ADAPTIVE BACKTEST FOR EXPECTED SHORTFALL

October 2017

msci.com

MSCI

EXECUTIVE SUMMARY

Carlo Acerbi
Balazs Szekely

Expected Shortfall (ES) replaced Value at Risk (VaR) under the Basel Committee's Minimum Capital Requirements for Market Risk in 2016, also known as the Fundamental Review of the Trading Book (FRTB). However, whether ES can be backtested (thus validating ES-based models) is still an open question. This issue is critical as regulation must be based on a testable model.

Backtesting is an "apples to oranges" comparison, as it entails checking predicted risk against realized returns on a daily basis, as in the standard "exceedances counting" VaR backtest. We have recently shown that ES can be backtested but only approximately, because any ES backtest is inevitably sensitive also to VaR predictions. This paper proposes a new backtest that minimizes such sensitivity to VaR. The bias is small and prudent, qualifying the backtest as an appropriate validation tool for ES-based models, both for financial risk management and regulation.

We find the ES backtest superior to the VaR backtest in some respects. The ES backtest estimates not only the model validation probability, but also measures the size of prediction discrepancy. It automatically generates Basel capital multiplier factors, in a model independent way. At a broader level, the backtest provides constructive feedback for self-adaptive risk predictive models. The backtest can in fact be expressed as an apples-to-apples comparison between predicted ES and a new emerging notion of "realized ES," which per se opens the way to a variety of applications, such as margin methodologies and model-free risk disclosure. These applications can be used in all segments of the financial industry — not only banks.

Adaptive Backtest for Expected Shortfall

Carlo Acerbi[*] *and Balazs Szekely*[†]

MSCI Inc.

October 16, 2017

Abstract

Recent results have shown that a backtest of Expected Shortfall (**ES**) is unavoidably affected by sensitivity to predictions of Value at Risk (**VaR**). We introduce the backtest for **ES** that minimizes such sensitivity, which is also of prudential nature: any imperfect **VaR** prediction results in a more punitive test against **ES** and the effect is negligible for small **VaR** discrepancies. This qualifies the backtest as an appropriate validation tool for **ES**–based models, notably within Basel regulation, where this is still an open question. As an important byproduct, the **ES** backtest, as opposed to the common **VaR** backtest, estimates not only model acceptance probability, but also the prediction discrepancy magnitude. A trailing backtest acts as a running indicator of the **ES** prediction error. In Basel jargon, this means that the backtest automatically measures portfolio–specific capital multipliers and provides the realized **ES** of a portfolio. In addition to banking regulatory applications, these features open new perspectives for internal risk management, such as the creation of self–adaptive risk forecast models.

[*] carlo.acerbi@msci.com
[†] balazs.szekely@msci.com

1 Apples to apples

The job of risk models is to predict risk. However, Risk, as opposed to return, is not observable a posteriori. This poses a model validation challenge: how can one check if risk predictions were correct? Backtesting is a consolidated industry practice exploiting the fact that sometimes risk predictions can be tested against return outcomes. This "apples to oranges" comparison, however, is possible only with risk measures, like Value at Risk (**VaR**), with special properties. Backtesting other risk measures can sometimes prove harder if not impossible.

Recent general results [8, 7, 2] have shed light on characterizing backtestable statistics, as well as classifying their backtests. The question is generally relevant in theoretical statistics, but paramount in financial risk management. The decision of the Basel Committee [5, 6] to replace **VaR** with Expected Shortfall (**ES**) as an international standard for banking supervision triggered much of this research, since backtestability of **ES** has long been an open and controversial question, and regulation based on untestable models is unacceptable.

Conclusive results on **ES** backtestability have now been achieved [2]. **ES** admits only approximate backtests, which are inevitably sensitive to **VaR** predictions too. However, we will show that there is one particular **ES** backtest whose sensitivity reduces to a small and prudential bias, which makes it suitable for financial risk management and regulation.

Not only is backtesting **ES** possible, but even superior in some aspects. The **ES** backtest, as opposed to the **VaR** backtest, not only extracts model acceptance probabilities, but also measures the amplitude of prediction discrepancy. This important fact opens the way to new applications such as dynamic, portfolio–specific capital charge adjustments or self–adaptive models that learn from past errors.

For the same reason, a notion of "realized **ES**" emerges, similar in all respects to realized variance, which makes it possible to estimate **ES** a posteriori, if not directly observe it. The **ES** backtest is actually the difference between predicted and realized **ES**, which takes us closer to comparing apples to apples.

2 Understanding backtestability

A statistic $\mathbf{y}(F)$ of a random variable $X \sim F$ is said to be *backtestable* [2] if there exists a *test function* $Z_{\mathbf{y}}(y, x)$ whose expected value $\mathbb{E}_F[Z_{\mathbf{y}}(y, X)]$ is strictly increasing in the *prediction* y and is zero if the prediction equals the *true value* $\mathbf{y}(F)$ of the statistic. To fix ideas, we will think of X as a portfolio return[1] and $\mathbf{y}$ as a risk measure. In the following, when we omit specifying the distribution, we will always understand the real distribution F, writing for instance just $\mathbf{VaR}_\alpha$ for $\mathbf{VaR}_\alpha(F)$ or $\mathbb{E}[\cdot]$ for $\mathbb{E}_F[\cdot]$.

[1] Return here means absolute, currency denominated profit and loss, $X_t = \pi_t - \pi_{t-1}$ where π is the portfolio value.

A *backtest* over an independent series $t = 1, \ldots, T$ of predictions y_t of the risk measure and realizations x_t of the return will then consist on evaluating the *realized test statistic*

$$\bar{z}_{\mathbf{y}}(y_t, x_t) = \frac{1}{T} \sum_{t=1}^{T} Z_{\mathbf{y}}(y_t, x_t) \tag{1}$$

which is expected to be negative (positive) in the case of under– (over–) predictions of $\mathbf{y}$.

To test the hypothesis that a realization $\bar{z}_{\mathbf{y}}$ is acceptable, the predictions $y_t = \mathbf{y}(P_t)$ need to come from model forecasts P_t of the entire distribution of X; the *p–value* $p = P_{\bar{Z}_{\mathbf{y}}}(\bar{z}_{\mathbf{y}})$ of $\bar{z}_{\mathbf{y}}$ can then be obtained by numerical re–sampling of the model distribution $P_{\bar{Z}_{\mathbf{y}}}$ of the test statistic $\bar{Z}_{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^{T} Z_{\mathbf{y}}(y_t, X_t)$, where $X_t \sim P_t$. The model will be accepted or rejected with some *significance level* $\eta \lesssim 1$ (say $\eta = 95\%$); if the test is designed to exclude underestimations, for instance, the rejection condition will be $p < 1 - \eta$ or equivalently $\bar{z}_{\mathbf{y}} < c_\eta$ with *critical threshold* $c_\eta = P^{-1}_{\bar{Z}_{\mathbf{y}}}(1 - \eta)$.

2.1 The VaR backtest, revisited

A classic example of a backtestable statistic[2] is $\mathbf{VaR}_\alpha(F) = -F^{-1}(\alpha)$. Its test function $Z_{\mathbf{VaR}_\alpha}(v, x) = \alpha - (x + v < 0)$ for predictions v has an expected value

$$\mathbb{E}[Z_{\mathbf{VaR}_\alpha}(v, X)] = \alpha - F(-v) \tag{2}$$

and realized test statistic

$$\bar{z}_{\mathbf{VaR}_\alpha}(v_t, x_t) = \frac{1}{T} \sum_{t=1}^{T} [\alpha - (x_t + v_t < 0)] = \alpha - \frac{\#\text{ of } \mathbf{VaR} \text{ exceptions}}{T} \tag{3}$$

which is essentially equivalent to the Basel **VaR** backtest (for $\alpha = 1\%$, $T = 250$ days) [4, 6]. An anomalously high number of **VaR** "exceptions" $\{X_t < -v_t\}$ yields a strongly negative $\bar{z}_{\mathbf{VaR}_\alpha}$ which warns of model underprediction.

The **VaR** backtest is unique because the distribution of the test statistic is binomial, $\bar{Z}_{\mathbf{VaR}_\alpha} \sim \alpha - B(T, T\alpha)/T$, independent of the model P, so that the p–value of a realization requires no re–sampling. Only the point predictions v_t are important, not the associated model details. This makes the implementation of the **VaR** backtest far simpler than any other, because it is sufficient to store two numbers (v_t, x_t) per day and not the entire model distributions P_t.

In Basel regulation, since 1996 [4], two critical thresholds at 5 and 10 exceptions ($\eta \simeq 95\%, 99.99\%$ respectively) over $T = 250$ days define the yellow and red zones of a traffic light system for capital adequacy of $\mathbf{VaR}_{1\%}$ models (Table 1). Since 2016 [6], with the adoption of $\mathbf{ES}_{2.5\%}$ as the new regulatory risk standard and the absence of a corresponding backtest, two distinct $\mathbf{VaR}_\alpha$ backtests (for $\alpha = 1\%, 2.5\%$) have been prescribed along with a controversial P&L attribution test [10], resulting in a contorted model eligibility test.

[2] **VaR** backtestability requires that distribution functions be continuous at the α–quantile and overall strictly increasing, see [2].

Basel $\mathbf{VaR}_{1\%}$ backtest over $T = 250$ days			
	Number of exceptions	Multiplier	Cumulative probability
Green zone	0	1.50	8.106%
	1	1.50	28.575%
	2	1.50	54.317%
	3	1.50	75.812%
	4	1.50	89.219%
Yellow zone	5	1.70	95.882%
	6	1.76	98.630%
	7	1.83	99.597%
	8	1.88	99.894%
	9	1.92	99.975%
Red zone	10 or more	2.00	99.995%

Table 1: Traffic light framework for internal models eligibility in Basel regulation [6]. When there are more than 5 exceptions, the model **VaR** is corrected by an increasingly larger multiplier.

2.2 Sharp backtests

Although the mean $\boldsymbol{\mu} = \mathbb{E}[X]$ is obviously backtestable with test statistic $Z_{\boldsymbol{\mu}}(m, x) = m - x$, the example is still interesting because $\mathbb{E}[Z_{\boldsymbol{\mu}}(m, X)] = m - \boldsymbol{\mu}$ is also a direct measure of the amplitude of the prediction discrepancy. Backtests with this property are called *sharp* in [2] because the expected value of the test statistic determines the unique true value (in this case $\boldsymbol{\mu}$).

Notice that $\bar{z}_{\boldsymbol{\mu}} = \frac{1}{T}\sum_{i=1}^{T} m_t - \hat{\boldsymbol{\mu}}$ is nothing but the difference between the average prediction and the *realized mean* over the period $\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{i=1}^{T} x_t$, which is in fact an unbiased estimator of the average true value $\mathbb{E}[\hat{\boldsymbol{\mu}}] = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{\mu}_t$.

2.2.1 Blindness of the VaR backtest to prediction discrepancy

The **VaR** backtest is not sharp, as indicated by (2) which is a function of the prediction v but not of the true value $\mathbf{VaR}_\alpha$. The **VaR** backtest does not distinguish between large and small exceedances, measuring the probability of the number of exceptions but not the amplitude of prediction discrepancies. This is a very serious limitation because for instance it penalizes a number of small exceedances more than a fewer but much larger exceedances. The true $\mathbf{VaR}_\alpha$ of the portfolio could be much smaller in the former case, predictions being equal. In fact, one can show (see [2]) that without making strong distributional assumptions, there is no relationship between the result of a **VaR** backtest and the true value $\mathbf{VaR}_\alpha$. Even if the exact expected value $\mathbb{E}[Z_{\mathbf{VaR}_\alpha}(v, X)] < 0$ is known, it is still compatible with any true $\mathbf{VaR}_\alpha > v$, namely any underprediction ranging from infinitely small to infinitely large. And yet, Basel regulation prescribes penalizing scaling factors to **VaR** predictions on the sole basis of the number of exceedances (see Table 1), as if such a relationship were there. These

"multipliers" were in fact determined in the yellow and red zones under generic and questionable gaussian assumptions [6].

3 The ES backtest

Backtesting $\mathbf{ES}_\alpha = \alpha^{-1} \int_0^\alpha \mathbf{VaR}_p \, dp$ is less straightforward and has been the subject of research debate over the recent past. Previously proposed **ES** backtests such as Z_2 in [1], turned out to be affected by substantial sensitivity to **VaR** prediction. Recently, it was proved [2] that **ES** is not rigorously backtestable because it is not "elicitable" as proven in [8]. As a consequence, for $\mathbf{ES}_\alpha$ backtests, some degree of sensitivity to $\mathbf{VaR}_\alpha$ (same α) predictions is unavoidable. $\mathbf{ES}_\alpha$ backtests are exact only if the $\mathbf{VaR}_\alpha$ prediction is perfect, something that no separate hypothesis test will ever ascertain with precision. Conclusions on $\mathbf{ES}_\alpha$ predictions can therefore be distorted by the unknown precision of $\mathbf{VaR}_\alpha$ predictions.

Despite these negative results, however, in [2] it was noted that because of the well–known representation[3] [9, 3]

$$\begin{cases} \mathbf{ES}_\alpha = \min_v \mathbb{E}\left[v + \frac{1}{\alpha}(X+v)_-\right] \\ \\ \mathbf{VaR}_\alpha = \arg\min_v \mathbb{E}\left[v + \frac{1}{\alpha}(X+v)_-\right] \end{cases} \tag{4}$$

one can write a **ES** backtest function for a prediction e

$$Z_{\mathbf{ES}_\alpha}(e, v, x) = e - v - \frac{1}{\alpha}(x+v)_- \tag{5}$$

whose sensitivity to **VaR** predictions v is small and one–sided. In fact, from (4) one obtains

$$\mathbb{E}[Z_{\mathbf{ES}_\alpha}(e, v, X)] = e - \mathbf{ES}_\alpha - B(v) \tag{6}$$

where the bias

$$B(v) = \mathbb{E}\left[v + \frac{1}{\alpha}(X+v)_-\right] - \mathbf{ES}_\alpha \geq 0 \tag{7}$$

is positive for any v, it vanishes if and only if the **VaR** prediction is correct, $v = \mathbf{VaR}_\alpha$ and is small[4] for small discrepancies $v \sim \mathbf{VaR}_\alpha$. Equation (6) tells us that the sensitivity to small **VaR** prediction errors is negligible. Moreover, in any case, the introduced bias is *prudential*[5], as it leads to a comparison against a true value $\mathbf{ES}_\alpha$ augmented by $B(v) \geq 0$.

[3] We denote the negative part of a number as $(a)_- = -\min(a, 0)$

[4] $B(v)$ is quadratic if the distribution function $F(x)$ admits a density $f(x)$ in $x = -\mathbf{VaR}_\alpha$

$$B(v) = \frac{f(-\mathbf{VaR}_\alpha)}{2\alpha}(v - \mathbf{VaR}_\alpha)^2 + \mathcal{O}(v - \mathbf{VaR}_\alpha)^3 \tag{8}$$

[5] A clear violation of Murphy's law. Had it been the other way around there would have been no viable prudential backtest for **ES** intended as a risk measure.

Example 3.1. To illustrate this fact, we fix a predictive P and compute the expected value of the test statistic[6] as a function of the **VaR** prediction discrepancy $(v - \mathbf{VaR})/v$ using[7] different real distributions F having the same **ES** but different **VaR**. We compare the backtest $Z_{\mathbf{ES}}$ with Z_2 proposed[8] in [1]. In the example shown in Figure 1, the **ES** prediction is correct. The mild, quadratic, prudential bias of $Z_{\mathbf{ES}}$ contrasts with the more pronounced and locally linear bias of Z_2, which leads more easily to Type I errors (rejection of correct models). Figure 2 shows instead a deliberate case of model underestimation: the prudential nature of the $Z_{\mathbf{ES}}$ bias will never cause a Type II error (acceptance of wrong models), which is something that the Z_2 linear bias could clearly produce.

3.1 How big is the bias?

Extensive analyses of the size of the bias (7) show that it is typically negligible when **VaR** predictions are accurate within $\pm 15\%$ and still small within $\pm 40\%$ approximately. This means that the accuracy of **VaR** predictions need not be very high. At the same time, some care must be exercised if the **VaR** prediction accuracy is completely unknown, because outside these ranges the bias term could have dominant – yet still prudential – distortive effects on the **ES** backtest.

Figure 3 shows the ratio $B(v)/\mathbf{ES}_\alpha$ versus the **VaR** prediction discrepancy for different values of α and degrees of freedom ν of Student–t real distributions. We notice a somewhat similar pattern across all cases, with more pronounced bias for small α and thin tails.

The hypothetical case of a correct **ES** prediction rejected for bias reasons (i.e. for a completely wrong **VaR** prediction) is interesting. The backtest embodies in a sense a regulator who wants to penalize not only an underprediction of **ES** but also an accidentally correct prediction issued by a model with a completely wrong tail shape.

3.2 Just like variance?

The reader may feel a déjà vu: also the common *variance* $\boldsymbol{\sigma}^2$ is backtestable only at the price of some sensitivity to predictions of the mean $\boldsymbol{\mu}$. In fact, similarly to (4) one can write

$$\begin{cases} \boldsymbol{\sigma}^2 = \min_m \mathbb{E}\left[(X-m)^2\right] \\ \\ \boldsymbol{\mu} = \arg\min_m \mathbb{E}\left[(X-m)^2\right] \end{cases} \tag{9}$$

[6] We use the relative backtest version, eq. (15).

[7] In the examples, P is a fixed Student–t centered in zero and F is also Student–t centered in zero, but with varying degrees of freedom and adjusted standard deviation so as to keep **ES** fixed. We used $\alpha = 2.5\%$ and $T = 250$.

[8] Z_2 has test function $Z_2(e, v, x) = e + \frac{1}{\alpha} x(x + v < 0)$. It requires distributions to be continuous at the α quantile.

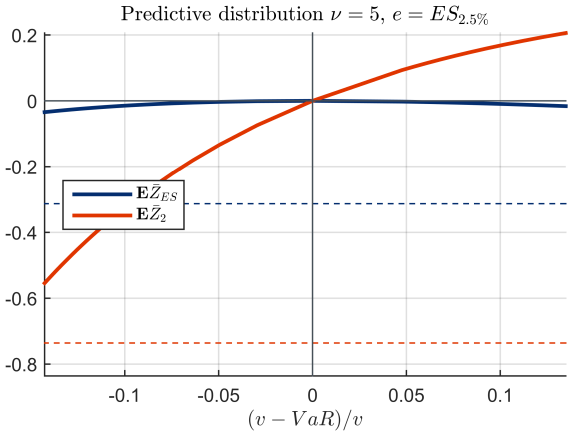

Figure 1: Sensitivity to **VaR** predictions for tests $Z_{\mathbf{ES}}$ and Z_2 in the case of a correct **ES** prediction. Dotted lines represent critical values at 5% for the two tests. Notice the linear sensitivity of the latter and the muted, prudential sensitivity of the former. Z_2 can easily generate a Type I error.

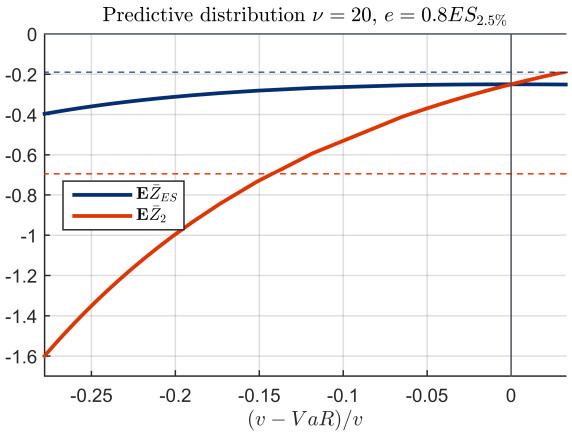

Figure 2: Similar example in the case of an underestimation $e = 0.8\mathbf{ES}$. As opposed to Z_2, the prudential bias of $Z_{\mathbf{ES}}$ can never cause a Type II error.

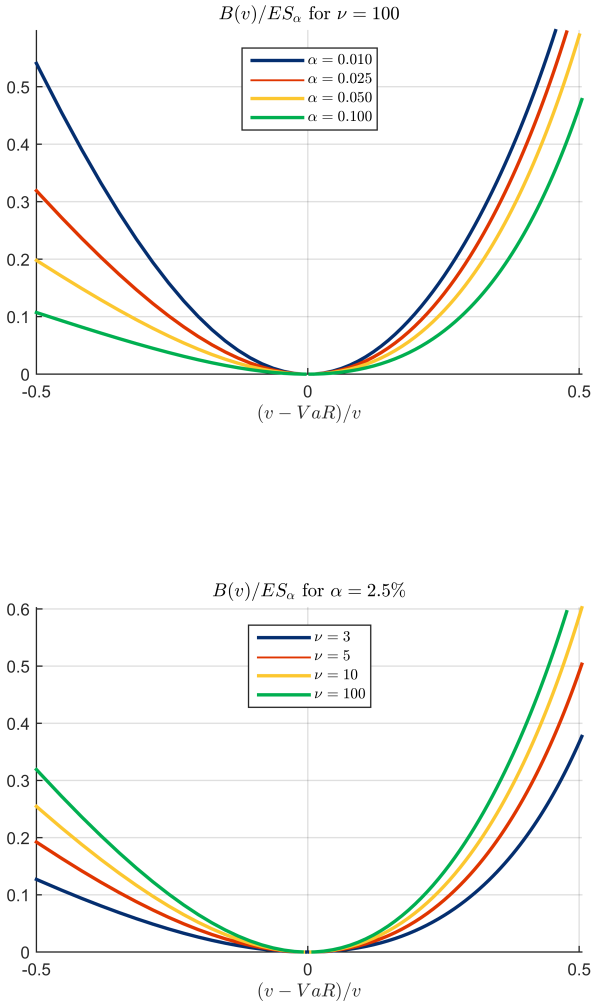

Figure 3: Relative size of the $\mathbf{ES}_\alpha$ backtest bias as a function of $\mathbf{VaR}_\alpha$ prediction error, for different α and Student–$t(\nu)$ distributions.

so that the variance backtest

$$Z_{\boldsymbol{\sigma}^2}(v, m, x) = v - (x - m)^2 \tag{10}$$

with expected value

$$\mathbb{E}[Z_{\boldsymbol{\sigma}^2}(v, m, X)] = v - \boldsymbol{\sigma}^2 - B(m) \tag{11}$$

is exposed to just a small, quadratic and prudential bias $B(m) = (m - \boldsymbol{\mu})^2$ for mean predictions $m \neq \boldsymbol{\mu}$.

The strong analogy is not accidental. In [2] it was shown that **ES** and $\boldsymbol{\sigma}^2$ are in fact two instances of a general mechanism (*ridge backtest*) involving statistics that can be expressed as the minimum of the expected scoring function of other backtestable statistics (here **VaR** and $\boldsymbol{\mu}$ respectively). This incidentally also shows that the backtest $Z_{\mathbf{ES}}$ is completely unique, in the sense that no other **ES** backtest can be found with mild and prudential bias.

3.3 Sharpness of $Z_{\mathbf{ES}}$ and *realized* ES

A glance at (6) reveals that, although biased, the **ES** backtest (5) is actually sharp, as it directly measures the difference between true and predicted value. This fact is extremely important because it provides corrective model feedback and allows estimating the true value **ES** itself.

3.3.1 Absolute backtest

The realized test statistic of (6) can be simply expressed as the difference

$$\bar{z}_{\mathbf{ES}_\alpha}(e_t, v_t, x_t) = \frac{1}{T}\sum_{t=1}^{T} e_t - \widehat{\mathbf{ES}}_\alpha \tag{12}$$

between the average prediction and the *realized* **ES** defined by

$$\widehat{\mathbf{ES}}_\alpha \equiv \frac{1}{T}\sum_{t=1}^{T}\left[v_t + \frac{1}{\alpha}(x_t + v_t)_-\right] \tag{13}$$

This is a positively biased estimator of the average true **ES**, because of (7)

$$\mathbb{E}[\widehat{\mathbf{ES}}_\alpha] = \frac{1}{T}\sum_{t=1}^{T}\mathbf{ES}_{\alpha,t} + \frac{1}{T}\sum_{t=1}^{T}B(v_t) \geq \frac{1}{T}\sum_{t=1}^{T}\mathbf{ES}_{\alpha,t} \tag{14}$$

The definition is completely analogous to realized variance $\widehat{\boldsymbol{\sigma}^2} = \frac{1}{T}\sum_t (x_t - m_t)^2$, which is also a positively–biased estimator of the average true variance because of (11).

The notion of realized **ES** is new. Notice that a meaningful realized version of a statistic (be it $\widehat{\mathbf{ES}}$, $\hat{\mu}$ or $\hat{\boldsymbol{\sigma}}^2$) exists exactly because the corresponding backtest

is sharp, allowing estimation of prediction discrepancy and hence of the true value (possibly with a prudential bias). This is the reason why no meaningful notion of realized **VaR** is known. We can affirm that **ES** is a statistic that in some sense can be observed a posteriori, while **VaR** is inherently unobservable. Sharp backtests are apple to apple comparisons, between ex–ante predictions and ex–post realizations.

As explained in section 2, the realized test statistic $\bar{z}_{\mathbf{ES}_\alpha}$ serves to compute a p–value for model validation. But it also represents a direct indicator – in currency units – of the prediction discrepancy (see Example 3.3). It is important to note that for the latter purpose no model distribution re–sampling is needed as $\bar{z}_{\mathbf{ES}_\alpha}$ can be directly computed storing just three numbers $\{e_t, v_t, x_t\}$ per day.

3.3.2 Relative backtest

A dimensionless version of the **ES** backtest can be obtained by normalizing the test function by the predicted value

$$Z^{rel}_{\mathbf{ES}_\alpha}(e, v, x) \equiv \frac{Z_{\mathbf{ES}_\alpha}(e, v, x)}{e}. \tag{15}$$

It can be shown [2] that this remains[9] a valid test function for **ES** in the assumption – never restrictive for small α and real financial portfolios – that both the true and the predicted **ES** are strictly positive[10]. From (6) we have

$$\mathbb{E}[Z^{rel}_{\mathbf{ES}_\alpha}(e, v, X)] = \frac{e - \mathbf{ES}_\alpha - B(v)}{e} \tag{16}$$

which is a prudentially (i.e. negatively) biased measure of the relative prediction discrepancy $(e - \mathbf{ES}_\alpha)/e$.

Tests $Z^{rel}_{\mathbf{ES}_\alpha}$ and $Z_{\mathbf{ES}_\alpha}$ reject in general[11] different models, as they measure average relative and absolute prediction errors respectively. The former is a test more suited for scale–independent model validation; the latter for measuring currency denominated average under–capitalization.

We can express the realized relative test statistic as

$$\bar{z}^{rel}_{\mathbf{ES}_\alpha} = 1 - \widehat{\phi}_{\mathbf{ES}} \tag{17}$$

where we define the *realized prediction ratio*

$$\widehat{\phi}_{\mathbf{ES}} = \frac{1}{T}\sum_{t=1}^{T} \frac{v_t + \frac{1}{\alpha}(x_t + v_t)_-}{e_t} \tag{18}$$

which is a positively biased estimator of the average prediction ratio $\mathbf{ES}_\alpha/e$ between true and predicted **ES**

$$\mathbb{E}[\widehat{\phi}] = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{ES}_{\alpha,t}}{e_t} + \frac{1}{T}\sum_{t=1}^{T} \frac{B(v_t)}{e_t} \geq \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{ES}_{\alpha,t}}{e_t} \tag{19}$$

[9] Normalizing by e is less innocuous than it may seem, as it remains to be checked whether the expected test function is strictly monotonic with respect to e under any distribution F.

[10] Another sufficient condition is $e \geq v > 0$.

[11] The two tests coincide only if the predictions e_t are constant.

Risk scaling factors										
		$\mathbf{VaR}_{1\%}$	$\mathbf{ES}_{2.5\%}$							
Model distribution			$\nu=\infty$	$\nu=10$	$\nu=5$	$\nu=3$	$\nu=\infty$	$\nu=10$	$\nu=5$	$\nu=3$
	$\eta=1-p$	Basel	$\widehat{\phi}_{\mathbf{ES}}$				γ			
Green zone	8.106%	1.00	0.89	0.85	0.81	0.72	0.85	0.80	0.72	0.60
	28.575%	1.00	0.94	0.92	0.89	0.83	0.94	0.91	0.87	0.80
	54.317%	1.00	1.00	1.00	0.99	0.96	1.00	0.99	0.99	0.96
	75.812%	1.00	1.05	1.07	1.09	1.12	1.05	1.07	1.08	1.11
	89.219%	1.00	1.11	1.15	1.21	1.31	1.09	1.13	1.18	1.26
Yellow zone	95.882%	1.13	1.17	1.24	1.34	1.55	1.13	1.19	1.26	1.41
	98.630%	1.17	1.23	1.33	1.49	1.86	1.17	1.24	1.35	1.59
	99.597%	1.22	1.29	1.42	1.66	2.31	1.21	1.29	1.44	1.82
	99.894%	1.25	1.35	1.52	1.86	3.01	1.24	1.34	1.54	2.13
	99.975%	1.28	1.42	1.62	2.14	4.24	1.27	1.39	1.65	2.61
Red zone	99.995%	1.33	1.48	1.72	2.52	6.36	1.30	1.43	1.78	3.34

Table 2: Scaling factors $\widehat{\phi}_{\mathbf{ES}_{2.5\%}}$ and bias–neutralized scaling factors γ for different model distributions compared with Basel prescription for $\mathbf{VaR}_{1\%}$. See Example 3.2.

The ratio $\widehat{\phi}_{\mathbf{ES}}$ is a prudential estimate of the scaling factor that should be applied to model **ES** predictions to compensate the backtested relative discrepancy. It is a portfolio–specific version of the scaling factors implicit in the **VaR** multipliers in Table 1.

It is important to observe that when $\widehat{\phi}_{\mathbf{ES}}$ differs from 1 by more than around $\pm 60\%$, bias effects are no longer negligible and lead to a large overestimation of the ratio $\mathbf{ES}/e$. Supposing for instance that $\mathbf{ES}_\alpha$ is underestimated by a factor of 2, $\mathbf{VaR}_\alpha$ will typically also be strongly underestimated generating a large bias and hence a $\widehat{\phi}_{\mathbf{ES}}$ well above 2.

Example 3.2. In Table 2, we compare the $\mathbf{VaR}_{1\%}$ scaling factors implicit in Table 1 with the scaling factors $\widehat{\phi}_{\mathbf{ES}_{2.5\%}}$. To match the test significance η, we assume different Student–$t(\nu)$ model distributions for the latter scaling factors. Notice that whereas the **VaR** backtest produces only p–values that are conventionally mapped onto scaling factors under generic gaussian assumptions, the factors $\widehat{\phi}_{\mathbf{ES}}$ are generated directly, specific to the true (not model!) portfolio distribution.

For progressively fatter tails and higher significance η, $\widehat{\phi}_{\mathbf{ES}}$ becomes larger and larger, but arguably also more and more biased. A possible way to remove the bias effect is to assume that the result is the expected value from an alternative hypothesis H_1 in which the real distribution $F_\gamma(x) = P(x/\gamma)$ differs from the predictive distribution P by an overall scale factor γ that applies to both **VaR** and **ES**. We then determine γ by matching

$$\mathbb{E}_{F_\gamma}[Z^{rel}_{\mathbf{ES}_\alpha}(e, v, X)] = \mathbb{E}_P[Z^{rel}_{\mathbf{ES}_\alpha}(e, v, \gamma X)] = 1 - \widehat{\phi}_{\mathbf{ES}}.$$

and we report it in the right half of Table 2. While the $\widehat{\phi}_{\mathbf{ES}}$'s represent a (hypothesis free) theoretical upper limit for the ratio $\mathbf{ES}_\alpha/e$, the γ's give a more realistic (but hypothesis specific) estimate.

Notice that for gaussian ($\nu = \infty$) distributions, γ approaches closely the

Basel scaling factors for $\mathbf{VaR}_{1\%}$. Fatter tails exhibit a significantly higher γ. This shows that the Basel multipliers provide too optimistic capital adequacy scaling factors for general, non–gaussian portfolios.

We illustrate the use of realized prediction ratio as a trailing indicator of under/overprediction magnitude in the following example.

Example 3.3. Figure 4, shows the $T = 250$ backtest history over 2 years for different real portfolios and various models. We compare the evolution of $\widehat{\phi}_{\mathbf{ES}_{2.5\%}}$ (solid line) with $\mathbf{VaR}_{1\%}$ Basel scaling factors (dotted). The line color shows the traffic light zone for each day. Remember that the trajectory of $\widehat{\phi}_{\mathbf{ES}_{2.5\%}}$ depends only on the true portfolio realizations x_t besides predictions $\{e_t, v_t\}$; it is only the associated traffic light color that depends on the model distribution.

Notice the general wider dynamics of **ES**, which should be taken seriously as long as it is not in the red zone where the bias is no longer negligible. The **ES** scaling factor is continuous whereas **VaR**'s is discrete valued. Furthermore, it may get smaller than one, providing corrections also to overcapitalization. The estimate is prudential, which means that the real ratio is typically even lower than $\widehat{\phi} < 1$.

3.4 Power of the test

We know that the test penalizes underestimations of $\mathbf{ES}_\alpha$ as well as bilateral mis–estimations of $\mathbf{VaR}_\alpha$. In general, this prudential bias will contribute positively to the power[12] of the test, at the cost of some Type I error probability, when **ES** is predicted correctly but the model tail shape is seriously wrong.

For a quantitative analysis of the power, we compare the tests for $\mathbf{ES}_{2.5\%}$ and $\mathbf{VaR}_{1\%}$ on a selected number of different alternative hypotheses H_1.

		Power	
ν	Scale γ	$Z_{\mathbf{ES}_{2.5\%}}$ (%)	$\mathbf{VaR}_{1\%}$ (%)
5	1.20	30.0	33.6
	1.50	90.3	90.9
100	1.20	71.3	62.6
	1.50	99.9	99.7

Table 3: Power of $\mathbf{ES}_{2.5\%}$ and $\mathbf{VaR}_{1\%}$ tests for $\nu = 5, 100$ Student–t distributions and alternative hypotheses overall rescaled by a factor γ.

In Table 3, we analyze the case in which the H_1 distribution $F(x) = P(x/\gamma)$ differs from the H_0 model distribution P by an overall scaling factor $\gamma = 1.2, 1.5$. For H_0 we choose a Student–t with either fat ($\nu = 5$) or thin tail ($\nu = 100$). The significance η is fixed[13] at 95.882%. We can see that the power of the two

[12] We recall that the power $= F_{\bar{Z}}(c_\eta)$ is the rejection probability of an alternative assumption H_1. $F_{\bar{Z}}$ is the distribution function of the test variable under H_1 and c_η is the critical threshold determined by H_0.

[13] This is the closest value to standard significance of 95% attained by the $\mathbf{VaR}_{1\%}$ backtest, and the threshold of the yellow zone in Table 1.

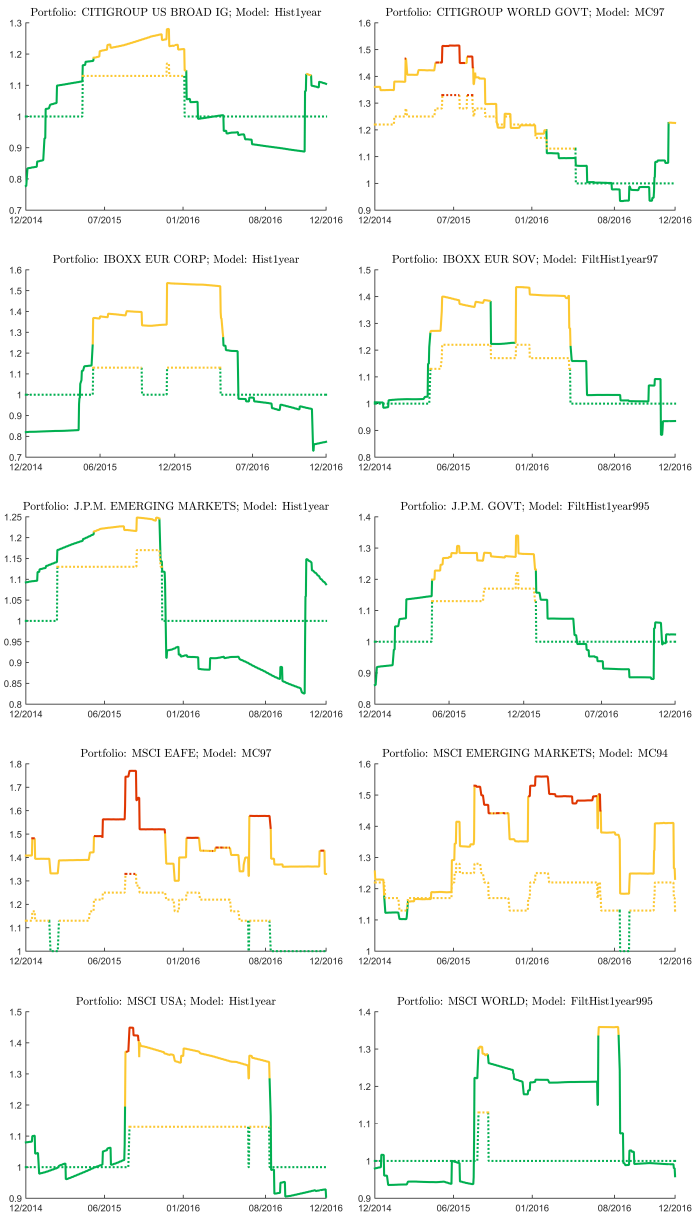

Figure 4: Time evolution of $\mathbf{ES}_{2.5\%}$ (solid line) and $\mathbf{VaR}_{1\%}$ (dotted) scaling factors, across diverse portfolios and models. See Ex. 3.3.

tests is comparable: the one of **ES** is slightly lower for fat tails and higher for thin tails.

	ν	$\mathbf{ES}_{2.5\%}$	$\mathbf{VaR}_{2.5\%}$	$\mathbf{VaR}_{1\%}$
Normalized	100	2.35	1.96	2.34
Student–t	10	2.52	1.99	2.47
	5	2.73	1.99	2.61
	3	2.91	1.83	2.62

			Power	
	ν		$Z_{\mathbf{ES}_{2.5\%}}$	$\mathbf{VaR}_{1\%}$
	H_0	H_1	(%)	(%)
Normalized	10	5	16.7	9.0
Student–t		3	27.5	7.4
	100	10	19.6	10.5
		3	48.0	12.0

Table 4: Power of $\mathbf{ES}_{2.5\%}$ and $\mathbf{VaR}_{1\%}$ tests for $\nu = 10, 100$ normalized Student–t distributions and alternative hypotheses with fatter tail.

In Table 4 we analyze the power relative to underestimations due to fatter tails in the H_1's. For this purpose, we adopt a family of normalized Student–t distributions, namely rescaled to have unit standard deviation. We assume an H_0 with $\nu = 10, 100$ and let H_1's span lower ν's. In this case the **ES** test displays much higher power. Notice that the value of $\mathbf{VaR}_{2.5\%}$ in the alternatives does not deviate much, so that the bias contribution to the power should be negligible.

Test $Z_{\mathbf{ES}}$ displays superior power also with respect to Z_2 [1]. It is clear from Figures 1 and 2 that the stronger and linear bias of Z_2 may lead to an accidentally high (low) power for alternatives in which **VaR** is underestimated (resp. overestimated). For alternatives where the power of Z_2 is not favored by a strong bias, $Z_{\mathbf{ES}}$ has been extensively tested to show higher power. Figure 1 shows such an example, where in a broad range of wrong alternatives $Z_{\mathbf{ES}}$ rejects the model whereas Z_2 does not.

4 Conclusions

Model validation for **ES**–based risk models is not only possible, but is also far more informative than traditional model acceptance on the basis of **VaR** exception counting. Risk managers and regulators can track model prediction performance and use backtest feedback to adjust risk models outcomes. The property of having a small and prudential sensitivity to **VaR** predictions is unique to the shown **ES** backtest, which exploits an extremal representation of **ES** (eq. 4) involving **VaR**.

The notion of realized **ES** emerging from the sharpness of the backtest opens the way to countless possible applications such as rule–based margins, risk budgeting, "stop risk" rules and so on, deserving separate research, because it is the first appearance of an ex–post version of a tail risk measure. The question

"how risky has a portfolio been?" has now a model–independent answer. Tail risk is no more as unobservable as it used to be.

References

[1] Acerbi, C. and Szekely, B. (2014) Backtesting Expected Shortfall, RISK Magazine, December

[2] Acerbi, C. and Szekely, B. (2017) General Properties of Backtestable Statistics, working paper, available on `ssrn.com`

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

[4] Basel Committee on Banking Supervision (1996) Amendment to the capital accord to incorporate market risks, `http://www.bis.org/publ/bcbs24.pdf`

[5] Basel Committee on Banking Supervision (2012) Fundamental Review of the Trading Book, `http://www.bis.org/publ/bcbs219.pdf`

[6] Basel Committee on Banking Supervision (2016) Minimum capital requirements for market risk, `http://www.bis.org/bcbs/publ/d352.pdf`

[7] Fissler, T. and Ziegel, J.F. (2016) Higher order elicitability and Osband's principle, The Annals of Statistics, 8, 4, 1680–1707.

[8] Gneiting, T. (2011) Making and Evaluating Point Forecasts, Journal of the American Statistical Association

[9] Rockafellar, R.T., Uryasev, S. (2002) Conditional Value–at–Risk for general loss distributions. Journal of Banking and Finance, **26** (7)

[10] Wood, D. (2016) The P&L attribution mess, RISK Magazine, September