Backtesting Expected Shortfall

Marcin Pitera

Institute of Mathematics,
Faculty of Mathematics and Computer Science,
Jagiellonian University,
Krakow, Poland

Krakow QuantHub

http://www.im.uj.edu.pl/quanthub

January 14, 2020

based on a joint work with F. Moldenhauer

Outline

Introduction
- A few words about backtesting
- General notation

VaR backtesting
- Regulatory backtesting
- Backtesting statistic as performance measure

ES backtesting
- Definition of ES backtest
- Financial intuition & practical computations
- Embedding ES backtests into regulations
- Traffic-light table and backtesting statistic distribution

Mathematical justification of ES backtest: dual-link quick summary

Live implementation demo

Introduction

Introduction

A few words about backtesting

- The are many backtesting techniques with different purposes, e.g.
  - **Conditional coverage backtests**, where one assess independence of observed breaches. This might be formally quantified e.g. by Kupiec Test but is typically done through visual inspection.
  - **General adequacy point-forecast backtests**, where realised values are compared with theoretical risks with two-way penalisation. Those are typically formulated in elicitability-based framework.[1]
  - **Density-forecast backtests**, where, one considers the adequacy of the whole projected distribution of the $\mathrm{P\&L}$ moves. Most backtests rely on the framework introduced in Berkowitz (2001); they are sometimes called *p*-value backtests. Typically, they use PIT transforms with realised quantiles as input.
- The backtesting framework we introduce is focused on assessing **point-forecast conservativeness**, where realised portfolio $\mathrm{P\&Ls}$ are compared with projected capital reserves. Following regulatory guidelines, we focus on so called **unconditional coverage** backtesting.

[1] these backtest penalise capital overestimation. This is not aligned with standard regulatory approach: the VaR model with zero-breaches is still classified into the green zone

Introduction

Before we begin...

- Currently, there are **no ES backtesting market standards**. In FRTB, despite VaR 1% to ES 2.5% shift, VaR's cumulative exception (breach) rate is used for backtesting.
- During the discussions, a lot of (academia) focus has been set on **elicitability** and related backtests. Recall that those are **not aligned** with **regulatory framework** (as they assess overall fit rather than conservativeness).
- We propose a **novel framework** based on a **dual link** between risk and performance measures, the same one that is used to define VAR.
- First, we will define the **backtest** statistic for VaR and ES, and discuss its **rationality**. Next, we will show how easy it is to **implement** it, and then focus on it's underlying **mathematical properties**.

Introduction

General notation

- We use ρ to denote a law-invariant risk measure (VAR or ES) and $\mathrm{P\&L}$ to denote a random variable associated with the future portfolio profits and losses.
- For transparency, we set the holding period to 1-day.
- We assume that we have an Internal Model (IM) that is used to compute/estimate the capital reserve to protect against fluctuations in the future value of a financial portfolio; this could refer to Historical Simulation, Gaussian, or Monte Carlo risk estimation models.
- For day i we use $\mathrm{P\&L}_i$ to denote the ith day realised portfolio $\mathrm{P\&L}$ and $\hat{\rho}_i$ to denote the projected (estimated) capital reserve for that portfolio that was computed using historical data (up to day $i-1$) combined with the IM methodology,

Introduction

Notation (backtest input)

- For a pre-specified period n (typically one year data, i.e. $n = 250$) the regulatory backtesting is focused on assessing IM methodology conservativeness based on two main inputs:
  1. **Realised portfolio P&Ls**: $(\mathrm{P\&L}_i)_{i=1}^n$;
  2. **Projected capital reserves for the porfolio**: $(\hat{\rho}_i)_{i=1}^n$.
- To ease the notation, for $(i = 1, 2, \ldots, n)$, we use $y_i := P\&L_i + \hat{\rho}_i$ to denote a **secured position sample**.

Normalization: If the portfolio profile or market volatility is changing (through time) one could introduce additional normalisation scheme and consider modified secured position sample $(\widetilde{y}_i)$ given by

$$\widetilde{y}_i := \frac{P\&L_i + \hat{\rho}_i}{\hat{\rho}_i} = \frac{P\&L_i}{\hat{\rho}_i} + 1.$$

For positively homogeneous risk measure, we scale our portfolio so that the estimated risk is equal to one. Note that this is aligned with regulatory expectations and linked e.g. to so called **Loss Overshooting Ratios** as defined in TRIM [Guideline 93(c)].[a]

[a] note this transformation does not impact breach count for VaR breach test.

VaR backtesting

Regulatory (breach) backtest

VAR exception rate statistic
Given secured sample $y_i = P\&L_i + \hat{\rho}_i$ and assuming the total number of backtesting days is n, the VaR exception rate backtesting statistic is given by

$$T_n := \frac{1}{n} \sum_{i=1}^{n} 1_{\{y_i < 0\}}.$$

- We count the number of capital breaches (exceptions) in the sample to quantify IM methodology performance. We divide by n to get *exception rate*.
- In **regulatory** backtesting the window length is fixed ($n = 250$) and a nominal number of breaches (nT_n) is used.
- For VaR at level 1% the model is said to be in:
  - **green zone**, if there are less then 5 breaches: this corresponds to $T_n \in [0.00, 0.02)$;
  - **yellow zone**, if the number of breaches is 5 to 9: this corresponds to $T_n \in [0.02, 0.04)$;
  - **red zone**, if there are 10 or more breaches: this corresponds to $T_n \in [0.04, 1.00]$.

Regulatory (breach) backtest

$$T_n := \frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i<0\}}.$$

- Test statistic T_n is used by the regulator to **quantify the performance** (conservativeness) of the *secured sample*. Counting the number of breaches is very tightly connected to VAR measurement philosophy.
- Indeed, we can rewrite T_n as

$$T_n = \inf\{\alpha \in (0,1] : \hat{\mathrm{V@R}}_\alpha(y) \leq 0\}, \qquad (1)$$

where $\hat{\mathrm{V@R}}_\alpha(y) := -y_{(\lfloor n\alpha \rfloor + 1)}$ is the family of empirical VAR estimators.

- From (1) we see that T_n is a natural estimator of performance measure dual to the VAR family; see e.g. [Cherny and Madan, 2009].
- We look for the minimal confidence level α which makes the secured position sample acceptable (i.e. having non-positive risk); traffic-light approach (with 0.02 and 0.04 thresholds) allows some variability due to various biases, misspecifications, etc.
- Note that in (1) we use empirical VAR estimator to quantify the performance of secured position. IM methodology is used to construct y, and does not impact definition of T_n.

ES backtesting

Definition of ES backtesting statistic

ES cumulative exception rate statistic
Given secured sample $y_i = P\&L_i + \hat{\rho}_i$ and assuming the total number of backtesting days is n, the ES cumulative exception rate backtesting statistic is given by

$$G_n := \frac{1}{n} \sum_{i=1}^{n} 1_{\{y_{(1)} + \ldots + y_{(i)} < 0\}},$$

where $y_{(k)}$ is the kth order statistics of (y_i).

- We look for the smallest number of worst realisations of the secured position that add up to a positive total, and then we divide the outcome by n. Indeed, we can rewrite G_n as

$$G_n = \frac{1}{n} \inf \left\{ k \in \mathbb{N} : \sum_{i=1}^{k+1} y_{(i)} \geq 0 \right\},$$

- Alternatively speaking, We look for the smallest (possible) number of scenarios so that **the aggregated risk reserve** is sufficient to cover **the aggregated loss**.

Intuitive financial interpretation

VaR framework: we identify the minimal index i, when the capital reserve is sufficient to cover losses, i.e when we get

$$-P\&L_{(i)} < \hat{\rho}_{(i)}.$$

The value $(i-1)$ defines T_n.

ES framework: we identify the minimal index k, when the **aggregated** capital reserve is sufficient to cover **aggregated** losses, i.e. when we get

$$-\sum_{i=1}^{k} P\&L_{(i)} < \sum_{i=1}^{k} \hat{\rho}_{(i)}.$$

The value $(k-1)$ determines defines G_n.

- This change of paradigm is very natural to VaR to ES migration. One needs to consider conditional scenario sums instead of single scenarios.

Practical computations

$$G_n = \frac{1}{n} \sum_{i=1}^{n} 1_{\{y_{(1)}+\ldots+y_{(i)}<0\}},$$

- It is very easy to implement the backtest and compute G_n:

1. Given inputs $(P\&L_i)$ and (y_i) we construct secured position
$$y_i = P\&L_i + \hat{\rho}_i.$$
2. We sort the values and produce a cumulative sum vector
$$(y_{(1)},\ y_{(1)} + y_{(2)},\ y_{(1)} + y_{(2)} + y_{(3)},\ \ldots).$$
3. We identify the smallest number of scenarios such that the aggregated secured position is non-negative, i.e. find minimal $k \in \mathbb{N}$ for which we get
$$y_{(1)} + \ldots + y_{(k)} \geq 0.$$
4. We set $G_n = \frac{k-1}{n}$.

- $n \cdot G_n$ could be computed in one line of R code: `sum(cumsum(sort(y))<0)`.
- $n \cdot T_n$ could be computed using R code: `sum(sort(y)<0)`.

Embedding backtests into regulations

- One can easily embed ES backtest into regulations. The language could be simplified to streamline the financial interpretation.
- To be consistent with VaR framework we keep backtesting period fixed ($n = 250$) and consider nominal number of scenarios ($n \cdot G_n$).
- For ES at level 2.5% and $n = 250$ the model is said to be in:
  - **green zone**, if the sum of the 12 smallest values of y is positive: this corresponds to $G_n \in [0.00, 0.05)$
  - **yellow zone**, if the sum of the 12 smallest values of y is negative but the sum of the 25 smallest values of y is positive: this corresponds to $G_n \in [0.05, 0.10)$;
  - **red zone**, if the sum of the 25 smallest values of y is negative: this corresponds to $G_n \in [0.10, 1.00]$.
- Note that $n \cdot G_n$ could be used to easily define multiplicative penalisation add-on as in VAR framework.
- Our choice of threshold is robust and consistent with the old framework. Why?

Rationale behind threshold choice

The consistency is achieved on multiple layers:

1. **First**, thresholds are obtained by multiplying base levels by 2 and 4:
   - VAR: base level is 1%, yellow zone threshold is 2%, and red zone threshold is 4%.
   - ES: base level is 2.5%, yellow zone threshold is 5%, and red zone threshold is 10%.
2. **Second**, assuming normal model, the theoretical thresholds correspond to approximately the same capital reserve values. Let $X \sim N(0,1)$. Then:
   - $\mathrm{V@R}_{1\%}(X) = 2.33$, $\mathrm{V@R}_{2\%}(X) = 2.05$, $\mathrm{V@R}_{4\%}(X) = 1.75$.
   - $ES_{2.5\%}(X) = 2.34$, $ES_{5\%}(X) = 2.06$, $ES_{10\%}(X) = 1.75$.
3. **Third**, proposed thresholds lead to statistical framework that is aligned with VAR backtest. Even under extreme specification imposed on the null distribution the proposed thresholds lead to statistical confidence thresholds close to 95% and 99.99%. The thresholds could be considered as (almost) model-independent.
4. **Fourth**, we will show later that the ES thresholds could be treated as maximal acceptable risk level misspecification thresholds as in VaR framework. In fact, G_n is a performance measure dual to the ES risk measure family in a same way that T_n is a performance measure dual to the VaR family. Arguably, this is the most important point as it shows **very deep mathematical link** between G_n and T_n, and both backtests.

Rationale behind threshold choice

The consistency is achieved on multiple layers:

1. **First**, thresholds are obtained by multiplying base levels by 2 and 4:
   - VAR: base level is 1%, yellow zone threshold is 2%, and red zone threshold is 4%.
   - ES: base level is 2.5%, yellow zone threshold is 5%, and red zone threshold is 10%.
2. **Second**, assuming normal model, the theoretical thresholds correspond to approximately the same capital reserve values. Let $X \sim N(0,1)$. Then:
   - $\mathrm{V@R}_{1\%}(X) = 2.33$, $\mathrm{V@R}_{2\%}(X) = 2.05$, $\mathrm{V@R}_{4\%}(X) = 1.75$.
   - $ES_{2.5\%}(X) = 2.34$, $ES_{5\%}(X) = 2.06$, $ES_{10\%}(X) = 1.75$.
3. **Third**, proposed thresholds lead to statistical framework that is aligned with VAR backtest. Even under extreme specification imposed on the null distribution the proposed thresholds lead to statistical confidence thresholds close to 95% and 99.99%. The thresholds could be considered as (almost) model-independent.
4. **Fourth**, we will show later that the ES thresholds could be treated as maximal acceptable risk level misspecification thresholds as in VaR framework. In fact, G_n is a performance measure dual to the ES risk measure family in a same way that T_n is a performance measure dual to the VaR family. Arguably, this is the most important point as it shows **very deep mathematical link** between G_n and T_n, and both backtests.

Rationale behind threshold choice

The consistency is achieved on multiple layers:

1. **First**, thresholds are obtained by multiplying base levels by 2 and 4:
   - VAR: base level is 1%, yellow zone threshold is 2%, and red zone threshold is 4%.
   - ES: base level is 2.5%, yellow zone threshold is 5%, and red zone threshold is 10%.
2. **Second**, assuming normal model, the theoretical thresholds correspond to approximately the same capital reserve values. Let $X \sim N(0,1)$. Then:
   - $\mathrm{V@R}_{1\%}(X) = 2.33$, $\mathrm{V@R}_{2\%}(X) = 2.05$, $\mathrm{V@R}_{4\%}(X) = 1.75$.
   - $ES_{2.5\%}(X) = 2.34$, $ES_{5\%}(X) = 2.06$, $ES_{10\%}(X) = 1.75$.
3. **Third**, proposed thresholds lead to statistical framework that is aligned with VAR backtest. Even under extreme specification imposed on the null distribution the proposed thresholds lead to statistical confidence thresholds close to 95% and 99.99%. The thresholds could be considered as (almost) model-independent.
4. **Fourth**, we will show later that the ES thresholds could be treated as maximal acceptable risk level misspecification thresholds as in VaR framework. In fact, G_n is a performance measure dual to the ES risk measure family in a same way that T_n is a performance measure dual to the VaR family. Arguably, this is the most important point as it shows **very deep mathematical link** between G_n and T_n, and both backtests.

Rationale behind threshold choice

The consistency is achieved on multiple layers:

1. **First**, thresholds are obtained by multiplying base levels by 2 and 4:
   - VAR: base level is 1%, yellow zone threshold is 2%, and red zone threshold is 4%.
   - ES: base level is 2.5%, yellow zone threshold is 5%, and red zone threshold is 10%.
2. **Second**, assuming normal model, the theoretical thresholds correspond to approximately the same capital reserve values. Let $X \sim N(0,1)$. Then:
   - $\mathrm{V@R}_{1\%}(X) = 2.33$, $\mathrm{V@R}_{2\%}(X) = 2.05$, $\mathrm{V@R}_{4\%}(X) = 1.75$.
   - $ES_{2.5\%}(X) = 2.34$, $ES_{5\%}(X) = 2.06$, $ES_{10\%}(X) = 1.75$.
3. **Third**, proposed thresholds lead to statistical framework that is aligned with VAR backtest. Even under extreme specification imposed on the null distribution the proposed thresholds lead to statistical confidence thresholds close to 95% and 99.99%. The thresholds could be considered as (almost) model-independent.
4. **Fourth**, we will show later that the ES thresholds could be treated as maximal acceptable risk level misspecification thresholds as in VaR framework. In fact, G_n is a performance measure dual to the ES risk measure family in a same way that T_n is a performance measure dual to the VaR family. Arguably, this is the most important point as it shows **very deep mathematical link** between G_n and T_n, and both backtests.

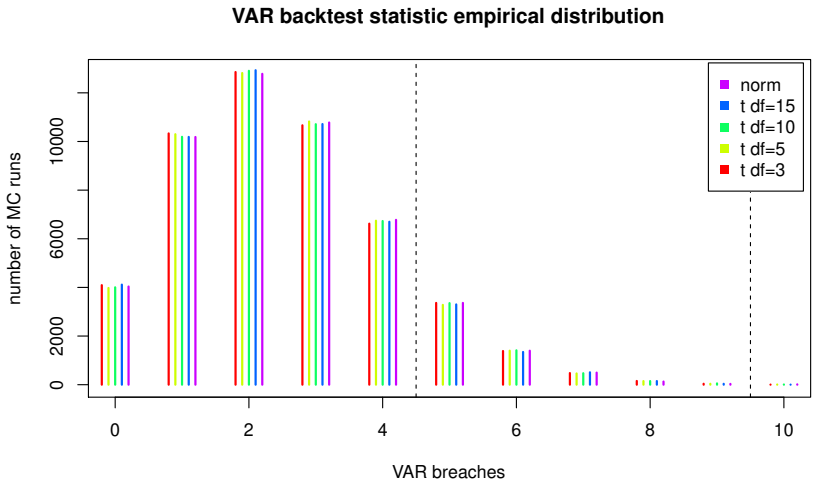
Summary: traffic-light table and p-values

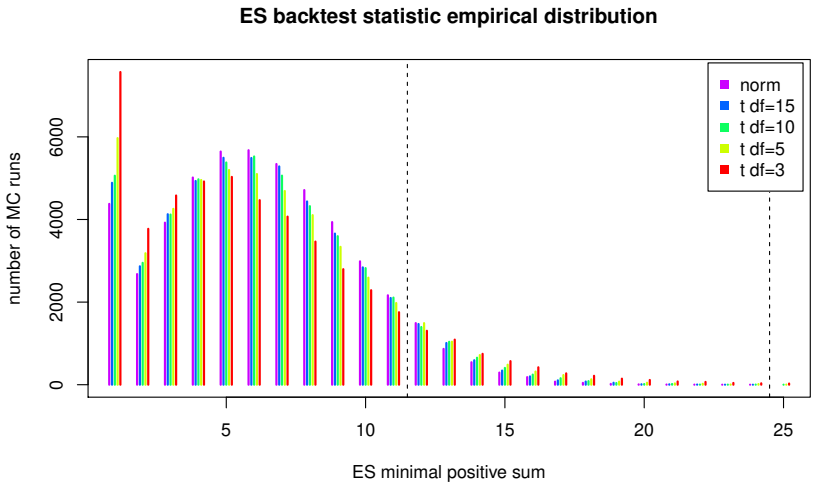
Zone (color)	VAR (number of exceptions)	ES (worst-case scenarios with negative sum)
Green	0–4	0–11
Yellow	5–9	12–24
Red	10+	25+

Risk metric		VAR				ES			
Statistic value		4	5	9	10	11	12	24	25
t-student	$\nu = 3$	0.8914	0.9586	0.9998	0.9999	0.8944	0.9205	0.9967	0.9973
	$\nu = 5$	0.8931	0.9588	0.9998	1.0000	0.9074	0.9372	0.9998	0.9999
	$\nu = 10$	0.8909	0.9580	0.9998	1.0000	0.9185	0.9464	1.0000	1.0000
	$\nu = 15$	0.8930	0.9590	0.9999	1.0000	0.9224	0.9518	1.0000	1.0000
normal		0.8913	0.9585	0.9997	1.0000	0.9292	0.9591	1.0000	1.0000

The second table presents the cumulative (empirical) distribution values of the nominal backtesting statistics for VAR and ES, for large samples from various pre-defined distributions. The distribution of VAR test statistics correspond to Bernoulli distribution with $p = 0.99$, and the theoretical threshold values are 0.9588, and 0.9999 (for 5 and 10 exceeds, respectively). One can see that ES backtesting statistic is stable even in extreme conditions (for $\nu = 3$) and the cumulative probability for the thresholds is comparable to VAR (for values 12 and 25). The values were obtained using a 50 000 Monte Carlo run.

Backtest statistic distribution under various distributions





Empirical probability mass functions of the nominal VAR backtest statistics $n \cdot T_n$ (left) and nominal ES backtest statistic $n \cdot G_n$ (right) for $n = 250$. We consider five different a priori distributions, and construct the secured samples using true risk capital reserve add-ons. Note that the probability mass function for VAR corresponds to Bernoulli probability mass function with $p = 0.99$. We can see that ES backtest statistic is remarkably stable under extreme conditions, i.e under t-student distribution with ? = 3 degrees of freedom. Dotted lines indicate the proposed traffic-light thresholds. The values were obtained using a 50 000 Monte Carlo run.

Mathematical justification of ES backtest: dual-link quick summary

Mathematical justification of ES backtest: dual-link quick summary

- We already shown a lot of arguments why our backtest is a solid choice.
- Now, we want to quickly explain why we believe that our backtest is the best choice.
- The argument is based on the dual-link theorem that was established in [Cherny and Madan, 2009, Theorem 1] as which links the family of risk measures to a *performance measure* (acceptability index).

Duality between risk measures and performance measures

Theorem [Cherny and Madan, 2009, Theorem 1]
A map ϕ is a *regular* (scale-invariant) performance measure if and only if there exists a family of (coherent) risk measures $(\rho_\alpha)_{\alpha\in\mathbb{R}_+}$ decreasing in α such that

$$\phi(X) = \inf\{\alpha \in \mathbb{R}\colon \rho_\alpha(X) \leq 0\}, \quad X \in \mathcal{X}.$$

Remarks

- We look for the minimal value of α which makes the position X acceptable (i.e. having non-positive risk).
- Typically, we assume that the parameter space is $\mathbb{R}_+$ but it could be different. For example, for VaR it is more reasonable to use $(0, 1)$ instead. One can apply some standard (parameter) distortion function to recover one from another (e.g. $g(x) = 1/(1 + x)$).
- If family of VAR maps is so popular if finance, where we can find the corresponding dual acceptability index?

Duality between risk measures and performance measures

- As already pointed out, it is exactly T_n backtest statistic, i.e. it's empirical estimator. Indeed, we have:

$$T_n = \inf\{\alpha \in (0,1] : \hat{\mathrm{V@R}}_\alpha(y) \leq 0\},$$

- I hope you know now how we obtained our ES backtesting statistic G_n?
- In exactly the same way! We get

$$G_n = \inf\{\alpha \in (0,1] : \hat{\mathrm{ES}}_\alpha(y) \leq 0\},$$

where $\hat{\mathrm{ES}}$ is the empirical estimator of ES.

- G_n is dual to the family of empirical ES estimators in a same way that T_n is dual to the family of empirical VAR estimators. If we want to be consistent, this is the way.

Duality between risk measures and performance measures

- As already pointed out, it is exactly T_n backtest statistic, i.e. it's empirical estimator. Indeed, we have:

$$T_n = \inf\{\alpha \in (0,1] : \hat{\mathrm{V@R}}_\alpha(y) \leq 0\},$$

- I hope you know now how we obtained our ES backtesting statistic G_n?
- In exactly the same way! We get

$$G_n = \inf\{\alpha \in (0,1] : \hat{\mathrm{ES}}_\alpha(y) \leq 0\},$$

where $\hat{\mathrm{ES}}$ is the empirical estimator of ES.

- G_n is dual to the family of empirical ES estimators in a same way that T_n is dual to the family of empirical VAR estimators. If we want to be consistent, this is the way.

References I

Bielecki, T. R., Cialenco, I., Pitera, M., and Schmidt, T. (2019).

Fair Estimation of Capital Risk Allocation.
Forthcoming in *Statistics & Risk Modeling*.

Cherny, A. S. and Madan, D. B. (2009).

New measures for performance evaluation.
The Review of Financial Studies, 22(7):2571–2606.

Moldenhauer, F. and Pitera, M. (2017).

Backtesting expected shortfall: a simple recipe
Journal of Risk, 22(1):17–42.

Pitera, M. and Schmidt, T. (2018).

Unbiased estimation of risk.
Journal of Banking & Finance, 91:133–145.

Live implementation demo

Now, using **R**, we are going to perform the live implementation demo. We will do the following:

1. Assuming we get $(P\&L_i)$ and $(\hat{\rho}_i)$ samples we will compute both T_n and G_n.
2. We will perform the backtest on simulated data for various setting (e.g. risk underestimated by 10% or 20%).
3. We will compare T_n outcomes to G_n outcomes to see the consistency between those frameworks.
4. We will compute the distribution of G_n under various null-distributions.

Let us switch to **R**.

The End

Thank you!