.

Expected Shortfall's attractiveness is not only due to being sensitive to tail events but also due to satisfying the useful properties of coherence. To classify as coherent (Artzner et al., 1999; Delbaen, 2002) a statistic must be monotonous, sub-additive, positively homogenous, and translation invariant. By standardizing the definition of ES, Acerbi and Tasche (2002) prove that this measure is coherent, as opposed to VaR, which is not sub-additive. They take it one step further by extending the concept of coherence to the very definition of a risk measure, claiming that any function, such as VaR, not satisfying the subadditivity or any of the other three axioms, should not be classified as a risk measure at all. It is at this point, when ES seems to be a favorite, that the problem of backtesting arises.

Backtesting VaR is a rather simple process, consisting essentially of counting the number of exceedances, which occur when a loss beyond VaR is recorded. Since VaR is a quantile, it is expected that the exceedance rate is close to the level of significance. This is the rationale behind Kupiec's (1995) Unconditional Coverage test. Under this hypothesis, the exceedances should follow an independent and identically distributed (i.i.d.) Bernoulli process. Christoffersen's (1998) Conditional coverage test tests for both the accuracy of the predictions and first-order independence of exceedances (Berkowitz et al., 2011, present a higher-order autocorrelation test).

For ES, it gets more complicated, as we would be dealing not with the number of exceedances, but with their average magnitude. The first attempts to backtest ES came from McNeil and Frey (2000), Berkowitz (2001), and Kerkhof and Melenberg (2004), each with their particularities. These methods, however, are somewhat model-dependent, as they rely on strong parametric assumptions or large exception samples. Both these factors are very problematic since the tests should be able to be performed on different models and ideally the sample size, that is the number of exceedances within a year, should be less than a dozen. Wong (2008) and Rigghi and Ceretta (2013) propose alternatives, based on the so-called saddle point technique and truncated distribution, respectively, that can backtest ES with smaller samples. However, these still require considerable assumptions, while also bearing a heavy computational and analytical burden.

A strong argument against backtesting ES came when Gneiting (2011) demonstrated that ES, contrary to VaR, lacks a mathematical property called elicitability. A statistic is elicitable if it minimizes the expected value of some scoring function. This was first received as a long-awaited confirmation that ES cannot be backtested at all. However, this idea has since been refuted by many authors. Acerbi and Szekely (2014) argue that "elicitability has in fact nothing to do with backtesting" (Acerbi and Szekely, 2014, p. 9) but is instead related to relative model selection. They also introduce three non-parametric testing methods for ES that show more power than the VaR tests but are, however, reliant on Monte Carlo simulations or resampling. Finally, this paper also introduces the concept of joint elicitability of ES and VaR into the debate. Fissler et al. (2016) and Nolde and Ziegel (2017) further explored this idea by proposing a new comparative approach to backtesting ES and VaR simultaneously, based on the scoring functions of the joint elicitability. This procedure redefines the traditional hypothesis by adding a benchmark model against which the "internal" model is to be tested. Essentially, this method works for comparison between models (model selection) and not for individual validation. This is in line with the idea of Acerbi and Szekely (2014). In another paper (Acerbi and Szekely, 2017) the same authors dwell on the definition of backtestable statistics (of which ES is not a part), and its relation with elicitability. A backtest is defined as "a null expected value involving only the statistic and its random variable, strictly monotonic in the former" (Acerbi and Szekely, 2017, p. 1).

Another interesting approach comes from approximating ES from VaR. Acerbi and Tasche (2002) were the first to define ES as an integrated VaR, that is, as the continuous average of VaRs at different significance levels. Emmer et al. (2015) proposed approximating the integral as a Riemann sum of VaR with four terms (significance levels) and then backtest each VaR individually. Kratz et al. (2018) extend this idea to N quantiles and propose a single multinomial unconditional coverage test for ES. Despite this method being applicable to any number of quantiles, the authors conducted numerical studies that point to four or eight different levels as the most powerful approximation.

The integrated VaR representation allowed Du and Escanciano (2016) to develop both unconditional and conditional coverage tests for ES, based on the so-called cumulative violation process. These tests are particularly interesting because not only they are analogous to Kupiec's and Christoffersen's VaR tests, but the test statistics, obtained via the Probability Integral Transform (PIT), are asymptotically normal and Chi-squared distributed, respectively. This allows for more direct testing, with no need for resampling or discrete approximations. Constanzino and Curran (2015) propose a similar unconditional coverage test with an equivalent statistic, constructed around the concept of spectral risk measures, introduced by Acerbi (2002).

Deng and Diu (2021) reviewed the most prominent of these procedures and compared them in terms of testing power (the probability of correctly rejecting the null hypothesis when it is false) and size (the probability of incorrectly rejecting the null hypothesis when it is true), relating them to the duality between backtesting error and estimation error. The authors study each method's performance for different estimating and backtesting sample sizes. For a fixed backtesting sample of 250 observations, they conclude that testing size and power generally improve as the estimation sample increases since the estimation error becomes relatively small. For this backtesting sample size and an estimation sample above 1000 observations, the authors point to the use of unconditional instead of conditional coverage ES tests.

Finally, Moldenhauer and Pitera (2018) suggest a different, non-parametric approach to backtesting ES using the so-called secured position, which consists of the observed P&L backed by the risk estimate. This procedure is essentially a test of the coverage of the capital requirements over some period instead of a test of the adequacy of daily estimates to observed returns. This comes as a clear response to the fact that divergences between $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ are likely to occur, meaning that the FRTB's method is not a proper approximation to the coverage of capital requirements determined by $\text{ES}_{2.5\%}$. Nevertheless, the authors claim that their approach is a "natural extension of the regulatory traffic-light approach for Value-at-Risk" (Moldenhauer and Pitera, 2018, p.1).

The search for an efficient way of backtesting ES directly has preoccupied both researchers and regulators for the past 20 years. However, the connection between the two, that is the adaptation and applicability of the developed backtesting procedures into the regulation's standards and requirements, is something that is not vastly covered in the literature. With this dissertation, I aim to explore these topics and conclude which of the procedures, if any, are better suited to be included in the regulatory framework, and which are better suitable for internal model validation.

3 Value at Risk vs Expected Shortfall

3.1 Value at Risk

Let X_t represent the daily uncertain P&L of some risky position and $F_t(x)$ the real unknown cumulative distribution of X_t. For a given significance level α, the 1-day $\text{VaR}_{\alpha,t}$ is defined as the worst expected loss with $1-\alpha$ confidence, which corresponds to the absolute value of the α-quantile of X_t, that is:

$$VaR_{\alpha,t} = -q_{\alpha}(X_t) = -\text{inf}\,\{x : F_t(x) \geq \alpha\} \tag{1}$$

Naturally, the distribution of X_t is not known. Therefore, VaR must be estimated from some predictive model P_t. There are many approaches to this, both parametric and non-parametric. The standard parametric approach consists of assigning some theoretical distribution $P_t(X_t, \vec{\Theta})$, where $\vec{\Theta} \in \mathbb{R}^n$ is a set of parameters, to the P&L and computing its quantile. The simplest alternative is using a normal distribution with mean μ_t and volatility σ_t. This allows us to compute VaR as[1]:

$$VaR_{\alpha,t} = -\mu_t - \sigma_t \Phi^{-1}(\alpha) = -\mu_t + \sigma_t \Phi^{-1}(1-\alpha) \tag{2}$$

where Φ is the cumulative distribution function (c.d.f.) of the standard normal distribution. For a one-day measurement horizon, the expected return should be close to zero, making μ_t negligible[2]. Regarding σ_t, there are many ways to estimate it, other than simply using the standard deviation of past returns. These include moving average models such as EWMA[3] (Exponentially Weighted Moving Averages).

The normality assumption, however, is usually flawed. Negative skewness and excess kurtosis (fat tails) are typically present in observed data, meaning that the normal distribution underestimates the presence of extreme losses. For this reason, alternative distributions are common, such as generalized versions of Student-t or Pareto distributions.

More common is the use of non-parametric models. The most popular is the Historical Simulation method (HS), which consists of retrieving the quantile directly from a sufficiently large sample of past returns. This is essentially estimating $P_t(X)$ from the empirical distribution of past returns. This method has the advantage of not relying on uncertain parametric assumptions, thus accounting for observable particularities such as fat tails. The challenge of the HS is its sensitivity to the sample size, which poses the following dilemma. To obtain a good estimate, a large sample is necessary. However, by

[1] This result is possible due to the equivariance property of quantiles.

[2] For longer horizons, ignoring the drift adjustment ($-\mu_t$) relates to Benchmark VaR (BVaR), relative to the expected return.

[3] This is the model used in RiskMetrics, a methodology developed by JP Morgan in 1996, a reference for risk managers.

going too far into past returns, the estimates would be influenced by past events that should not have the same effect as more recent ones. It is logical to presume that more recent information, regarding returns and volatility, has a higher impact on the P&L distribution. To accommodate this, adjustments to the sample of past returns are recommended, namely weighting them based on their "age", making the estimate more sensitive to recent events. Another alternative is to improve the sample to account for volatility changes. Hull and White (1998) propose adjusting the returns to the difference between the historical volatility and the current or predicted volatility (this methodology will be better detailed in section 4). With these adjustments, it becomes possible to combine the advantages of HS with the structure of econometric models like GARCH processes. Barone-Adesi and Giannopoulos (1998) extend a similar approach to h-day ahead VaR, in a methodology called Filtered Historical Simulation (FHS).

Other relevant methodologies include computing VaR from Monte Carlo simulations or using Quantile Regressions (Engle and Manganelli, 2004; Xiao et al., 2014).

VaR's attractiveness is mostly reliant on its simplicity, easy interpretation, and practical application. Its main disadvantage is the fact that it is limited to its confidence level and says nothing about the potential severity of the worst α of scenarios.

3.1.1 Backtesting Value at Risk

The standard VaR backtesting procedure begins by identifying VaR violations, or exceptions. One can define the violation function as the indicator function of a VaR violation, that is:

$$I_{\alpha,t} = \mathbb{1}_{\{r_t < -VaR_{\alpha,t})\}} \tag{3}$$

where r_t represents the return (P&L) observed at time t, that is, the realized X_t.

If the model is well specified, $I_{\alpha,t}$ should follow an i.i.d. Bernoulli process with probability of success (a violation) α, meaning the total number of violations in a sample of T days would be distributed along a Binomial(T, α) distribution. From here, it is already possible to assess the accuracy of a VaR model, by comparing the number of VaR exceptions recorded over some time with the probability of such an amount occurring in that binomial distribution.

This is the rationale behind the VaR backtesting requirements present in the FRTB. The procedure is performed over a sample of 250 trading days, shortly under 1 year, and is summarized in Table 3.1.

Models are classified in one of three zones, determined from the cumulative probability of the recorded exceptions. They are allocated to the green zone when the cumulative probability is below 95%, in the amber zone if it exceeds 95% but is below 99.99%, and in the red zone for cumulative probabilities equal to or higher than 99.99%. For models placed in the amber and red zones, a capital add-on is enforced based on the number of exceptions recorded.

Backtesting Zone	Number of exceptions	Backtesting Multiplier	Cumulative Probability
Green	0	1.50	8.11%
	1	1.50	28.58%
	2	1.50	54.32%
	3	1.50	75.81%
	4	1.50	89.22%
Amber	5	1.70	95.88%
	6	1.76	98.63%
	7	1.83	99.60%
	8	1.88	99.89%
	9	1.92	99.97%
Red	≥ 10	2.00	99.99%

Table 3.1. FRTB's traffic-light backtesting zones. The cumulative probability is that of the Binomial Distribution with 250 trials and probability of success of 0.01.

There are other more statistically evolved methods for VaR backtesting. The most prominent are the unconditional coverage test (Kupiec, 1995) and the conditional coverage test (Christoffersen, 1998), the latter of which tests both the frequency of exceptions and their (1 lag) independence. Considering the regulatory framework's disregard for conditional coverage testing, as well as the conclusion of Deng and Diu (2021) that unconditional coverage tests for ES are often more reliable than conditional coverage, I will focus on unconditional coverage testing in this study.

Kupiec's test also sets from the null assumption that exceptions are i.i.d. Bernoulli variables. It then checks if the observed exception rate (π_{obs}) is equal to the expected exception rate ($\pi_{exp} \equiv \alpha$), via the likelihood ratio (LR):

$$LR = \left(\frac{\pi_{exp}}{\pi_{obs}}\right)^{n_1} \left(\frac{1-\pi_{exp}}{1-\pi_{obs}}\right)^{n_0} \tag{4}$$

where $n_1 = \sum_{t=1}^{T} I_{\alpha,t},\ n_0 = T - n_1,\ \pi_{obs} = n_1/T,$ and $\pi_{exp} = \alpha$.

With the logarithm of LR, it is possible to obtain a Chi-squared distributed test statistic, as follows:

$$-2\ln(LR) \sim \chi_1^2 \tag{5}$$

which allows one to test whether the observed exception rate is statistically equal to α, which for the purpose of the analysis in this dissertation is hypothesized as :

Hypothesis 1. Kupiec's unconditional coverage test:

$H_0: \ \pi_{obs} = \alpha$

$H_1: \ \pi_{obs} \neq \alpha$

This is the standard conditional coverage testing method and, along with the FRTB's procedure, it will be included in the analysis ahead, in comparison with existing Expected Shortfall backtesting methodologies.

3.2 Expected Shortfall

Expected Shortfall (ES) with significance level α is defined as the expected loss of a VaR$_\alpha$ exception, that is, losses that equal or exceed VaR$_\alpha$:

$$ES_{\alpha,t} = -E[X_t | X_t \leq -VaR_{\alpha,t}] \tag{6}$$

To estimate ES, the same methodologies as VaR can usually be applied. In the case of a normal distribution parametric model, it will be (ignoring the drift adjustment)[4]:

$$ES_{\alpha,t} = \frac{\sigma_t}{\alpha} \int_0^{\Phi^{-1}(\alpha)} x\phi_t(x)dx = \frac{\sigma_t}{\alpha}\phi(\Phi^{-1}(\alpha)) \tag{7}$$

where ϕ is the standar normal probability density function (p.d.f.).

In the presence of an HS model, ES$_\alpha$ is derived from the observed average of the α-tail of the sample.

The reason for the significance level when changing from VaR$_{1\%}$ to ES$_{2.5\%}$ is that the values are nearly identical for Gaussian returns, with ES$_{2.5\%}$ increasing over VaR$_{1\%}$ for distributions with heavier tails. This serves as a penalization for the latter in the form of higher capital requirements. Table 3.2. shows the different discrepancies between ES$_{2.5\%}$ and VaR$_{1\%}$ for the normal and Student-t distributions with different degrees of freedom (different tail weights).

	Std.	Student-t		
	Normal	$\nu = 10$	$\nu = 5$	$\nu = 2.5$
VaR$_{1\%}$	2.3263	2.7638	3.3649	5.3531
ES$_{2.5\%}$	2.3378	2.8190	3.5216	6.2057

Table 3.2. ES$_{2.5\%}$ and VaR$_{1\%}$ for different-tailed Normal and Student-t distributions. ν represents the degrees of confidence. A lower ν means heavier tails.

ES is a natural response to VaR's unaccountability of extreme losses. It is, however, not so simplistic as VaR, and necessarily requires the calculation of VaR, posing a slightly

[4] The second equality in Equation (7) is exclusive to the normal distribution.

higher computational burden (which is not a big problem). On the other hand, unlike VaR, counting ES exceptions has no significance whatsoever. Instead of the number of exceptions observed, an ES backtest would be based on their average magnitude, which is not as direct.

A useful representation of ES, first introduced by Acerbi and Tasche (2002), comes by performing a variable change and defining ES from an integrated VaR, yielding:

$$ES_{\alpha,t} = \frac{1}{\alpha}\int_0^{\alpha} VaR(u)du \tag{8}$$

where VaR(u) $\equiv$ VaR$_u$. From Equation (8) it is possible to define the cumulative violation process (Constanzino and Curran, 2015, Escanciano and Du, 2016), analogous to Equation (3):

$$H_{\alpha,t} = \frac{1}{\alpha}\int_0^{\alpha} \mathbb{1}_{\{r_t < -VaR(u)\}} du \tag{9}$$

It is possible to construct a test statistic from this process, that would allow for coverage tests of ES. However, as I will show later on, it is not as straightforward or as easily interpretable as for VaR.

The next section will present and analyze the properties and singularities of both measures that relate to their advantages and disadvantages, to better understand the backtesting problem.

3.3 Attractive properties of risk measures

Before entering into the backtestability question, I will start by covering the aforementioned property of coherence.

3.3.1 Coherence

Introduced by Artzner et al. (1999), coherence distinguishes the class of risk measures that satisfy the four axioms of monotonicity, positive homogeneity, translation invariance, and sub-additivity. For a risk measure $\mathbf{y}(\cdot)$, these are defined as:

(i) Monotonicity: $\mathbf{y}(X) \geq \mathbf{y}(W) \quad \text{iff} \quad X \leq W$

(ii) Positive homogeneity: $\mathbf{y}(cX) = c\mathbf{y}(X)$

(iii) Translation invariance: $\mathbf{y}(X + c) = \mathbf{y}(X) - c$

(iv) Subadditivity: $\mathbf{y}(X + W) \leq \mathbf{y}(X) + \mathbf{y}(W)$

Where W and X are random variables and $c \in \mathbb{R}$ is a constant. The interpretation of these properties in the context of financial markets is both simple and important. If W

and X represent the uncertain P&L distributions of two risky positions, (i) means that lower-valued positions will bear higher risk, (ii) and (iii) relate to the effect on the risk measure of the scale of the position and of risk-free assets, respectively and (iv) ensures diversification benefits, stating that the overall risk of the portfolio shall not be greater than the sum of individual positions' risk. The latter takes a particularly important role in this discussion because it establishes a disadvantage of VaR compared to ES.

While ES satisfies the four properties of coherence (Acerbi & Tasche, 2002[5]), VaR does not generally satisfy the property of sub-additivity and therefore is not coherent. Nonsubadditivity of VaR is demonstrated when it comes to portfolios with nonlinear payoffs, such as including options (Artzner et al., 1999). This constitutes a problem regarding the admissibility of VaR.

3.3.2 Elicitabilty

Coined by Lambert et al. (2008), elicitability of a statistic $\mathbf{y}(X)$ of a random variable X is defined as the existence of some scoring function S, whose expected value is minimized by $\mathbf{y}$. That is, $\mathbf{y}$ is elicitable if there is a function S such that:

$$\mathbf{y} = \arg\min_{y} E[S_{\mathbf{y}}(y, X)] \tag{10}$$

The α-quantile of the random variable's distribution (the symmetric of VaR) minimizes the average absolute error between the observed variable and the estimator, weighted by α for observations above the quantile and $1-\alpha$ for observations below it. This means that the α-quantile is elicited by the scoring function[6]

$$S_{q_\alpha}(y, x) = \alpha(x-y)^+ + (1-\alpha)(x-y)^- \tag{11}$$

On the other hand, it is not possible to construct a scoring function whose expected value is minimized by ES alone, making it not elicitable (Gneiting, 2011).

The importance of elicitability stems from the fact that the mean score (realized mean of the score function) allows us to qualify a predictive model for a statistic using only a sample of model estimates and realizations of the random variable, establishing a clear relation with backtesting. However, elicitability alone does not relate to the desired backtesting procedure as the value of the realized mean score has no absolute significance. One can only use it in a comparison between two or more models.

Although useful, elicitability is not sufficient to guarantee backtestability for model validation.

[5] To guarantee sub-additivity for discrete distributions, ES must be defined slightly differently from Equation (6), with no impact whenever $F_t(VaR_{\alpha,t}) = \alpha$, which is the generality of cases. For simplicity, I will make that assumption.

[6] The notation used is the same as the authors', that is: $(a)^+ = \max(a, 0)$ and $(a)^- = -\min(a, 0)$.

3.3.3 Backtestability

According to Acerbi and Szekely (2017), a statistic $\mathbf{y}$ of a random variable X is backtestable if there exists a function $Z(y, x)$ that has a null expected value and is strictly monotonic in y, that is:

$$\begin{gathered} E[Z_{\mathbf{y}}(y, X)] = 0, \\ E[Z_{\mathbf{y}}(y_1, X)] < \; E[Z_{\mathbf{y}}(y_2, X)] \;\text{ iff }\; y_1 < y_2 \end{gathered} \tag{12}$$

In the financial context, Acerbi and Szekely's backtestability refers to the existence of a test function of the estimated statistic and the realized P&L, that qualifies predictions based on their distance from 0 (the further away from 0, the worse a prediction is), distinguishing overestimations from underestimations by the sign of the function's value.

This definition is closely related to elicitability. In fact, not only is the latter a necessary condition for backtestability, but any backtestable function is elicited by the (convex) antiderivative of the test function $Z_{\mathbf{y}}$:

$$S_{\mathbf{y}}(y, x) = \int Z_{\mathbf{y}}(y, x)\, dy \tag{13}$$

On the other hand, the convexity of the scoring function of an elicitable statistic is a sufficient condition for backtestability (and the backtest function will naturally be equal to the derivative of the scoring function). This means that non-elicitable statistics such as ES are not single-handedly backtestable. However, ES allows for what the authors call a *Ridge Backtest*, using a function of ES and VaR as an auxiliary statistic, taking advantage of ES-VaR joint-elicitability. This will be detailed in section 3.

At this point, the backtest function still has no absolute value, meaning it is not enough to perform absolute model validation, that is, hypothesis testing. To do that, one must know the distribution P_Z of $Z_{\mathbf{y}}$ under some null hypothesis. Except for VaR, P_Z must be approximated via simulation or resampling, which requires the daily recording of the entire P&L predictive distribution P_t.

VaR is special in this, due to the Bernoulli distribution of $I_{\alpha,t}$ (Equation 3), which allows for the computation of the p-value without recording the entire predictive model to simulate the test statistic's distribution. This is a particularity of the quantile inherent to its nature, and unrelated to the fact that it is a backtestable statistic. In fact, other backtestable statistics, like the expectile (Newey and Powell, 1987; Ziegel, 2014), do not show this advantage and require the predictive distribution to determine the p-value of the test.

When it comes to backtesting, the ideal procedure is based on a model-free test function, such that has no parametric assumptions and does not need the recording model information, be it for resampling purposes (determining the p-value) or for the computation of the test statistic itself.

3.4 Backtesting Expected Shorfall

In practice, what separates ES from VaR is not just the fact that the former is elicitable and backtestable and the latter is neither, but the fact that VaR allows for absolute, model-independent hypothesis testing using only the estimates and the realized P&L. For ES, this is, strictly speaking, impossible. To perform model validation on ES one can either approximate it from VaR (this is essentially the case for the FRTB's procedure) or relax the model information requirement. Kratz et al. (2016) follow the first option to propose approximating ES from VaR at N significance levels, while other authors propose different approaches that fall on the second option. The next sections will detail the most interesting and prominent existing methodologies.

3.4.1 Multinomial tests

This method comes as an extension of that by Emmer et al. (2015) where the authors suggest approximating ES_α as the average of VaR at four confidence levels, that is:

$$ES_{1-\beta}(L) \approx \frac{1}{4}\left[q_\beta(L) + q_{0.75\beta+0.25}(L) + q_{0.5\beta+0.5}(L) + q_{0.25\beta+0.75}(L)\right] \tag{14}$$

where $\beta = 1 - \alpha$[7] and $L = (X)^-$ represents the distribution of portfolio losses.

Kratz et al. (2016) extend this approximation to N discrete equally spaced levels, between β and 1, yielding:

$$\beta_j = \beta + \frac{(j-1)}{N}(1-\beta), \quad j = 1, 2, \ldots N \tag{15}$$

which means that $\beta_1 \equiv \beta = 1 - \alpha$. Additionally, $\beta_0 = 0$ and $\beta_{N+1} = 1$. In this study, I will focus on $\beta = 0.975$, as that is the ES confidence level used in the FRTB.

Using $I_{j,t} \equiv I_{1-\beta_j,t}$, as defined in Equation (3), would lead to N binomial VaR backtests. Defining V_t as

$$V_t = \sum_{j=1}^{N} I_{t,j}, \quad t = 1, \ldots, T \tag{16}$$

$V_t \in \{0, 1, \ldots N\}$ counts the levels that are violated by each observation. If the model is

[7] The authors use α to refer to the confidence and not the significance level. I will maintain the usual notation.

accurate, we should have, analogously to VaR:

$$Pr(V_t \leq j) = \beta_{j+1} \quad \text{and} \quad V_s, V_t \text{ independent for all } s \neq t \tag{17}$$

which is equivalent to saying that V_t follows a multinomial distribution with 1 trial and N+1 possible outcomes. To test this statement over a sample of T observations, one must define the random vector $\vec{O}$ such that:

$$O_j = \sum_{t=1}^{T} \mathbb{1}_{\{V_t=j\}} \quad j = 0, 1, \dots N. \tag{18}$$

O_j can be seen as the exception rate for each cell, that is the number of occasions in T days in which X_t is between $\text{VaR}_{1-\beta_j}$ and $\text{VaR}_{1-\beta_{j+1}}$. By definition, $\vec{O} = \{O_0, \dots, O_N\}$ should follow the same muntinomial distribution as V_t but with T trials. That is:

$$\vec{O} \sim \text{MN}(T, (\beta_1 - \beta_0, \dots, \beta_{N+1} - \beta_N)) \tag{19}$$

Under the assumption that $\vec{O} \sim \text{MN}(T, (\theta_1 - \theta_0, \dots, \theta_{N+1} - \theta_N))$, this procedure tests the hypotheses given by:

Hypothesis 2. Multinomial tests:
$H_0: \ \beta_j = \theta_j$ for all $j \in \{1, \dots, N\}$
$H_1: \ \beta_j \neq \theta_j$ for some $j \in \{1, \dots, N\}$

The null states that each cell's exception rate is equal to $(1-\beta)/N = \alpha/N$. To assess that, the authors propose 3 tests: The Pearson test, the Nass corrected test, and an LRT (likelihood ratio test). The first two are rather simple, while the latter is a more complex procedure that involves estimating the parameters via maximum likelihood. Since this is not a standard procedure in backtesting and adds a new out-of-model estimation step, this test will not be included in the analysis. I will now detail the Pearson and Nass test statistics.

(i) Pearson chi-squared test (Pearson, 1900):

$$S_N = \sum_{j=0}^{N} \frac{(O_j - T(\beta_{j+1} - \beta_j))^2}{T(\beta_{j+1} - \beta_j)} \underset{H0}{\overset{a}{\sim}} \chi^2_N \tag{20}$$

The Pearson test is a standard procedure to assess the inference of a random variable's parameters. In this case, it measures the mean squared difference between the observed and the expected cell violations.

(ii) Nass size-corrected test (Nass, 1959):

$$cS_N \underset{H0}{\overset{a}{\sim}} \chi^2_\nu \tag{21}$$

where $c = \frac{2\mathrm{E}[S_N]}{\mathrm{var}[S_N]}$, $\nu = c\mathrm{E}[S_N]$, $\mathrm{E}[S_N] = N$ and $\mathrm{var}[S_N] = 2N - \frac{N^2+4N+1}{T} + \frac{1}{T}\sum_{j=0}^{N} \frac{1}{\beta_{j+1}-\beta_j}$.

This correction comes because the Pearson test reveals excessive size when the cell probabilities are small, that is, as N increases. This does however come with a price, as the Nass test shows slightly less power. Kratz et al. (2016) state that the Nass test with $N = 8$ should be preferable to the Pearson test with $N = 4$.

3.4.2 Acerbi and Szekely's Z tests

Along with some research on elicitability, Acerbi and Szelely (2014) suggest 3 methods for unconditional coverage of ES. The first originates in the conditional expectation

$$\mathrm{E}\left[\frac{X_t}{ES_{\alpha,t}} + 1 \,\middle|\, X_t + VaR_{\alpha,t} < 0\right] = 0 \tag{22}$$

from which the authors define the test statistic Z_1 as:

$$Z_1 = \frac{\sum_{t=1}^{T}(r_t I_{\alpha,t}/ES_{\alpha,t})}{n_1} + 1 \tag{23}$$

where $n_1 = \sum_{t=1}^{T} I_{\alpha,t}$, as in Equation (4). This naturally implies that $n_1 > 0$ (if it isn't, there would be no need for exceptions backtesting). Hypothesis 3 describes the null and alternative of this test.

Hypothesis 3. Acerbi and Szekely's Z1:

$H_0: \; P_t^{[\alpha]} = F_t^{[\alpha]}$

$H_1: \; ES_{\alpha,t}^F \geq ES_{\alpha,t},$ for all t and $>$ for some t

$\quad\;\; VaR_{\alpha,t}^F = VaR_{\alpha,t},$ for all t

where $P_t^{[\alpha]}$ and $F_t^{[\alpha]}$ are the α-tail distributions of P_t and F_t (introduced in section 3.1).

Note that VaR_α is still correct under the alternative hypothesis. This method assumes that VaR_α is properly specified, testing only the coverage of ES_α, and therefore requires previous VaR_α backtesting. The second test backtests VaR_α and ES_α simultaneously. Z_2 comes from the unconditional expectation

$$ES_{\alpha,t} = -\mathrm{E}\left[\frac{X_t I_{\alpha,t}}{\alpha}\right] \tag{24}$$

and is defined as:

$$Z_2 = \sum_{t=1}^{T} \frac{r_t I_{\alpha,t}}{T\alpha ES_{\alpha,t}} + 1 \tag{25}$$

The hypotheses are then:

Hypothesis 4. Acerbi and Szekely's Z2:

$H_0: \; P_t^{[\alpha]} = F_t^{[\alpha]}$

$H_1: \; ES^F_{\alpha,t} \geq ES_{\alpha,t}, \text{ for all } t \text{ and } > \text{ for some } t$

$\quad\;\; VaR^F_{\alpha,t} \geq VaR_{\alpha,t}, \text{ for all } t$

While Z1 tests only the average magnitude of the exceptions, completely insensitive to their frequency, Z2 jointly measures both and therefore does not require previous VaR$_\alpha$ backtesting. From the hypotheses of the tests, it is clear that they are one-sided. However, their direction is the opposite of the VaR backtests. Both test statistics Z_1 and Z_2 have expected value 0 under the null, and illustrate underestimation of ES$_\alpha$ when they are negative.

Z1 and Z2 test whether, on average, the observed losses exceeding VaR$_{\alpha,t}$ are equal to the estimated ES$_{\alpha,t}$. However, both null hypotheses regard the whole tail distribution under the model. This is done because the test statistics' null distributions need to be simulated. To do so, one must assume that the predictive model is accurate and iteratively simulate the distribution (the process is better detailed ahead). If it weren't for that necessity, weaker null hypotheses regarding only the ES$_\alpha$ and VaR$_\alpha$ coverage would suffice. The same will apply to the Ridge test.

Regarding the third test (Z3), it is considerably more complicated and involves additional estimations, much like the LRT test of the multinomial approach. Moreover, the authors conclude that Z3 is not as adequate as Z1 and Z2 for model validation. Z3 will therefore be excluded from this study.

Finally, after concluding that ES is not strictly backtestable, Acerbi and Szekely (2017) state that it does however allow for a *Ridge Backtest*. This is possible when a statistic can be expressed as the attained minimum of the expected value of some scoring function (in the sense of the scoring functions mentioned in section 3.3) of an auxiliary statistic. To apply this logic to ES, aided by VaR, the authors use the results in Rockafellar and Uryasev (2002):

$$\begin{aligned} VaR_{\alpha,t} &= \arg\min_y E\left[y + \frac{1}{\alpha}(X_t + y)^-\right] \\ ES_{\alpha,t} &= \min_y E\left[y + \frac{1}{\alpha}(X_t + y)^-\right] \end{aligned} \tag{26}$$

From which it is possible to write a test statistic $\overline{Z}_{ES}$ that is no more than the difference between the average prediction and the realized ES ($\widehat{ES}_\alpha$), that is[8]:

$$\overline{Z}_{ES} = \frac{1}{T}\sum_{t=1}^{T} ES_{\alpha,t} - \widehat{ES}_\alpha \tag{27}$$

[8] $\overline{Z}_{ES}$ is the average of the daily backtest $Z_{ES,t} = ES_{\alpha,t} - VaR_{\alpha,t} - \frac{1}{\alpha}(r_t + VaR_{\alpha,t})^-$

where

$$\widehat{ES}_\alpha = \frac{1}{T}\sum_{t=1}^{T}\left[VaR_{\alpha,t} + \frac{1}{\alpha}(r_t + VaR_{\alpha,t})^-\right] \tag{28}$$

The authors conclude that the ridge backtest, henceforth referred to as ZR, should constitute an improvement on Z1 and Z2, in the sense that it shows small sensitivity to the VaR predictions. The Z1 strategy of previously backtesting VaR is imperfect since one can never be sure that the predictions are correct (the backtesting procedure can only reject large discrepancies). The Z2 test shows significant sensitivity to even relatively small VaR mispredictions, which can lead to both type I and type II errors, by rejecting correct ES predictions due to small VaR underestimation and not rejecting ES underestimations due to VaR overestimation, respectively. On the other hand, the sensitivity of the ZR test is not only very small for minor VaR mispredictions, it is prudentially biased, which means that large VaR mispredictions have a negative (prudential) impact on the test statistic.

Since the authors do not state the hypotheses, I will write them equally to Hypothesis 4:

Hypothesis 5. Acerbi and Szekely's ZR

$H_0:\ P_t^{[\alpha]} = F_t^{[\alpha]}$

$H_1:\ ES_{\alpha,t}^F \geq ES_{\alpha,t},$ for all t and $>$ for some t

$VaR_{\alpha,t}^F \geq VaR_{\alpha,t},$ for all t

Another relevant result in Acerbi and Szekely (2017) is the property of *Sharpness*, satisfied by ZR and not by the VaR tests. A test is said to be sharp if it is strictly decreasing in the real value of the statistic. What this means is that a sharp backtest will provide information on the discrepancy between the observed and the estimated values of a statistic.

The number of VaR exceptions says nothing about the real VaR, it only makes it possible to assess the statistical compatibility of that number with a correctly specified model. As for the ZR test, the test statistic will be more negative for predictions that underestimate the realized ES (Equation 28) by a larger difference, that is, for bigger discrepancies between the predicted and the observed values of ES. This property will prove very important in the context of this study, particularly for the determination of capital requirements.

Despite the usefulness of the discrepancy information, the p-values of the tests still need to be computed. As mentioned before, Acerbi and Szekely's test statistics lead to no statistical results (convergence to some theoretical distribution), so the distribution must be simulated under the null assumption that the predictive P&L distribution is correct. After computing the observed test statistic Z_{obs}, the process is the following:

1. Simulate a vector $\vec{X}^i$ of T independent realizations of $X_t \sim P_t, \quad t = 1, ..., T$;
2. Compute the simulated test statistic Z^i_{sim} from $\vec{X}^i$);
3. Repeat M times to obtain the empirical distribution of Z;
4. Estimate the p-value of the realized test statistic Z_{obs}:

$$p = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{Z^i_{sim} \geq Z_{obs}\}} \tag{29}$$

This way, as in the VaR test, the model is rejected with 95% confidence (amber zone) when $p \geq 0.95$ and with 99.99% (red zone) when $p \geq 0.9999$.

This process is not limited to parametric models. If one is dealing with a historical model, then P_t is the empirical distribution of previous returns. In this case, instead of simulating $\vec{X}^i$ from some assumed distribution, the random sample is drawn directly from the past returns (age or volatility standardized if implied by the model), from which Z^i_{sim} is computed.

In this study, T is fixed at 250. M must be sufficiently large for significant results. Usual settings are 10 000, 20 000, or 50 000, the same as for Monte Carlo simulations.

3.4.3 Cumulative violations unconditional coverage test

As I previously mentioned, it is possible to backtest ES using the cumulative violation process (Equation 9). This approach was first presented by Constanzino and Curran (2015), by defining a coverage test for the class of spectral measures of risk (Acerbi, 2002), which includes ES. Du and Escanciano (2016) suggest an almost identical methodology (specific for ES). I will use the inputs of both authors but simply refer to this as the C&C test (Constanzino and Curran).

Recovering the predictive cumulative distribution P_t, Equation (9) is equivalent to:

$$H_{\alpha,t} = \frac{1}{\alpha} \int_0^{\alpha} \mathbb{1}_{\{r_t < -VaR(u,t)\}} du = \left(1 - \frac{P_t(r_t)}{\alpha}\right) I_{\alpha,t} \tag{30}$$

Much like VaR violations, cumulative violations are i.i.d. under correctly specified models. Therefore, following the Lindeberg–Lévy Central Limit Theorem:

$$\sqrt{N}(\overline{H}_{\alpha,t} - \mu_H) \underset{H0}{\overset{a}{\sim}} N(0, \sigma^2_H) \tag{31}$$

where $\overline{H}_{\alpha,t}$ is the average $H_{\alpha,t}$ over T trading days and μ_H and σ^2_H are the expected value and variance of $H_{\alpha,t}$ under the null. Both authors then show that $\mu_H = \alpha/2$ and $\sigma^2_H = \alpha(4 - 3\alpha)/12$, and consequently $H_{\alpha,t}$ admits a standard Z-test (not to be confused with Acerbi and Szekely's), with standard normal distributed test function Z_H:

$$Z_H = \frac{\sqrt{N}(\overline{H}_{\alpha,t} - \mu_H)}{\sigma_H} = \frac{\sqrt{3N} - (2\overline{H}_{\alpha,t} - \alpha)}{\sqrt{\alpha(4-3\alpha)}} \tag{32}$$

The hypotheses are then:

Hypothesis 5. Cumulative violations (C&C) test:
$H_0: \ E[H_{\alpha,t}] = \alpha/2$
$H_1: \ E[H_{\alpha,t}] > \alpha/2$

Along with a conditional coverage test, Du and Escanciano (2016) also derive a corrected test statistic to cover estimation errors, which will not be presented here. Not only is the correction reliant on assumptions that are not met by non-parametric models, but it is also insignificant when the estimation period is considerably larger than the backtesting period.

3.4.4 Moldenhauer and Pitera's secured position test

Moldenhauer and Pitera (2018) suggest a revamped approach to backtesting. Instead of looking at the risk measure and P&L separately, analyzing discrepancies from one to the other, the authors bind them into what they call the *secured position*, which consists of no more than adding the risk estimate to the returns. In the case of ES, the day t realization of the secured position s will be:

$$s_t = r_t + ES_{\alpha,t} \tag{33}$$

As the name points out, this is simply the daily P&L covered by the capital reserves. Logically, the secured position should be positive if the requirements are prudential enough. Before detailing the test procedure, Moldenhauer and Pitera suggest correcting s_t for volatility changes by standardizing it by the risk estimates, yielding:

$$\widetilde{s}_t = \frac{r_t + ES_{\alpha,t}}{ES_{\alpha,t}} = \frac{r_t}{ES_{\alpha,t}} + 1 \tag{34}$$

Using either s_t or $\widetilde{s}_t$ the proposed test statistic will consist of "the biggest number of worst realizations of the secured position that add up to a negative total" (Moldenhauer and Pitera, 2018, p.6) that is[9]:

$$G = \sum_{t=1}^{T} \mathbb{1}_{\{\widetilde{s}_{(1)} + ... + \widetilde{s}_{(t)} < 0\}} \tag{35}$$

where $\widetilde{s}_{(k)}$ is the k-th order statistic of the secured position $\widetilde{s} = \{\widetilde{s}_1, ..., \widetilde{s}_T\}$. Much like

[9] The authors also divide G by T for a sample general statistic. I will stick to nominal values

Acerbi and Szekely's tests, the secured position test (henceforth named M&P test) lacks statistical results regarding the test statistic distribution. Conveniently, the authors define this test specifically for a traffic-light approach intended to be analogous to the Basel procedure.

To establish the amber and red zone thresholds, it is first useful to note that under a properly specified $\mathrm{ES}_{2.5\%}$ model, the expected value of G coincides with the expected amount of $\mathrm{VaR}_{2.5\%}$ exceptions, that is, 6.25 (6 considering the integer part). From here, the authors set the thresholds, visible in Table 3.3. in a way that its interpretation is aligned with that of the $\mathrm{VaR}_{1\%}$ test.

Table 3.3. Basel and M&P tests traffic-light thresholds

Backtesting Zone	$\mathrm{VaR}_{1\%}$ exceptions	G (M&P test)
Green	≤ 4	≤ 11
Amber	5 - 9	12 - 24
Red	≥ 10	≥ 25

The reason for the thresholds at 12 and 25 is to make the implied coverage decreases equal to those of the VaR test thresholds. That is, 5 and 10 exceptions are associated with 98% and 96% coverage[10] respectively, which is an implicit twofold and fourfold increase in α, from 0.01 to 0.02 and 0.04. Applying the same multiplication to $\mathrm{ES}_{2.5\%}$ leads to 95% and 90% confidence levels, associated with an expected (integer) G of 12 and 25, respectively. Additionally, the values of $\mathrm{ES}_{5\%}$ and $\mathrm{ES}_{10\%}$ are approximately equal to $\mathrm{VaR}_{2\%}$ and $\mathrm{VaR}_{4\%}$ under Gaussian returns, just like $\mathrm{ES}_{2.5\%}$ and $\mathrm{VaR}_{1\%}$.

The authors don't present a null hypothesis for this test, which is reasonable given the ad-hoc nature and lack of statistical conclusions of the procedure. The relevance of this test relies mostly on the alignment with the FRTB's traffic-light backtest, including the possibility of analogously determining capital ad-ons. Nevertheless, the analysis ahead will also focus on the statistical implications of it.

[10] 5 exceptions imply a coverage of 1 - 5/250 = 0.98, 10 exceptions imply 1 - 10/250 = 0.96.

4 Empirical analysis

4.1 Sample description

The hypothetical portfolio used for this study is meant to be simple yet representative of banks' assets. This way, only long spot positions on equity were included. To focus on systematic, undiversifiable risk, disregarding any idiosyncrasies from the analysis, without the need to perform risk factor mappings, the positions will be exclusively on stock indexes, featuring a few of the most relevant international indexes, allowing for a widespread analysis. This way, 4 indexes were selected in an attempt to significantly cover international markets. These are the S&P500, STOXX Europe 600, Nikkei 225 and the Hang Seng Index.

The exposure to each index was determined in a way that it is proportional to its market cap (relative to the total market cap of the portfolio in April 2023), as it is set out in Table 4.1. No portfolio value was considered, as the statistics (VaR and ES) will be computed in % USD. Since three out of four indexes are denominated in different currencies, the exposure to exchange rate risk must also be included, particularly to the EUR-USD, JPY-USD, and HKD-USD exchange rates.

Table 4.1. Portfolio exposures

Risk Factor	Exposure
S&P500	0.64
Stoxx 600	0.24
Nikkei 225	0.07
Hang Seng	0.05
EUR - USD	0.24
JPY - USD	0.07
HKD - USD	0.05

The daily closing prices of the four indexes and the daily exchange rates were downloaded from Bloomberg so that the VaR and ES at the portfolio level can be computed, as set out in the next section.

4.2 Methodology

After the portfolio P&L is obtained, VaR and ES estimates must be computed. I will use two different models in this study, one parametric and one historical. The results of both will be compared to assess the applicability of the backtesting methods to different types of models, as well as the stability of the conclusions. The parametric model used is the standard RiskMetrics VaR, a simple yet well-implemented approach, and the HS model will consist of adjusting the sample of realized returns to the most recent volatility conditions as in Hull and White (1998) using GJR-GARCH estimates (Glosten, Jagannathan and Runkle, 1993).

4.2.1 RiskMetrics parametric model

The RiskMetrics (RM) model consists of a parametric model that assumes that returns follow a normal distribution with EWMA volatilities with a smoothing factor (λ) of 0.94.

The EWMA variance estimates are given by:

$$\hat{\sigma}_t^2 = (1-\lambda) r_{t-1}^2 + \lambda \hat{\sigma}_{t-1}^2 \quad \text{with } 0 < \lambda < 1 \tag{36}$$

EWMA models are a simple yet effective way to increase the sensitivity of variance (and volatility) estimates to the most recent market conditions. By weighing the previous day's return with $1-\lambda$, the preponderance given to one particular observation decreases exponentially as the days pass. The parameter λ determines how fast this decrease happens, hence the name smoothing factor. It is straightforward to tell that the lower λ is, the higher the weight of the most recent conditions.

This way, the daily VaR_α is computed as:

$$VaR_{\alpha,t} = \hat{\sigma}_t \Phi^{-1}(1-\alpha) \tag{37}$$

and ES_α as:

$$ES_{\alpha,t} = \frac{\hat{\sigma}_t}{\alpha} \phi(\Phi^{-1}(\alpha)) \tag{38}$$

From its development in 1994 and upgrade in 1996, the RM VaR approach grew from a pioneer to an industry standard in risk management. However, as I mentioned before, the normality assumption has been proven to be inaccurate to model most P&L distributions. Regardless of that fact, I will still use it in this study, not only due to its significance, but as a standard representation of a parametric approach. Additionally, since the main topic of this study is backtesting, it is also useful generality to have an imperfect model. This way, the RM model will constitute a starting point, and the HS model will serve as an improvement in a way that it should provide better backtesting results and be more representative of the models used by banks that qualify for the IMA.

4.2.2 GJR-GARCH Historical Simulation

As I have previously mentioned, the HS method is currently the most used method for VaR and ES estimation[11]. To address the sample dependence issue, each $\text{VaR}_{\alpha,t}$ and

[11] See Pérignon and Smith (2009) and Mehta et al. (2012) for studies on this topic.

$\text{ES}_{\alpha,t}$ will be computed from a sample of volatility-adjusted returns from the previous 1 000 days. The volatility adjustment works in the following way:

1. Each daily r_t is standardized by the most recent volatility estimate at the end of t, which is $\hat{\sigma}_{t+1}$ (as soon as r_t is known, $\hat{\sigma}_{t+1}$ can be estimated).

2. To compute $\text{VaR}_{\alpha,T+1}(t < T)$ at the end of T, the series of previous returns is scaled by $\hat{\sigma}_{T+1}$, yielding a series of volatility adjusted returns $\hat{r}_t$, from which the quantile is computed[12]:

$$\hat{r}_t = \frac{\hat{\sigma}_{T+1}}{\hat{\sigma}_{t+1}} r_t \tag{39}$$

This approach allows us to compute VaR and ES, directly from a sample that matches the current market conditions, still without making any parametric assumptions. Essentially, the assumption is that the P&L at time $T+1$ will follow the empirical distribution of the previous 1000 volatility-standardized returns (including T), scaled by the volatility estimate for $T+1$.

In addition to the model itself, the volatility estimation is also altered in this model. For the HS approach, they are assumed to follow a GJR-GARCH process.

A GARCH model is a stochastic process used to describe time series where the residuals (in this study these will be equivalent to the demeaned returns) are heteroskedastic (non-constant variance). This way, a model for the conditional variance, dependent on past residuals and variance estimates, is set up.

Firstly, some stochastic model must be assumed for the series of returns. Typical choices are AR(1,1) or ARMA(1,1) processes. However, the choice of model tends to have little impact on the volatility estimates, being more important for P&L forecasting purposes. For this reason, and for the sake of simplicity, I will simply demean the returns, that is:

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2) \tag{40}$$

A standard GARCH(1,1) is defined by the following equation:

$$\text{GARCH(1,1)}: \quad \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{41}$$

with $\omega, \alpha, \beta > 0$ and $\alpha + \beta < 1$, where α (unrelated to the significance and confidence levels α and β are the past residuals and past variance parameters, respectively. This model can be generalized for (p, q) lags, in which case there would naturally be p residuals parameters and q variance parameters.

There are many possible GARCH specifications. In this study, I will use one slightly

[12] The quantile can be derived from the standardized returns and then multiplied by $\hat{\sigma}_{T+1}$ (equivariance property).

more elaborated than Equation (41), entitled GJR-GARCH. A GJR-GARCH(1,1) is defined by the equation:

$$\text{GJR-GARCH(1,1):} \quad \sigma_t^2 = \omega + (\alpha + \gamma I_{t-1})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{42}$$

with $\omega, \alpha, \beta > 0,\ \alpha + \gamma > 0$ and $\alpha + \beta + \gamma/2 < 1$.

The difference between Equations (41) and (42) is simply the parameter γ, which is introduced to assign different impacts to negative shocks as compared to positive ones. This way, the coefficient associated with a negative return is now $(\alpha + \gamma)$ while the coefficient of a positive return remains α. Typically negative shocks have a stronger influence in increasing volatility, a phenomenon called asymmetry (Caporin and McAleer, 2012)[13].

Having established the model, the parameters are then estimated via maximum likelihood, and from the equation along with the realized returns, it is now possible to obtain a series of volatility estimates. As soon as r_t is known the residual ε_t is drawn as:

$$\varepsilon_t = r_t - \hat{\mu} \tag{43}$$

from where $\hat{\sigma}_{t+1}$ is obtained:

$$\hat{\sigma}_{t+1}^2 = \hat{\omega} + (\hat{\alpha} + \hat{\gamma} I_{t-1})\varepsilon_t^2 + \hat{\beta}\hat{\sigma}_{t-1}^2 \tag{44}$$

From the daily P&L of the portfolio described earlier, I repeated the estimation process once every 65 business days (roughly 3 months), using a sample of the previous 1000 observations to perform the likelihood optimization. This is in line with the FRTB's request to update the model data sets every 3 months.

The series of volatility estimates is now complete and I can perform the adjustments as detailed in Equation (39) and compute the HS VaR and ES starting from day 1001 (early 2007).

[13] A study by Stavroyiannis (2018) explains the parameter constraints in the presence of asymmetry and leverage.

5 Results and discussion

5.1 Rolling sample

Based on the $\mathrm{VaR}_{1\%}$ and $\mathrm{ES}_{2.5\%}$ estimates detailed above, each of the following backtesting methods was applied daily to a rolling sample of the previous 250 days, complying with the FRTB procedure:

(i) $\mathrm{VaR}_{1\%}$: Binomial test as recommended in the FRTB (henceforth entitled Basel test) and Kupiec's unconditional coverage test.

(ii) $\mathrm{ES}_{2.5\%}$: Pearson and Nass tests, each with $N = 2, 4, 8$, Cumulative Violations test (C&C) and Secured Position test (M&P).

Due to the computational burden of simulating the test statistics' distributions daily, the Z tests of Acerbi and Szekely were excluded from the rolling sample analysis. Acerbi and Szekely (2014) argue that the Z2 test leads to stable 95% and 99.99% thresholds, around the values of -0.7 and -1.8, respectively, which would make the simulation process unnecessary. However, these thresholds will not be considered for this study, and only simulated p-values will be used further ahead. A comparative analysis of these values against simulated thresholds under the RM and HS models was conducted, concluding that convergence to the -1.8 (99.99%) threshold is uncertain, especially in the HS model[14].

Each day, the RM and HS models were placed in the traffic-light zones, according to the same criteria as in the FRTB backtesting procedure extended to each of our alternative methods. Table 5.1. summarizes the results.

		Parametric model (RM)			Historical model (HS)		
	Test	G (%)	A (%)	R (%)	G (%)	A (%)	R (%)
$\mathrm{VaR}_{1\%}$	Basel	21.41	66.54	12.05	88.13	11.87	0.00
	Kupiec	47.43	46.57	6.00	93.14	6.86	0.00
$\mathrm{ES}_{2.5\%}$	Pearson N=2	39.36	49.48	11.16	96.23	3.77	0.00
	Pearson N=4	42.60	44.75	12.66	94.99	5.01	0.00
	Pearson N=8	33.54	53.66	12.81	86.41	12.98	0.61
	Nass N=2	45.28	45.05	9.67	97.54	2.46	0.00
	Nass N=4	42.62	46.22	11.16	95.87	4.13	0.00
	Nass N=8	41.66	51.46	6.88	92.41	7.59	0.00
	C&C	24.50	56.09	19.41	94.46	5.54	0.00
	M&P	16.38	78.54	5.09	93.57	6.43	0.00

Table 5.1. Rolling sample backtesting results (2007 - 2022).
The values displayed are the percentage of placements in each zone.

The first thing Table 5.1. tells us is that, as expected, the HS model performs much better than the RM, in all tests. In fact, not only would these results imply a higher capital burden for the RM model, but the model would at one point (between March and

[14] The results of this analysis are presented in Annex C.

April 2020) not even qualify for the IMA, due to the occurrence of too many exceptions[15]. On the other hand, the HS model proves to be much more adequate, never falling in the Basel red zone.

Compared to the Basel binomial test, all but one of the multinomial ES tests lead to considerably fewer 95% level rejections, that is, more green zone allocations. Despite being significant in both models, this increase is much larger in the RM model, peaking at the Nass test with $N = 2$, more than doubling the Basel test's result (the percentage of green zone allocations from the latter was as low as 21.41%). Still regarding the RM model, in particular the Nass test, there is also a considerable decrease in red zones for all N. This means that a change to the Nass test from the Basel test would be less prudential by decreasing the capital add-ons of poorly performing models. As for the Pearson test, the percentage of red zone allocations does not change dramatically. The changes in the HS model are much lighter but still relevant, also peaking at the Nass test with $N = 2$.

Looking closely at the multinomial ES backtesting results, it is possible to tell that increasing N in both test statistics of the multinomial approach generally leads to an increase in 95% rejections. In the HS model, the percentage of green zones from the Pearson test decreases by nearly 10 p.p. from $N = 2$ to $N = 8$, while the Nass test decrease is around 5 p.p. This difference in sensitivity is in line with Pearson's decrease in effectiveness with increasing N, associated with higher test size, as mentioned in section 3.4. This means that the Pearson test's rejections with N=8 are less trustworthy and more likely to be incorrect than in the Nass test. In fact, the Pearson test with $N = 8$ is the only one to have 99.99% rejections of the HS model. In the RM model, the zone shifts are less stable, as the green zones even increase slightly in the Pearson test from $N = 2$ to $N = 4$, decreasing back about 9 p.p. with $N = 8$, reflecting again a big rise in (possibly incorrect) rejections with the largest N. Looking at the red zone in the Nass test, these reach a maximum with $N = 4$ and minimum with $N = 8$, exposing the correction properties of this test with higher N.

It is interesting to note that the multinomial test results, while differing a lot from the Basel test, are very close to Kupiec's, especially in the case of the Nass test. These procedures have in common the fact that they are based on asymptotic convergences to the Chi-Square distribution that lead to less restrictive critical values. In particular, Kupiec's test rejects the null with 95% confidence only in the presence of 7 or more $\mathrm{VaR}_{1\%}$ exceptions, and with 99.99% only from 11 exceptions, as opposed to Basel's 5 and 10. This essentially means that the Basel Committee prioritized simplicity and prudentiality when defining the backtesting method. This exercise grows more elaborate with the multinomial tests, as the thresholds are dependent on N[16].

[15] Models are disqualified if more than 12 exceptions are observed.

[16] Kratz et al. (2016) display a figure illustrating the amber and red zone thresholds of the Nass test with $N = 2$, dependent on the combinations of the two cells's exceptions.

On the other hand, the C&C, and M&P tests present outcomes that are more aligned with the Basel VaR test, particularly in the RM model. Given their nature, this was an expected conclusion, however, they show some remarkable differences between each other. In the MP model, the M&P test outcomes are more centered in the amber zone. It has the lowest amount of both green and red zones out of all the tests displayed, which can be worrying. While it shows prudentiality by rejecting more models on the 95% level, it is the opposite when it comes to the 99.99% threshold, which has dubious implications in the capital requirements. For a procedure intended to mimic the output of the Basel test, 5.09% is a surprisingly low amount of red zone allocations, a decrease of nearly 7 p.p.. When it comes to the HS model, it is less prudent in both the 95% and 99.99% thresholds. The C&C leads to more dispersed results in the MP model, with more green and red zones than both the Basel and the M&P tests. In fact, this is the test with the most red zone allocations overall, with over 19%. In the HS model, it does not drift much from the other tests.

The conclusions from this analysis must take into account the fact that they come from a rolling sample, meaning each day the sample changes only one observation, maintaining the other 249. This naturally leads to the repetition of exceptions. In the real world, a red or even amber zone allocation would lead to changes in the model that should prevent it from being placed in the same zone in the future. To better understand the time dynamics of the backtesting results, they should be drawn from more "discrete" 250-day samples.

5.2 Discrete samples

For this analysis, 31 samples of the previous 250 trading days were drawn every 130 days (roughly a half year). Tables 5.2. and 5.3. detail the results using the RM and HS models, respectively. This time, Acerbi and Szekely's Z tests were included in the analysis, with a simulation sample (test statistic distribution under the null) of 20 000 observations. Regarding the Z1 test, a $\text{VaR}_{2.5\%}$ backtest, identical to the Basel test[17], was performed beforehand. In both tables, in the Z1 column, A* means that the $\text{VaR}_{2.5\%}$ was rejected with 95% confidence but $\text{ES}_{2.5\%}$ was not.

For the Multinomial, C&C, and M&P tests, the conclusions from the discrete sample analysis are mostly in line with those of the rolling sample. That being said, with fewer backtesting samples, the multinomial tests move closer to the Basel test. As for Acerbi and Szekely's tests, Z1 stands out with only two green zones and 29 amber zones. This is of course not a good trait, for the same reasons that were pointed out before regarding the M&P test. However, on five occasions, the Z1 test did not reject the hypothesis with 95% confidence, with the model being placed in the Amber zone due to VaR 97.5% backtesting. Nevertheless, the absence of red zones perfectly illustrates one of the weaknesses of the

[17] The 95% and 99.99% thresholds are, in this case, 11 and 17, respectively.

Testing period	Basel	Kupiec	Pearson N=2	Nass N=8	C&C	M&P	ZR	Z1	Z2
Jan/07 - Jan/08	A	A	R	A	R	A	A	A	A
Jul/07 - Jul/08	R	A	A	A	R	A	A	A*	A
Jan/08 - Jan/09	A	G	G	A	A	A	G	A	A
Jul/08 - Jul/09	G	G	A	A	A	A	G	A*	A
Jan/09 - Jan/10	G	G	G	G	G	G	G	G	G
Jul/09 - Jul/10	A	A	A	A	A	A	A	A	A
Jan/10 - Jan/11	A	A	A	A	A	A	A	A	A
Jul/10 - Jul/11	A	A	A	G	A	A	A	A*	A
Jan/11 - Jan/12	R	A	R	A	R	A	A	A	A
Jul/11 - Jul/12	A	G	A	A	A	A	A	A*	A
Jan/12 - Jan/13	G	G	G	G	G	G	G	G	G
Jul/12 - Jul/13	A	G	G	G	G	A	A	A	G
Jan/13 - Jan/14	A	G	G	G	G	A	A	A	G
Jul/13 - Jul/14	G	G	G	G	G	G	G	A	G
Jan/14 - Jan/15	A	A	A	A	R	A	A	A	A
Jul/14 - Jul/15	A	A	A	A	R	A	A	A	A
Jan/15 - Jan/16	A	A	A	A	A	A	A	A	A
Jul/15 - Jul/16	A	A	G	G	A	A	A	A	A
Jan/16 - Jan/17	G	G	G	G	G	A	A	A	G
Jul/16 - Jul/17	G	G	G	G	G	G	G	A	G
Jan/17 - Jan/18	A	G	G	G	G	G	G	A	G
Jul/17 - Jul/18	A	A	A	A	A	A	R	A	A
Jan/18 - Jan/19	R	R	R	A	R	R	R	A	R
Jul/18 - Jul/19	A	A	A	A	A	A	A	A	A
Jan/19 - Jan/20	A	A	A	A	A	A	A	A	G
Jul/19 - Jul/20	R	R	R	R	R	R	R	A	R
Jan/20 - Jan/21	R	A	R	R	R	R	R	A	R
Jul/20 - Jul/21	A	G	A	G	A	A	A	A	A
Jan/21 - Jan/22	A	A	A	A	A	A	A	A	A
Jul/21 - Jun/22	A	G	A	A	A	A	A	A	A
Jan/22 - Dec/22	G	G	A	A	A	A	G	A*	A
Total (%) G	22.58	45.16	32.26	35.48	25.81	16.13	25.81	6.45	29.03
A	61.29	48.39	51.61	58.06	48.39	74.19	61.29	93.55	61.29
R	16.13	6.45	16.13	6.45	25.81	9.68	12.90	0.00	9.68

Table 5.2. Discrete samples traffic-light backtesting results - RM model (2007 - 2022). The bottom rows display the percentage of placements in each zone over the 31 samples.

Z-tests, which is the absence of statistical results and the necessity to simulate the p-value. Not only does this make the results dependent on the number of simulations (M), but it also creates uncertainty when the simulated p-values are close to the thresholds since nothing guarantees a strong convergence of the test statistic distribution. To better illustrate this point, the procedure was repeated with $M = 50\ 000$, which changed the amber zone of the Jan/13 - Jan/14 sample to green (no further changes were recorded among the three tests). Even with the same M, infinitesimal changes may occur in the p-value from one run to another that can impact the tests' conclusion.

Regarding the ZR and Z2 tests, these present similar results, that are also very close to the Basel test. It is safe to say that the ZR is the test that is closer to the Basel test when it comes to the traffic-light results, with the same amount of amber zones, one more green, and consequently one fewer red. These are also similarly distributed in time, although not identically.

To assess different tests' behavior at periods of more stress, it is useful to analyze the ES tests' results in the periods that led to Basel red zone allocations. The Nass and Z1 tests seem the least suitable in these periods. The former fails to recognize red zones in all but one of the stress periods, that being the year 2020 (marked by the COVID-19 pandemic), while the latter has no red zones at all (although inconclusive in the 2015 and 2020 periods, as mentioned above). The M&P, ZR, and Z2 tests revealed red zones in 2018 and 2020 but failed to do so in 2008 and 2011. The remaining tests, Pearson and C&C, show themselves as the most prudential in periods of stress, assigning red zones in about the same periods as the Basel tests. Additionally, the C&C test also led to red zones in 2015, a period of remarkable losses that yielded "just" amber zones in all of the other tests, reinforcing detail that this is the test with the most red zone allocations. On the other hand, it is also relevant to look at amber zones and mention periods such as 2013, in which the Basel test leads to amber zones, accompanied by the M&P and ZR tests, while Pearson and C&C yield green zones. Finally, a mention to the year 2022, in which, surprisingly given the economic context, the VaR tests revealed green zones, while all but the ZR test resulted in amber zones. The same happened in the Jul/08 - Jul/09 period.

Conclusions from the HS model are of course much more limited, due to the preponderance of green and total absence of red in Table 5.3. Still, it is relevant to note that the ZR test is the only one to match Basel's two amber zones from 2020. It is interesting to note none of the remaining tests reject the model in this period, while in the RM, model all but the Z1 test resulted in red zones. This shows that, once again, ZR presents the closest results to the Basel test. It differs only in the January 2009 sample, when it leads to a green zone, as opposed to Basel's amber. The Nass test is the only ES test to present an amber zone in this period. On the other hand, in the first backtesting sample it surprisingly yields an amber zone, contrary to the remaining tests. In the January 2016 sample, it is the only test with an amber zone.

Testing period	Basel	Kupiec	Pearson N=2	Nass N=8	C&C	M&P	ZR	Z1	Z2
Jan/07 - Jan/08	A	A	A	G	A	A	A	A*	A
Jul 07 - Jul/08	A	A	A	A	A	A	A	A*	A
Jan/08 - Jan/09	A	G	G	A	G	G	G	G	G
Jul 08 - Jul/09	G	G	G	G	G	G	G	G	G
Jan/09 - Jan/10	G	G	G	G	G	G	G	G	G
Jul 09 - Jul/10	G	G	G	G	G	G	G	G	G
Jan/10 - Jan/11	G	G	G	G	G	G	G	G	G
Jul 10 - Jul/11	G	G	G	G	G	G	G	G	G
Jan/11 - Jan/12	G	G	G	G	G	G	G	G	G
Jul 11 - Jul/12	G	G	G	G	G	G	G	G	G
Jan/12 - Jan/13	G	G	G	G	G	G	G	G	G
Jul/12 - Jul/13	G	G	G	G	G	G	G	G	G
Jan/13 - Jan/14	G	G	G	G	G	G	G	G	G
Jul/13 - Jul/14	G	G	G	G	G	G	G	A	G
Jan/14 - Jan/15	G	G	G	G	G	G	G	G	G
Jul/14 - Jul/15	G	G	G	G	G	G	G	G	G
Jan/15 - Jan/16	G	G	G	A	G	G	G	G	G
Jul/15 - Jul/16	G	G	G	G	G	G	G	G	G
Jan/16 - Jan/17	G	G	G	G	G	G	G	G	G
Jul/16 - Jul/17	G	G	G	G	G	G	G	G	G
Jan/17 - Jan/18	G	G	G	G	G	G	G	G	G
Jul/17 - Jul/18	G	G	G	G	G	G	G	G	G
Jan/18 - Jan/19	G	G	G	G	G	G	G	G	G
Jul/18 - Jul/19	G	G	G	G	G	G	G	G	G
Jan/19 - Jan/20	G	G	G	G	G	G	G	G	G
Jul/19 - Jul/20	A	G	G	G	G	G	A	G	G
Jan/20 - Jan/21	A	G	G	G	G	G	A	G	G
Jul/20 - Jul/21	G	G	G	G	G	G	G	G	G
Jan/21 - Jan/22	G	G	G	G	G	G	G	G	G
Jul/22 - Jun/22	G	G	G	G	G	G	G	G	G
Jan/22 - Dec/22	G	G	G	G	G	G	G	G	G
Total (%) G	83.87	93.55	93.55	90.32	93.55	87.10	93.55	90.32	93.55
A	16.13	6.45	6.45	9.68	6.45	6.45	12.90	9.68	6.45

Table 5.3. Discrete samples traffic-light backtesting results - HS model (2007 - 2022). The bottom rows display the percentage of placements in each zone over the 31 samples.

5.3 Capital multipliers

Other than analyzing model rejections, it is interesting to assess the impact that a change to ES backtesting could have on the capital add-ons. To do so, it would be useful to extend the rationale that is currently used to determine the VaR backtesting multipliers (Table 3.1.) to the ES backtesting results.

The idea behind the multipliers in Table 3.1. is that "the increase in the multiplication factor should be sufficient to return the model to a 99th percentile standard" (BCBS, 2019, p. 128). What this means is that based on the coverage levels implied by the amount of

VaR exceptions, the multiplier is set so that the final coverage is 99%. This requires some statistical assumptions that are not detailed in the document[18]. The multiplier (m_{Basel}) is then determined from the ratio (under said assumptions) of the 99% VaR to the VaR with a significance level equal to the observed exception rate (π_{obs}), that is:

$$m_{Basel} = 1.5 \, \frac{P^{-1}(0.01)}{P^{-1}(\pi_{obs})} \tag{45}$$

where P is the cdf of the P&L distribution assumed in the FRTB and π_{obs} bears the same meaning as in Equation (4). As Table 3.1. shows, the factor of 1.5 is always applied, regardless of backtesting results.

In the case of the M&P test, one can apply the same multipliers to the biggest number of worst-case observations with negative sum. Extending the relation presented in Table 3.3. to the remaining multiplier levels results in Table 5.4., from which it is possible to run capital multipliers for the M&P test.

Number of Exceptions	Observed Rate (π_{obs})	G (M&P test)	Backtesting Multiplier
5	2%	12, 13, 14	1.70
6	2.4%	15, 16	1.76
7	2.8%	17, 18, 19	1.83
8	3.2%	20, 21	1.88
9	3.6%	22, 23, 24	1.92
$\geq$ 10	4%	$\geq$ 25	2.00

Table 5.4. Basel and M&P backtesting multipliers. When the exception rate (π_{obs}) leads to a decimal G, its integer value is considered for that multiplier, for prudentiality.

Figures 5.1. and 5.2. detail the percentual difference from the Basel to the M&P capital multipliers across the rolling sample periods, in the RM and HS models, respectively. Additionally, to better understand where the differences come from, Tables 5.5. and 5.6. show the allocations of both the Basel and the M&P tests simultaneously.

Although the average difference in the capital add-on is around zero in the RM model (0.1573 % to be precise), figure 5.1. shows that there are large shifts in given moments, both positive and negative. These can reach up to 20% increases or decreases in the capital reserves, which is of course extremely relevant. This means that substituting the Basel with the M&P test could have a very significant but unpredictable impact on the capital levels.

[18] For all but 9 and 10 exceptions, the multipliers seem to be in line with a standard normal distribution assumption

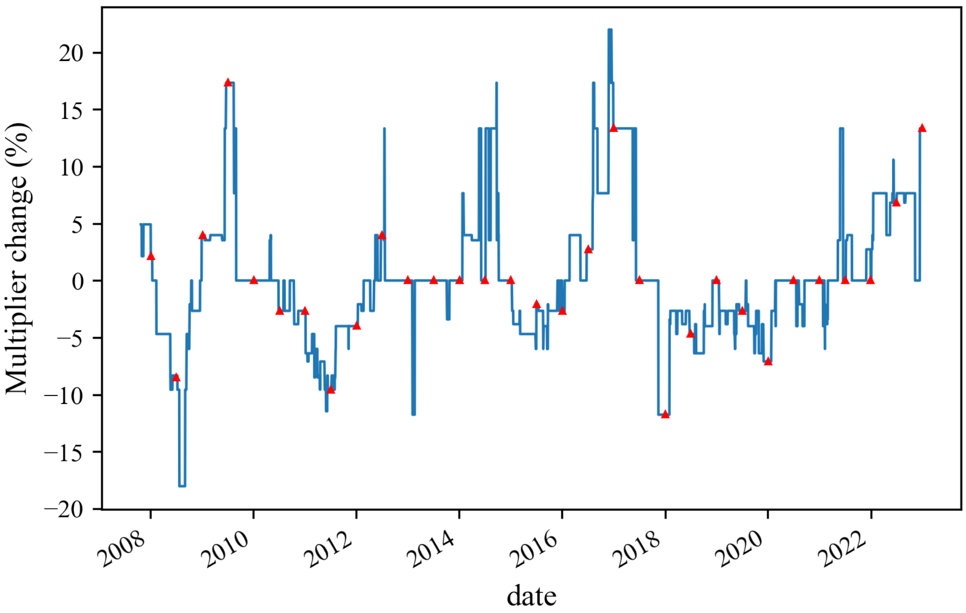

Figure 5.1. % change in capital requirements from the Basel to the M&P test - RM model. The red marks refer to the discrete sample periods used before.

		M&P		
		Green	Amber	Red
Basel	Green	550	296	
	Amber	97	2532	0
	Red		275	201

Table 5.5. Rolling sample M&P backtesting results - RM model (2007 - 2022). The diagonal cells are the occasions when both tests result in the same zone.

In the HS model, the differences are naturally less common, given the amount of green zones. However, when capital add-ons are appliable (amber zones), there are again remarkable discrepancies in their value, but this time always negative, surpassing -18% in 2008. This occurs when the Basel test yields amber zones and the M&P test returns green or less significant amber zones (there is no occasion where the Basel test is green and the M&P test amber, as Table 5.6. shows), reinforcing the idea that the M&P test, as it is presented, is less prudential than the Basel test in the better performing HS model. Assuming that this model is unchanged over time, the presented change would lead to a decrease in the total capital add-on (sum of all periods, including green zones) of 1.49%. Excluding the green zones, the total decrease would be 10.68% and on average 10.73% a day.

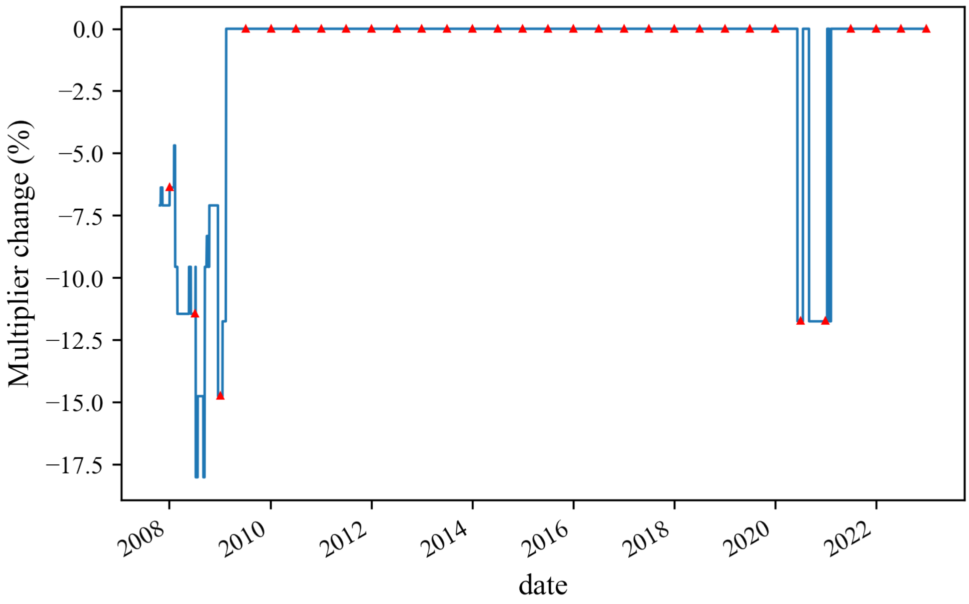

Figure 5.2. % change in capital requirements from the Basel to the M&P test - HS mode. The red marks refer to the discrete sample periods used before.

		M&P		
		Green	Amber	Red
Basel	Green	3482	0	
	Amber	215	254	0
	Red		0	0

Table 5.6. Rolling sample M&P backtesting results - HS model (2007 - 2022). The diagonal cells are the occasions when both tests result in the same zone.

The M&P results seem somewhat unsatisfactory, given the authors' conclusion that it would allow for a smooth transition from the Basel backtesting procedure. To understand why, it is necessary to take a closer look at the thresholds used. The choice of thresholds at 12 and 25 shares the ad-hoc nature of the Basel test, however, the latter does add some statistical reasoning with the use of the binomial distribution's 95% and 99.99% thresholds. If the M&P statistical framework were aligned with the Basel one, we would not be seeing such discrepancies in both the zone allocations and the capital multipliers. Naturally, it comes down to the fact that the M&P test statistic does not converge to any distribution and no conclusion can be made in that regard without assumptions on the real P&L distribution, upon which the 95% and 99% rejection thresholds will then be dependent. Figure 5.3. shows how the test statistic's null distribution can vary according to the real, unknown, P&L distribution.

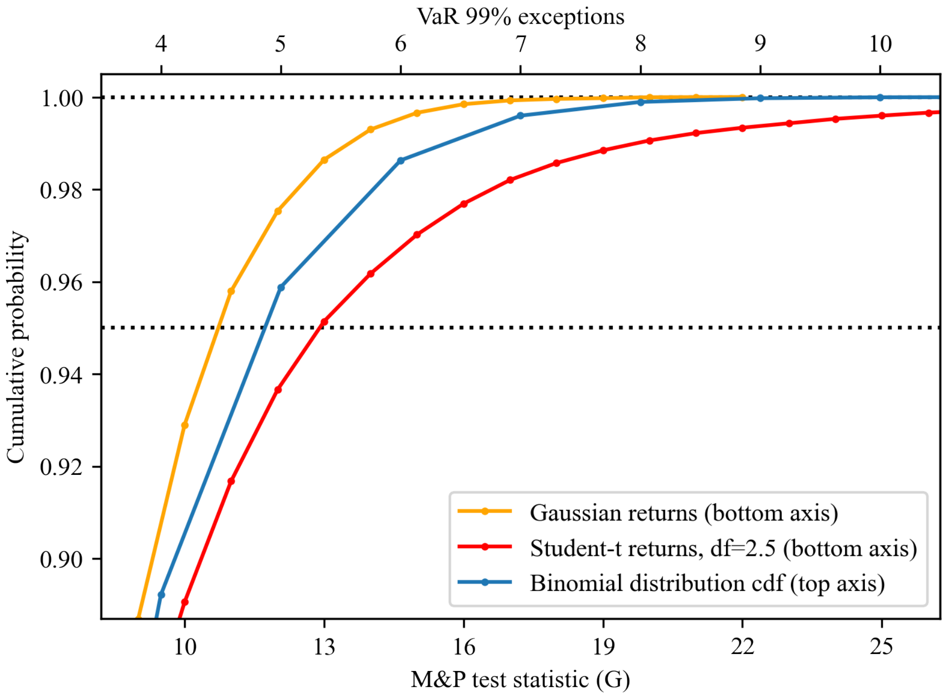

Figure 5.3. Tails of the cumulative distribution of the M&P test statistic under the null and of the FRTB's test's binomial distribution. The M&P statistic's distribution is simulated from Gaussian and student-t (2.5 degrees of freedom) returns. The axes are set so that 5 exceptions are aligned with a G of 12 and 10 with 25.

The empirical distributions in Figure 5.3. are obtained from 50 000 Gaussian and student-t (2.5 degrees of freedom) simulations and their respective $\mathrm{ES}_{2.5\%}$. A student-t distribution was fitted to the entire P&L sample, resulting in an estimated d.f. of 2.54, approximated to 2.5. The dotted lines mark the 0.95 and 0.9999 thresholds, making it clear that the values of 12 and 25 are not guaranteed to be statistically aligned with the Basel framework. The bottom line is that, despite its useful and simple interpretation, the test lacks a strong, general statistical background.

Regarding the ZR test, it is logical to use the ratio of the realized ES to the average ES estimate as the ES analogous to Equation (45). The ZR multiplier would then be given by:

$$m_{ZR} = 1.5\,\frac{\widehat{ES}_\alpha}{\overline{ES}_\alpha} \tag{46}$$

The most interesting aspect of this approach is that it requires no parametric assumptions to calculate the capital multiplier. However, it is still necessary to conduct the simulation process to determine the p-value of the ZR test statistic, since the multipliers are meant to be applied only on amber and red zones. Figure 5.4. illustrates the potential effect on the capital requirements of changing from the Basel to the ZR test.

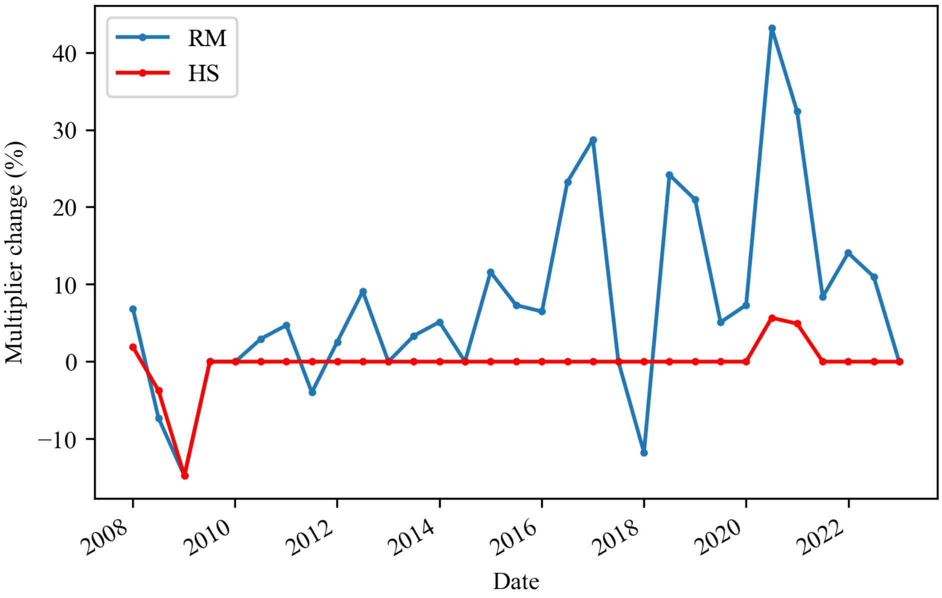

Figure 5.4. % change in capital requirements from the Basel to the ZR test (discrete samples). When the p-value is lower than 0.95, the multipliers are set at 1.5.

Figure 5.4. shows that the impact of the ZR test can be considerably greater than the M&P test. In the RM model, the change is mostly positive, averaging 7.77%. The two largest increases come in the 2020 period (Jul/19 - Jul/20 and Jan/20 - Jan/21 samples) surpassing 40%. Interestingly, these happen in a period in which both the Basel and the ZR test result in red zones, however, it is relevant to note that the Basel multiplier is capped at 2.00 (10 or more exceptions), meaning that 11, 12, or more exceptions lead to the same add-on as 10 exceptions. On the other hand, no limit was set for the ZR multiplier. If it had been, the differences would naturally be smaller in some occasions, in particular the Jul/19 - Jul/20 and Jan/18 - Jan/19 samples, in which 12 and 11 $\mathrm{VaR}_{1\%}$ exceptions were recorded, respectively. In the HS model, there are both positive and negative changes, contrary to the M&P which produces only decreases. Increases peak at 5.67% during the 2020 stress period, while the largest decrease is 14.77% in the Jan/08 - Jan/09 sample, the only period of the HS model in which the ZR test result (green zone) is different from the Basel test (Amber zone). In each model, there is only one decrease occurring when both tests place the model in the amber zone, which are in the Jul/10 - Jul/11 period in the RM model and Jul/07 - Jul/08 in the HS model. This points to the idea that the ZR test, as presented, leads to generally more prudent (higher) capital multipliers. Figure 5.5. illustrates this conclusion by plotting the multipliers resulting from the ZR test in the RM model along with their associated p-values (considering only p-values above 0.95).

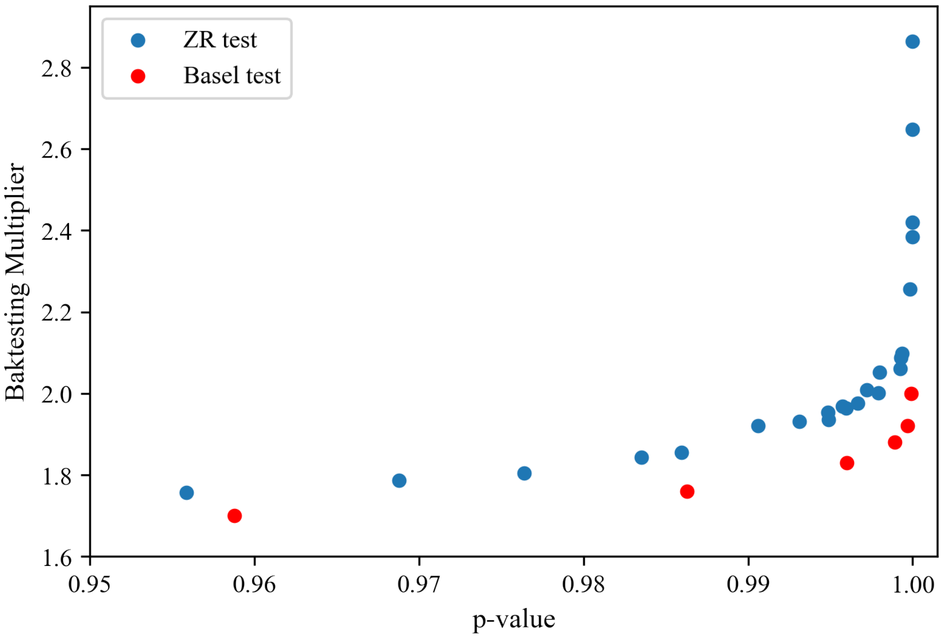

Figure 5.5. Relation between obtained p-values and capital multipliers. The blue dots are the multipliers obtained from the ZR test in the RM model. The red dots are the ones from the Basel test (Table 3.1.)

Figure 5.5. shows that, for the same p-values, the ZR test multipliers are higher than the Basel test ones. Naturally, the information in Figure 5.5. is dependent on the model, due to different test statistic distributions. A similar analysis with the HS model results is not useful due to the preponderance of green zones.

The implementation of the test does not need to be executed in this exact manner, leaving space for changes that mitigate its impact, if such is the desire of regulators. What these results show is that it is possible to use the Ridge test to determine ES backtesting dependent capital add-ons, without statistical assumptions. The add-ons are not derived from the test statistic (Equation 27) but rather from two observable values that relate to it. Since the test statistic measures absolute discrepancies, computing the multipliers directly from it would require statistical assumptions regarding not only the shape but also the scale of the P&L distribution. With assumptions only on the former, it may be possible to use the p-value instead of the test statistic. These conclusions are consequences of the sharpness of the test.

6 Conclusion

Backtesting Expected Shortfall is possible. In this study, I reviewed and analyzed the most prominent methods to do so and compared them with the standard VaR procedures. Setting from the unanimous inference that VaR backtesting is far simpler and easier to interpret, the idea was to understand if using a VaR-based procedure to test for ES coverage (as recommended in the most recent regulation) is more adequate than backtesting ES by itself.

To do so, two differently constructed models were estimated for both VaR and ES and the backtests performed daily from 2007 to 2022 following the (adapted) regulatory guidelines. This allowed me to compare the rejection rates of the tests and their implication on capital requirements, as well as their statistical robustness and computational burden, to distinguish their possible implementation between internal and external model validation.

The results show that every test procedure has its advantages and shortcomings. The multinomial tests are the only ES tests that satisfy both requisites of not requiring model information and being sustained by statistical results. However, these are shown to be less prudential than the VaR test currently recommended by the regulation, being closer to Kupiec's widely used Unconditional Coverage test. Judging by the Basel Committee's choice in favor of a simpler yet more prudent binomial test instead of Kupiec's, a multinomial test is not likely to be included in the framework, at least without further changes. Nevertheless, it should be appropriate for internal validation, in particular the Nass test with 8 levels, for more significant results.

The cumulative violations test (C&C) and the Z tests share the disadvantage of requiring the storage of model information, although in different steps of the process. Out of these, the ZR test shows the most promising results, especially for internal validation, as the simulation requirement should pose little to no difficulties to banks. When it comes to external validation, the challenge would naturally be to include a detailed unequivocal simulation process in Basel's backtesting requirements. One that is suitable for all kinds of models. If this challenge is overcome, this test brings along a simple, model-independent way of determining capital add-ons, that may result in higher capital requirements.

Finally, M&P's secured position test has some very interesting aspects, mostly due to its convenience and similar nature to the Basel test. The results, however, show that the two are not as aligned as one could hope, which means that a change to the M&P test would have a significant and uncertain impact on the capital amounts. This is due to the absence of robust statistical conclusions, which also means it is not particularly interesting for internal validation.

Naturally, the results of this study are bounded by the methodology used. The simplicity of the portfolio and non-inclusion of different assets such as bonds of derivatives, as well as the disregard of other VaR/ES models (both different specifications within parametric and historical approaches, as well as different models entirely, such as Quantile Regressions) are some limitations of this dissertation that naturally affect the generality of the conclusions. Additionally, fixing the backtesting sample at 250 throughout the analysis is also limiting. Although this is the duration recommended in the framework, when it comes to internal model validation, it is naturally possible to backtest ES over larger periods, with an impact on the tests' results.

Further changes to the standards introduced by the FRTB are not expected in the near future, as those very standards are just starting to be implemented. Nevertheless, it is not too early to debate the inclusion of ES backtesting in the framework. To do so, further research is required, mainly regarding the ZR test, given its interesting results. Additional studies should cover the properties of the ZR test statistic under different models, in the search for a general way of performing (or avoiding) the simulation process and a widespread relation between the statistic's distribution and the capital multipliers. The possibility of using a relative version of the ZR test, by standardizing the daily test statistic by the predicted ES is also an interesting aspect to be further analyzed. Other useful studies include more extensive analyses of the statistical properties of the M&P test and possible alternative backtesting thresholds.

There is no perfect risk measure and the fact is that the change from VaR to ES brought along the question of backtesting. This a question which, as standard in risk management, has no definitive answer. Be that as it may, backtesting ES internally will surely be a reality for banks, and including it in future regulation changes should be seriously considered.

7 References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26(7), 1505–1518.

Acerbi, C. and Szekely, B. (2014). Backtesting expected shortfall. *Risk Magazine.*

Acerbi, C. and Szekely, B. (2017) General properties of backtestable statistics. *Working Paper.*

Acerbi, C., Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance*, 26(7), 1487–1503.

Artzner, P., Delbaen, F., Eber, J. M., Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3), 203–228.

Barone-Adesi et al. (1998). VaR Without Correlations for Nonlinear Portfolios. *Working Paper.*

BCBS – Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: a revised market risk framework. *Bank for International Settlements.* `https://www.bis.org/publ/bcbs265.pdf`

BCBS – Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk. *Bank for International Settlements.* `https://www.bis.org/bcbs/publ/d457.pdf`

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4), 465–474.

Berkowitz, J., Christoffersen, P., Pelletier, D. (2011). Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science*, 57(12), 2213–2227.

Caporin, M., McAleer, M. (2010). Model selection and testing of conditional and stochastic volatility models. *RePEc: Research Papers in Economics.*

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

Costanzino, N., Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *The Journal of Risk Model Validation*, 9(1), 21–31.

Costanzino, N., Curran, M. (2018). A Simple traffic-light Approach to Backtesting Expected Shortfall. *Risks*, 6(1), 2.

Delbaen, F. (2002). Coherent Risk Measures on General Probability Spaces. *Advances in Finance and Stochastics*, 1–37.

Deng, K., Qiu, J. (2021). Backtesting expected shortfall and beyond. *Quantitative Finance*, 21(7), 1109–1125.

Du, Z., Escanciano, J. C. (2017). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 63(4), 940–958.

Emmer, S., Kratz, M., Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *The Journal of Risk*, 18(2), 31–60.

Engle, R. F., Manganelli, S. (1999). CAVIAR: Conditional Autoregressive Value at Risk by regression quantiles. *RePEc: Research Papers in Economics.*

Fissler, T., Ziegel, J. F., Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting. *RePEc: Research Papers in Economics.*

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

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

Kerkhof, J., Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking and Finance*, 28(8), 1845–1865.

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

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives, 3(2), 73–84.*

Lambert, N. S., Pennock, D. M., Shoham, Y. (2008). Eliciting properties of probability distributions. *SIGecom Exchanges*, 7(3), 1–5.

Löser, R., Wied, D., Ziggel, D. (2018). New backtests for unconditional coverage of expected shortfall. *Journal of Risk.*

McNeil, A. J., Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3–4), 271–300.

Mehta, A., Neukirchen, M., Pfetsch, S., Poppensieker T. (2012). Managing market risk: today and tomorrow. *McKinsey Working Papers on Risk*, 32: 1-16.

Moldenhauer, F., Pitera, M. (2019). Backtesting expected shortfall: a simple recipe? *The Journal of Risk.*

Newey, W. K., Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819.

Nolde, N., Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4).

Pérignon, C., Smith, D. R. (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking and Finance*, 34(2), 362–377.

Righi, M. B., Ceretta, P. S. (2013). Individual and flexible expected shortfall backtesting.*The Journal of Risk Model Validation*, 7(3), 3

Rodrigues, M. (2020). *Fundamental Review of the Trading Book – Impact assessment on banks' capital requirements under the Internal Models Approach* [Master thesis, Iscte Business School].

Stavroyiannis, S. (2018). A note on the Nelson-Cao inequality constraints in the GJR-GARCH model: Is there a leverage effect? *International Journal of Economics and*

Business Research, 16(4), 442.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking and Finance*, 32(7), 1404–1415.

Rockafellar, R. T., Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7), 1443–1471.

Xiao, Z., Guo, H., Lam, M. S. (2014). Quantile regression and value at risk. *Handbook of Financial Econometrics and Statistics*, 1143–1167.

Ziegel, J. F. (2014). Coherence and Elicitability. *Mathematical Finance*, 26(4), 901–918.

Annex A Descriptive Statistics of the Portfolio

Risk Factor	Mean	Max	Min	Median	Standard Deviation	Skewness	Excess Kurtosis
S&P 500	0.0003	0.1158	-0.1198	0.0004	0.0119	-0.2487	13.0442
STOXX 600	0.0002	0.0987	-0.1148	0.0005	0.0115	-0.2746	8.7522
Nikkei 225	0.0003	0.1415	-0.1140	0.0000	0.0139	-0.2529	7.9188
Hang Seng	0.0002	0.1435	-0.1270	0.0000	0.0141	0.2782	9.4152
EUR - USD	0.0000	0.0351	-0.0240	0.0001	0.0058	0.0996	1.9504
JPY - USD	0.0000	0.0393	-0.0533	0.0000	0.0061	0.1812	5.0551
HKD - USD	0.0000	0.0062	-0.0027	0.0000	0.0004	1.8523	31.4243
Portfolio	0.0003	0.1054	-0.0954	0.0006	0.0103	-0.3238	12.3017

Table A.1. Descriptive statistics of the Portfolio. Daily returns of each risk factor and the of the combined portfolio.

Annex B VaR/ES estimation results

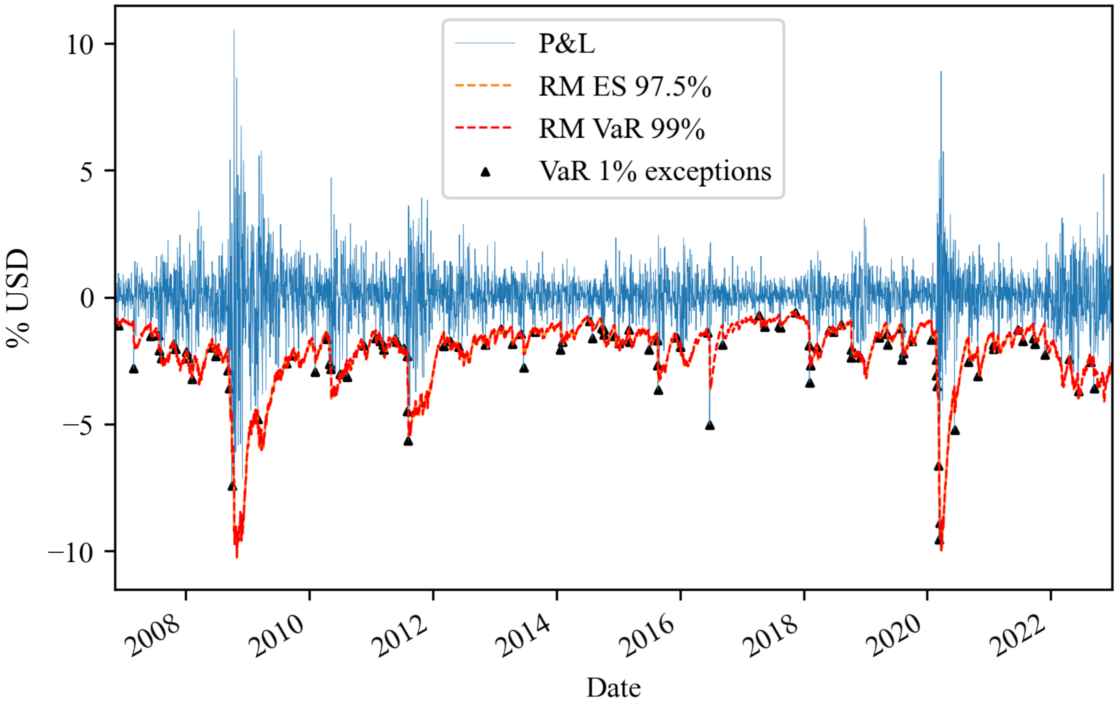

Figure B.1. Portfolio P&L combined with $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ estimates - RM model. The black marks represent the P&L in the presence of $\text{VaR}_{1\%}$ exceptions.

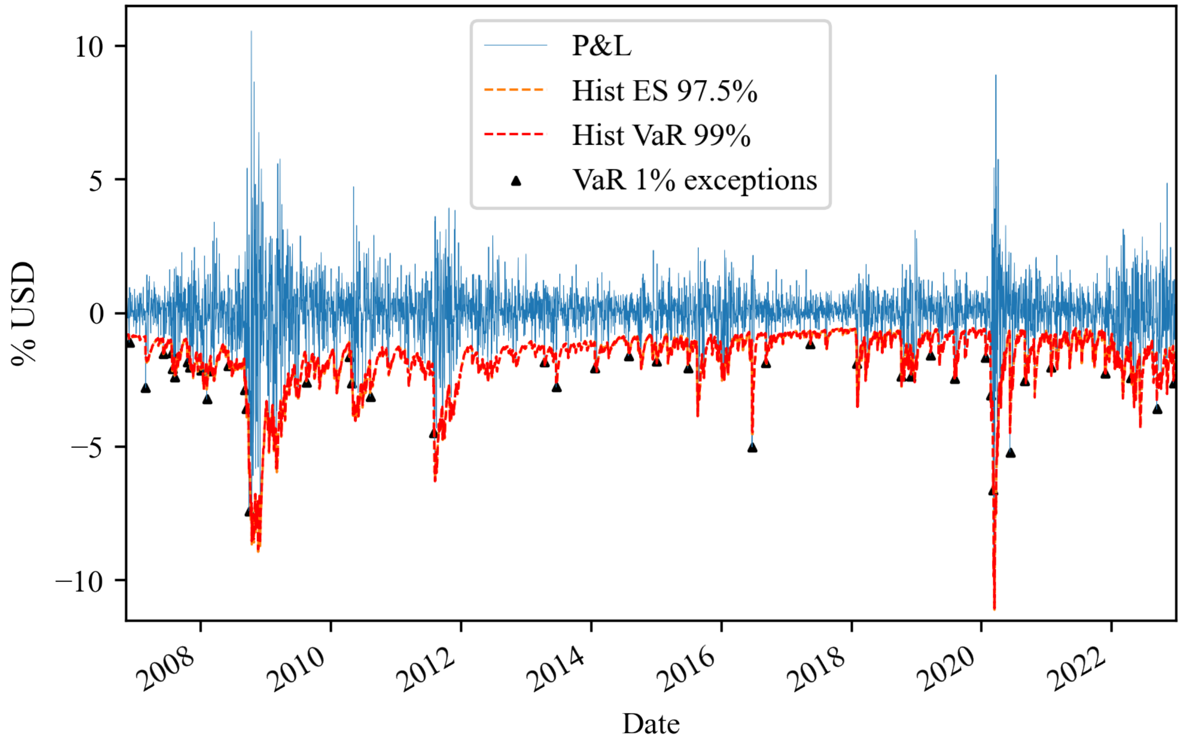

Figure B.2. Portfolio P&L combined with $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ estimates - HS model. The black marks represent the P&L in the presence of $\text{VaR}_{1\%}$ exceptions.

Annex C Z2 test 95% and 99.99% thresholds

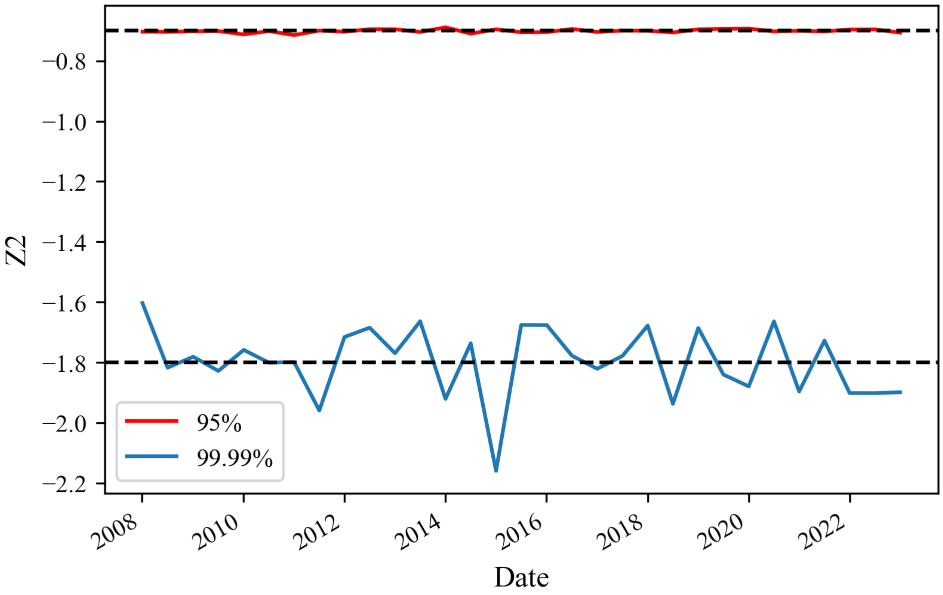

Figure C.1. Simulated Z2 test thresholds (discrete samples) - RM model, 20 000 simulations. The dotted lines mark the -0.7 and -1.8 thresholds recommended by Acerbi and Szekely (2014)

	Mean	Std. Deviation	RMSD
95%	-0.701181	0.005706	0.005827
99.99%	-1.797774	0.114312	0.114333

Table C.1. Statistics of the simulated thresholds (2007 - 2022) - RM model, 20 000 simulations. RMSD is the root of the mean squared deviation between the recommended and the simulated thresholds.

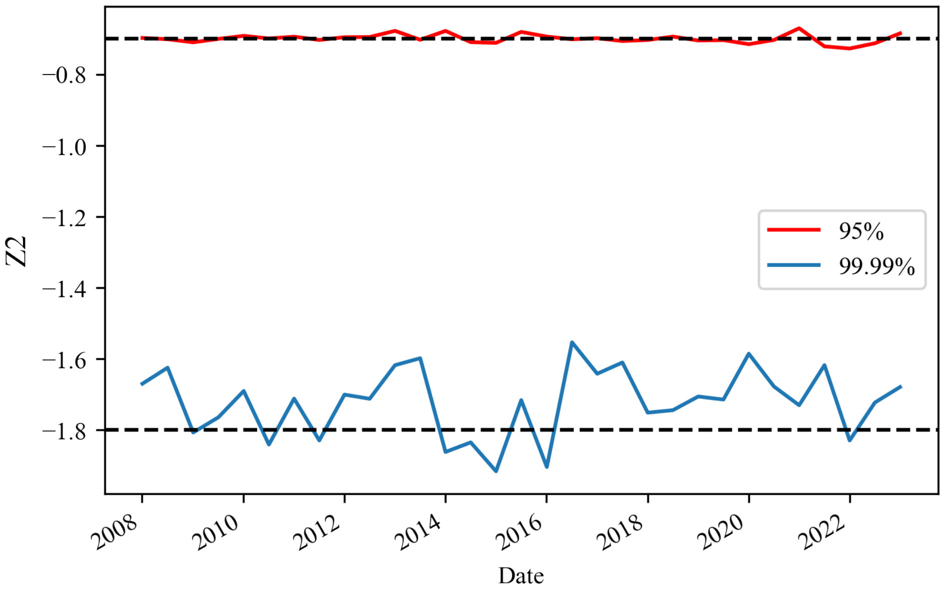

Figure C.2. Simulated Z2 test thresholds (discrete samples) - HS model, 20 000 simulations. The dotted lines mark the -0.7 and -1.8 thresholds recommended by Acerbi and Szekely (2014)

	Mean	Std. Deviation	RMSD
95%	-0.699150	0.012356	0.012385
99.99%	-1.721564	0.093692	0.122190

Table C.2. Statistics of the simulated thresholds (2007 - 2022) - HS model, 20 000 simulations. RMSD is the root of the mean squared deviation between the recommended and the simulated thresholds.

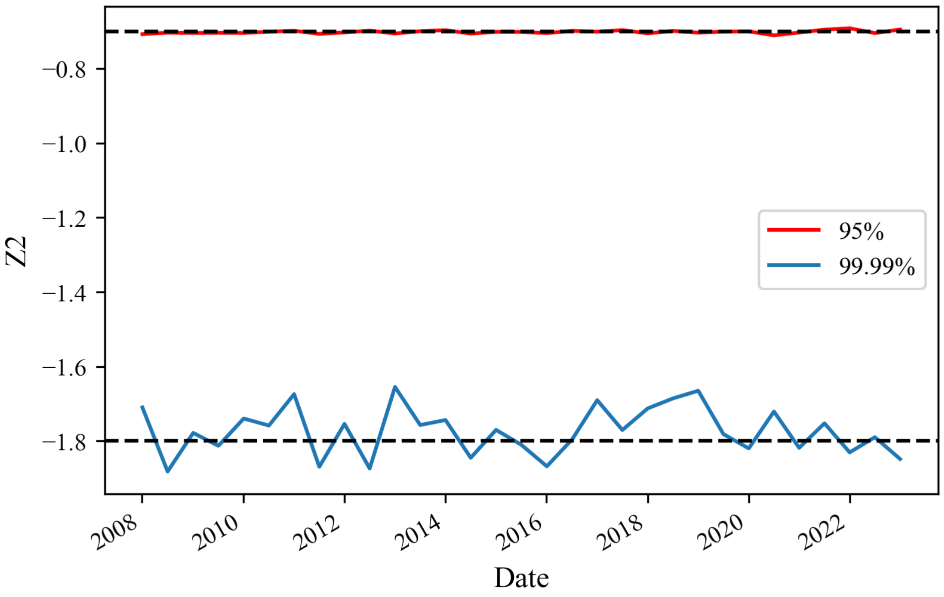

Figure C.3. Simulated Z2 test thresholds (discrete samples) - RM model, 50 000 simulations. The dotted lines mark the -0.7 and -1.8 thresholds recommended by Acerbi and Szekely (2014)

	Mean	Std. Deviation	RMSD
95%	-0.701220	0.004081	0.004260
99.99%	-1.774058	0.063802	0.068874

Table C.3. Statistics of the simulated thresholds (2007 - 2022) - RM model, 50 000 simulations. RMSD is the root of the mean squared deviation between the recommended and the simulated thresholds.

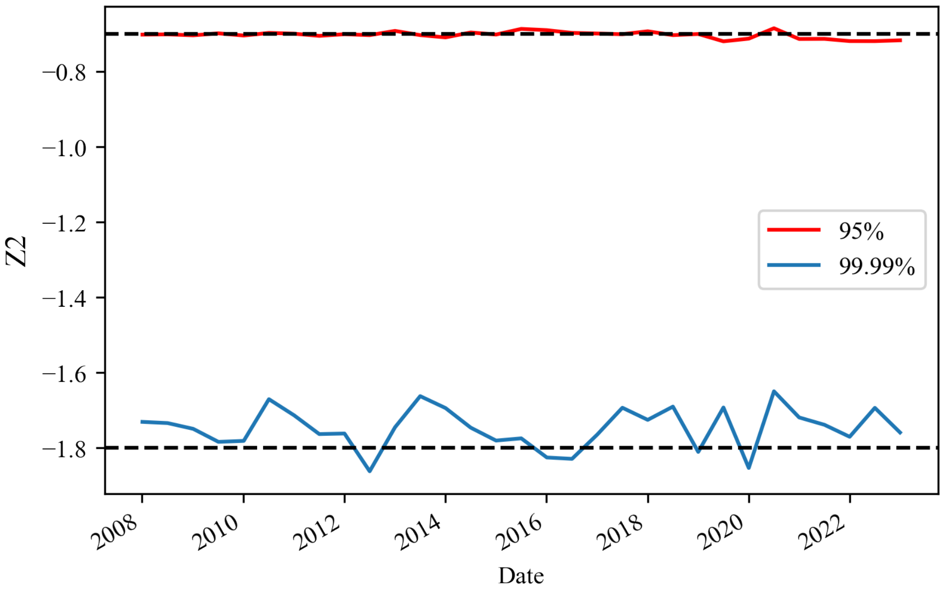

Figure C.4. Simulated Z2 test thresholds (discrete samples) - HS model, 50 000 simulations. The dotted lines mark the -0.7 and -1.8 thresholds recommended by Acerbi and Szekely (2014)

	Mean	Std. Deviation	RMSD
95%	-0.702905	0.009002	0.009460
99.99%	-1.747567	0.053129	0.074646

Table C.4. Statistics of the simulated thresholds (2007 - 2022) - HS model, 50 000 simulations. RMSD is the root of the mean squared deviation between the recommended and the simulated thresholds.