Zeliade Systems

Backtesting Expected Shortfall

Zeliade Systems

TITLE:	Backtesting Expected Shortfall
AUTHOR:	Zeliade Systems
NUMBER OF PAGES:	23
DATE:	2020-12-16
VERSION:	2.0

Contents

1 Abstract **2**

2 Expected Shortfall vs Value-at-Risk **3**

3 Backtesting ES **6**

3.1 Moldenhauer and Pitera test statistic . . . 7

3.1.1 Theoretical presentation . . . 7

3.1.2 Theoretical misspecification of the backtest . . . 8

3.1.3 How to avoid simulations . . . 9

3.2 Acerbi and Szekely $\bar{Z}_2$ statistic . . . 9

3.2.1 Theoretical presentation . . . 9

3.2.2 How to avoid simulations . . . 10

3.3 Acerbi and Szekely $\bar{Z}_{\text{MB}}$ statistic . . . 10

3.3.1 Theoretical presentation . . . 11

3.4 Acerbi and Szekely $\bar{Z}_3$ statistic . . . 11

3.4.1 Theoretical presentation . . . 11

3.5 Conclusion . . . 12

4 Quantitative assessment **13**

4.1 Tests on simulated data . . . 13

4.2 Tests on historical data . . . 15

4.3 Tests on historical data with approximated thresholds . . . 17

4.4 Conclusion . . . 19

5 Appendix I **20**

5.1 Moldenhauer and Pitera counterexamples . . . 20

5.1.1 Example 1: $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$ does not imply $\hat{ES}^\alpha < ES^\alpha$ . . . 20

5.1.2 Example 2: $E\hat{S}^\alpha < ES^\alpha$ does not imply $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$ . . . 21

5.2 Moldenhauer and Pitera alternative hypothesis . . . 22

1. Abstract

Since its introduction in 2001, the Expected Shortfall (ES) quickly became the standard risk measure used by financial institutions including central clearing counterparties (CCPs). Indeed, many CCPs switched from the Value at Risk (VaR) to the more conservative ES to compute their initial margins. The need of a sound backtest for the ES arose then naturally. In 2011, the proof that the Expected Shortfall (ES) lacks a property called elicitability has led to the incorrect conclusion that the ES is not backtestable. Three years later, Acerbi and Szekely designed three possible backtests for the ES and, since then, many other backtests have been proposed in the practitioner literature. In this work we study four of these test statistics from both a theoretical and practical point of view and eventually give some advice for CCPs in search of a good backtest for ES.

2. Expected Shortfall vs Value-at-Risk

Value-at-Risk (VaR) has become a standard risk measure for financial risk management due to its conceptual simplicity, ease of computation, and immediate applicability. VaR measures the maximum potential change in the value of a portfolio with a given probability over a pre-set horizon:

$$\mathrm{VaR}_\alpha\left(\mathrm{PnL}_d\right) = -\inf\{x|\mathbb{P}\left(\mathrm{PnL}_d \leq x\right) > \alpha\}$$

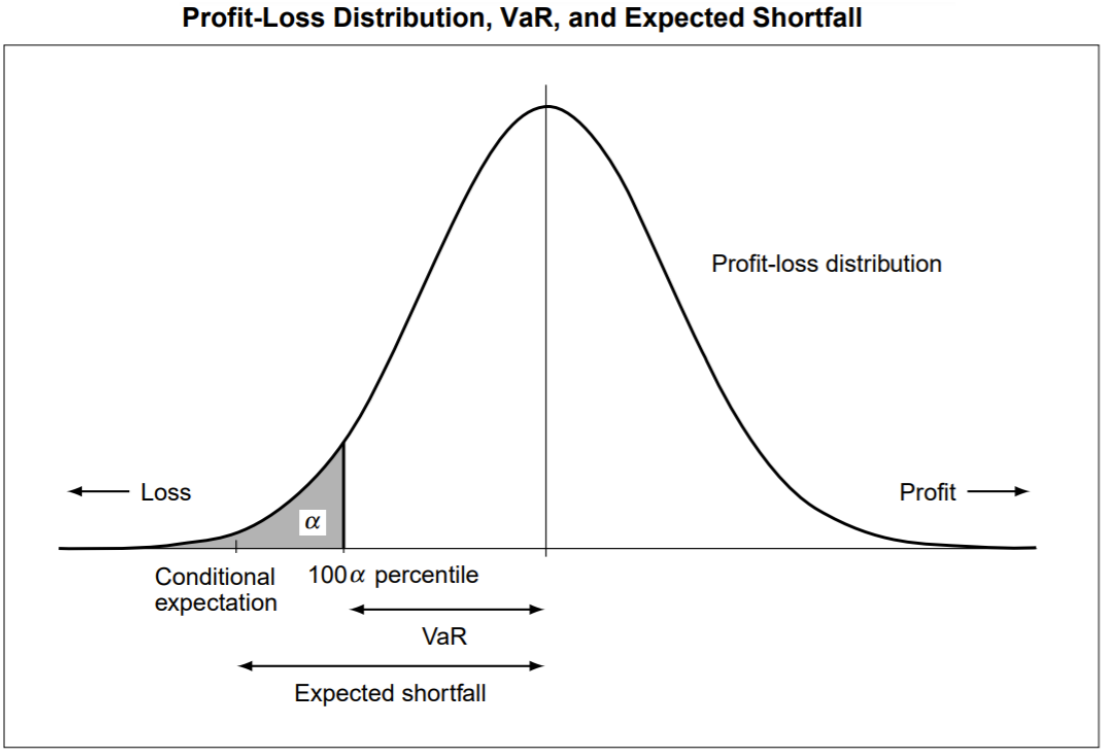
Nevertheless, VaR has several conceptual problems:

- VaR measures only a quantile of PnL distributions and does not account for the losses beyond this level;
- VaR is not coherent, since it is not subadditive, a property that implies that the sum of sub-VaRs is not necessarily conservative.

The latter item means that if we split a portfolio into two sub-portfolios and compute the VaR for each sub-portfolio then the sum of the two VaRs can be smaller than the true VaR of the global portfolio.

As an alternative to the VaR risk measure, Artzner et al. (1997) [4] proposed Expected Shortfall (ES shortly, also called "conditional VaR", "mean excess loss", "beyond VaR", or "tail VaR"). ES is the conditional expectation of loss given that the loss is beyond the VaR level; that is, the expected shortfall is defined as follows:

$$\mathrm{ES}_\alpha\left(\mathrm{PnL}_d\right) = \mathbb{E}\left[-\mathrm{PnL}_d \mid \mathrm{PnL}_d \leq \mathrm{VaR}_\alpha\left(\mathrm{PnL}_d\right)\right]$$



ES is generally considered a more useful risk measure than VaR thanks to its robustness and to the fact that the ES verifies the subadditivity property, as opposed to the VaR. This means that the sum of two sub-ESs is greater than the global ES, entailing an inherent conservativeness.

The ES takes into account, by definition, the severity of the tail observations beyond the VaR. This makes the ES a more conservative risk measure than the VaR, for the same confidence level:

$$\mathrm{ES}_\alpha\left(\mathrm{PnL}_d\right) \geq \mathrm{VaR}_\alpha\left(\mathrm{PnL}_d\right).$$

Moreover, the ES is more robust: the fact that the ES is an average of all PnLs beyond the VaR makes its estimation more stable, since a change in a single observation would be mitigated by the rest of the values in the average. In the VaR case, its estimation is driven by a single value (or at most two values if a linear interpolation is used), which means that the VaR would suffer from large jumps when the time window moves and extreme observations are included or excluded. This robustness/stability plays an important role in diminishing the procyclicality of the margins, since when extreme market moves happen, the margins would increase slower than for VaR. This prevents, partially, from exacerbating the market stress events.

All these advantages of the ES explain its use by the majors CCPs for their margin computations, to the point where it became an industry standard.

The only weak feature of the ES was the lack of backtesting tests, while the VaR has several robust statistical tests such as the Kupiec and Christoffersen tests. This weakness was remedied by the recently proposed statistical tests, starting from the work of Acerbi and Szekely [1].

We illustrate in the following graph an example of a VaR and ES computation for the Brazil Stock Market Index (BOVESPA):

The use of the ES allows to reduce the number of breaches from 8 to 0.

3. Backtesting ES

The computation of the ES needs to be backtested, which means that one should check a posteriori whether the risk prediction was correct. This check, when it leads to a negative result, is a good indicator that the ES computation method should be revised. In the past years the backtestability of the ES has been questioned: since Gneiting proved in 2011 ([6]) that this risk measure lacks a property called elicitability (whereas the pair (VaR, ES) is elicitable, [5]), some mathematicians concluded that the ES was not backtestable. However, in 2014 Acerbi and Szekely proved in [1] that this is not the case, by simply finding some backtest statistics for the ES. Since then, many articles regarding the backtestability of the ES have been published but a lot of them lack of a proper definition of backtestability and, as a consequence, the tests proposed are not theoretically rigorous. For this reason we start the current section by mathematically defining what backtesting the ES means.

When we talk about backtesting a statistic, we refer to Definition 3.1 of [2]. In particular, we say that

Definition 3.1. *The statistic ES is* backtestable *if there exists a backtest function $Z(e, v, x)$ such that*

- $E_H[Z(e, v, X)] = 0$ *iff* $e = -E_H[X|X \leq -v]$;
- $E_H[Z(e_1, v, X)] < E_H[Z(e_2, v, X)]$ *if* $e_1 < e_2$

for a fixed v.

This means that when we underestimate the value of the ES, the sign of the backtest function will be negative on average.

With these tools, we can then define the backtest test as

$$\bar{Z}(X) = \frac{1}{N} \sum_{d=1}^{N} Z(\hat{ES}_d^\alpha, \hat{VaR}_d^\alpha, X_d)$$

where X, $\hat{ES}^\alpha$ and $\hat{VaR}^\alpha$ are the vectors of respectively the valuations of a portfolio (e.g. PnLs), the estimated ESs and the estimated VaRs. From the previous definition, we have that under the null hypothesis $H_0 : ES_d^\alpha = \hat{ES}_d^\alpha$, it holds $E_{H_0}[Z(\hat{ES}^\alpha, \hat{VaR}^\alpha, X)] = 0$ while under the alternative hypothesis of underestimation of the ES, $H_1 : ES_d^\alpha \geq \hat{ES}_d^\alpha \; \forall d \wedge \exists d : ES_d^\alpha > \hat{ES}_d^\alpha$, it holds

$$E_{H_1}[Z(\hat{ES}^\alpha, \hat{VaR}^\alpha, X)] < E_{H_1}[Z(ES^\alpha, VaR^\alpha, X)] = 0 = E_{H_0}[Z(\hat{ES}^\alpha, \hat{VaR}^\alpha, X)]$$

(note that here we are supposing $\hat{VaR}^\alpha = VaR^\alpha$ but this is not needed if we ask in the definition of backtestability that $E_H[Z(e, v_1, X)] < E_H[Z(e, v_2, X)]$ if $v_1 < v_2$ and add to the alternative hypothesis the requirement $VaR_d^\alpha \geq \hat{VaR}_d^\alpha \; \forall d$).

Then, in order to backtest the ES, one has to compute the value of $\bar{Z}(X)$ and compare it with a threshold value. The latter can be chosen as the ϕ-quantile of the distribution of $\bar{Z}(X)$ under the null hypothesis and can be empirically obtained by repeating M times the following steps:

1. Simulate a N-vector of X's under the distribution of the null hypothesis;
2. Calculate $\bar{Z}(X)$ using the already computed $\hat{ES}_d^\alpha$, $\hat{VaR}_d^\alpha$ for $d = 1, \ldots, N$.

The ϕ-quantile can be calculated as $\bar{Z}(X)_{([x])} + (x - [x])(\bar{Z}(X)_{([x]+1)} - \bar{Z}(X)_{([x])})$ where $x = \frac{\phi}{100}(M-1)+1$. At this point, the computed value $\bar{Z}(X)$ is accepted iff it is greater than the threshold.

We will refer to type I and type II errors as

1. Type I error: when $\hat{ES}^{\alpha}$ is correct but it is rejected;
2. Type II error: when $\hat{ES}^{\alpha}$ underestimates the real ES but it is accepted.

In the following subsections we will discuss and compare different possible Expected Shortfall backtest procedures. We start with the analysis of the test statistic proposed by Moldenhauer and Pitera in [7]. We try to explain why this statistic is not strictly a proper backtest for ES, in the sense of definition 3.1, but rather tests the distribution of the PnLs. Then, from a theoretical point of view, one should prefer $\bar{Z}_2$ and the minimally biased statistic, which we denote by $\bar{Z}_{\text{MB}}$, proposed by Acerbi and Szekely in the articles [1] and [3]. We also have a look at the so called $\bar{Z}_3$ statistic in [1] and see that it has the same theoretical issues than the Moldenhauer and Pitera statistic. From a theoretical point of view, we end up by suggesting the use of $\bar{Z}_{\text{MB}}$ as its own authors do. On the other hand, we will see in the next section that from a practical point of view the Moldenhauer and Pitera statistic is as good as $\bar{Z}_{\text{MB}}$, at least on the tests we performed.

3.1 Moldenhauer and Pitera test statistic

This test was proposed by Moldenhauer and Pitera in their article *Backtesting expected shortfall: a simple recipe?* [7].

3.1.1 Theoretical presentation

Let X_d denote the random process of the valuation of a portfolio (e.g. the PnL) and let ES_d^{α} denote the computed Expected Shortfall value for the probability α at day d. We define the random process

$$Y = X + ES^{\alpha}$$

to be the secured position. Alternatively, the definition $Y = \frac{X+ES^{\alpha}}{ES^{\alpha}}$ can be used. With this choice, the whole following discussion does not change and also the paragraphs 'Theoretical misspecification of the backtest' and 'How to avoid simulations' are still valid.

The ES test statistic used by Moldenhauer and Pitera is

$$G(X, ES^{\alpha}) = \sum_{k=1}^{N} \mathbb{1}_{(Y_{[1]}+\cdots+Y_{[k]}<0)}$$

where N is the number of days in the observation window and the random process $Y_{[d]}$ denotes the ordered statistic of Y_d. Note that the authors define this statistic divided by N but for stability properties, we do not do this division (see paragraph How to avoid simulations).

In order to check whether this is a good backtest for ES, we need to see what happens to the statistic when the ES is underestimated. Suppose that the calculated value of the ES at time d is $E\hat{S}_d^{\alpha}$. We set the null hypothesis to be $H_0 : ES_d^{\alpha} = E\hat{S}_d^{\alpha}\ \forall d$ while the alternative hypothesis will be $H_1 : ES_d^{\alpha} \geq E\hat{S}_d^{\alpha}\ \forall d \wedge \exists d :$

$ES_d^\alpha > \hat{ES}_d^\alpha$. Note that if we use a unique method to evaluate the ES, the distributions of X are different in the two hypothesis.

Under the alternative hypothesis note that $X_d + \hat{ES}_d^\alpha \leq X_d + ES_d^\alpha$ for every d so that $(X_d + \hat{ES}_d^\alpha)_{(k)} \leq (X_d + ES_d^\alpha)_{[k]}$ for every k and $(X_d + \hat{ES}_d^\alpha)_{[1]} + \cdots + (X_d + \hat{ES}_d^\alpha)_{[k]} \leq (X_d + ES_d^\alpha)_{[1]} + \cdots + (X_d + ES_d^\alpha)_{[k]}$ for every k and the inequality is strict for some k. So

$$\begin{aligned} E_{H_1}[G(X, \hat{ES}^\alpha)] &= \sum_{k=1}^{N} P_{H_1}((X_d + \hat{ES}_d^\alpha)_{[1]} + \cdots + (X_d + \hat{ES}_d^\alpha)_{[k]} < 0) \\ &> \sum_{k=1}^{N} P_{H_1}((X_d + ES_d^\alpha)_{[1]} + \cdots + (X_d + ES_d^\alpha)_{[k]} < 0) = E_{H_1}[G(X, ES^\alpha)]. \end{aligned}$$

What we found is that if we underestimate the ES value, the statistic G will have on average a value greater than its true value, which depends of course also on the distribution of X. In order to find a threshold, however, we cannot use the value $E_{H_1}[G(X, ES^\alpha)]$, since we do not know it. What is fundamental to prove is rather that $E_{H_1}[G(X, \hat{ES}^\alpha)] > E_{H_0}[G(X, \hat{ES}^\alpha)]$. This could be done if, for example, the statistic G is constructed in such a way that $G(X, ES^\alpha) = 0$ iff ES^α is the true value of the ES. In this case then $E_{H_1}[G(X, ES^\alpha)] = 0 = E_{H_0}[G(X, \hat{ES}^\alpha)]$ and the required inequality would be automatically achieved. After, we could proceed with setting the threshold value to be the empirical ϕ-quantile obtained by simulations of $G(X, \hat{ES}^\alpha)$. The requirement that a test statistic is null at the true value of the backtested quantity is exactly what we have reported in the definition of backtestability.

3.1.2 Theoretical misspecification of the backtest

In Appendix I we present two counterexamples which show that from a strict theoretical prospective, the G statistic is not the best choice for backtesting the ES, according to our definition 3.1.

In the first example, we show that $E_{H_1}[G(X, \hat{ES}^\alpha)] > E_{H_0}[G(X, \hat{ES}^\alpha)]$ does not imply $\hat{ES}^\alpha < ES^\alpha$. This means that if we calculate the threshold as the ϕ-quantile of the simulated vector of G's and then accept $\hat{ES}^\alpha$ iff $G(X, \hat{ES}^\alpha)$ is smaller than the threshold, then we could be rejecting $\hat{ES}^\alpha$ even if it correctly overestimates the real ES^α. This causes an higher probability to make type I errors.

The second example proves that the very general hypothesis $H_1 : ES_d^\alpha \geq \hat{ES}_d^\alpha \ \forall d \wedge \exists d : ES_d^\alpha > \hat{ES}_d^\alpha$ does not imply that $E_{H_1}[G(X, \hat{ES}^\alpha)] > E_{H_0}[G(X, \hat{ES}^\alpha)]$. This fact could cause the error of accepting $\hat{ES}^\alpha$ even if it underestimates the real ES^α and so an error of type II.

These are very easy examples since we take $N = 1$ but even with only one variable X it is possible to show that the statistic G does not satisfy definition 3.1 of a backtest function. Why then does it seem to work properly in the article of Moldenhauer and Pitera? We think that rather than being a backtest for the ES, it is a backtest for the generic distribution of the X's, with similar hypothesis as in the paragraph 'Acerbi and Szekely $\bar{Z}_3$ statistic'. Indeed, we prove this fact in Appendix, paragraph 'Moldenhauer and Pitera alternative hypothesis'.

The hypothesis used do not lead to the desired result in the previous examples. Then, if we restricted ourselves to the strict theoretical aspect, the G statistic would not be considered. On the other hand, from a practical point of view, the PnLs' distributions are generally approximated by Student-t, whose tails (in particular for negative values) can be compared by stochastic dominance. In particular, if P_{ν_1} and P_{ν_2} are

the distributions of two Student-t with $\nu_1 < \nu_2$ degrees of freedom, then $P_{\nu_1} \preceq P_{\nu_2}$. This means that if we set the degrees of freedom of the X's distributions to be higher than in reality, the statistic G will correctly signal it.

3.1.3 How to avoid simulations

The G statistic, whatever it backtests, is rather robust with respect to the underlying distribution for N small enough and ϕ not too high and so the threshold can be simulated by standard normal distributions for the X's. The threshold can be calculated as follows:

1. Compute ES^α for the standard normal distribution (with closed formulas);
2. Iterate M times the following steps:
   a. Simulate a N-vector of X's under the standard normal distribution;
   b. Calculate $G(X, ES^\alpha)$ using the already computed ES^α.
3. Take the ϕ-quantile of the $G(X, ES^\alpha)$'s.

In particular, for $\alpha = 0.5\%$ and $\phi = 95\%$, the threshold can be set at 6. Note that in this case, if we use a Student-t distribution for the X's, the result is still 6.

It must be remarked that increasing N or ϕ, the statistic G is not stable anymore and its threshold cannot be approximated in this way but it must be computed as explained in the beginning of this section. Indeed, from the following table we can see that the thresholds for G under a Student-t distribution with 5 degrees of freedom or a standard normal distribution for the X's can drastically change (we set $\alpha = 0.5\%$):

	N	$\phi(\%)$	Normal	Student-t
0	500	95.00	6	6
1	500	99.99	12	17
2	1000	95.00	10	10
3	1000	99.99	18	24
4	2000	95.00	17	18
5	2000	99.99	27	35

Table 1: Thresholds of G for $\alpha = 0.5\%$

3.2 Acerbi and Szekely $\bar{Z}_2$ statistic

This test was proposed by Acerbi and Szekely in their 2014 article *Backtesting Expected Shortfall* [1].

3.2.1 Theoretical presentation

Define the backtest function

$$Z_2(e, v, x) = \frac{x\mathbb{1}_{\{x+v<0\}}}{\alpha e} + 1.$$

Then, under the hypothesis that $VaR^\alpha(X) = v$ and $ES^\alpha(X) = e$ it holds $E[Z_2(e, v, X)] = 0$. Furthermore Z_2 is strictly increasing with v and with e, meaning that when $E[Z_2(e, v, X)] < 0$, the computed VaR v and/or the computed Expected Shortfall e underestimate the real ones.

A natural test statistic for the calculated value $\hat{ES}^\alpha$ can then be chosen as

$$\bar{Z}_2(X) = \frac{1}{N} \sum_{d=1}^{N} Z_2(\hat{ES}_d^\alpha, \hat{VaR}_d^\alpha, X_d).$$

It is easy to see that, under the null hypothesis of correctly chosen $\hat{ES}_d^\alpha$, the mean value of $\bar{Z}_2(X)$ is 0. Otherwise, under the alternative hypothesis of underestimation of the risk

$$\begin{aligned} H_1 : & ES_d^\alpha \geq \hat{ES}_d^\alpha \ \forall d \wedge \exists d : ES_d^\alpha > \hat{ES}_d^\alpha \\ & VaR_d^\alpha \geq \hat{VaR}_d^\alpha \ \forall d, \end{aligned}$$

it holds $E_{H_1}[\bar{Z}_2(X)] < 0 = E_{H_0}[\bar{Z}_2(X)]$.

This means that in contrast with the Moldenhauer and Pitera test, the $\bar{Z}_2$ statistic correctly backtests the ES, following our definition of backtestability.

3.2.2 How to avoid simulations

For fixed α and ϕ, it is possible to numerically check that the thresholds for the $\bar{Z}_2$ statistic in case of Student-t distributions are quite stable through the ν's. The threshold values for $\alpha = 0.5\%$ and $\phi = 5\%$ are for example (here we do 500000 simulations):

	Threshold
ν	
3	-1.3
5	-1.2
10	-1.2
100	-1.1
1000	-1.1

Table 2: Thresholds of $\bar{Z}_2$ for $\alpha = 0.5\%$ and $\phi = 5\%$

It follows that for this test statistic one can take as fixed threshold a value of -1.2 avoiding to calculate it.

3.3 Acerbi and Szekely $\bar{Z}_{\mathbf{MB}}$ statistic

This test was proposed by Acerbi and Szekely in their 2017 article *General properties of backtestable statistics* [2].

3.3.1 Theoretical presentation

Following the steps for $\bar{Z}_2$, the authors define a different test statistic. This time the backtest function is

$$Z_{\text{MB}}(e, v, x) = e - v + \frac{(x+v)\mathbb{1}_{\{x+v<0\}}}{\alpha}.$$

As before, if $VaR^\alpha(X) = v$ and $ES^\alpha(X) = e$, then $E[Z_{\text{MB}}(e, v, X)] = 0$ and Acerbi and Szekely show in section 4.2 of [2] that $E[Z_{\text{MB}}(e, v, X)] < 0$ when the calculated Expected Shortfall e underestimates the real one, no matter the value of v.

The corresponding test statistic for $\hat{ES}^\alpha$ is

$$\bar{Z}_{\text{MB}}(X) = \frac{1}{N}\sum_{d=1}^{N} Z_{\text{MB}}(\hat{ES}_d^\alpha, \hat{VaR}_d^\alpha, X_d).$$

Setting the less strict alternative hypothesis $H_1 : ES_d^\alpha \geq \hat{ES}_d^\alpha \ \forall d \wedge \exists d : ES_d^\alpha > \hat{ES}_d^\alpha$, it holds again $E_{H_1}[\bar{Z}_{\text{MB}}(X)] < 0 = E_{H_0}[\bar{Z}_{\text{MB}}(X)]$ and the ES can be backtested as in the previous example.

This statistic is preferred by Acerbi and Szekely since it presents a smaller sensitivity to VaR predictions. In particular, the test statistic $\bar{Z}_2$ could face type I and type II errors with more probability than the test statistic $\bar{Z}_{\text{MB}}$ if the prediction $\hat{VaR}^\alpha$ is not correct.

3.4 Acerbi and Szekely $\bar{Z}_3$ statistic

We now consider another statistic which does not directly backtest the computed value of the ES but which rather backtests the distribution of the X's used to evaluate the ES. This test was proposed by Acerbi and Szekely in their 2014 article *Backtesting Expected Shortfall* [1].

3.4.1 Theoretical presentation

In particular call P_d the predicted distribution of X_d used to evaluate the VaR and the ES and call F_d the real unknown distribution of X_d. We put

$$\begin{aligned} H_0 &: F_d = P_d \ \forall d \\ H_1 &: F_d \preceq P_d \ \forall d \wedge \exists d : F_d \prec P_d \end{aligned}$$

where $\preceq$ denotes that the left side is first order stochastically dominated by the right side. This is equivalent to say that the cdf of F_d is no smaller than the cdf of P_d at every point and that for every non-decreasing function u it holds $\int u(x)\, dF_d(x) \leq \int u(x)\, dP_d(x)$. As a consequence, both VaR and ES are underestimated under P_d.

If the test ends up to accept the null hypothesis, then it is possible to evaluate $\hat{ES}_d^\alpha$ through the formula

$$\hat{ES}_d^\alpha = \hat{ES}_M^\alpha(Y^d) = -\frac{1}{[M\alpha]}\sum_{i=1}^{[M\alpha]} Y_{(i)}^d$$

where M is a big number (e.g. $M = N$ if an historical simulation is used) and Y^d is an M-vector of simulated variables distributed as P_d.

The test statistic used is

$$\bar{Z}_3 = -\frac{1}{N}\sum_{d=1}^{N}\frac{\hat{ES}_N^{\alpha}(P_d^{-1}(U))}{E_V[\hat{ES}_N^{\alpha}(P_d^{-1}(V))]} + 1$$

where U is an iid N-vector such that $U_d = P_d(X_d)$ while V is an iid N-vector of variables $U([0,1])$. Denoting a regularized incomplete beta function as $I_x(a,b)$, the denominator can be analytically computed as

$$E_V[\hat{ES}_N^{\alpha}(P_d^{-1}(V))] = -\frac{N}{[N\alpha]}\int_0^1 I_{1-p}(N-[N\alpha],[N\alpha])P_d^{-1}(p)\,dp.$$

This entails that $E_{H_0}[\bar{Z}_3] = 0$ and $E_{H_1}[\bar{Z}_3] < 0$.

This test statistic is very general and its alternative hypothesis does not directly involve the computed ES: this means that it is not a backtest for the ES. Furthermore, it is not as straightforward as the other statistics considered so we do not suggest its use for the precise purpose of backtesting the ES at least.

3.5 Conclusion

We can sum up the pros and cons of each test statistic:

- Moldenhauer and Pitera test statistic G:
  - Pros: the threshold can be calculated taking a standard normal distribution for the X's.
  - Cons: it does not satisfy definition 3.1 of a backtest function;
- Acerbi and Szekely $\bar{Z}_2$:
  - Pros: extremely easy to be implemented, the threshold can be calculated taking a Student-t (with e.g. $\nu = 5$ degrees of freedom) distribution for the X's.
  - Cons: it could face type II errors;
- Acerbi and Szekely $\bar{Z}_{\text{MB}}$:
  - Pros: extremely easy to be implemented, it is very little influenced by the VaR predictions.
  - Cons: the threshold must be evaluated through simulation of the X's distribution;
- Acerbi and Szekely $\bar{Z}_3$:
  - Pros: not many.
  - Cons: it is the most difficult to be implemented, it is not a proper backtest for the ES.

The best theoretical choice is the $\bar{Z}_{\text{MB}}$ statistic because it correctly tests the ES, it is very easy to be implemented and it is little sensitive to VaR predictions. From the practical point of view, however, the fact that the threshold cannot be approximated by standard distribution requires, if using a Filtered Historical Simulation method, to store all the simulation of the X's used to compute the ES. The statistics G and $\bar{Z}_2$ do not face this problem, although G lacks some theoretical justifications and it is a little bit more difficult to be implemented, while $\bar{Z}_2$ is not as precise as $\bar{Z}_{\text{MB}}$ in the choice of accepting or rejecting the computed ES.

4. Quantitative assessment

4.1 Tests on simulated data

We compare here the power of the four statistics G, $\bar{Z}_2$, $\bar{Z}_{\text{MB}}$ and $\bar{Z}_3$. The simulated distribution of the PnL process are Student-t with ν degrees of freedom and the null and alternative hypothesis change for the number of degrees of freedom of the distribution. In particular, $H_0 : \nu = \nu_0$ while $H_1 : \nu = \nu_1$. The power of a statistic is the probability to reject the null hypothesis when indeed the alternative hypothesis is correct. This means that the higher the power, the better is the test statistic in terms of avoiding type II errors. To evaluate the power of the tests under an alternative hypothesis H_1 which underestimates the risk, it is necessary to have $\nu_1 < \nu_0$.

We can use the function power for two purposes:

1. Evaluation of the probability to commit type I errors: this error arises when the null hypothesis is rejected even if it is true and the probability at which it arises is equal to the significance level ϕ. In order to check whether the function power is correctly written, we put $\nu_1 = \nu_0$ and see if its value is actually ϕ.
2. Evaluation of the probability to commit type II errors: this probability is the difference between 1 and the power.

We set the level of the Expected Shortfall at $\alpha = 99.5\%$, the significance level of the test at $\phi = 5\%$ (which means that if the p-value is less than ϕ, the statistic is rejected) and the number of days in the observation window at $N = 500$. To calculate the threshold level for the statistic we compute 250000 simulations while to calculate the power, that is the rate of rejected statistics, we do 100000 simulation.

In order to use the same input data as the Acerbi and Szekely's statistics, instead of computing the G statistic, we calculate $-G$. Furthermore, as Moldenhauer and Pitera suggest, we use the relative secured positions: $Y = \frac{X+ES^\alpha}{ES^\alpha}$.

We add also the power column for the $\bar{Z}_2$ statistic with a precomputed threshold equal to -1.2, calling it $\bar{Z}_2$ bis.

	ν in H_0	ν in H_1	G	$\bar{Z}_2$	$\bar{Z}_{\text{MB}}$	$\bar{Z}_3$	$\bar{Z}_2$ bis
0	3	3	7.4	4.9	4.9	5.1	6.1
1	5	3	76.0	76.7	68.8	55.4	76.4
2	10	3	99.5	99.5	99.3	97.2	99.5
3	100	3	100.0	100.0	100.0	100.0	100.0
4	5	5	6.9	5.0	5.0	5.0	5.1
5	10	5	71.0	67.7	66.2	54.8	66.7
6	100	5	99.4	99.0	99.2	97.7	98.7
7	10	10	5.9	5.0	5.1	5.0	4.6
8	100	10	75.0	70.0	73.4	64.4	66.9
9	100	100	5.3	5.0	5.0	5.2	4.3

Table 3: Power of G, $\bar{Z}_2$, $\bar{Z}_{MB}$, $\bar{Z}_3$ and $\bar{Z}_2$ bis (%)

We can see that G, $\bar{Z}_2$ and $\bar{Z}_{\text{MB}}$'s powers are definitely higher than $\bar{Z}_3$'s one. Also, the evaluation of the latter statistic requires much more time than the formers, whose computation times are similar.

The computation time of $\bar{Z}_2$ bis is definitely lower since it does not require the calculation of the threshold. Its power cannot actually be compared with the other statistics' ones since it is evaluated on significance levels which differ from 5%. In particular, the significance level is 6.1% for $\nu = 3$, 4.5% for $\nu = 10$ and 4.3% for $\nu = 100$. A higher (lower) significance level leads to a higher (lower) power so setting these levels of significance for the other statistics would also increase (decrease) their power. For this reason we suggest the use of a precomputed threshold statistic only in the case of very long computation times, which in these examples do not actually arise.

This argument holds also for G, whose significance levels are somehow higher than 5% for small ν's, which means that the corresponding powers will also be higher. Then, the real power of G is not as high as it seems. The reason why G faces a higher probability of type I error is explained in Example 1 of paragraph Moldenhauer and Pitera test statistic, section 3-2.

For a matter of completeness we report in the following two tables the power of $\bar{Z}_2$ and $\bar{Z}_{\text{MB}}$ for the actual significance levels used by $\bar{Z}_2$ bis (first table) and by G (second table). We see that their power changes as predicted.

	ϕ	$\bar{Z}_2$	$\bar{Z}_{\text{MB}}$	$\bar{Z}_2$ bis
0	6.1	6.0	6.2	5.1
1	6.1	78.6	72.1	55.4
2	6.1	99.6	99.5	97.2
3	6.1	100.0	100.0	100.0
4	5.1	5.1	5.2	5.0
5	5.1	68.2	66.9	54.8
6	5.1	99.0	99.2	97.7
7	4.6	4.5	4.7	5.0
8	4.6	68.3	72.3	64.4
9	4.3	4.2	4.2	5.2

Table 4: Power of $\bar{Z}_2$, $\bar{Z}_{MB}$ and $\bar{Z}_2$ bis with different ϕs (%)

	ϕ	$\bar{Z}_2$	$\bar{Z}_{\text{MB}}$	G
0	7.4	7.4	7.3	7.4
1	7.4	80.8	75.0	76.0
2	7.4	99.7	99.5	99.5
3	7.4	100.0	100.0	100.0
4	6.9	6.9	6.9	6.9
5	6.9	72.6	71.2	71.0
6	6.9	99.2	99.4	99.4
7	5.9	5.9	5.8	5.9
8	5.9	72.1	75.4	75.0
9	5.3	5.2	5.2	5.3

Table 5: Power of $\bar{Z}_2$, $\bar{Z}_{MB}$ and G with different ϕs (%)

4.2 Tests on historical data

The evaluation of the ES is done as described in section 2 and as before we denote $\hat{ES}_d^\alpha$ its estimated value at day d. With the notations of section 3.2, we have $X_d = P_{d+1} - P_d$ where P_d is the value of the asset at day d. The statistic $\bar{Z}$ will be evaluated on these realised values of X:

$$\bar{Z}(X) = \frac{1}{N}\sum_{d=1}^{N} Z(\hat{ES}_d^\alpha, \hat{VaR}_d^\alpha, X_d).$$

In order to calculate the threshold, it is necessary to simulate the test statistic M times and then to compare the ϕ-quantile with $\bar{Z}(X)$. How can $\bar{Z}$ be simulated if we use an historical distribution? In order to calculate $\hat{ES}_d^\alpha$ through an historical method as HS or FHS, we needed to simulate M scenario of X_d for every d, taking into account the history of X. Since we do an historical simulation, M corresponds to the number of data available (for us, ten years so $M = 2500$). We can then use the same simulations to compute

$$\bar{Z}_k = \frac{1}{N}\sum_{d=1}^{N} Z(\hat{ES}_d^\alpha, \hat{VaR}_d^\alpha, X_{d,k})$$

for each $k = 1, \dots, M$ and finally take the ϕ-quantile of the $\bar{Z}$'s vector.

We now let run the G, $\bar{Z}_2$ and $\bar{Z}_{\text{MB}}$ backtests on some portfolios on equity products obtained from Yahoo Finance in the period from 02/12/2005 to 02/12/2020.

	G	Threshold	Accepted
AXJO	-9	-9	No
BVSP	0	-9	Yes
FCHI	-3	-12	Yes
GDAXI	-2	-14	Yes
GSPC	-9	-11	Yes
GSPTSE	-3	-14	Yes
KS11	-7	-15	Yes
MXX	-8	-10	Yes
SSMI	-3	-11	Yes
TWII	-12	-11	No

Table 6: Accepted ES for G

	$\bar{Z}_2$	Threshold	Accepted
AXJO	-2.5397	-2.76928	Yes
BVSP	-1.67839	-2.27365	Yes
FCHI	-0.527363	-3.96901	Yes
GDAXI	-0.458829	-4.07022	Yes
GSPC	-2.1588	-2.95039	Yes
GSPTSE	-0.160738	-3.80838	Yes
KS11	-1.77876	-4.18092	Yes
MXX	-1.52592	-2.83534	Yes
SSMI	-0.481356	-2.68431	Yes
TWII	-1.39444	-3.85184	Yes

Table 7: Accepted ES for $\bar{Z}_2$

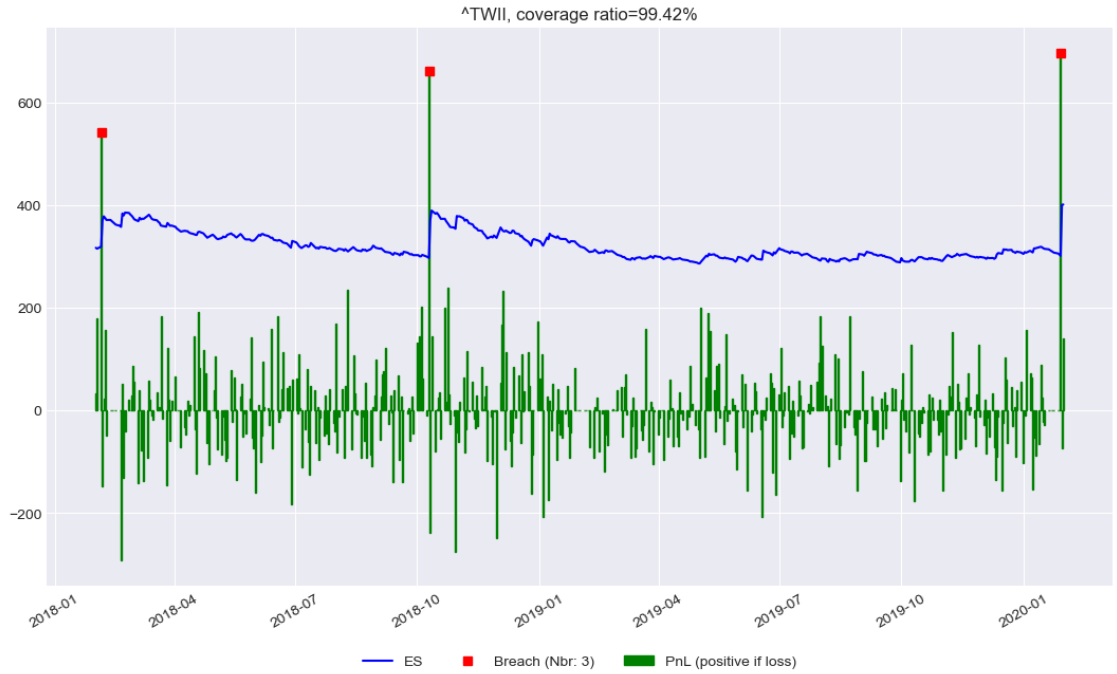
	$\bar{Z}_{\text{MB}}$	Threshold	Accepted
AXJO	-82.1309	-76.2563	No
BVSP	-396.507	-2533.57	Yes
FCHI	-15.1573	-132.957	Yes
GDAXI	-5.50087	-318.857	Yes
GSPC	-46.6548	-64.55	Yes
GSPTSE	4.73223	-459.015	Yes
KS11	-26.4939	-59.7472	Yes
MXX	-799.427	-904.342	Yes
SSMI	-15.3335	-228.596	Yes
TWII	-381.284	-197.07	No

Table 8: Accepted ES for $\bar{Z}_{MB}$

It can be seen that the backtests G and $\bar{Z}_{\text{MB}}$ lead to the same results. This is somehow surprising since from the theoretical point of view we have proven that G does not satisfy theoretical conditions of definition 3.1 of a backtest function. However, our observations suggest that G is a backtest for the whole distribution of the PnLs and so it could have the same results of a backtest for the ES, when the distributions employed for the evaluation of the ES are misspecified.

The backtest $\bar{Z}_2$ accepts the estimated ES for a higher number of portfolios and if G or $\bar{Z}_{\text{MB}}$ accept the ES, then also $\bar{Z}_2$ does. This could be explained by Example 4.3 and Figure 5 of [2], where it is shown that when the ES is underestimated, $\bar{Z}_2$ could fail to reject it causing a type II error, while this would never happen for $\bar{Z}_{\text{MB}}$. Which of the three statistics is then right?

Let us plot the realized PnLs versus the ES estimated level for the Taiwan Weighted Index.



From this plot, we can see that there are only three breaches but they are huge, so G and $\bar{Z}_{\text{MB}}$ take into account also the magnitude of the breaches while $\bar{Z}_2$ seems to slightly neglect it. Once again, we suggest the use of the $\bar{Z}_{\text{MB}}$ statistic.

4.3 Tests on historical data with approximated thresholds

We repeat the tests done in the previous session for G and $\bar{Z}_2$ but approximating the thresholds with precomputed ones. This will save a lot of memory since it does not require the storage of all the PnLs simulations but it will affect the results. This time the test cannot be done with the statistic $\bar{Z}_{\text{MB}}$ because it is not stable in the distribution of the underlying. We stress the fact that this approximation can be done on G when N is not too large and ϕ is not too small.

For G we precompute the threshold with standard normal distributions or, equivalently, with Student-t distributions for the PnLs. Setting $\alpha = 99.5\%$ and $\phi = 5\%$, we find that the threshold value is -6. For $\bar{Z}_2$ we use a Student-t with 5 degrees of freedom distributions. In this case the threshold is -1.2.

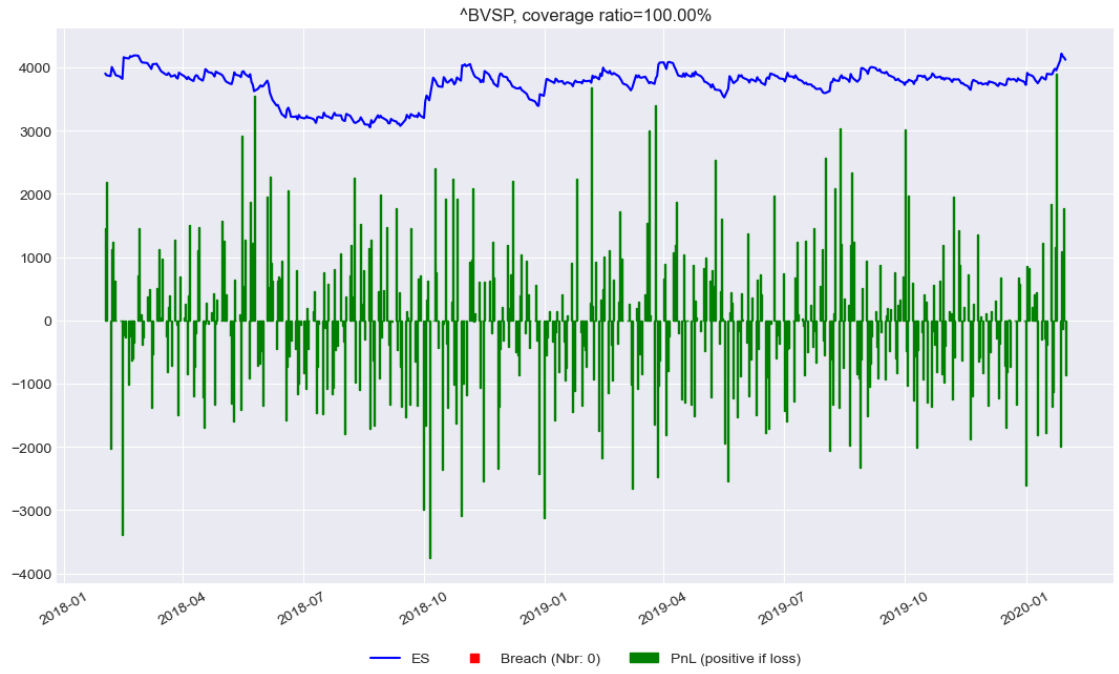
	G	Threshold	Accepted
AXJO	-9	-6	No
BVSP	0	-6	Yes
FCHI	-3	-6	Yes
GDAXI	-2	-6	Yes
GSPC	-9	-6	No
GSPTSE	-3	-6	Yes
KS11	-7	-6	No
MXX	-8	-6	No
SSMI	-3	-6	Yes
TWII	-12	-6	No

Table 9: Accepted ES for G

	$\bar{Z}_2$	Threshold	Accepted
AXJO	-2.5397	-1.2	No
BVSP	-1.67839	-1.2	No
FCHI	-0.527363	-1.2	Yes
GDAXI	-0.458829	-1.2	Yes
GSPC	-2.1588	-1.2	No
GSPTSE	-0.160738	-1.2	Yes
KS11	-1.77876	-1.2	No
MXX	-1.52592	-1.2	No
SSMI	-0.481356	-1.2	Yes
TWII	-1.39444	-1.2	No

Table 10: Accepted ES for $\bar{Z}_2$

Of course, the values of the statistics are the same as in the previous tests but the output results regarding the acceptance of the ES are different. We can see that in this case, both statistics become more conservative and that $\bar{Z}_2$ seems more conservative than G, because the calculated ES for the Brazil Stock Market Index is accepted by G and rejected by $\bar{Z}_2$. However, we can see from the following graph that there are no breaches in the portfolio.



The reason why $\bar{Z}_2$ rejects the computed ES is that this statistic is very sensitive to VaR misspecifications. Since the number of breaches for the computed VaR amounts to 8, then there is a rejection, even if there should not be. Then G gives better results regarding the acceptance of $\hat{ES}^{\alpha}$.

4.4 Conclusion

To sum up, we found that the best statistic from the theoretical and numeric point of view is $\bar{Z}_{\text{MB}}$ since it is the most conservative one as it correctly accepts or rejects the computed values of ES and it is not influenced by VaR misspecifications. However, the evaluation of the threshold requires the storage of the historical simulations used to calculate the ES and this slows down computations. The time required for the evaluation of the threshold is in any case of some seconds so the additional storage is not computationally demanding.

If one prefers not having to deal with the storage of the PnLs simulations, both the $\bar{Z}_2$ and G statistics can be used. From a practical point of view, G gives better results (the same results as $\bar{Z}_{\text{MB}}$) on the portfolios that we tested, although it lacks some theoretical justifications.

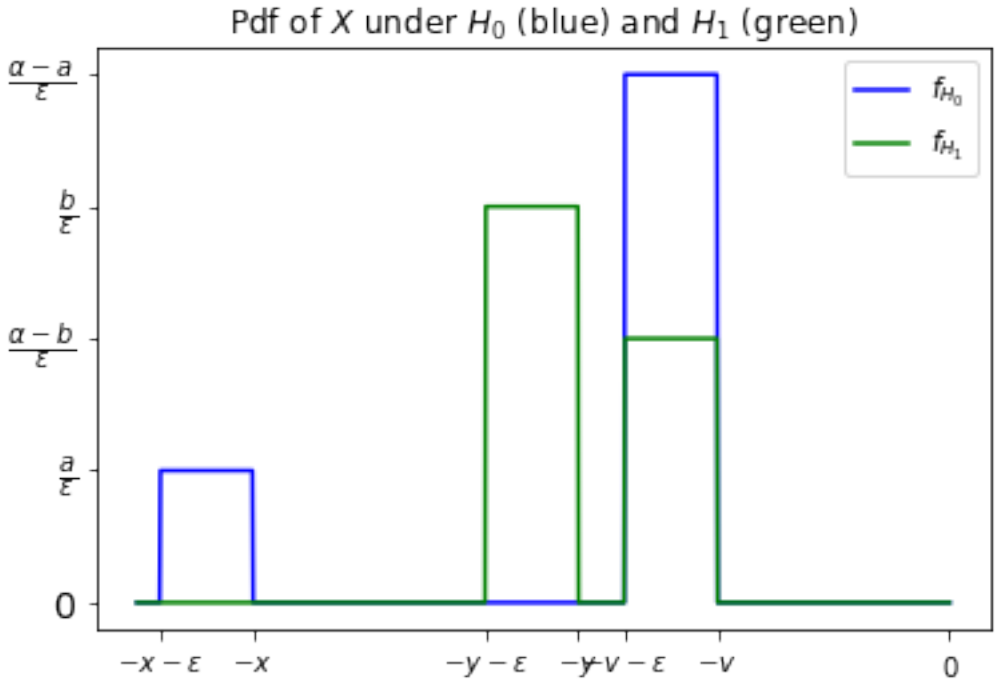
5. Appendix I

5.1 Moldenhauer and Pitera counterexamples

5.1.1 Example 1: $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$ does not imply $\hat{ES}^\alpha < ES^\alpha$

In this example we will prove that $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$ does not imply that we are underestimating the ES or in other words that $E\hat{S}^\alpha < ES^\alpha$.

Consider in particular a toy example with $N = 1$. In the following we plot the pdf of the unique X under the null hypothesis, denoted by f_{H_0}, and under the alternative hypothesis, denoted by f_{H_1}. We consider only the part regarding extreme losses of X, the distribution for $X > v$ can be arbitrarily chosen.



Then the VaRs under H_0 and H_1 are both equal to v. The ES under H_0 is

$$\hat{ES}^\alpha = -\frac{\frac{a}{\varepsilon}\frac{x^2}{2}\Big|_{-x-\varepsilon}^{-x} + \frac{\alpha-a}{\varepsilon}\frac{x^2}{2}\Big|_{-v-\varepsilon}^{-v}}{\alpha} = \frac{\frac{\alpha}{2}\varepsilon + (ax + (\alpha-a)v)}{\alpha}$$

and similarly the ES under H_1 is

$$ES^\alpha = \frac{\frac{\alpha}{2}\varepsilon + (by + (\alpha-b)v)}{\alpha}.$$

Note that $\hat{ES}^\alpha > v$ iff $\frac{\alpha}{2}\varepsilon + (ax + (\alpha-a)v) > \alpha v$ iff $\frac{\alpha}{2}\varepsilon + a(x-v) > 0$ which holds true.

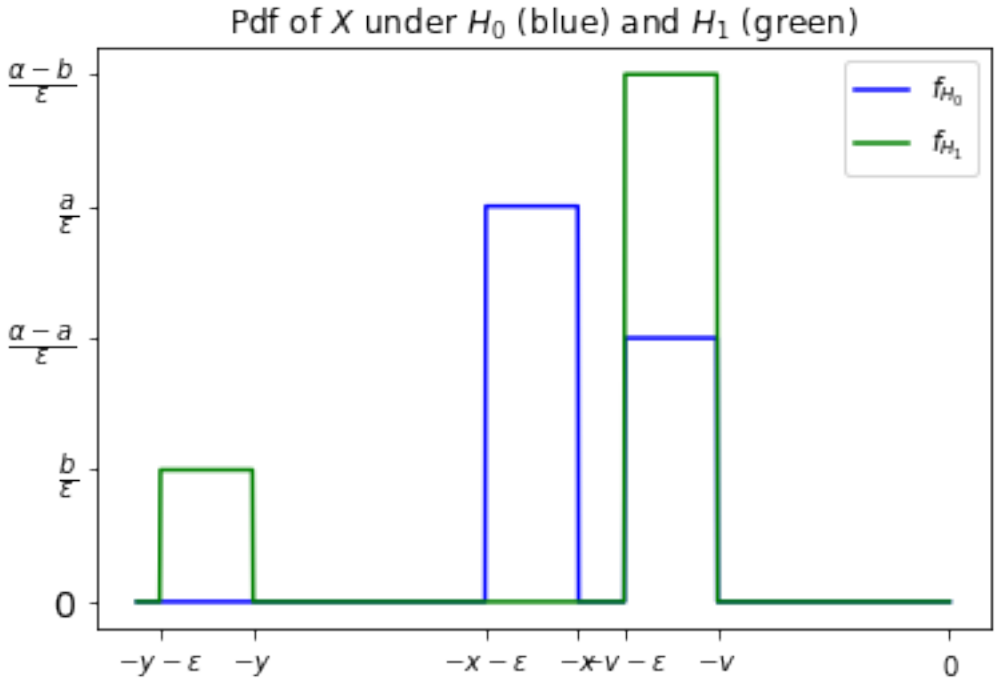
Set $\varepsilon < \frac{2a(\alpha-b)}{\alpha b}(x-v)$. We have $\hat{ES}^\alpha < x$ iff $\frac{\alpha}{2}\varepsilon + (ax + (\alpha-a)v) < \alpha x$ iff $\varepsilon < \frac{2(\alpha-a)}{\alpha}(x-v)$ and this is true for the chosen ε since $a < b$. We can then choose $y = \hat{ES}^\alpha$. In this way $\hat{ES}^\alpha > ES^\alpha$, iff $ax + (\alpha-a)v > \frac{b}{\alpha}(\frac{\alpha}{2}\varepsilon + (ax + (\alpha-a)v)) + (\alpha-b)v$ iff $a(\alpha-b)(x-v) > \frac{\alpha b}{2}\varepsilon$ which is true for the chosen ε.

This means that $ES^\alpha < \hat{ES}^\alpha$ so that we are overestimating the real ES. Let us see what happens to the statitic $G = \mathbb{1}_{X+ES^\alpha<0}$. We have, $E_{H_1}[G(X, E\hat{S}^\alpha)] = P_{H_1}(X + E\hat{S}^\alpha < 0) = b$ while $E_{H_0}[G(X, E\hat{S}^\alpha)] = P_{H_0}(X + E\hat{S}^\alpha < 0) = a$ so $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$, even if we are overestimating the ES.

5.1.2 Example 2: $E\hat{S}^\alpha < ES^\alpha$ does not imply $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$

On the other hand, we can construct an example which shows that the very general hyphothesis $H_1 : ES_d^\alpha \geq E\hat{S}_d^\alpha \; \forall d \wedge \exists d : ES_d^\alpha > E\hat{S}_d^\alpha$ does not imply that $E_{H_1}[G(X, E\hat{S}^\alpha)] > E_{H_0}[G(X, E\hat{S}^\alpha)]$.

As before, we take $N = 1$ and we plot the tail pdfs under the null and the alternative hypothesis:



As before, we have $\hat{ES}^\alpha = \frac{\frac{\alpha}{2}\varepsilon + (ax + (\alpha-a)v)}{\alpha}$ and $ES^\alpha = \frac{\frac{\alpha}{2}\varepsilon + (by + (\alpha-b)v)}{\alpha}$. For $\varepsilon < \frac{2b(\alpha-b)}{\alpha a}(y-v)$, it holds $v < ES^\alpha < y$ so we can set $x = ES^\alpha$. In this way it can be shown that $\hat{ES}^\alpha < ES^\alpha$ and $E_{H_1}[G(X, E\hat{S}^\alpha)] < E_{H_0}[G(X, E\hat{S}^\alpha)]$, which was our aim.

5.2 Moldenhauer and Pitera alternative hypothesis

For the G statistic, let us consider the same hypothesis used for $\bar{Z}_3$, which are:

$$H_0 : F_d = P_d \ \forall d$$
$$H_1 : F_d \preceq P_d \ \forall d \wedge \exists d : F_d \prec P_d$$

where P_d is the predicted distribution of X_d used to evaluate the ES and F_d is the real unknown distribution of X_d. In particular, for every non-increasing function u, we have $E_{P_d}[u(X_d)] \leq E_{F_d}[u(X_d)]$.

Let us consider the function $\mathbb{1}_{((X+\hat{ES}^\alpha)_{[1]}+\cdots+(X+\hat{ES}^\alpha)_{[k]}<0)}$. We prove that it is non-increasing as a function of each X_i. We have that the function $Y \to \mathbb{1}_{(Y<0)}$ is non-increasing. Call f the function $f(X_i) = (X + \hat{ES}^\alpha)_{[1]} + \cdots + (X + \hat{ES}^\alpha)_{[k]}$ where X_d is fixed for $d \neq i$. Then, it is enough to prove that f is increasing or equivalently that $f(X_i) < f(X_i + \Delta X_i)$ for every ΔX_i). Let us suppose to increase X_i to $X_i + \Delta X_i$. It follows that

- if $X_i + \hat{ES}_i^\alpha \leq (X + \hat{ES}^\alpha)_{[k]}$ and $X_i + \Delta X_i + \hat{ES}_i^\alpha \leq (X + \hat{ES}^\alpha)_{[k+1]}$, then $f(X_i + \Delta X_i) = f(X_i) + \Delta X_i > f(X_i)$;
- if $X_i + \hat{ES}_i^\alpha \leq (X + \hat{ES}^\alpha)_{[k]}$ and $X_i + \Delta X_i + \hat{ES}_i^\alpha > (X + \hat{ES}^\alpha)_{[k+1]}$, then $f(X_i + \Delta X_i) = f(X_i) - (X_i + \hat{ES}_i^\alpha) + (X + \hat{ES}^\alpha)_{[k+1]} > f(X_i)$ since $X_i + \hat{ES}_i^\alpha < (X + \hat{ES}^\alpha)_{[k+1]}$;
- if $X_i + \hat{ES}_i^\alpha > (X + \hat{ES}^\alpha)_{[k]}$, then also $X_i + \Delta X_i + \hat{ES}_i^\alpha > (X + \hat{ES}^\alpha)_{[k]}$ and $f(X_i + \Delta X_i) = f(X_i)$.

So f is an increasing function and $X_i \to \mathbb{1}_{(f(X_i)<0)}$ is decreasing for every $i = 1, \ldots, N$. We also recall that the expected value of a decreasing function is still a decreasing function.

Applying Fubini's Theorem and sequentially using the fact that $F_d \preceq P_d$, we have

$$\begin{aligned} E_{H_0}[\mathbb{1}_{((X+\hat{ES}^\alpha)_{[1]}+\cdots+(X+\hat{ES}^\alpha)_{[k]}<0)}] &= E_{P_1}[E_{P_2}[\ldots E_{P_N}[\mathbb{1}_{((X+\hat{ES}^\alpha)_{[1]}+\cdots+(X+\hat{ES}^\alpha)_{[k]}<0)}]]] \\ &\leq E_{F_1}[E_{F_2}[\ldots E_{F_N}[\mathbb{1}_{((X+\hat{ES}^\alpha)_{[1]}+\cdots+(X+\hat{ES}^\alpha)_{[k]}<0)}]]] \\ &= E_{H_1}[\mathbb{1}_{((X+\hat{ES}^\alpha)_{[1]}+\cdots+(X+\hat{ES}^\alpha)_{[k]}<0)}]. \end{aligned}$$

From this, it follows that $E_{H_0}[G(X, \hat{ES}^\alpha)] < E_{H_1}[G(X, \hat{ES}^\alpha)]$ where the inequality is strict since $F_d \prec P_d$ for some d.

References

[1] Carlo Acerbi and Balázs Székely. Backtesting Expected Shortfall. *RISK Magazine*, December 2014.

[2] Carlo Acerbi and Balázs Székely. General properties of backtestable statistics. *Available at SSRN 2905109*, 2017.

[3] Carlo Acerbi and Balázs Székely. The minimally biased backtest for es. *Risk. net*, 29, 2019.

[4] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Thinking coherently. *Risk*, pages 68–71, 1997.

[5] Tobias Fissler, Johanna F Ziegel, and Tilmann Gneiting. Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*, 2015.

[6] Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[7] Felix Moldenhauer and Marcin Pitera. Backtesting expected shortfall: a simple recipe? *Journal of Risk*, 22(1), 2019.