Into the Trading Book: Estimating Expected Shortfall

Robin Eric Schmutz, Leonard Schneider
Lund University
School of Economics and Management
Department of Economics

MSc in Finance, Thesis
Supervisor: Anders Vilhelmsson
June 2023

Abstract

In light of the revised 2019 proposals constituting the Fundamental Review of the Trading Book, which amend the third Basel Accord, expected shortfall is set to replace value at risk as the risk measure dictating banks' capital reserving requirements for exposure to market risk. This paper examines how best to accurately estimate expected shortfall from a regulatory perspective by carrying out an array of non-parametric as well as parametric methods over the recent years of financial instability. While previous research has predominantly made use of stock indexes to proxy bank's trading books, we not only employ the S&P500 Index, but also real profit-and-loss data of three large European banks. Through backtesting we identify a GARCH-Generalized Pareto distribution model (rooted in the peaks-over-threshold model) as yielding the most satisfactory ES forecasts for both the index and bank data sets, with the age-weighted historical simulation method also showcasing an all-around strong performance.

Keywords: Expected shortfall; Trading book; Historical simulation; Parametric estimation; Backtesting

Acknowledgements

The realization of this research project would not have been possible without the valuable guidance of our supervisor, Associate Professor Anders Vilhelmsson, whose passion for teaching and risk management course design within this Master's program inspired this essay, and whose advice and feedback we are grateful for.

Contents

1 Introduction

Risk management is crucial to ensure the stability of financial institutions. For the past three decades, Value at Risk (VaR) was the predominant risk measure for estimating market risk. After the financial crisis in 2008, however, VaR received a lot of criticism, mainly due to its inability to capture tail risk (Hull, 2018). VaR does not take the size of losses beyond a certain confidence level into account, which means that it is completely insensitive to losses which happen with a very small probability. For example, if the size of the losses, which occur with a probability of less than 1% increases, VaR at a 99% confidence interval will nevertheless remain the same, as it does not consider the losses beyond the confidence level. Furthermore, VaR is no coherent risk measure since it violates the assumption of subadditivity. Practically speaking, subadditivity means that a risk measure should never discourage diversification, which makes sense from an intuitive point of view. VaR, however, only fulfills this property under certain conditions (e.g., if the losses are normally distributed) (Hull, 2018).

Due to the mentioned shortcomings of VaR, the Basel committee proposed the Fundamental Review of the Trading Book (FRTB) in 2012 as a new framework for market risk and capital standards, which introduced Expected Shortfall (ES) as a new risk measure (BIS, 2012). ES considers all losses beyond VaR and computes the average of those losses. Therefore, ES takes losses which occur with a very low probability into account and is better suited to capture tail risk (Yamai & Yoshiba, 2005). Furthermore, ES has superior properties compared to VaR, as it always acknowledges the advantages of diversification (Hull, 2018). Initially, the shift to ES was planned to be implemented in 2019, but it got delayed several times and finally became active in January 2023 (BIS, 2023). Regarding the new standards, banks are required to use a 10-day horizon, a 97.5th percentile and a one-tailed confidence interval in order to compute ES, which will subsequently determine their capital requirements (BIS, 2023). However, there is no strict guideline which models should be used for calculating ES. Therefore, the question which methods deliver the best results regarding the estimation of ES remains a pivotal topic of discourse. Academics most commonly use stock indexes as a proxy for a bank's trading book in order to conduct empirical analyses contributing to this discussion. Yet, it is unclear whether the same estimation methods that work for stock indexes, commodities or exchange rates are also delivering the best results for an actual bank's trading book. Thus, we will use real-world profit-and-loss (P/L) data of banks' trading books as well as the returns of the S&P500 in order to investigate if the same methodical approaches for estimating ES hold strong for both a stock index and a bank's trading book.

1.1 Purpose

The aim of this thesis is to contribute to the empirical literature on estimating ES by applying both parametric and non-parametric approaches on the returns of the S&P500 as well as on P/L data of banks' trading books. First, this analysis will evaluate which methods are best suited for estimating ES. Identifying a model that generates precise ES forecasts is highly relevant since ES is a recently implemented risk measure that determines banks' capital requirements. Second, this paper will also critically reconsider the relevance of previous empirical studies within this area. Previous research has mainly used stock indexes as a proxy for a bank's trading book, while this study uses real P/L data of banks' trading books. Banks typically do not provide their P/L data for research; however, they include plots comparing their P/L data with their VaR estimates within their pillar 3 disclosures under the rubric MR4. Those plots include data points of their daily P/L, which we manually extracted. Thus, we are able to use real bank data to assess which methodical approach is best suited for estimating ES. Furthermore, by comparing the results for the banks' P/L data with the ones for the S&P500, we can evaluate whether a common stock index is a suitable proxy for a bank's trading book. If we would be to find out that the results are very different, this would imply that the results of previous research using stock indexes are only limitedly applicable to real-world banks' trading books.

1.2 Literature Overview

One of the most wide-ranging empirical study comparing different ES estimation models was conducted by Righi and Ceretta in 2015. Not only examined the study 17 different ES estimation models, but it also used multiple data sets from different financial asset classes (equity, fixed income, exchange rates and commodities). Further the authors investigated different sizes of rolling windows and significance levels. Their results suggest that conditional models deliver the best results, particularly the GARCH methods and Filtered Historical Simulation methods. Furthermore, they found that incorporating leptokurtic asymmetry, by using the skewed Student's t distribution, leads to better estimation results. These two findings indicate that the consideration of stylized facts of financial returns, mainly volatility clusters, heavy tails and asymmetry, are critical for a correct estimation of ES. Asymmetry was especially prevalent for the equity and commodity market, whereas the fixed income and exchange rate market seemed to be relatively symmetric. Regarding the length of the estimation windows, large rolling windows generally delivered better results than smaller windows, which might miss relevant information.

An example for a recent empirical paper on estimating ES is given by Sobreira and Louro (2020). They used Historical Simulation, several parametric volatility models of different GARCH classes as well as the Extreme Value Theory approach (EVT) with Peak-over-Threshold (POT). They performed their analysis on the stock returns of large Portuguese companies individually, instead of using an index as a proxy. The results show that the EVT approach delivered the best estimates overall. The best performing model was the asymmetric GARCH with EVT. However, in contrast to Righi and Ceretta (2015), they found that skewed distributions generally do not outperform their symmetric counterparts. The paper applied different backtests and concludes that the results were relatively similar across most backtests. Regarding the size the of the rolling window, their analysis was ambiguous. The authors generally found modest advantage for larger sample sizes, but smaller sample sizes performed better when a low confidence level was chosen.

One of the earlier studies on estimating ES was carried out by McNeil and Frey (2000), who used different GARCH approaches and EVT to model return data. After comparing their results to those of several other estimation methods, they conclude that taking leptokurtosis into account is of major importance and suggest using a fat-tailed distribution and an EVT approach. A very similar analysis was done by Bah, Munga'tu and Waititu (2016), who fitted GARCH models to return data of the Nairobi 20 Share Index and showed how EVT can be used in order to model tail risk.

Harmantzis, Miao and Chien (2006) empirically evaluated different models for estimating VaR and ES by using returns of six major stock indexes as well as different exchange rates. Their analysis suggests that the historical method as well as the EVT and POT approach lead to the most accurate estimations. Further, the Gaussian approach was found to underestimate ES whereas the Stable Paretian approach resulted in an overestimation of ES.

1.3 Outline

The thesis is structured in the following manner: Section 2 covers the definition of ES and the backtesting approach. In section 3 the utilized data and the estimation methods are discussed. Section 4 presents the empirical results and section 5 concludes.

❈

2 Expected Shortfall

In this part, the concept of expected shortfall is briefly touched upon, laying the foundation for estimation in the methodology section by starting out with some necessary definitions before introducing the process of backtesting and reviewing the test chosen to perform this function within this essay.

2.1 Definition

For some loss distribution F_L, expected shortfall (ES) is a risk measure defined in the following manner by McNeil, Frey and Embrechts (2015), encompassing both discrete and continuous cases:

$$ES_\alpha = \frac{\mathbb{E}(L \cdot \mathbb{1}_{L>VaR_\alpha}) + VaR_\alpha(1 - \alpha - \mathbb{P}(L > VaR_\alpha))}{1 - \alpha} \tag{1}$$

with $\alpha \in (0, 1)$ denoting an arbitrary confidence level.

Expected shortfall itself rests on another risk measure, which it is a function of - value at risk (VaR). Both measures reflect loss tail risk, the latter defined in its general form as the minimum possible loss realization such that the likelihood of a loss being superior to said value is no greater than $1 - \alpha$:

$$VaR_\alpha = \min(l : \mathbb{P}(L > l) \leqslant 1 - \alpha). \tag{2}$$

When F_L is continuous, one is able to derive the relevant nested definition in Equation 2:

$$VaR_\alpha = \min(l : \mathbb{P}(L > l) = 1 - \alpha) \tag{3}$$

or, equivalently:

$$\mathbb{P}(L > VaR_\alpha) = 1 - \alpha \tag{4}$$

that is, the α-quantile of the loss distribution (taking the c.d.f.'s inverse):

$$VaR_\alpha = F_L^{-1}(\alpha). \tag{5}$$

In light of Equation 4, Equation 1 can consequently be simplified for continuous distributions:

$$ES_\alpha = \frac{\mathbb{E}(L \cdot \mathbb{1}_{L>VaR_\alpha})}{1 - \alpha}. \tag{6}$$

In order to rewrite Equation 6 differently, first denote random variables L and $\mathbb{1}$ by X and Y respectively. Under this setting, the joint distribution for two such random variables can be decomposed into one's conditional and the other's marginal distributions:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y) \tag{7}$$

which can in turn be used to evaluate the expectation of X conditional on Y:

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \int_{\mathbb{R}} x \cdot f_{X|Y}(x|y)dx \\ &= \int_{\mathbb{R}} x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)}dx = \frac{1}{f_Y(y)} \int_{\mathbb{R}} x \cdot f_{X,Y}(x,y)dx\end{aligned} \tag{8}$$

making use of Equation 7 for the second equality. Given the indicator function

$$\mathbb{1} = \begin{cases} 1 & L > VaR_\alpha \\ 0 & L \leqslant VaR_\alpha \end{cases} \tag{9}$$

the RHS of Equation 6 can be separated into two parts based on cases where L is and is not greater than value at risk:

$$\begin{aligned}ES_\alpha = \frac{\mathbb{E}(L \cdot (\mathbb{1}_{L>VaR_\alpha} = 1))}{1-\alpha} &+ \frac{\mathbb{E}(L \cdot (\mathbb{1}_{L>VaR_\alpha} = 0))}{1-\alpha} \\ &= \frac{\mathbb{E}(L \cdot (\mathbb{1}_{L>VaR_\alpha} = 1))}{1-\alpha}\end{aligned} \tag{10}$$

the last equality holding due to the indicator function's state rendering the expectation in the second term null.

In Equation 9, $\mathbb{1}$ is 1 with probability $\mathbb{P}(L > VaR_\alpha) \equiv 1-\alpha$. This corresponds to the denominator in Equation 10, which can be associated with $f_Y(y) = \mathbb{P}(Y=y)$ in Equation 8 if $y = 1$. For this value of y alone, the integral equates to $\mathbb{E}(XY) = \mathbb{E}(X \cdot 1)$, the expression found in Equation 10's numerator. As such, Equation 6 can finally be re-expressed as $\mathbb{E}(X|Y=y) = \mathbb{E}(X|Y=1)$:

$$\boxed{ES_\alpha} = \mathbb{E}(L|\mathbb{1}=1)\boxed{= \mathbb{E}(L|L > VaR_\alpha)} \tag{11}$$

the expected value of all possible losses realizations larger than value at risk, thus offering better clarity as to why expected shortfall is sometimes referred to as expected tail loss (ETL).

Expected shortfall may alternatively be presented as:

$$ES_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 VaR_x \mathrm{d}x \tag{12}$$

and is commonly termed conditional value at risk (CVaR), the acronym of which should not be confused with component value at risk (a VaR disaggregation measure pertaining to individual sub-portfolios/portfolio constituents).

Given the definition in Equation 11, it is natural to deduce that expected shortfall is always at least as large as its associated value at risk.

To illustrate this graphically, assume that losses within a given time frame follow a standard normal distribution. Using Equation 5 and Equation 12, VaR and ES at a chosen confidence level of 95% have been computed. As shown in Figure 1 below, ES is located farther out in the right tail of the distribution, where more extreme losses might occur, giving a more prudent assessment of the risk exposure faced.

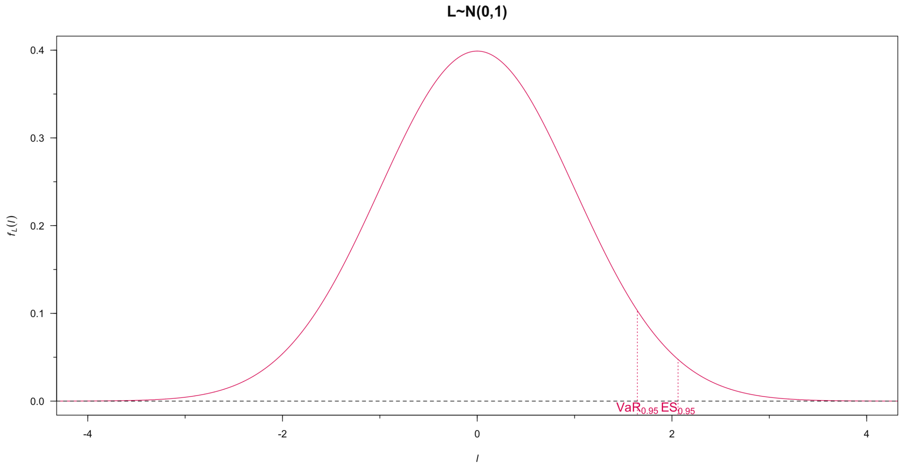

Figure 1: 95% value at risk and expected shortfall for standard normal losses.

2.2 Backtesting ES Severity

In order to assess the adequacy of daily ES estimates obtained via methods described further into this essay from a regulatory perspective (placing sole emphasis on avoiding grossly underestimating ES - from which minimum capital provisions needing to be set aside to account for market risk are to be derived), backtesting results is, in addition to thorough risk factor identification, of the utmost importance in the validation process, not to mention when it comes to gauging proper incentive alignment/lack of negligence. As the Basel Committee on Banking Supervision (BCBS) puts it, a bank trading desk using an internal model as opposed to the standardised approach is required to ensure its quantification of risk is ”sufficiently conservative” (Bank for International Settlements (2019, p.6)).

Within the context of the latest proposals constituting the imminent Fundamental Review of the Trading Book (FRTB) reform which will amend the third Basel Accord - with a staggered implementation across countries over the coming years - a shift from VaR to ES (the latter more apt at capturing tail risk, as is particularly flagrant in heavy-tailed distributions) is imposed, albeit the confidence level lowered from 99% to 97.5% generally making this change practically immaterial according to the BCBS. While still performed on VaR as of to date, backtesting is done on ES in this essay. In keeping with occurrences but not the timing of VaR exceptions (or violations) being the crux of current regulation, their severity (and not independence) is focused on here.

Before proceeding, let B denote a given evaluation period (the set of consecutive time periods over which an ES estimate-generating model is considered) with cardinality $|B| = b \approx 250$ (trading days in a calendar year) under the Basel framework. This is the standard interval subject to backtesting.

2.2.1 Acerbi and Szekely (2014)

One such way of testing the adequacy of an ES model (or estimation method) is by executing Acerbi and Szekely's (2014) so-called unconditional (based on the unconditional expectation: Equation 6) **second test**, a one-sided test wherein whether the model in question underestimates ES too often is sought after. This test, in contrast to their so-called conditional (based on the conditional expectation: Equation 11) first test, does not call for a preliminary VaR test such as the Kupiec frequency test to support it (the first test's alternative hypothesis keeps predicted and true VaRs strictly equal to ensure Equation 17 in the case of its test statistic Z_1 (not covered here)).

Its null hypothesis is formulated as follows:

$$\begin{aligned} H_0 : \forall t.\ \hat{ES}_{\alpha,t} &= ES_{\alpha,t} \\ \forall t.\ V\hat{a}R_{\alpha,t} &= VaR_{\alpha,t} \end{aligned} \tag{13}$$

effectively postulating that the ES model pending validation always perfectly estimates (or predicts) ES. More specifically, the formulation above is implied from stating that the (upper) tail of the predictive (given the information set Ω_{t-1}) loss distribution for any time t is correctly specified, as in equal to that of the inherently unknown data generating process (DGP):

$$\forall t.\ \hat{F}_{L,t}(l) = F_{L,t}(l) \quad \forall l > VaR_{\alpha,t}$$

The alternative hypothesis, given the preamble to this section, is:

$$\begin{aligned} H_1 : \exists t.\ \hat{ES}_{\alpha,t} &< ES_{\alpha,t}\ (\forall t.\ \hat{ES}_{\alpha,t} \leqslant ES_{\alpha,t}) \\ \forall t.\ V\hat{a}R_{\alpha,t} &\leqslant VaR_{\alpha,t} \end{aligned} \tag{14}$$

The authors then define the test statistic (here in sample counterpart representation) as:

$$\boxed{Z_2 = -\frac{1}{b}\sum_{t=1}^{b} \frac{l_t \cdot \mathbb{1}_{l_t > V\hat{a}R_{\alpha,t}}/(1-\alpha)}{\hat{ES}_{\alpha,t}} + 1.} \tag{15}$$

This random variable is able to take a maximum value of 1 when there are no VaR exceptions within a given year. As the authors mention and as is noticeable in Z_2, the magnitude of VaR violations contributes positively in lowering the test statistic's value. Additionally, by taking $1-\alpha$ out of the sum, we effectively divide said sum by the fixed $b(1-\alpha)$, the expected number of VaR violations under H_0.

The test statistic applicable to Acerbi and Szekely's (2014) first test, Z_1, differs from Z_2 exclusively in that this aforementioned $b(1-\alpha)$ term dividing the sum is replaced by the actual number of VaR violations in B: let this number be denoted by v. While the number of non-zero terms in the sums in both Z_1 and Z_2 is always v, Z_1 does not reward low vs ($< b(1-\alpha)$), nor does it punish high vs ($> b(1-\alpha)$), and instead renders the negative term in itself essentially nothing but an average violation over predicted ES quotient. Z_2 on the other hand, is sensitive to the frequency of VaR violations - which also positively contributes in lowering the test statistic's value - and always defined.

Under H_0, Z_2 has has expectation:

$$\begin{aligned}\mathbb{E}_{H_0}(Z_2) &= \mathbb{E}_{H_0}\left(-\frac{1}{b}\sum_{t=1}^{b}\frac{l_t \cdot \mathbb{1}_{l_t > V\hat{a}R_{\alpha,t}}/(1-\alpha)}{\hat{ES}_{\alpha,t}} + \frac{1}{b}\sum_{t=1}^{b}1\right) \\ &= \frac{1}{b}\sum_{t=1}^{b}\left(\mathbb{E}_{H_0}\left(-\frac{L_t \cdot \mathbb{1}_{L_t > VaR_{\alpha,t}}/(1-\alpha)}{ES_{\alpha,t}} + 1\right)\right) \\ &= \frac{1}{b}\sum_{t=1}^{b}\left(-\frac{\mathbb{E}(L_t \cdot \mathbb{1}_{L_t > VaR_{\alpha}})/(1-\alpha)}{ES_{\alpha}} + 1\right) = 0. \qquad (16)\end{aligned}$$

The penultimate equality is explained by the fact that the true (unknown) distribution's VaR/ES are time-invariant, whereas the last equality simply follows from the definition of ES in Equation 6 (subtracting the RHS from both sides before dividing them by the initial LHS).

Naturally, the denominator in the final sum above under H_1 would be $\mathbb{E}_{H_1}(\hat{ES}_{\alpha,t})$ ($< ES_\alpha$ as for some t: $\hat{ES}_{\alpha,t} < ES_{\alpha,t}$), and the indicator function in the numerator $\mathbb{1}_{L_t > V\hat{a}R_{\alpha,t}}$ ($\geqslant \mathbb{1}_{L_t > VaR_\alpha}$ as $\forall t.\ V\hat{a}R_{\alpha,t} \leqslant VaR_{\alpha,t}$), inducing $\mathbb{E}_{H_1}(L_t \cdot \mathbb{1}_{L_t > V\hat{a}R_{\alpha,t}}) \geqslant \mathbb{E}_{H_1}(L_t \cdot \mathbb{1}_{L_t > VaR_\alpha})$ since we are working with portfolio losses (assumed to be positive beyond sufficiently extreme VaRs) as opposed to payoffs. As a result:

$$\mathbb{E}_{H_1}(Z_2) < 0. \qquad (17)$$

What precedes goes to show that a model that is subject to few VaR violations within an evaluation period and/or whose violations are not egregious, and this on a consistent basis (across evaluation periods) is pivotal in it performing well.

2.2.2 Simulating Test 2 Critical Values

To determine the appropriate significance thresholds (or critical values) below which one might reject the null hypothesis (in this case when Z_2 is sufficiently negative), a **Monte Carlo (MC) simulation** is performed.

By generating a sequence of 250·1,000,000 pseudo-random draws from a N(0,1) distribution - assuming that this is the theoretical distribution characterizing the underlying DGP for losses - and computing Z_2 in concordance with Equation 15 (note that in this designed setting we implicitly now have deterministic $\hat{ES}_{\alpha,t} = ES_{\alpha}^{N(0,1)}$, $\forall t$ and $V\hat{a}R_{\alpha,t} = VaR_{\alpha}^{N(0,1)}$, $\forall t$) for each of the 1,000,000 sets, we arrive at Z_2's simulated distribution under this specific theoretical loss distribution. In general:

$$\hat{ES}_{\alpha,t} = ES_{\alpha}^{D}, \quad \forall t$$
$$V\hat{a}R_{\alpha,t} = VaR_{\alpha}^{D}, \quad \forall t$$

with $D = F_L^{th.}$. This can be done numerically with Equation 12 and Equation 5. Next, critical values are approximated by extracting the α'-quantile of interest (α' for this test *coincidentally* corresponding to the arbitrary significance level of interest since we probe the left-tail). Note that here we denote the significance level by α' so as to avoid confusion with the predefined confidence level α (relating to VaR/ES) while maintaining conventional notation.

To clarify notation, let these quantiles be denoted by:

$$z_{2,\alpha'}^{MC,D} = F^{-1}_{Z_2^{MC,D}}\left(\alpha'\right), \quad D = F_L^{th.}$$

with *th.* shorthand for 'theoretical'. That is, approximately,

$$z_{2(s)} + \left(z_{2(s+1)} - z_{2(s)}\right)\left(\alpha' - \frac{s-1}{n-1}\right)(n-1), \quad s = \sum_{i=1}^{n} \mathbb{1}_{\frac{i-1}{n-1} \leqslant \alpha'}$$

with $z_{2(s)}$ the s^{th} order statistic of the simulated sample $z_{2,1}, ..., z_{2,n}$ (dropping 'MC, D' notation superscripts for convenience). Linear interpolation intervenes merely when the s^{th} order statistic and its successor are non-identical, and α' is neither 0 nor a multiple of $\frac{1}{n-1}$. Note that for $n \gg 1$:

$$\alpha' - \frac{s-1}{n-1} < \frac{1}{n-1} \approx 0$$
$$\mathbb{P}\left(Z_2^{MC,D} \leqslant z_{2,\alpha'}^{MC,D}\right) = \frac{s}{n} \approx \frac{s-1}{n-1}, \quad \frac{s}{n} \geqslant \frac{s-1}{n-1}$$

This difference is negligible here.

In Acerbi and Szekely (2014), a table (#4 in the paper) is provided to this effect, detailing said values when losses follow (not limited to) standardized Student's t-distributions of varying degrees of freedom (including the Gaussian case). For instance, for the N(0,1), the authors report values of -0.70 (5% significance level) and -1.80 (0.01% significance level).

Running this experiment ourselves, we get $z_{2,5\%}^{MC,N(0,1)} = -0.702 \approx -0.70$ and $z_{2,0.01\%}^{MC,N(0,1)} = -1.798 \approx -1.80$ (rounded to 3 d.p.). Note that these results are derived from choosing a VaR/ES confidence level of 97.5%. From this we deduce that n =1,000,000 sets is likely large enough to yield satisfactory results when extrapolated to other theoretical loss distributions.

When simulating a Z_2 distribution, all n realizations $z_2^{MC,D}$ stem from the theoretical distribution, and hence the probability of committing a type I error (falsely rejecting H_0) is α', the proportion of simulated test statistics $\leqslant z_{2,\alpha'}^{MC,D}$.

H_0 rejection: critical value criterion

H_0 is rejected in favor of H_1 at the α' level when the observed test statistic is inferior to the critical value of interest: $z_2 < z_{2,\alpha'}^{MC,D}$.

Quantile-quantile (QQ) plots (Figure 5/Figure 6) for two of the loss data sets tackled in this essay ('S&P500': Section 3.1.1/'Bank': Section 3.1.2) suggest a fitted Normal distribution may not constitute an appropriate choice for a theoretical distribution when observing what in this essay is termed the 'test period' (the period of time spanning all (successive) evaluation periods considered) *ex-post*, as the sample quantiles at the extremes are much farther out than their theoretical counterparts (what would be expected of said fitted distribution).

This is affirmed when carrying out the (one-sample) Kolmogorov-Smirnov (KS) goodness-of-fit test, where the largest absolute discrepancy between empirical and theoretical quantile pairs (the test statistic) by far exceeds the relevant KS critical value at the 10% significance level in both data sets. As such, the null hypothesis entailing that the empirical distribution may be assimilated to the theoretical one is rejected in both cases.

While these results are reversed for fitted Student's t-distributions (borderline for the S&P500 data), which provide superior mappings onto the test period samples, the simulated Z_2 critical values at low significance levels resulting from these fitted t-distributions (the degrees of freedom of which are $\leqslant 3$ in both cases) were deemed too lenient (low) for practical usage.

Therefore, fitting a normal-inverse Gaussian (NIG(α,β,δ,μ)) distribution to each data set was favored, and in doing so enabled to generally better encompass data, the matching undoubtedly aided to some extent by a supplementary asymmetry parameter in β as compared to the Student's t. Parameter estimates obtained from maximum likelihood estimation optimization along with the KS p-values for NIG fits when pitted against the various data sets are shown in Appendix 7.4.

To illustrate what precedes, we take the S&P500 data set, for which Figure

2 portrays the simulated critical values at some of the common statistical significance levels. Said values at the first and last of these levels are, following the Basel traffic light system logic, equated in this essay to the amber and red zone thresholds, as is urged in Acerbi and Szekely (2014).

❉

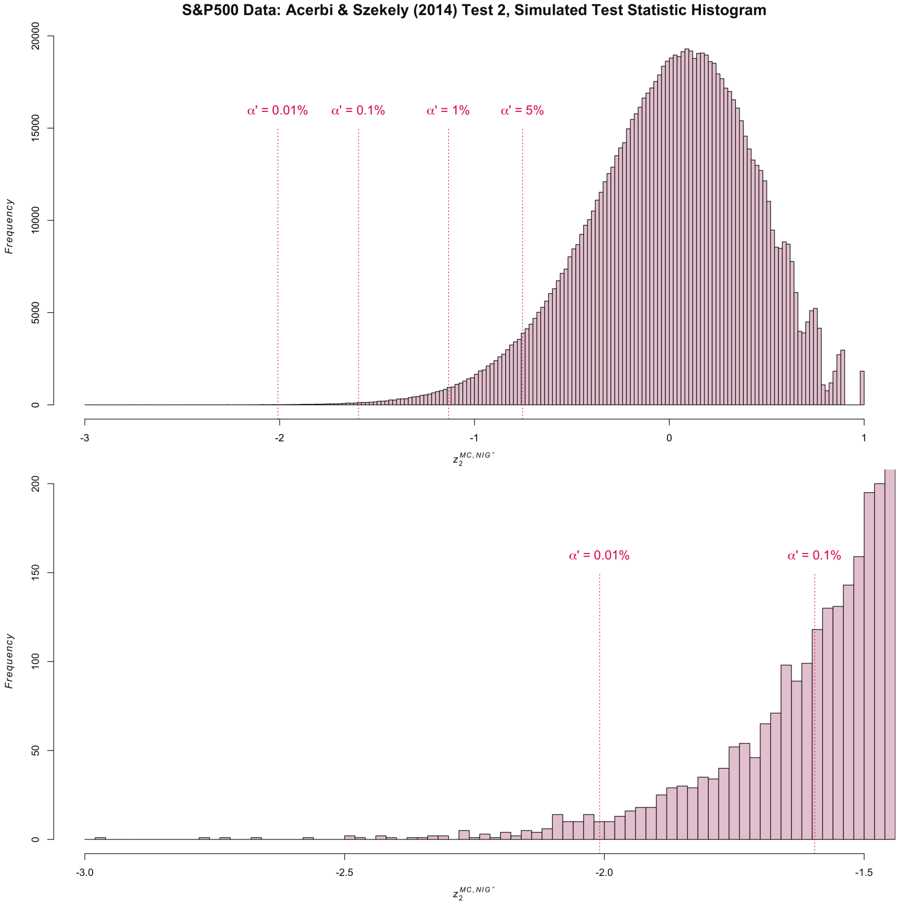

Figure 2: **S&P500.** Histogram (with bin widths of 0.02) of 1,000,000 realizations of the Z_2 test statistic obtained through a Monte Carlo simulation (full distribution (top), zoom of left-tail (bottom)). Max = 1.00, min ≈ -2.97, mean $\approx -0.055\%$. More detail are supplied in Appendix 7.4. The theoretical loss distribution is assumed to be NIG*, the asterisk signifying its parameters were chosen as those yielded by maximum likelihood estimation ex-post on the test period.

3 Data and Methodology

The intent behind this portion of this text is twofold.

Firstly, the various data sets examined are described together with the ways in which they are envisaged to be employed. Broadly, we classify these data sets linked with financial losses into two distinct categories: those (here one) originating from an equity index's movements, and those of trading book profit and loss (P/L) provenance with more concrete implications for the banking sector.

Secondly, the backbone of this work is laid out by clarifying the ES estimation methods that are retroactively applied to these data sets under the lens of the upcoming FRTB regulation.

3.1 Loss Data

In the sub-chapters that follow, the sources of data are revealed, and the specifics of the index data set and one bank data set investigated.

3.1.1 S&P500 Data

To commence with, daily data of the value of the Standard and Poor's 500 (S&P500) index was procured from the S&P Capital IQ database, from which one-period holding returns since the turn of the 21^{st} century (2000 through 2022) were calculated. These returns (Figure 3) served as an initial proxy for a bank's trading book, the index's constituents being established firms with deep investor bases by virtue of their elevated market capitalizations (their relatively risk-free features attractive to institutional investors).

In addition to the S&P500 representing a well-diversified portfolio of US equities, its longstanding use as a common investment benchmark, along with its precedent in the finance literature, make it a fine choice for this very purpose.

For this data set, the test period (ensemble of all evaluation periods/the concatenation of all chronological Bs) consists of its last 21 years.

Min	Q1	Median	Q3	Max
−0.116	−0.577%	0.066%	0.452%	0.120
N. Obs	Mean	Std	Skew (g_1)	Ex. Kurt (g_2)
5287	0.031%	1.24%	0.177	11.2

Table 1: Descriptive statistics for S&P500 daily losses (negative returns) across the test period of the data set.

3.1.2 Bank Data

To cross-check certain conclusions ascertained from the initial testing performed on the S&P500 data (namely whether results on said stock index that is prevalent in this area of research may reasonably be extended to a more practical context), realized (actual) gains/losses together with their hypothetical counterparts (Figure 4) relating to trading book positions of a large Danish bank (Danske Bank) were made use of. This bank is given the plain denomination ’Bank’ in the methodology portion of this essay, where it acts as exemplar (steps described are analogous for other banks considered later down the line). For said bank, raw data points were made publicly available in its (Basel) Pillar 3 disclosure reports under the EU MR4 rubric, enabling the retrieval of precise values spanning the most recent five years (2018 through 2022) and the portrayal of a real-world financial institution’s exposure to market risk factors.

Note that for the Bank data, VaR/ES measures derived later in this section are estimated off of the hypothetical (sometimes referred to as ’clean’/’net’) P/L values (were assets held until the close of day without trading adjustments). Actual (sometimes referred to as ’dirty’/’gross’) P/L is thus not required for the non-test (beginning) period.

Importantly, seeing as asset managers have the ability to mitigate losses and accentuate profits through intervention, hypothetical losses present the critical case as far as backtesting is concerned. Comparing actual and hypothetical losses in the test period is enough to showcase this generalization, where one notices that $l^{hyp} > l^{act}$ in 544 of 764 instances. Therefore, despite both actual and hypothetical P/L values needing to be used to validate a VaR/ES model in practice through backtesting, ES backtesting in this essay is limited to the hypothetical (without intervention) set of values, with underestimation (as opposed to overestimation) of lower (as opposed to upper) VaR/ES being the primary legislative concern.

Min	Q1	Median	Q3	Max
−322	−19.3	2.94	23.8	444
N. Obs	Mean	Std	Skew (g_1)	Ex. Kurt (g_2)
764	4.71	49.6	1.33	17.0

Table 2: Descriptive statistics for Bank daily hypothetical losses (negative profits, in M DKK) across the test period of the data set.

The lessened data availability for this type of information has induced our analysis to be restricted to 2019 onward, effectively placing a particular emphasis on the COVID-19 recession.

The test period for this data set thus consists of the last 3 years.

3.1.3 Other Banks

In addition to Danske Bank A/S (DK), Nykredit A/S (DK) and Commerzbank AG (DE) were also studied. These financial institutions all place among the top 50 largest European banks by total assets as of April 14^{th} 2023: rankings 24, 39 and 25 respectively (Mones and Taqi (2023)). Granted handpicked, this subset may be viewed as a random sample from said larger category of 50 banks, as no predispositions were tangled up in what was an extensive search process. Rather, factors pertaining to data availability, including:

- banks not using the standardized approach;
- lack of missing or corrupt data;

and presentation quality, not limited to:

- graph type (scatter/bar plots are favored over line plots), resolution (higher is preferable) and degree of clutter (lower is preferable);
- scale fragmentation (higher is preferable) and compression (lower is preferable);

dictated considerations.

As should come as no surprise given the mildly lax disclosure requirements constituting the EU MR4, the task of getting ahold of worthy P/L data is a tedious one, demanding - on top of making certain the aforementioned prerequisites are met - painstaking work be done to ensure data published in graphical form is accurately extracted.

3.2 Non-parametric Estimation of ES

To estimate value at risk and expected shortfall for the data at hand, three non-parametric approaches were initially employed, the first of which being **basic historical simulation** (henceforth BHS). For such methods, estimates are derived from empirical loss distributions obtained through samples of historical losses on an iterative basis. More precisely, a one-period-ahead (periods here corresponding to trading days) forecast is obtained for the coming period t given the n portfolio losses in successive preceding periods, this for each period in the test period. Put differently, the ensemble of *observed losses* in periods $t-i$:

$$l_{t-i}, \quad i = 1, ..., n$$

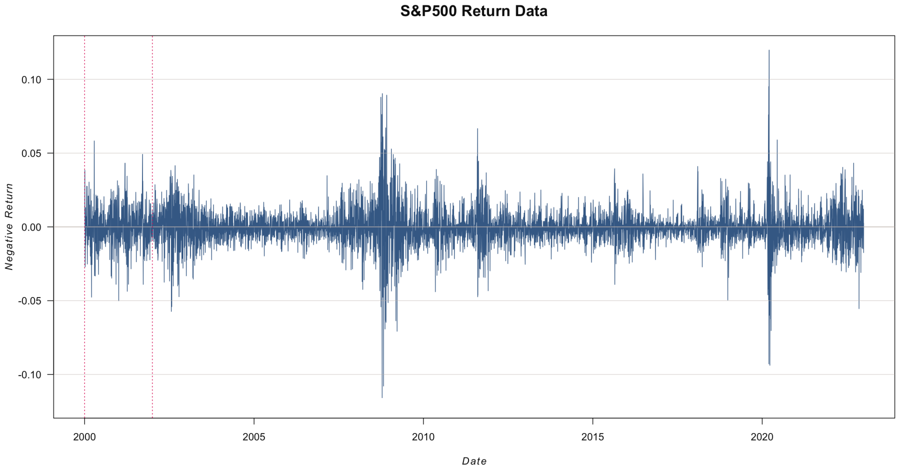

Figure 3: Time series of 5787 observations of the Standard and Poor's 500 index (SPX) daily returns, for trading days from 3^{rd} January 2000 to 30^{th} December 2022. Values are closing values. The period between red dotted lines constitutes the first training set used for estimation.

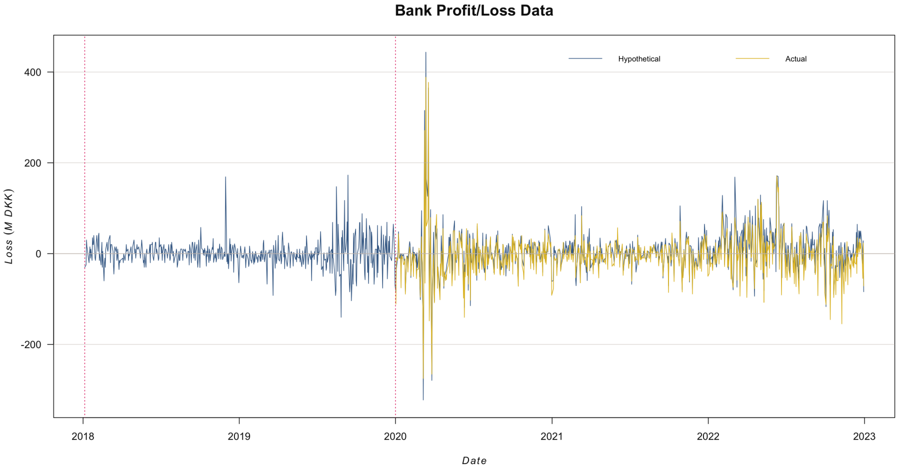

Figure 4: Time series of 1264 observations of a large Danish bank's hypothetical daily profit and losses (764 actual - provided for illustrative purposes), for trading days from 5^{th} January 2018 to 30^{th} December 2022. The period between red dotted lines constitutes the first training set used for estimation.

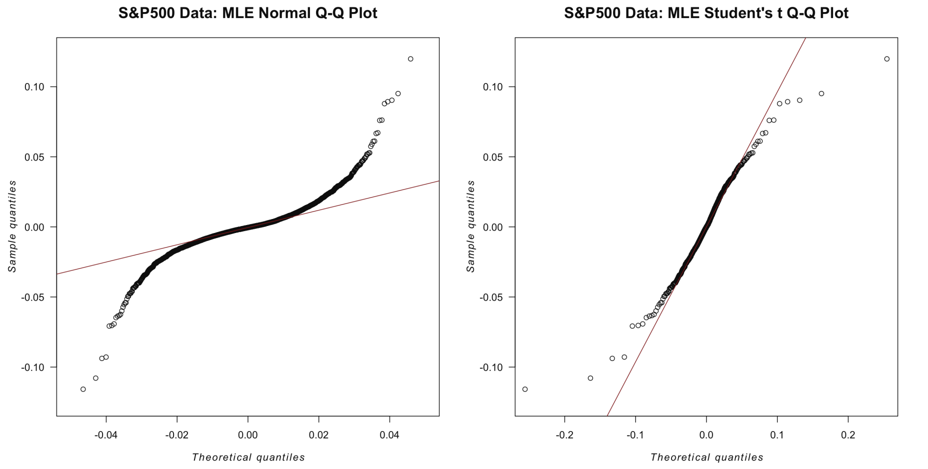

Figure 5: QQ plots for S&P500 daily losses in the test period (5287 observations, 2002-2023) against fitted Normal (left chart)/Student's t (right chart) distributions. KS test p-value: $1.55 \cdot 10^{-46}$ (MLE Normal: H_0 rejected at $\alpha' = 0.10$); 0.099 (MLE Student's t: H_0 rejected at $\alpha' = 0.10$). *Note:* 2.45 degrees of freedom for the Student's t output from the fit (useful for Section 2.2.2).

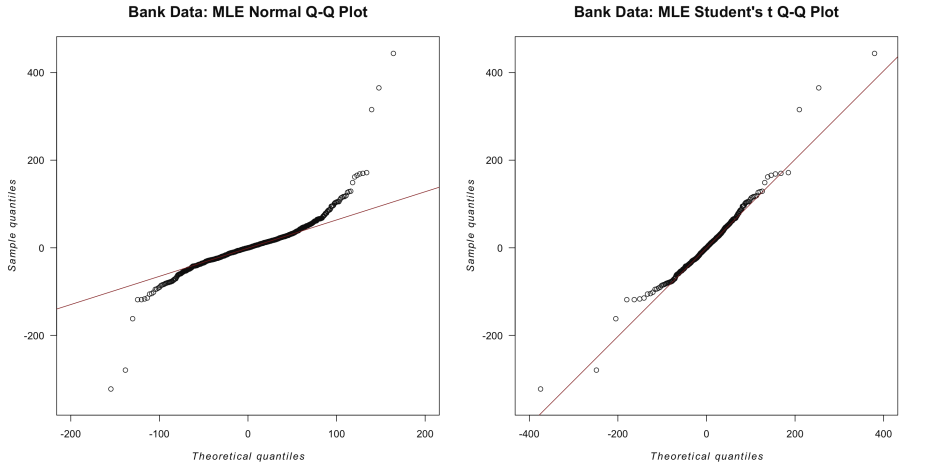

Figure 6: QQ plots for Bank daily hypothetical losses in the test period (764 observations, 2020-2023) against fitted Normal (left chart)/Student's t (right chart) distributions. KS test p-value: $4.58 \cdot 10^{-9}$ (MLE Normal: H_0 rejected at $\alpha' = 0.10$); 0.949 (MLE Student's t: H_0 not rejected at $\alpha' = 0.10$). *Note:* 2.75 degrees of freedom for the Student's t output from the fit (useful for Section 2.2.2).

Graph lines run through 25th and 75th theoretical-sample percentile pairs.

constitutes the available conditioning information up to time $t-1$ (end of period $t-1$) that is relied upon to forecast both $VaR_{\alpha,t}$ and $ES_{\alpha,t}$ in period t; this procedure is repeated $\forall t \in [2002, 2022]$ for S&P500 data ($\forall t \in [2020, 2022]$ for Bank data): the *test period*. Hence, predictions under this specification call for the implementation of a rolling window for *training sets* (in-sample data sets) of size n. Consequently, forecasting and backtesting are not executed for the period coinciding with the first rolling window.

For the S&P500 data, n was set as 500, corresponding to the number of trading days in 2000 and 2001 combined. Regarding Bank data, n so too was fixed at 500 for consistency's sake. However due to the bank's limited divulging pre-2020, data points prior to Q4 2019 (for which exact values - which generally are seldom made available - were not published) were estimated by visual inspection to the nearest 5M (concerns the first 185 observations in the data set).

Additionally, in-line with the latest regulatory changes imposed under the soon-to-be-enacted FRTB, an α of 0.975 was chosen for running all experiments.

3.2.1 BHS

To conduct BHS, each sample of portfolio losses (training set) is ordered by magnitude in descending fashion to get *sorted losses*

$$l_j^s, \quad j = 1, ..., n$$

$$l_1^s \geqslant ... \geqslant l_n^s$$

with l_j^s denoting the j^{th} largest observed loss. From this resulting discrete distribution, VaR is consequently:

$$\boxed{VaR_{\alpha,t} = l^s_{\lfloor (1-\alpha)\cdot n+1 \rfloor}} \tag{18}$$

as the proportion of observations larger than this one is $1-\alpha$. The floor function applied to the index in the aforementioned formula ensures the inequality in Equation 2 is adhered to when the index is not an integer.

$ES_{\alpha,t}$ is finally calculated based on Equation 11, underpinned by the implicit assumption that the observations in question are a realization of a random sample from some underlying continuous distribution.

3.2.2 AWHS

The second method - **age-weighted historical simulation** (henceforth AWHS) - differs from the first in that it takes into consideration how recent historical losses are in relation to the estimation period t, attributing them distinctive weights, starting with the latest one, which is given weight

$$w_{t-1} = \frac{1-\lambda}{1-\lambda^n}, \quad 0 < \lambda < 1 \tag{19}$$

and decreasing exponentially with age beyond that such that in general we have:

$$w_{t-i} = \lambda^{i-1} \cdot w_{t-1}, \quad i = 1, ..., n \tag{20}$$

with λ the decay factor. It is worth noting that w_{t-1}^{-1} is by construction equal to the n^{th} partial sum of the geometric series specified by coefficient $a = 1$ and common ratio $r = \lambda$, i.e. $w_{t-1}^{-1} = s_n = \sum_{i=1}^{n} \lambda^{i-1}$ (valid provided $r \neq 1$). The reasoning for such a choice is that summing over all weights in the sample yields $w_{t-1} \cdot s_n$, which reduces to 1 under this parametrization and enables the interpretation of said weights as probabilities. It is thus clear that this method places less likelihood on observations having occurred longer ago.

Only once weights have been tied to their respective losses is the sample sorted. VaR stemming from AWHS is then:

$$\boxed{VaR_{\alpha,t} = l_k^s, \quad k = \min\left(j : \sum_{j<k} w_j^s \leqslant 1 - \alpha\right)} \tag{21}$$

the sorted weights evidently also being arranged according to loss magnitude. Contrary to BHS, weights in AWHS are no longer equal to one another and of value $\frac{1}{n}$.

$ES_{\alpha,t}$ is subsequently also obtained using Equation 11.

3.2.3 VWHS

The ultimate non-parametric method applied was **volatility-weighted historical simulation** (henceforth VWHS), whereby historical losses are re-scaled prior to being sorted to account for volatility clustering in financial markets, an empirical phenomena affecting asset returns which has been documented extensively in the finance literature and has led to the advent of time-varying volatility models such as the generalized auto-regressive conditional heteroskedasticity model (dealt with in Section 3.3.1).

Ding and Clive (1996) compute the sample auto-correlation function $\hat{\rho}_k$ pertaining to the time series of S&P500 daily returns from 03/01/1928 to 30/08/1991 at different lags k, and find that return magnitudes appear to be persistent, in-line with swings between periods with varying degrees of financial turmoil. As a matter of fact, they report $k = 2705$ as the first lag at which $\hat{\rho}_k < 0$ for $|r_t|$ (a similar observation is noted for r_t^2), while for r_t this number is already $k = 2$. The SACF for $|r_t|$ is also characterized by the authors as not exhibiting exponential decay insofar as statistical significance is concerned.

Given what precedes, the logic behind this approach is to adjust loss observations in the training set in such a manner that their sizes become representative of the estimation period t's volatility, and hence reflect current market conditions. A way one could contemplate this is as though the observed losses were to have occurred during the (future) period succeeding the last observation. *Re-scaled losses* are in this context defined as:

$$l_{t-i}^r = \frac{\sigma_t}{\sigma_{t-i}} \cdot l_{t-i}, \quad i = 1, ..., n.$$

Once the re-scaling completed, the remaining procedure follows that used in BHS, beginning with the sorting process to end up with the $l_j^{r,s}$s.

3.2.4 EWMA Conditional Volatility Model

In order to implement VWHS in practice, a volatility model is needed to output volatilities for each period in the loss sample and forecast next period's volatility. For this essay, the **exponentially-weighted moving average** (henceforth EWMA) model was chosen for this purpose.

As for the initial setup, a constant mean model was assumed:

$$l_t = \mu + \epsilon_t, \quad t \in \mathbb{Z} \tag{22}$$

with μ estimated as $\overline{x}_l$, the sample mean of the observed losses.

In the same vein as AWHS, the EWMA model makes use of the exponentially decreasing weights in Equation 20, here to improve upon simple historical variance by defining the one-period-ahead conditional variance as the age-weighted average of the past n squared deviations from the mean:

$$\sigma_t^2 = \sum_{i=1}^{n} w_{t-i}\epsilon_{t-i}^2 \tag{23}$$

which can be shown to be approximated as:

$$\sigma_t^2 \approx (1-\lambda)\cdot\epsilon_{t-1}^2 + \lambda\cdot\sigma_{t-1}^2 \tag{24}$$

provided $n \gg 1$:

$$\begin{aligned}
&\sigma_t^2 - \lambda\cdot\sigma_{t-1}^2 = \\
&\frac{1-\lambda}{1-\lambda^n}\left[\epsilon_{t-1}^2 + \lambda\cdot\epsilon_{t-2}^2 + ... + \lambda^{n-1}\cdot\epsilon_{t-n}^2 - (\lambda\cdot\epsilon_{t-2}^2 + \lambda^2\cdot\epsilon_{t-3}^2 + ... + \lambda^n\cdot\epsilon_{t-n-1}^2)\right] \\
&\qquad = \frac{1-\lambda}{1-\lambda^n}\left[\epsilon_{t-1}^2 - \lambda^n\cdot\epsilon_{t-n-1}^2\right] \approx (1-\lambda)\cdot\epsilon_{t-1}^2.
\end{aligned}$$

Equation 24 is used recursively to arrive at $\sigma_{t-n}^2, ..., \sigma_t^2$ successively, with $\epsilon_{t-n-1} = 0$ and $\sigma_{t-n-1}^2 = s_l^2$ (sample variance of the observed losses) having been selected to begin the recursion. Said recursion was started anew for every forecast in the test period (performed on the same rolling window used for each VaR/ES forecast pair), meaning μ and σ_{t-n-1}^2 were repeatedly updated.

Equation 23 is the exponentially-weighted variance estimator by virtue of the error variance in Equation 22 being equivalent to the loss variance, and $\overline{\epsilon}_{t-i} = 0$ by construction.

Based on RiskMetrics (Longerstaey & Spencer (1996)), a value of $\lambda \approx 0.9908$ was chosen in this essay ((see Appendix 7.1)).

The ES estimates derived from the three methods are presented graphically on the following page in Figure 7.

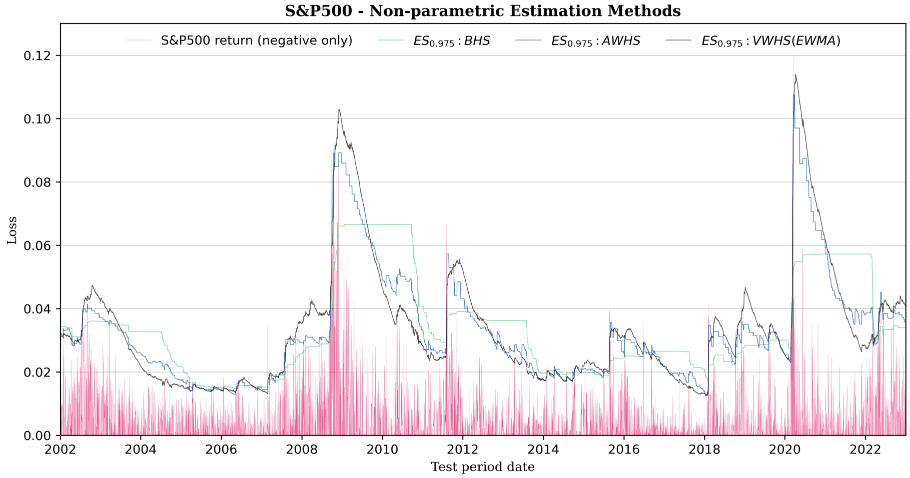

Figure 7: **FRTB.** S&P500 negative returns, *97.5% expected shortfall* one-period-ahead forecasts spanning over the test period for non-parametric estimation methods - based on a rolling window comprising of 500 trading days.

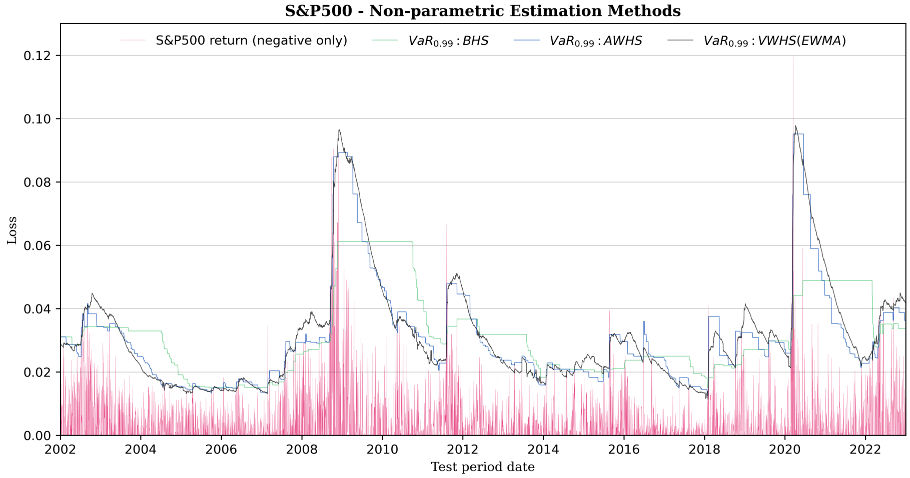

Figure 8: **Basel 2.5.** S&P500 negative returns, *99% value at risk* one-period-ahead forecasts spanning over the test period for non-parametric estimation methods - based on a rolling window comprising of 500 trading days. Figure for comparative purposes.

3.2.5 VIX Conditional Volatility Model

Implied volatility models may of course be used in lieu of historical ones such as EWMA, and the (inherently latent) volatility backed out from the Black-Scholes option pricing model by setting a given option's current market price as the Black-Scholes price. An example of one such relative measure that is commonly used would be the value of the CBOE Volatility Index (VIX), serving as a proxy for stock market participants' anticipation of volatility based on S&P500 index options.

To implement the model in a manner akin to that proposed in the context of VWHS by Nossman and Vilhelmsson (2014), we define (estimation period) volatility forecasts for re-scaling losses in a training set as

$$\sigma_t = VIX_{t-1}^{Close} \quad t \in [2002, 2022]$$

and past volatilities as

$$\sigma_{t-i} = VIX_{t-i}^{Open} \quad t \in [2002, 2022],\ i = 1, ..., n.$$

More precisely, some calculation (e.g. assuming independence so as not to have to deal with any covariance terms) to get from Index value to annualized volatility to daily volatility such as

$$\frac{VIX/100}{\sqrt{250}}$$

would constitute a given σ; needless to say these computations are rendered trivial considering they cancel out in the re-scaling ratio that is applied to losses in VWHS.

Opening values are thus required from the first trading day of 2000 to the penultimate trading day of 2022, while closing values (Figure 9) are needed from the ultimate trading day of 2001 to the penultimate trading day of 2022.

In 32 instances for the Bank data (0 for the S&P500 data) did a Bank trading day t not have a VIX counterpart. In these limited cases, the last available VIX^{Close} was used for both VIX_t^{Open} and VIX_t^{Close}.

ES estimates generated with this alternative to an EWMA model are illustrated with Figure 10.

3.3 Parametric Estimation of ES

To complement the non-parametric methods, a selection of parametric ones are elaborated upon in what follows, all of which are made sensitive to volatility clustering by infusing them with a Gaussian-GARCH(1,1) model, the summary of which we preface the methods with.

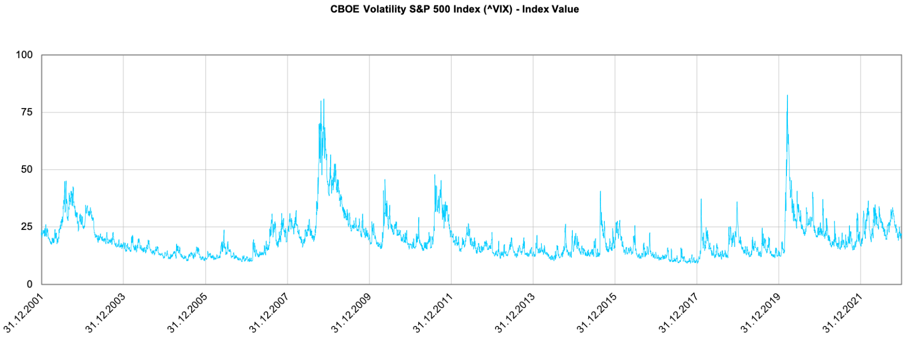

Figure 9: The Chicago Board Option Exchange's CBOE Volatility Index (VIX) closing values used as next-day forecasts, 31^{st} December 2001 - 29^{th} December 2022. *Data source: Yahoo Finance.*

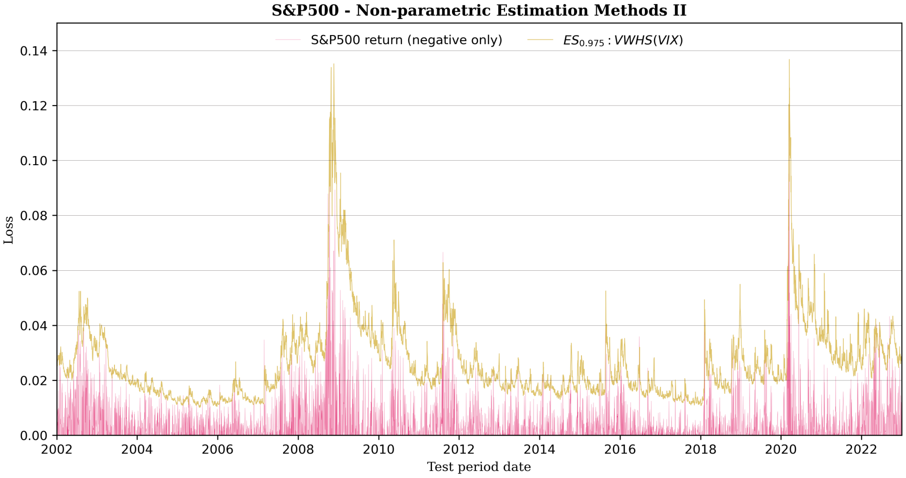

Figure 10: **FRTB.** S&P500 negative returns, *97.5% expected shortfall* one-period-ahead forecasts spanning over the test period for the VIX-VWHS estimation method - based on a rolling window comprising of 500 trading days.

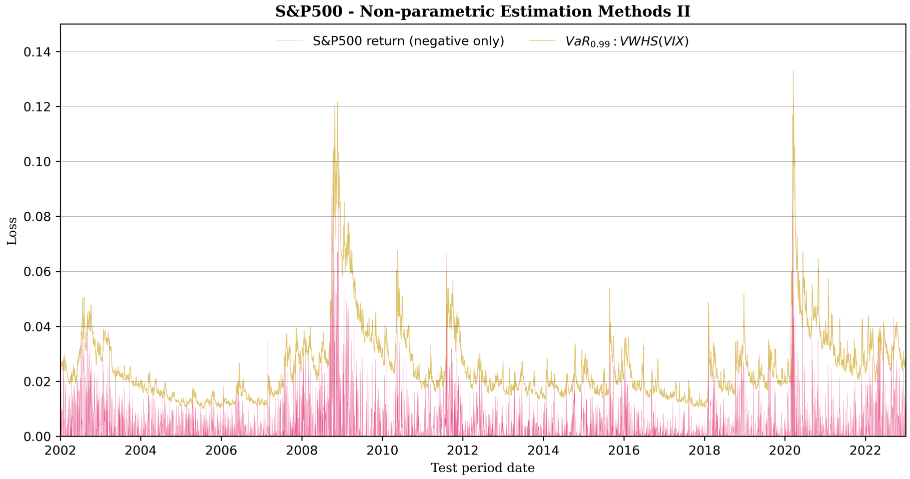

Figure 11: **Basel 2.5.** S&P500 negative returns, *99% value at risk* one-period-ahead forecasts spanning over the test period for the VIX-VWHS estimation method - based on a rolling window comprising of 500 trading days. Figure for comparative purposes.

3.3.1 Gaussian-GARCH(1,1) Conditional Volatility Model

Before covering the parametric methods that were undertaken, we briefly circle back to the **generalized auto-regressive conditional heteroskedasticity** (henceforth GARCH) model, which was used to refine certain methods in order to capture the volatility clustering effect touched upon in Section 3.2.3. This model of changing volatility, put forth by Bollerslev (1986) as a natural extension of Engle's 1982 ARCH model, is utilized in this essay in its uni-variate constant mean form, where the linear regression model for losses is (note that GARCH is estimated by maximum likelihood):

$$y_t = \mu + \epsilon_t, \quad t \in \mathbb{Z} \tag{25}$$

with a GARCH(1,1) process characterizing Gaussian innovations:

$$\begin{gathered} \epsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2) \\ \epsilon_t = \sigma_t \cdot \eta_t, \quad \eta_t \overset{iid}{\sim} N(0,1) \end{gathered} \tag{26}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \cdot \epsilon_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2 \tag{27}$$

with

$$s.t. \quad \alpha_0 > 0, \ \alpha_1, \beta_1 \geqslant 0 \tag{28}$$

needed to ensure positive variance and Ω_{t-1} the information set (the training sets to estimate the model in this essay). The error (or innovation) variance in GARCH is thus written as an ARMA model instead of Engle's original AR specification. Further, Bollerslev (1986) highlights GARCH's pragmatic utility: notably its less restrictive lag quantity requirement (lower orders being feasible) as well as increased preservation of historical information through its MA component.

The conditional variance is found as:

$$\sigma_t^2 = var(\epsilon_t|\Omega_{t-1}) \equiv \mathbb{E}_{t-1}(\epsilon_t^2) - \mathbb{E}_{t-1}(\epsilon_t)^2 = \mathbb{E}_{t-1}(\epsilon_t^2) \tag{29}$$

and exploiting the expectation operator's distributive property given that all components of σ_t^2 are known at time $t-1$, and the resulting expectation of the stochastic part $\mathbb{E}(\eta_t^2)$ dropping out by reducing to 1.

Substituting the unconditional expectation in place of its conditional counterpart in Equation 29, one arrives at:

$$\sigma^2 = \mathbb{E}(\epsilon_t^2) \equiv cov(\sigma_t^2, \eta_t^2) + \mathbb{E}(\sigma_t^2)\mathbb{E}(\eta_t^2) = \mathbb{E}(\sigma_t^2) \tag{30}$$

which if performed on Equation 27 yields the process' unconditional (∀ lags) variance

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}. \tag{31}$$

As Bollerslev (1986) states, making certain coefficients

$$s.t. \quad \alpha_1 + \beta_1 < 1 \tag{32}$$

suffices in guaranteeing ϵ_t be stationary with $\mathbb{E}(\epsilon_t), cov(\epsilon_t, \epsilon_s) = 0 < \infty, \ \forall t \neq s$. Equation 31 can be seen tending to infinity as this sum approaches 1, violating the necessary finite constant variance condition.

Analogously to the manner in which the EWMA model was conducted in Section 3.2.4, GARCH parameters were updated daily in rolling window-style estimation to obtain one-period-ahead analytical forecasts of error variances (in the constant mean model equating to return variances) for the entirety of the test period. The different $\epsilon_{t-n-1}, \sigma_{t-n-1}^2$ too were set as they had been before to start the manual recursions of Equation 27.

In 405 of 5287 instances in the test period for the S&P500 data (0 of 764 instances for the Bank data), the model fit output did not respect Equation 28 and/or Equation 32. In these cases, said fit was discarded, and the variance forecast made into a j-day-ahead one on the basis of the last available satisfactory fit from j training sets ago (with a maximum j of 28 recorded):

$$\mathbb{E}_{t-1}(\sigma_{t+j}^2) = (\alpha_1 + \beta_1)^j(\sigma_t^2 - \sigma^2) + \sigma^2 \xrightarrow[j\to\infty]{} \sigma^2 \quad j \in \mathbb{Z}_0^+ \tag{33}$$

which when $j = 0$ is simply the one-period-ahead forecast:

$$\mathbb{E}_{t-1}(\sigma_t^2) = \sigma_t^2. \tag{34}$$

Equation 33 can be derived by recognizing that $\mathbb{E}_{t-1}(\sigma^2_{t+j-1}) = \mathbb{E}_{t-1}(\epsilon^2_{t+j-1})$ (from Equation 30) and iterating backward the resulting $f(\mathbb{E}_{t-1}(\sigma^2_{t+j-1}))$ until one is left with $f(\mathbb{E}_{t-1}(\sigma^2_t))$ after that:

$$\mathbb{E}_{t-1}(\sigma^2_{t+j}) = \alpha_0 + (\alpha_1 + \beta_1)\mathbb{E}_{t-1}(\sigma^2_{t+j-1}) = ...$$
$$= \alpha_0 \sum_{i=1}^{j} (\alpha_1 + \beta_1)^{i-1} + (\alpha_1 + \beta_1)^j \sigma^2_t = \alpha_0 \cdot \frac{1 - (\alpha_1 + \beta_1)^j}{1 - (\alpha_1 + \beta_1)} + (\alpha_1 + \beta_1)^j \sigma^2_t$$

before spotting Equation 30 and rearranging. As mentioned when applying AWHS in Section 3.2.2, the final transition above re-expresses the j^{th} partial sum of the geometric series specified by $a = 1$ and $r = \alpha_1 + \beta_1 < 1$.

Variance forecast outputs from both the GARCH and EWMA model for S&P500 data (only) are displayed in Figure 12.

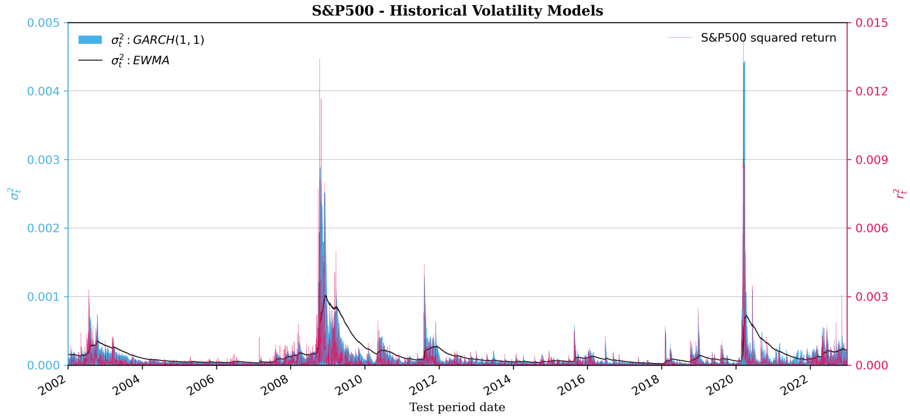

Figure 12: **S&P500.** Squared returns (provided for appreciation of return magnitude - irrespective of sign - over time), (one)-period-ahead variance forecasts spanning over the test period for GARCH(1,1) (Gaussian-η_t model) and EWMA models - based on a rolling window comprising of 500 trading days.

To close the discussion on GARCH, we overview the estimation procedure for it, namely **maximum likelihood estimation** (henceforth MLE), which optimizes distributional parameters maximizing the probability of a set of observations $x_i,\ i = 1, ..., n$ having been what they were. The so-called likelihood function (objective function) is given as

$$L(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \tag{35}$$

with f the p.d.f. (p.m.f. if discrete) of the underlying random variable X and $\boldsymbol{\theta}$ its vector of parameters. In practice, the log-likelihood function is of interest:

$$l(\boldsymbol{x}; \boldsymbol{\theta}) = \ln L(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta}) \tag{36}$$

this due to the natural logarithm's product, quotient and power properties (while the point at which the optimum is attained remains unaltered due to ln being strictly increasing). $\hat{\boldsymbol{\theta}}$ is thus obtained numerically by

$$\max_{\boldsymbol{\theta}} l(\boldsymbol{x};\boldsymbol{\theta}). \tag{37}$$

Given the error ϵ_t (here demeaned dependant variable) specification in Equation 26 (or equivalently $y_t|\Omega_{t-1} \sim N(\mu, \sigma_t^2)$), we have that

$$l(\boldsymbol{\epsilon};\boldsymbol{\theta}) = \sum_{t=1}^{n} \ln{(2\pi\sigma_t^2)^{-0.5}} \exp\left(-\frac{\epsilon_t^2}{2\sigma_t^2}\right) = -0.5\sum_{t=1}^{n} \ln 2\pi\sigma_t^2 - \sum_{t=1}^{n} \frac{\epsilon_t^2}{2\sigma_t^2}. \tag{38}$$

$$\boldsymbol{\theta} = (\mu, \alpha_0, \alpha_1, \beta_1)', \quad \epsilon_t = f(\mu),\ \sigma_t^2 = f(\alpha_0, \alpha_1, \beta_1) \quad \forall t.$$

Importantly, $\epsilon_t = \sigma_t \cdot \eta_t$ (Equation 26) by definition in the model, with the η_t (here standardized dependant variable) being IID. The entire GARCH procedure is repeated from scratch for each training set to get the one-day-ahead volatilities (standard deviations) $\hat{\sigma}_t,\ \forall t \in [2002, 2022]$.

η_t being the standardized dependant variable is rooted in the following property of a Normal random variable's $X \sim N(\mu, \sigma^2)$ linear transformation ((see Appendix 7.2)):

$$Y = a + bX \sim N(a + b \cdot \mu, b^2 \cdot \sigma^2).$$

Finally, it is worth noting that the forecasted mean for each training set is effectively $\hat{\mu}$ since constant mean is presumed (although this estimate is updated with each new fit, hence the $\hat{\mu}_t$ notation in VaR/ES formulae below).

3.3.2 GARCH-N, GARCH-ST Models

Outputs from the GARCH (with Gaussian standardized errors: N(0,1)) volatility model and the associated mean model it is nested in produced above for each training set served as scale and location parameters respectively to enhance ordinary VaR/ES estimates from both the **Normal** distribution (shorthand notation N(μ,σ^2)) and the **Student's t**-distribution (shorthand notation ST(μ,σ^{*2},ν)), the latter despite being symmetric like the former allowing for increased tail heaviness through its degrees of freedom parameter ν, undoubtedly more apt here given sample excess kurtosis (ex-post) results in Table 1/Table 2.

For the Normal distribution, from Equation 5 we have:

$$\boxed{V\hat{a}R_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t \cdot \Phi^{-1}(\alpha)} \tag{39}$$

with $\Phi^{-1}(\alpha)$ denoting z_α, the α-quantile of the standardized N. Developing Equation 10, ES is found to be (through integration by substitution):

$$\boxed{\hat{ES}_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}} \tag{40}$$

with ϕ the p.d.f. of N(0,1).

Similarly, for the Student's t-distribution:

$$\boxed{\hat{VaR}_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t^* \cdot t_{\alpha,\hat{\nu}}} \tag{41}$$

$$\boxed{\hat{ES}_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t^* \frac{f^*_{stdST}(t_{\alpha,\hat{\nu}})}{1-\alpha} \frac{\hat{\nu} + t^2_{\alpha,\hat{\nu}}}{\hat{\nu}-1}} \tag{42}$$

with

$$\hat{\sigma}_t^* = \sqrt{\frac{\hat{\nu}-2}{\hat{\nu}}} \cdot \hat{\sigma}_t.$$

This follows from the fact that the standardized ST has variance $\frac{\nu}{\nu-2}$ (for $\nu > 2$) and not 1.

ν (not given by a Gaussian-GARCH model) was estimated as:

$$\hat{\nu} = \frac{6}{x} + 4, \quad x = \max(k_{exc}, 10^{-4})$$

with $k_{exc} = g_2$ the training set's sample excess kurtosis. A minimum threshold for excess kurtosis (here a very small arbitrary number) was thus imposed on this measure for all training sets to ensure a positive degrees of freedom (here > 4), and in-line with the stylized empirical fact of returns tending to be leptokurtic. Such a constraint does not prove to be too stringent for the S&P500 data, where realized $k_{exc} \in [-0.21, 23.02]$ (211 of 5287 non-positive values), and inconsequential for the Bank data, where $k_{exc} \in [2.06, 34.59]$.

Had a GARCH model with Student's t standardized errors (ST$(0, 1, \nu)$) been implemented instead, this supplementary parameter could have been directly estimated along with σ^*, however this alternative was not deemed suitable (and was consequently not used for this essay) given that for at least the S&P500 data, the number of training sets where the fit did not satisfy stationarity requirements increased, approximately doubling vis-à-vis the Gaussian version (total: 405 to 812, longest streak: 28 to 52).

3.3.3 GARCH-GPD Model (POT)

To contrast previous models, the **Generalized Pareto** distribution (shorthand notation GPD(ξ,β)), used in the context of the **peaks-over-threshold** (henceforth POT) method, was also opted for, not least for its polished reputation in the field. POT does not require making an explicit assumption as to which distribution losses $L \sim F$ follow, nor is it overly data-intensive, and holds for a plethora of underlying distributions including the aforementioned N and ST (McNeil and Frey, 2000). Rather, it hinges on the unspecified distribution's extreme values through the conditional excess distribution function,

$$\begin{gathered} F_u(y) \equiv \mathbb{P}(L - u \leqslant y | L > u) = \frac{F(u+y) - F(u)}{S(u)}, \\ 0 < y \leqslant \sup dom(f) - u \end{gathered} \tag{43}$$

which characterizes losses exceeding some high threshold (or l quantile) $u \gg \bar{l}$ (S being the survival function of F, i.e. $S(x) \equiv 1 - F(x) \equiv \mathbb{P}(L > x)$). The last equality above follows from the conditional probability equating to the probability of the intersection of the two events over the conditioning event's likelihood. For this essay, u was set to the 95^{th} percentile of each training set/loss sample.

Leaning on the Pickands-Balkema-De Haan theorem, $F_u(y)$ is approximated by the GPD's c.d.f. with the same evaluation, said distribution commonly being used in modeling an underlying distribution's right-tail:

$$F_u(y) \approx G_{\xi,\beta}(y) = 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-\frac{1}{\xi}}. \tag{44}$$

$F(u)$ is estimated from the random sample at hand as $1 - \frac{N_u}{N}$, the proportion of observed losses below (not exceeding) the threshold, with $N_u = \sum_{i=1}^{n} \mathbb{1}_{l_{t-i} > u}$ and $N = n$.

The p.d.f. $g_{\xi,\beta} = \frac{\mathrm{d}}{\mathrm{d}l} G_{\xi,\beta}$ is subjected to MLE (Equation 37) to get parameter estimates $\hat{\xi}, \hat{\beta}$, but critically only over the N_u large loss observations and not N since the approximation in Equation 44 pertains to F_u and not F. Along with excesses over the threshold being $y = l - u$, we now effectively have after rewriting Equation 43:

$$G_{\hat{\xi},\hat{\beta}}(l - u) = \frac{\hat{F}(l) - 1 + \frac{N_u}{N}}{\frac{N_u}{N}}. \tag{45}$$

From Equation 4, it is clear that $F(l) \equiv \mathbb{P}(L \leqslant l) = \alpha$ when $l = VaR_\alpha$. Thus, rearranging Equation 45 under these conditions yields the VaR estimator:

$$\boxed{V\hat{a}R_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{N}{N_u}(1 - \alpha)\right)^{-\hat{\xi}} - 1\right).} \tag{46}$$

Applying Equation 12, the corresponding ES is (see Appendix 7.3):

$$\boxed{\hat{ES}_\alpha = \frac{V\hat{a}R_\alpha + \hat{\beta} - \hat{\xi} \cdot u}{1 - \hat{\xi}}.} \tag{47}$$

To arrive at dynamic VaR/ES estimates, the procedure proposed by McNeil and Frey (2000) within the conditional EVT framework was implemented.

What this entails in this work is taking the Gaussian-GARCH(1,1) model from Section 3.3.1 ('Gaussian' signifying that the IID standardized error η term follows a standard N) which was used to generate unconditional mean and conditional variance forecasts, and assuming the (white noise) standardized residuals themselves to be GPD-distributed beyond some u. From Equation 25 and

Equation 26 we can express the standardized residuals after the fit as

$$\hat{\eta}_{t-i} = \frac{y_{t-i} - \hat{\mu}}{\hat{\sigma}_{t-i}}, \quad i = 1, ..., n \tag{48}$$

with y as usual the portfolio loss. The final step involves re-scaling Equation 46 and Equation 47 (each of which being applied to the standardized residuals) as follows:

$$\boxed{V\hat{a}R_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t \cdot V\hat{a}R_{\alpha}^{GPD}(\hat{\eta})} \tag{49}$$

$$\boxed{\hat{ES}_{\alpha,t} = \hat{\mu}_t + \hat{\sigma}_t \cdot \hat{ES}_{\alpha}^{GPD}(\hat{\eta}).} \tag{50}$$

As a closing remark, the ξ parameter caused some inconveniences when it came to the denominator of Equation 47, this in 4 instances bunched in the first half of March 2022 for the Bank data (0 instances for the S&P500 data). In two cases, parameter values slightly greater than 1 resulted in very large (not to mention negative) ES estimates, while in the two others values slightly less than 1 led to aberrant spikes in ES.

To remedy this minor issue, any $\hat{\xi} \geqslant 0.8$ was therefore ignored, and the last appropriate fit's value used in place of it.

Figure 13 depicts the evolution of dynamic ES over the course of the test period, together with previous parametric results. In general, it is noticeable that the most conservative estimates are those from the GPD, while the least so are obtained using the N.

Figures for Bank data are not provided here, but all VaR and ES estimates were inspected to ensure no computational abnormalities were inadvertently encountered.

❉

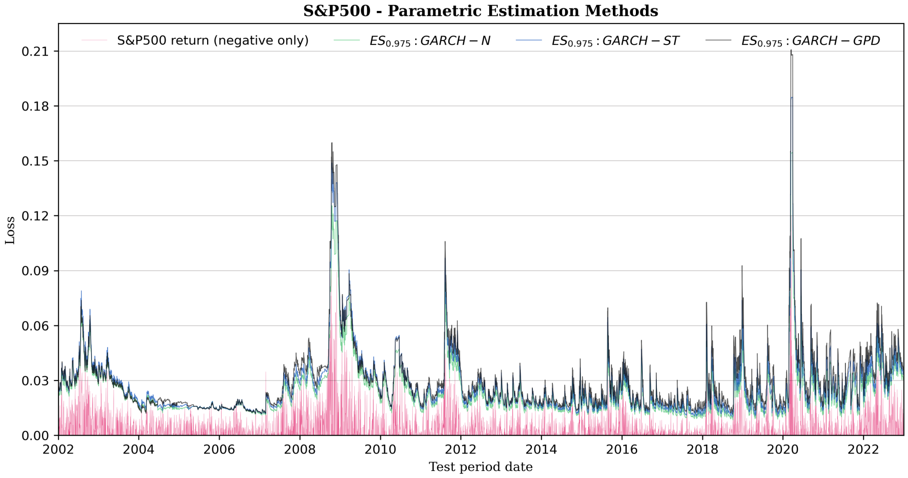

Figure 13: **FRTB.** S&P500 negative returns, *97.5% expected shortfall* one-period-ahead forecasts spanning over the test period for parametric estimation methods - based on a rolling window comprising of 500 trading days.

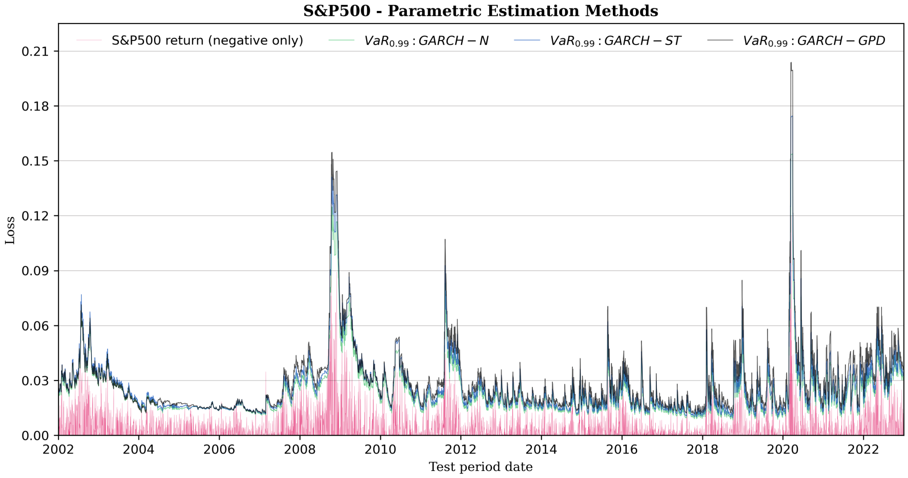

Figure 14: **Basel 2.5.** S&P500 negative returns, *99% value at risk* one-period-ahead forecasts spanning over the test period for parametric estimation methods - based on a rolling window comprising of 500 trading days. Figure for comparative purposes.

4 Results

In this section, ES backtesting results across VaR/ES estimation models (or methods) and years are presented, this for the stock index and multiple banks referenced previously.

In order to facilitate the interpretation of these backtesting results, we define a given estimation model m's relative score as follows:

$$\text{Score}_m = 0.5\left[1 - \min\left(\max\left(\frac{\bar{z}_{2,m}}{z^{\text{MC,NIG*}}_{2,0.01\%}}, 0\right), 1\right) + \left(1 - \min\left(\frac{sd(z_{2,m})}{\left|z^{\text{MC,NIG*}}_{2,0.01\%}\right|}, 1\right)\right)\right] \in [0,1]$$

placing equal emphasis on both the test statistic realizations' mean not straying too far below zero - the expectation of $Z_{2,m}$ under H_0 (fully penalized once beyond the relevant simulated critical value at the 0.01% significance level and not whatsoever if $\geqslant 0$), and low propensity for underestimation as captured by the sample's semi-deviation (fully penalized once as large as the arbitrary distance from 0 to the relevant critical value at the 0.01% level), thus incorporating both a measure of centrality and dispersion.

Needless to say, the higher the score, the better performing and the more desirable the model. Despite the critical value specified in the score formula varying between data sets here, these variations are slight. Moreover, their use in defining proportions (relative) in our estimation make scores comparable between data sets.

Starting with the S&P500 data set, Table 3 regroups the values of the Acerbi and Szekely (2014) Test 2 test statistic (cells tagged in pink do *not* relate to ES backtesting and are explained in Section 4.1). From these a key hallmark of any 'good' model revolves around avoiding very strong statistical significance (here in terms of underestimation). More specifically, this involves not ending up in the dreaded 'red zone' (as defined in the Basel traffic light test), something that would carry the most severe minimum capital requirement adjustment (imposed

	1.B	2.AW	3.E-VW	4.V-VW	5.G-N	6.G-S	7.G-GP
2002	−0.81*	−0.47	−0.45	−0.29	−0.01	0.08	0.05
2003	0.71	0.84	0.86	0.35	0.27	0.32	0.28
2004	1.00	0.66	0.29	0.10	−0.29	−0.17	0.07
2005	0.52	0.09	0.05	−0.52	0.22	0.24	0.40
2006	−0.01	−0.05	−0.18	−0.25	−0.07	−0.05	−0.06
2007	−2.38****	−1.39**	−1.50**	−1.40**	−2.29****	−1.95***	−1.61***
2008	−4.24****	−1.87***	−0.86*	−0.99*	−1.36**	−1.11*	−0.67
2009	0.76	1.00	1.00	1.00	−0.35	−0.05	0.42
2010	1.00	0.54	0.31	0.45	−1.03*	−0.75	0.22
2011	−0.77*	−0.61	−0.53	−0.85*	−1.42**	−1.02*	−0.12
2012	1.00	0.86	0.76	0.40	−0.19	−0.03	0.27
2013	0.89	0.47	0.55	−0.11	−0.34	−0.17	0.12
2014	−0.49	−0.45	−0.71	−0.47	−1.20**	−0.89*	0.00
2015	−0.85*	−0.56	−0.24	−0.44	−1.82***	−1.58**	−0.21
2016	0.17	0.33	0.49	0.19	−0.14	−0.02	0.44
2017	1.00	0.47	0.40	0.30	0.16	0.27	0.39
2018	−2.52****	−1.00*	−1.14**	−1.42**	−2.09****	−1.51**	−0.77*
2019	0.26	0.40	0.47	0.20	−0.62	−0.39	0.38
2020	−2.25****	−0.88*	−0.93*	−0.40	−2.10****	−1.51**	−0.51
2021	1.00	0.73	0.56	0.74	−0.75	−0.45	0.18
2022	−1.10*	−0.80*	−0.97*	−1.59**	−1.29**	−0.91*	0.12
Green	0.62	0.76	0.76	0.76	0.57	0.62	0.90
Amber	0.19	0.24	0.24	0.24	0.29	0.38	0.10
Red	0.19	0.00	0.00	0.00	0.14	0.00	0.00
$\bar{z}_2$	−0.34	−0.08	−0.08	−0.24	−0.80	−0.56	−0.03
sd(z_2)	1.80	0.93	0.77	0.72	0.93	0.79	0.73
Score	(0.47)	0.75	0.79	0.76	(0.57)	0.66	0.81

Table 3: **S&P500.** 2002-2022. ES severity backtesting. Ex-post test period MLE NIG loss fit yields, 5%, 1%, 0.1%, 0.01% - critical values (simulated Z_2 distribution quantiles): -0.75, -1.13, -1.60, -2.01 respectively (see Appendix 7.4).

Table contains Z_2 test statistic realizations across estimation methods and evaluation periods, with its sample mean and semi-deviation for BHS, AWHS, EWMS/VIX-VWHS, GARCH-N/ST/GPD. *Note concerning statistical significance:* (*),(**),(***),(****) denote p-values $p < 5\%$, $p < 1\%$, $p < 0.1\%$, $p < 0.01\%$ respectively. Invalidated models have scores shown in parentheses.

	1.B	2.AW	3.E-VW	4.V-VW	5.G-N	6.G-S	7.G-GP
2019	0.26	0.40	0.47	0.20	−0.62	−0.39	0.38
2020	−2.25****	−0.88*	−0.93*	−0.40	−2.10****	−1.51**	−0.51
2021	1.00	0.73	0.56	0.74	−0.75	−0.45	0.18
2022	−1.10*	−0.80*	−0.97*	−1.59**	−1.29**	−0.91*	0.12
Green	0.50	0.50	0.50	0.75	0.50	0.50	1.00
Amber	0.25	0.50	0.50	0.25	0.25	0.50	0.00
Red	0.25	0.00	0.00	0.00	0.25	0.00	0.00
$\bar{z}_2$	−0.52	−0.14	−0.22	−0.26	−1.19	−0.81	0.05
sd(z_2)	1.29	0.71	0.73	0.94	0.65	0.50	0.55
Score	(0.55)	0.79	0.76	0.70	(0.54)	0.67	0.86

Table 4: **S&P500.** 2019-2022. ES severity backtesting. Equivalent to Table 3 but on a restricted sample.

upscaling), while simultaneously putting the model's authorization in jeopardy (potential forced discontinuation).

Two models are found to effectively fail backtesting on the S&P500 sample: BHS (4 red zone occurrences) and GARCH-N (3 red zone occurrences), both of which grossly underestimated ES repeatedly. This is reflected in the comparatively low scores that they exhibit. As for the remaining models, those of non-parametric nature stack up relatively evenly, all superior to GARCH-ST but inferior to GARCH-GPD. However, two of said remaining models (AWHS, GARCH-ST) pose a potential liability when considering different threshold levels: this caveat is expanded upon in Section 4.1.

It is worthwhile noting that conclusions with regard to which models are deemed unfit for use and (approximate) rankings by score do not alter drastically when focusing on the most recent four (turbulent) years, as seen from Table 4.

4.1 Short Aside on VaR Severity Backtesting

The number of **97.5% VaR violations** for models in Table 3 were **1)164, 2)141, 3)142, 4)157, 5)209, 6)203, 7)132**. For comparative purposes (beyond the scope of this essay), had 99% VaR been backtested instead of 97.5% ES, as is required under current regulation for internal model approval, **99% VaR violations** would have been of interest, amounting to **1)87, 2)70, 3)68, 4)90, 5)115, 6)82, 7)59** for the same data, with too many violations reflecting systematic VaR underestimation.

To this effect, some cells in Table 3 are highlighted in magenta ($\geqslant$ 10 99% VaR exceptions recorded within year). This corresponds to the red zone, which begins at $\min(x : F(x) \geqslant 0.9999)$, $x \in \mathbb{Z}_0^+$, with F the c.d.f. of a Bin(n=250, p=0.01). These color labels demonstrates that for the S&P500 data, two additional models (AWHS, GARCH-ST) are effectively invalidated when performing a standard VaR backtest. Yet despite this glaring difference, the two backtesting approaches are largely congruent with one another, particularly when considering that Acerbi and Szekely (2014) state that using critical values $z_{2,5\%} = -0.7$ and $z_{2,0.01\%} = -1.8$ (which are more stringent than the ones identified through simulation in this essay) "would perfectly do in all occasions", owing to the apparent threshold levels' stability across distributions. Employing said more general/'safe' values as thresholds would also result in AWHS and GARCH-ST models failing the 97.5% ES backtest in 2008 and 2007 respectively. Note that GARCH-N would also fail in 2015.

Critically, one may notice that under these 'all purpose' critical values $z_{2,5\%} = -0.7$ and $z_{2,0.01\%} = -1.8$, a certain mismatch manifests itself between the Acerbi and Szekely (2014) Test 2 (on 97.5% ES) and the Basel traffic light test (on 99% VaR) for BHS (2018) and GARCH-N (2015), the latter not assigning the red zone designation to the models (less stringent). The reverse is not observed under this setting.

As a final comment, a compelling observation from the 99% VaR backtest lies in the maximum number of violations within a given year across the board being 11, ignoring one egregious exception suffered by the grossly inappropriate BHS model in 2008: 21 violations.

4.2 Results Pertaining to Trading Books

Next, results relating to individual banks are gathered after the processing of data having been performed. *A minor modification was nonetheless made:* In addition to what was asserted in Section 3.1.2 for Danske Bank, data for 2017 was also able to be collected, albeit needing to be estimated on the basis of a graph contrary to subsequent years (dates were set to those of VIX trading days for the year, seeing as the number of trading days in 2017 between the two matched - something that was not the norm for the following years), enabling the extension of the test period by one supplementary year (data farther back in time was not available). Thus, 2019 was able to be appended (Table 5).

	1.B	2.AW	3.E-VW	4.V-VW	5.G-N	6.G-S	7.G-GP
2019	-1.47**	-0.72	-0.47	-0.85*	-0.92*	-0.63	-0.77*
2020	-1.19**	-0.29	-0.08	0.50	-0.02	0.15	0.02
2021	0.82	0.39	0.19	0.43	0.06	0.22	0.11
2022	-2.04****	-0.70	-1.12**	-1.87***	-1.68***	-1.34**	-0.82*
Green	0.25	1.00	0.75	0.50	0.50	0.75	0.50
Amber	0.50	0.00	0.25	0.50	0.50	0.25	0.50
Red	0.25	0.00	0.00	0.00	0.00	0.00	0.00
$\bar{z}_2$	-0.97	-0.33	-0.37	-0.45	-0.64	-0.40	-0.37
sd(z_2)	0.69	0.38	0.53	1.05	0.76	0.68	0.43
Score	(0.58)	0.82	0.77	0.62	0.64	0.72	0.80

Table 5: **Danske Bank A/S (DK).** 2019-2022. ES severity backtesting. 5%, 1%, 0.1%, 0.01% - thresholds used: -0.74, -1.12, -1.57, -1.96 respectively.

	1.B	2.AW	3.E-VW	4.V-VW	5.G-N	6.G-S	7.G-GP
2019	0.75	0.70	0.74	0.77	0.85	0.86	0.73
2020	-2.94^{****}	-1.58^{**}	-1.53^{**}	-0.55	-1.17^{**}	-0.87^{*}	-0.95^{*}
2021	0.92	0.84	0.40	0.38	0.16	0.52	0.33
2022	-3.38^{****}	-0.96^{*}	-1.23^{**}	-3.94^{****}	0.32	0.42	0.15
Green	0.50	0.50	0.50	0.75	0.75	0.75	0.75
Amber	0.00	0.50	0.50	0.00	0.25	0.25	0.25
Red	0.50	0.00	0.00	0.25	0.00	0.00	0.00
$\bar{z}_2$	-1.16	-0.25	-0.40	-0.84	0.04	0.23	0.07
$\mathrm{sd}(z_2)$	2.01	1.07	0.99	3.11	1.21	1.11	1.01
Score	(0.24)	0.68	0.67	(0.30)	0.71	0.73	0.76

Table 6: **Nykredit A/S (DK).** 2019-2022. ES severity backtesting. 5%, 1%, 0.1%, 0.01% - thresholds used: -0.77, -1.16, -1.66, -2.08 respectively.

	1.B	2.AW	3.E-VW	4.V-VW	5.G-N	6.G-S	7.G-GP
2019	1.00	0.75	0.86	1.00	0.23	0.28	0.85
2020	-2.44^{****}	-1.30^{**}	-1.78^{***}	0.06	-1.63^{***}	-1.19^{**}	-1.19^{**}
2021	0.91	0.47	0.38	0.51	-0.12	0.18	0.15
2022	-1.09^{*}	0.21	-0.16	-1.15^{**}	0.30	0.52	0.65
Green	0.50	0.75	0.75	0.75	0.75	0.75	0.75
Amber	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Red	0.25	0.00	0.00	0.00	0.00	0.00	0.00
$\bar{z}_2$	-0.40	0.03	-0.17	0.11	-0.30	-0.05	0.12
$\mathrm{sd}(z_2)$	1.52	1.34	1.61	0.89	1.32	1.14	1.31
Score	(0.50)	0.65	0.54	0.77	0.58	0.69	0.66

Table 7: **Commerzbank AG (DE).** 2019-2022. ES severity backtesting. 5%, 1%, 0.1%, 0.01% - thresholds used: -0.74, -1.11, -1.56, -1.93 respectively.

Results for the two remaining banks are exhibited in Table 6/Table 7. Note that for Nykredit, a rolling window of size 488 (as opposed to 500) was used due to early data availability restrictions - the exponential decay factor was adjusted accordingly. In general, dates having had to be estimated were done so with software and some manual adjustments to ensure no repetitions. The accuracy of these could only affect the VIX-VWHS model, which is dependent on dates; however the extent of this impact should not be overwhelming.

For all three of these data sets, BHS is esteemed to be invalid once more, evidently not overly reactive/adaptable, swaying back and forth between underestimation and overestimation-ridden years. It also consistently posts the worst scores of the bunch. The only other model to fail is VIX-VWHS (yet paradoxically being the best-performing model on the Commerzbank data set), this on the Nykredit data set (although under the aforementioned 'all purpose' critical values it would not be satisfactory in 2022 for Danske Bank either).

Of the non-parametric models left, AWHS appears to edge out EWMA-VWHS and be quite a strong model overall. For parametric models, GARCH-GPD is generally found to be superior to GARCH-ST (first noted in the Index case, this result follows suit here, though the discrepancy between these two models is less marked here). However, contrary to the S&P500 data set, GARCH-ST tends to stand its ground score-wise against the two viable non-paramertic alternatives, perhaps not as poor as initially painted out to be. While GARCH-N holds for all bank data sets, it remains on the weaker end of the spectrum. The results for bank data can be summarized as done in Table 8:

	B	AW	E-VW	V-VW	G-N	G-S	G-GP
Avg. Score	(0.44)	0.72	0.66	(0.56)	0.64	0.71	0.74

Table 8: Average scores posted by models across bank data sets, illustrating that G-GP, G-S and AW (on these samples) are a step above the rest.

Lastly, we remark that once again, the backtest on 99% VaR looks less harsh than Test 2 (BHS in 2020 for Nykredit not classified as belonging to the red zone, likewise for VIX-VWHS in 2022 for Danske Bank if 'all purpose' critical values are used for Test 2).

❈

5 Conclusion

This essay has delved into the question of which ES estimation method is most apt at producing results in recent years that invariably do *not* underestimate 97.5% ES, thereby fully concentrating on regulatory concerns.

An initial observation stemming from this research is BHS's frequent failure to properly model risk (as characterized by not ending up in the Basel red zone), be it for the S&P500 Index or the trading books considered - a major cause for concern that calls into question its feasibility. This model should therefore not be deemed appropriate for any usage of serious practical intent, nor do we suggest it be focused on in potential future extensions to this research.

Interestingly, despite the GARCH-N model failing over the S&P500 sample much like BHS, it proves tenable for each bank data set, albeit lackluster and by and large being the weakest of valid models. The reverse is true of VIX-VWHS, which sees greater difficulty in adapting to some banks' losses (its underlying volatility model perhaps not entirely generalizable), yet still performing very well on one of them, this dichotomy making for somewhat inconclusive results.

EWMA-VWHS perceives a noticeable drop in performance when transferred from index to bank data, while the only other viable non-parametric model on both index and bank data - AWHS - does not, remaining high-scoring throughout.

It is however a parametric model that appears to be the safest alternative across the board: GARCH-GPD, although the difference between it and competing models is less marked on the bank data. As for the GARCH-ST, an increase in terms of performance and ranking is observed when applied to the bank data.

In short, generally speaking, GARCH-GPD and AWHS were found to be the most reliable models across sampled data sets for the time period that was examined (2019-2022).

Lastly, we note that results regarding the VaR backtest (based solely on the number of 99% VaR violations) suggest it is less inclined toward rejecting estimation models (and thus more forgiving) than Acerbi and Szekely's (2014) Test 2, something that is exacerbated when a stricter 'all purpose' 0.01% significance level critical value is used in Test 2 when backtesting (97.5%) ES.

One exception to this rule (for the S&P500 full sample (2002-2022)) nonetheless stands: AWHS and GARCH-ST being rejected by the VaR test but not the ES test (the ES test under the aforementioned 'all purpose' critical values however *also* rejects these models).

Yet while AWHS and GARCH-ST present a tangible concern under these conditions given their failure during the 2007-2008 financial crisis for the S&P500 data, one bad year amongst twenty-one each is nowhere near as worrying as the four suffered by both BHS and GARCH-N (Table 3). Furthermore, the severity of the backtesting 'breaches' are less pronounced for the former pair. This exception does not factor into the final analysis of this conclusion.

5.1 Limitations and Possible Extensions

To conclude this essay, we recap the principal delimitations this work was conducted under and provide some potential ideas and pointers as to how it could be expanded upon. Note that we recommend any effort to do so be focused on a subset of the following suggestions, as the implementation of any one specific extension can be time-consuming. These can be grouped into distinct buckets:

Data: For the Danish banks selected, data did not extend back farther than 2017, such a limited supply not anything out of the ordinary. From a time allocation perspective, a trade-off exists between obtaining bank data sets spanning longer time horizons and obtaining a greater number of data sets, with both of these variables aiding in generating more representative backtesting results. Having restricted our analysis to encapsulate a period of high uncertainty - where models tend to be shakier in general - we believe our results offer insight regardless.

Methods: As far as estimation is concerned, a relatively wide and balanced assortment of relevant methods was explored here. Still, it does not remotely come close to exhausting the options proposed in the literature. For instance, a range of more sophisticated variants of volatility/mean models could even constitute a focal point.

Backtests: This research piece only calls upon one backtest to evaluate ES predictions, the second test of Acerbi and Szekely (2014). Additional backtests may be run in order to assess whether different backtests produce similar outcomes. Alternatively, backtesting independence to complement severity testing and push already decent models (from an underestimation point of view) would be impactful (e.g. Du and Escanciano's (2017) conditional test, requiring the storage of daily predictive distributions).

Parameters: One example that immediately comes to mind is rolling window sizes, although plenty more could be tweaked. If model-related, the model in question should ideally be a solid one to begin with (one having cleared backtesting for underestimation).

⁕

6 References

Acerbi, C., & Szekely, B. (2014). Backtesting Expected Shortfall [pdf], *MSCI Inc*, Available at: https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b [Accessed 6 March 2023]

Bah, K., Munga'tu, J., & Waititu, A. (2016). Expected Shortfall Estimation Using Extreme Theory, *Global Journal of Finance and Management*, vol. 8, no. 1, pp.75-87

Bank for International Settlements. (2012). Fundamental review of the trading book [pdf], Available at: https://www.bis.org/publ/bcbs219.pdf [Accessed: 11 May 2023]

Bank for International Settlements. (2019). Explanatory note on minimum capital requirements for market risk [pdf], Available at: https://www.bis.org/bcbs/publ/d457_note.pdf [Accessed 6 March 2023]

Bank for International Settlements (BIS). (2023). MAR - Calculation of RWA for market risk, Available online: https://www.bis.org/basel_framework/standard/MAR.htm [Accessed: 11 May 2023]

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, vol. 31, no. 3, pp.307-327

Ding, Z., & Clive, C.W. (1996). Modeling volatility persistence of speculative returns: A new approach, *Journal of Econometrics*, vol. 73, no. 1, pp.185-215

Du, Z., & Escanciano, J.C. (2017). Backtesting Expected Shortfall: Accounting for Tail Risk, *Management Science*, vol. 63, no. 4, pp.940-958

Harmantzis, F.C., Miao, L., & Chien, Y. (2006). Empirical study of value-at-risk and expected shortfall models with heavy tails, *The Journal of Risk Finance*, vol. 7, no. 2, pp.117-135

Hull, J. (2018). Risk Management and Financial Institutions, 5th edn, Hoboken, New Jersey: John Wiley & Sons, Inc.

Longerstaey, J. & Spencer, H. (1996). RiskMetrics$^{\text{TM}}$ - Technical Document [pdf], *Morgan Guaranty Trust Company of New York*, Available at: https://www.msci.com/documents/10199/5915b101-4206-4ba0-aee2-3449d5c7e95a [Accessed 13 April 2023]

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*, vol. 7, no. 3-4, pp.271-300

McNeil, A.J., Frey, R., & Embrechts, P. (2015). Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition, Princeton University Press

Mones, D., & Taqi, M. (2023). Europe's 50 largest banks by assets, 2023, Available online: https://www.spglobal.com/marketintelligence/en/news-insights/research/europes-50-largest-banks-by-assets-2023 [Accessed 27 May 2023]

Nossman, M., & Vilhelmsson, A. (2014). Nonparametric forward-looking value-at-risk, *Journal of Risk*, vol. 16, no. 4

Righi, M.B., & Ceretta, P.S. (2015). A comparison of expected shortfall estimation models, *Journal of Economics and Business*, vol. 78, pp.14-47

Sobreira, N., & Louro, R. (2020). Evaluation of volatility models for forecasting Value-at-Risk and Expected Shortfall in the Portuguese stock market, *Finance Research Letters*, vol. 32, p.101098

Yamai, Y., & Yoshiba, T. (2005). Value-at-risk versus expected shortfall: A practical perspective, *Journal of Banking & Finance*, vol. 29, no. 4, pp.997–1015

❉

7 Appendix

7.1 Choice of Exponential Decay Factor (AWHS, VWHS)

RiskMetrics approximate

$$\sum_{i=1}^{n} w_{t-i} = \frac{1-\lambda}{1-\lambda^n} \sum_{i=1}^{n} \lambda^{i-1} = 1$$

in Equation 23 as

$$(1-\lambda) \sum_{i=1}^{n} \lambda^{i-1} \approx 1$$

but since in actuality

$$(1-\lambda) \sum_{i=1}^{\infty} \lambda^{i-1} = (1-\lambda) \sum_{i=0}^{\infty} \lambda^{i} = (1-\lambda) \frac{1}{1-\lambda} = 1$$

they define a tolerance level γ (i.e. approximation error)

$$\gamma = (1-\lambda) \left[\sum_{i=1}^{\infty} \lambda^{i-1} - \sum_{i=1}^{n} \lambda^{i-1} \right] = (1-\lambda) \sum_{i=n}^{\infty} \lambda^{i} = (1-\lambda) \cdot \lambda^n \sum_{i=0}^{\infty} \lambda^{i} = \lambda^n.$$

Taking ln on both sides and solving for λ leaves one with

$$\lambda = \exp\left(\frac{\ln \gamma}{n} \right).$$

Finally, γ was set to 1% (n being 500).

7.2 Linear Transformation of a Gaussian Random Variable

Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ a monotonic function with $Y = g(X)$. For g monotone increasing ($g'(x) \geqslant 0,\ \forall x \in dom(g)$):

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

by applying the chain rule, whereas for g monotone decreasing ($g'(x) \leqslant 0,\ \forall x \in dom(g)$) we get the corresponding survival function due to the reversal in inequality sign in the c.d.f.'s probability representation $\mathbb{P}(Y \leqslant y)$:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

and in general

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| .$$

From this result, for linear transformation $Y = a+bX,\ b \neq 0,\ X \sim N(\mu, \sigma^2)$:

$$g^{-1}(y) = \frac{y-a}{b}$$

$$\begin{aligned} f_Y(y) = \frac{1}{|b|} \cdot (2\pi\sigma^2)^{-0.5} \exp\left(- \frac{\left(\frac{y-a}{b} - \mu\right)^2}{2\sigma^2} \right) \\ = (2\pi b^2\sigma^2)^{-0.5} \exp\left(- \frac{(y-(a+b\mu))^2}{2b^2\sigma^2} \right). \end{aligned}$$

7.3 Derivation of ES for POT

$$\begin{aligned} ES_\alpha &= \frac{1}{1-\alpha} \int_\alpha^1 VaR_x \mathrm{d}x \\ &= \frac{1}{1-\alpha} \left[u \int_\alpha^1 1\mathrm{d}x + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{N}{N_u} \right)^{-\hat{\xi}} \int_\alpha^1 (1-x)^{-\hat{\xi}} \mathrm{d}x - \int_\alpha^1 1\mathrm{d}x \right] \right] \\ &= \frac{1}{1-\alpha} \left[u \cdot x \Big|_\alpha^1 + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{N}{N_u} \right)^{-\hat{\xi}} \cdot \frac{1}{\hat{\xi}-1} (1-x)^{-\hat{\xi}+1} \Big|_\alpha^1 - x \Big|_\alpha^1 \right] \right] \\ &= \frac{1}{1-\alpha} \left[u(1-\alpha) + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{N}{N_u} \right)^{-\hat{\xi}} \cdot \frac{1}{1-\hat{\xi}} (1-\alpha)^{-\hat{\xi}+1} - (1-\alpha) \right] \right] \\ &= u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{N}{N_u} \right)^{-\hat{\xi}} \cdot \frac{1}{1-\hat{\xi}} (1-\alpha)^{-\hat{\xi}} - 1 \right] \\ &= u \frac{1-\hat{\xi}}{1-\hat{\xi}} + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{N}{N_u} \right)^{-\hat{\xi}} \cdot \frac{1}{1-\hat{\xi}} (1-\alpha)^{-\hat{\xi}} - \frac{1-\hat{\xi}}{1-\hat{\xi}} \right] \\ &= \frac{VaR_\alpha}{1-\hat{\xi}} + \frac{\hat{\xi}\left(\frac{\hat{\beta}}{\hat{\xi}} - u\right)}{1-\hat{\xi}} = \frac{VaR_\alpha + \hat{\beta} - \hat{\xi} \cdot u}{1-\hat{\xi}} \end{aligned}$$

7.4 Backtesting Addendum

	5% level	1% level	0.1% level	0.01% level
S&P500	-0.75	-1.13	-1.60	-2.01
Danske Bank	-0.74	-1.12	-1.57	-1.96
Nykredit	-0.77	-1.16	-1.66	-2.08
Commerzbank	-0.74	-1.11	-1.56	-1.93
Acerbi & Szekely	-0.74	-	-	-2.00

MLE period	Dist.	KS test p
$[2002, 2022]$	NIG(0.335, 0.034, 0.007, -0.001)	0.46
$[2019, 2022]$	NIG(0.418, 0.043, 28.864, -0.307)	0.91
$[2019, 2022]$	NIG(0.181, -0.017, 5.285, -0.025)	0.29
$[2019, 2022]$	NIG(0.481, 0.032, 2.716, -0.004)	0.97
N/A	$\approx$ST(0, 1, 5)	N/A

Note on result replication: A seed value of 1 was set in R for each data set in simulations. The 'fBasics' package was used for NIG implementation.

❉