General properties of backtestable statistics

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

MSCI Inc.

January 23, 2017

Abstract

We propose a formal definition of backtestable statistic: a backtest is a null expected value involving only the statistic and its random variable, strictly monotonic in the former. We discuss the relationship with elicitability and identifiability which turn out being necessary conditions. The variance and the Expected Shortfall are not backtestable for this reason. We discuss (absolute) model validation in the context of one– or two–sided hypothesis tests, as well as (relative) model selection obtained by ranking realizations of the backtest statistic. We introduce the concept of sharpness which refers to whether a backtest is strictly monotonic with respect to the real value of the statistic and not only to its prediction. This decides whether the expected value of a backtest determines the extent of a prediction discrepancy and not only its likelihood. We show that the quantile backtest is not sharp and in fact provides no information whatsoever on the real value of the statistic. The Expectile is also not sharp; we provide bounds for its real value, which are looser for outer confidence levels. We then introduce ridge backtests, applicable to particular non–backtestable statistics, such as the variance and the Expected Shortfall, which coincide with the attained minimum of the scoring function of another elicitable auxiliary statistic. This allows to produce sharp backtest procedures in which the prediction for the auxiliary variable is also involved but with small sensitivity and known bias sign. The ridge mechanism explains why the variance has always been de–facto backtestable and allows for similar efficient ways to backtest the expected shortfall. We discuss the relevance of this result in the current debate of financial regulation (banking and insurance), where Value at Risk and Expected Shortfall are adopted as regulatory risk measures.

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

1 Motivation and objectives

Backtest procedures for testing predictions of a statistic are common in the statistical and probabilistic disciplines, notably in the field of financial risk management. Backtesting in general, however, remains to date a collection of disparate practices in the wait for a clear definition. Whether or not a statistic is backtestable, whether or not a methodology is a backtest, remain vague questions, prone to controversy.

1.1 Understanding backtestability

In this paper, we address the problem of formalizing the concept of backtestable statistic and backtesting procedure. We will try to distill the key features that make it possible to perform the common backtests of statistics such as the mean, the quantile and the variance. The inference of these features however is not so straightforward, because even finding what these simple cases have in common requires some care. For instance, we will see that the quantile is completely atypical in that it does not require the issuance of a predictive distribution; and the variance is not even backtestable, strictly speaking. It's probably because of subtleties like these that the concept of backtestability still awaits formalization after decades of industry practices.

Backtesting a statistic (a risk measure, for instance), is not so straightforward as it may seem. When you make a sequence of bets, say, through a season of horse races, you can easily check if your predictions were right or wrong simply because the winning horse is publicly declared at the end of any race. But if you make a prediction y on a statistic $\mathbf{y}$ (say the variance, $\boldsymbol{\sigma}^2$) of a distribution F of future events (say portfolio profit/losses X), neither the distribution F nor the true value of the statistic $\mathbf{y}(F)$ are publicly announced at the end of the day. What is revealed is just a single random draw x from an unknowable real world distribution F. The tricky point is here: how to compare a prediction of a statistic $\mathbf{y}$ against a single outcome of its random variable X? When this is possible it's because there exists a test function $Z(y, x)$ depending only on these observable quantities – prediction y of the statistic and realization x of the random variable – whose expected value somehow reveals if predictions are over/underestimated. For a given statistic, however, the existence of such a test function is not granted a priori: as a matter of fact, many statistics are not backtestable.

1.1.1 Quantile: the prototype of all backtests. Or maybe not

A common example of a backtesting procedure is provided by the quantile. Suppose that a prediction y_t is issued for a quantile $\mathbf{q}_\alpha$ with confidence level $\alpha \in (0, 1)$. If the prediction were correct, the realization x_t should be lower than y_t with probability[1] equal to α. Over multiple predictions at different times t, the expected number of quantile "shootouts" $x_t < y_t$ should be a fraction α

[1] Let's assume for the moment that F is continuous

Figure 1: Testing bets on a horse race is easy, because the winner is publicly announced.

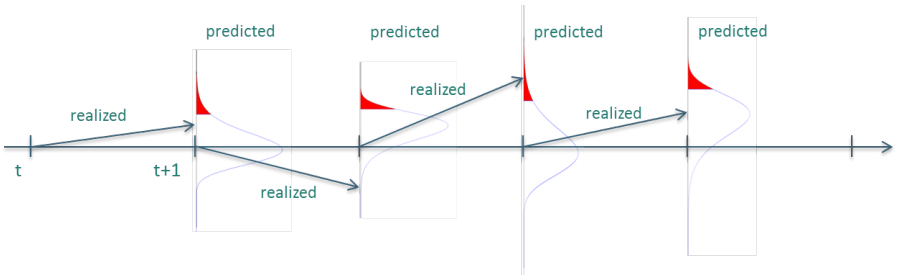

Figure 2: Testing predictions on a statistic is harder, because the true statistic will never be revealed. What is revealed is just one random draw.

of all predictions. A lower (bigger) number of shootouts would signal that the model tends to underestimate (overestimate) the quantile. The procedure has been adopted within Basel regulation for capital adequacy, for backtesting a bank's Value at Risk ($\mathbf{VaR} = -\mathbf{q}$) model, for over two decades [3, 4].

At the heart of this procedure, lies the fact that for quantiles there exists a test function $Z_{\mathbf{q}_\alpha}(y, x) = (x \leq y) - \alpha$ such that $\mathbb{E}_F[Z_{\mathbf{q}_\alpha}(y, X)] \gtreqless 0$ if $y \gtreqless \mathbf{q}_\alpha(F)$. The sign of the realized test function $\bar{z}_{\mathbf{q}} = (1/T) \sum_{t=1}^{T} Z_{\mathbf{q}_\alpha}(y_t, x_t)$ over successive predictions y_t, is used to detect the tendency to under- or over- estimations. We will take the existence of such a function as the definition of backtestability for a given statistic, with some additional requirements and distinctions.

Remark 1.1. The case of quantile backtest, although prototypical in some sense, is very special in some others. In the case of an exact prediction, the test variable $Z_{\mathbf{q}_\alpha}$ is distributed along a Bernoulli distribution of parameter α and zero mean and therefore $\bar{z}_{\mathbf{q}}$ follows a binomial distribution $B(T, \alpha)$, with zero mean, independently of the predictive model. For this reason, computing the p–value for a realization of $\bar{z}_{\mathbf{q}}$ in the null hypothesis that the predictions are correct, requires only generation and storage of the point predictions y_t but not of the predictive distributions P_t. We will see that in the case of all other backtestable statistics the distribution of $\bar{z}_{\mathbf{q}}$ under the predictive model is model–dependent, requiring the storage of the predictive distributions P_t to compute p–values.

Another peculiar feature of the quantile backtest is the discontinuity of the test function, which requires smoothness conditions on the distribution functions involved.

1.2 Backtesting in the financial regulation debate

Although the scope of the paper is more general, the initial motivation for the present publication comes from financial risk management.

The question over whether the Expected Shortfall (**ES**) [2, 13] can be backtested or not has been the subject of a lively debate over the past years, since it was found that this risk measure doesn't possess a property called elicitability [8]. Part of the controversy stems from the fact that in the absence of a formal definition of backtestability, its precise connection with elicitability remains unclear.

The subject is paramount in the current regulatory debate. In 2013, the Basel Committee of Banking Supervision (BCBS), after a proposal exposed to public consultation, took the decision [6] of replacing Value at Risk (**VaR**) with **ES** for capital adequacy internal models. The decision is now part of the final rules of Basel 4 [4] that banks are currently implementing.

A similar debate was the subject of a public consultation [9] from the International Association of Insurance Supervisors on Insurance Capital Standards in December 2014; the result of the consultation was in this case exactly the opposite, with IAIS finally opting for the adoption of **VaR** in the Capital Standard [10]. Among the motivations for this choice was the absence of a backtest for **ES**.

ES has a number of advantages over **VaR**, among which its ability to detect tail risks, respecting basic rules of risk diversification and convexity. But when it comes to financial regulation, backtestability – whatever it exactly means – is not a negotiable property. The term backtesting, as commonly intended in the risk industry, refers to the practice of validating ex–ante model predictions of some statistic against observed ex–post realizations of the random variable it refers to. It's the only way for verifying if a risk model makes acceptable predictions or not, a reality check for its output. Banks have been following backtesting practices for decades, because **VaR** – despite all its deficiencies – lends itself to straightforward backtesting. If it were ascertained that **ES** cannot be backtested in any possible way, related risk models could not be validated, and for instance the BCBS should seriously reconsider their decision.

Currently, Basel regulation imposes banks to compute the capital charge for their internal models based on **ES** but in absence of a valid backtest for this measure, requires the models to be validated based on a traditional **VaR** backtest. Whether and how **ES** can be backtested has become an urgent question.

1.3 Structure of the paper and main findings

In section 2 we define the notation and the experiment setting and we review the established concepts of elicitation and identification, which will get us very close to a proper definition of backtestability. In section 2.2.1 we discuss how elicitability serves to the purpose of conducting relative non–directional tests of goodness between competing models issuing point predictions on a statistic.

In section 3 we pose the central definition of backtestability, corresponding to strictly y–increasing identification and y–convex elicitation. Non–elicitable statistics, such as the variance and the expected shortfall, are not backtestable. We show that convexity forces the scoring and backtest functions of the mean and the quantile to be essentially unique (Corollaries 3.6 and 3.7). In section 3.1.1 we propose relative model tests for point predictions based on backtestability, similar to the ones based on elicitability. The proposed tests are however directional, namely able to detect under– and over– estimations. Section 3.1.2 describes the full blown hypothesis backtesting procedure that allows for absolute validation of a model issuing entire predictive distributions.

In section 3.2 we notice that only some backtests (such as the mean backtest) are monotonic also with respect to the real value of the statistic. We call them sharp because their expected value singles out one value for the real statistic, providing information on both the likelihood and the magnitude of a prediction discrepancy. Other backtests provide limited information on the value of the statistic, in the form of a possible range. We show (proposition 3.15) that the quantile backtest tells basically nothing on the position of the real statistic. Expectiles are not sharp either, unless $\alpha = 1/2$ and we provide bounds for the location of the real value (proposition 3.18).

In section 4 we show that some non–backtestable statistics admit what we call a ridge backtest. This is a test involving also predictions of a second, elicitable, auxiliary statistic, to which the test displays only limited and one–

sided sensitivity. Such mechanism allows for effective backtests when the bias side happens to be prudential for the purpose of the ridge backtest (proposition 4.2). The variance and the tail mean (or expected shortfall) are shown to admit a ridge backtest. The result opens the way for effective backtests of the expected shortfall.

2 Preliminary concepts

In this section we describe the setup, set the notation and briefly review a number of concepts and results in the related literature.

2.1 Backtest experiment setup

Let $X \in \mathbb{R}$ be a random variable distributed along an unknowable real–world distribution function F. Let $\mathbf{y}$ be a statistic of X, which we will alternatively denote as $\mathbf{y}(F)$, $\mathbf{y}(X)$ or simply $\mathbf{y}$, case by case, depending on the emphasis. $\mathbf{y}$ may be real–valued or interval–valued as in the case of quantiles (see section A.1).

We will consider a discrete sequence of times $t = 0, 1, \ldots, T$ and random variables X_t observed at every $t > 0$. Let F_t denote the distribution of X_t conditional to the information available at time $t-1$. We will assume that at any time $t-1$, a forecast y_t for the statistic $\mathbf{y}$ is issued by some model, as in the point forecast setting adopted in [8]. Sometimes we will assume however that an entire forecast distribution P_t for the random variable X_t is issued by the model, and used to generate the prediction $y_t = \mathbf{y}(P_t)$. We will abstract completely from the nature of the model, which can be any sort of algorithm, procedure, divination or guesswork.

To pave the ground for a definition of backtestability, we first review some useful concepts.

2.2 Elicitability

An established instrument for the evaluation of point forecasts is the notion of *elicitability* [8].

Definition 2.1. A statistic $\mathbf{y}$ is called *$\mathcal{F}$–elicitable*[2], if there exists a *scoring function* $S_{\mathbf{y}}(y, x)$ such that the statistic can be expressed as the minimizer of the expectation

$$\mathbf{y}(F) = \arg\min_y \mathbb{E}_F \left[S_{\mathbf{y}}(y, X) \right], \qquad \forall F \in \mathcal{F} \tag{1}$$

[2] Our definition of elicitability corresponds to strict elicitability in [8]. We are not interested in the non–strict case.

An interpretation of this definition is that the expected scoring function plays the role of a penalty function for possible forecasts y of the statistic in the sense that it is minimized in expectation by a perfect forecast $y = \mathbf{y}(F)$.

We are primarily interested in the case when the class of distribution functions $\mathcal{F}$ is *maximal*, in the sense that it contains all distributions F for which both the statistic $\mathbf{y}(F)$ and an expected scoring function $\mathbb{E}_F[S_{\mathbf{y}}]$ are finite. In this case, we will omit to denote $\mathcal{F}$ and simply speak of elicitable statistics.

Common examples of elicitable statistics include

- The mean $\boldsymbol{\mu}(F) = \mathbb{E}_F[X]$ which is the minimizer of the expected squared error and therefore has scoring function

$$S_{\boldsymbol{\mu}}(y, x) = (y - x)^2$$

- The median – the 1/2–quantile $\mathbf{q}_{1/2}(F)$ – which is the minimizer of the expected absolute error and therefore has scoring function

$$S_{\mathbf{q}_{1/2}}(y, x) = |y - x|$$

- More generally any α–quantile, for $\alpha \in (0, 1)$, (see A.1) is elicited by a scoring function

$$S_{\mathbf{q}_\alpha}(y, x) = \alpha(x - y)_+ + (1 - \alpha)(x - y)_-$$

  a well known fact in quantile regression. Equation (55) anticipates this fact, and can be easily shown to be based on essentially the same scoring function as the above.

- Expectiles $\mathbf{e}_\alpha(F)$ [12, 7] (see A.3) are elicitable with scoring function

$$S_{\mathbf{e}_\alpha}(y, x) = \alpha(x - y)_+^2 + (1 - \alpha)(x - y)_-^2$$

  Expectiles represent a remarkable example because for $\alpha \leq 1/2$ the associated risk measure (of opposite sign) was proven to be [14] the only elicitable statistic which is also a coherent measure in the sense of [5].

Remarkable examples of non–elicitable statistics include the tail mean **TM** [8] but also the popular variance $\boldsymbol{\sigma}^2$ of a distribution [11].

When the scoring function of a statistic exists, it is not unique. For example, it is immediate to check that a scoring function $S_{\mathbf{y}}(y, x)$ is defined up to inessential transformations

$$S_{\mathbf{y}}(y, x) \to a S_{\mathbf{y}}(y, x) + h(x) \tag{2}$$

where $a > 0$ and h any function. Less obvious alternative scoring functions for the same statistic however exist. For instance [8] reports the most general expression for the scoring function of the quantile

$$S_{\mathbf{q}_\alpha}(y, x) = ((y \geq x) - \alpha)(g(y) - g(x)) \qquad g \text{ nondecreasing} \tag{3}$$

$\mathbf{y}$	$S_{\mathbf{y}}(y,x)$	$\mathcal{F}_S$
$\boldsymbol{\mu}$	$(y-x)^2$	maximal
$\mathbf{q}_{1/2}$	$\lvert y-x\rvert$	maximal
$\mathbf{q}_\alpha$	$\alpha(x-y)_+ + (1-\alpha)(x-y)_-$	maximal
$\mathbf{e}_\alpha$	$\alpha(x-y)_+^2 + (1-\alpha)(x-y)_-^2$	maximal

Table 1: Common examples of canonical scoring functions

and for the mean

$$S_{\boldsymbol{\mu}}(y,x) = \phi(x) - \phi(y) - \phi'(y)(x-y) \qquad \phi \text{ convex} \tag{4}$$

A natural notion of simplest possible, "canonical" scoring function, however, seems to emerge, at least for the most common cases, defined up to inessential transformations only. Importantly, canonical scoring functions respect dimensional analysis rules, when x and y are not dimensionless, which is a crucial property for applications in finance, whereas other more general scoring functions typically do not. Interestingly, we will see (Corollary 3.6) that for the quantile and the mean, only the canonical cases are convex in the prediction variable y, a fact that will turn out to be crucial for backtestability.

It is immediate to notice [8] that elicitability is robust with respect to strictly monotonic transformations $g : \mathbb{R} \to \mathbb{R}$, in the sense that if $\mathbf{y}$ is elicitable, then also $\mathbf{y}_g = g(\mathbf{y})$ is elicitable with $S_{\mathbf{y}_g}(y,x) = S_{\mathbf{y}}(g^{-1}(y),x)$. In particular, a statistic is elicitable if and only if its opposite is elicitable, so that we can equivalently speak of the elicitablity of the quantile $\mathbf{q}$ or of $\mathbf{VaR} = -\mathbf{q}$, for instance. Similarly, we can equivalently speak of the non–elicitability of the variance $\boldsymbol{\sigma}^2$ and the standard deviation $\boldsymbol{\sigma}$.

2.2.1 Model selection via scoring functions

The importance of elicitability stems from the fact that it allows to say something about predictions of a statistic based only on realizations of the random variable, which is exactly the problem we posed ourselves in section 1. If a statistic is elicitable, the *realized mean score*,

$$\bar{s}_{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T} S_{\mathbf{y}}(y_t, x_t) \tag{5}$$

measured over a sequence of forecasts y_t for the statistic $\mathbf{y}(X)$ and realizations x_t of the random variable X can be used to set up a contest of *relative* quality between different models labeled by k issuing sequences of forecasts $y_t^{(k)}$. The idea is that the smaller the mean score $\bar{s}_{\mathbf{y}}^{(k)}$, the better the model k.

The big advantage of this type of test is that it doesn't require an entire predictive probability distribution P_t for issuing the predictions $y_t = \mathbf{y}(P_t)$; the numbers y_t are all what is needed, and the forecast model can be of whatever type.

Notice however that if there is a single model, the mean score will tell nothing on the quality of its predictions in *absolute* terms. For this reason, as we pointed out in [1], this type of procedure allows for *model selection* among multiple models rather than *model validation* which needs an absolute scale for testing even a single model, and is the goal of a proper backtesting procedure.

We also notice that a realized mean score test is not directional, in the sense that it will not reveal whether a model under- or overestimates the statistic. So you may end up preferring a model that underestimates a risk measure over another one that overestimates it by more, which may not be what you want if your goal is of prudential type. In the context of model validation for prudential financial regulation, it is clearly important to look for directional validation methods, which is what a backtesting procedure should provide.

2.3 Identifiability

A close concept to elicitability is *identifiability*.

Definition 2.2. A statistic $\mathbf{y}$ is called $\mathcal{F}$*–identifiable*[3], if there exists an *identification function* $I_{\mathbf{y}}(y, x)$ such that

$$\mathbb{E}_F\left[I_{\mathbf{y}}(y, X)\right] = 0, \quad \text{iff} \quad y \in \mathbf{y}(F), \qquad \forall F \in \mathcal{F} \tag{6}$$

In most common cases, identifiability and elicitability occur jointly, when equation (6) represents the first–order stationarity condition of a scoring function

$$S(y, x) = \int^{y} I(t, x)dt \tag{7}$$

whose expected value $\mathbb{E}_F[S(y, X)]$ has a global minimum in $y \in \mathbf{y}(F)$. Tables 1 and 2 summarize some (canonical) scoring and identification functions for the analyzed examples, all of which are related by eq. (7). Canonical identification functions (and backtest functions in the following) are defined up to an overall multiplicative constant; for convenience we conventionally set the sign of identification functions in accordance to (7), namely such that the function is increasing around $y = \mathbf{y}(F)$.

Similarly to elicitability, we will simply speak of identifiability, without specifying $\mathcal{F}$, when the class is maximal, namely it contains all distribution functions for which $\mathbf{y}(F)$ and $\mathbb{E}_F[I_{\mathbf{y}}]$ are finite. Equation (7) implies that the maximal class for identifiability includes the maximal class for elicitability.

Also identifiability (and later backtestability) is robust with respect to strictly monotonic transformations $g : \mathbb{R} \to \mathbb{R}$, in the sense that if $\mathbf{y}$ is identifiable, then also $\mathbf{y}_g = g(\mathbf{y})$ is identifiable with $I_{\mathbf{y}_g}(y, x) = I_{\mathbf{y}}(g^{-1}(y), x)$.

Remark 2.3. As opposed to all other examples in table 1, the quantile (and the median as a special case) is not identifiable under a maximal class $\mathcal{F}$ of distributions via the proposed (canonical) function, but requires further continuity

[3] Strictly identifiable in most literature.

$\mathbf{y}$	$I_{\mathbf{y}}(y,x)$	$\mathcal{F}_I$
$\boldsymbol{\mu}$	$y-x$	maximal
$\mathbf{q}_{1/2}$	$(y>x)-(y<x)+c(x=y)$	$F(x)$ cont. in $\mathbf{q}_{1/2}$
$\mathbf{q}_\alpha$	$(1-\alpha)(y>x)-\alpha(y<x)+c(x=y)$	$F(x)$ cont. in $\mathbf{q}_\alpha$
$\mathbf{e}_\alpha$	$(1-\alpha)(x-y)_- - \alpha(x-y)_+$	maximal

Table 2: Common examples of canonical identification functions. See remark 2.4.

conditions on $F(x)$, at least in $x=\mathbf{q}_\alpha$. To illustrate the point let $F(x)$ be discontinuous in $x=\mathbf{q}_\alpha\in\mathbb{R}$. We will have $F(\mathbf{q}_\alpha^-)<F(\mathbf{q}_\alpha^+)$ and $F(\mathbf{q}_\alpha^-)\le\alpha\le F(\mathbf{q}_\alpha^+)$. We can immediately check that the expected value of the identification function for a correct prediction $y=\mathbf{q}_\alpha$

$$\mathbb{E}_F[I_{\mathbf{q}_\alpha}(\mathbf{q}_\alpha,X)]=-\alpha(1-F(\mathbf{q}_\alpha^+))+(1-\alpha)F(\mathbf{q}_\alpha^-)$$

is in general different from zero. The problem is intrinsic to the discontinuity of the map $y\mapsto I_{\mathbf{q}}(y,x)$ and is common to any other identification function of the quantile derived from the most general scoring function [8].

Remark 2.4. The identification function of quantile in table 2 (and the backtest function in table 3) is defined up to a term concentrated in $(x=y)$. The most general form is

$$I^c_{\mathbf{q}_\alpha}(y,x)=(1-\alpha)(y>x)-\alpha(y<x)+c(x=y) \tag{8}$$

for any choice of $c\in[-\alpha,1-\alpha]$. For instance, the popular choices $I^{1-\alpha}_{\mathbf{q}_\alpha}(y,x)=(x\le y)-\alpha$ and $I^{-\alpha}_{\mathbf{q}_\alpha}(y,x)=(x<y)-\alpha$ are obtained by setting $c=1-\alpha$ and $c=-\alpha$ respectively.

The range for c comes from the fact that

$$\mathbb{E}_F[I^c_{\mathbf{q}_\alpha}(y,X)]=F(y)-d\Pr[X=y]-\alpha\equiv F_d(y)-\alpha \tag{9}$$

with $d=1-\alpha-c$. We need to require that $F_d(y)\equiv F(y)-d\Pr[X=y]\in[F(y^-),F(y^+)=F(y)]$, otherwise there can be solutions to $\mathbb{E}_F[I^c_{\mathbf{q}_\alpha}(y,X)]=0$ different from $y=\mathbf{q}_\alpha$. This forces $d\in[0,1]$ and therefore $c\in[-\alpha,1-\alpha]$. Notice that this restriction makes (8) nondecreasing in y.

3 Backtesting: definitions, methods and properties

We now have the elements for attempting a definition of backtestability. For a reason that will be clear from remark 3.3, we consider in this section statistics $\mathbf{y}$ that are single–valued and we adopt the correspondingly simpler notation.

3.1 Backtestability

We define backtestability in such a way to be able to rank predictions on a scale where positive (negative) values denote over– (under–) estimation and worse predictions rank further from zero. This will allow us also to set up a meaningful hypothesis test for the correctness of a model prediction in probabilistic terms.

Definition 3.1. We define a statistic $\mathbf{y}$ to be $\mathcal{F}$*–backtestable*, if there exists a *backtest function* $Z_{\mathbf{y}}(y, x)$ such that

$$\mathbb{E}_F\left[Z_{\mathbf{y}}(y, X)\right] = 0, \quad \text{iff} \quad y = \mathbf{y}(F), \qquad \forall F \in \mathcal{F} \tag{10}$$

which is strictly increasing in the prediction y, $\forall F \in \mathcal{F}$.

$$\mathbb{E}_F\left[Z_{\mathbf{y}}(y_1, X)\right] < \mathbb{E}_F\left[Z_{\mathbf{y}}(y_2, X)\right] \quad \text{if} \quad y_1 < y_2 \tag{11}$$

The requirement is very natural: posing this definition, the sign of $\mathbb{E}_F\left[Z_{\mathbf{y}}(y, X)\right]$ will coincide with the sign of the prediction discrepancy $y - \mathbf{y}(F)$

$$\begin{cases} \mathbb{E}_F\left[Z_{\mathbf{y}}(y, X)\right] < 0 & \text{if} \quad y < \mathbf{y}(F) \\ \mathbb{E}_F\left[Z_{\mathbf{y}}(y, X)\right] > 0 & \text{if} \quad y > \mathbf{y}(F) \end{cases} \tag{12}$$

and given two predictions that both overestimate (or underestimate) the statistics, the worse prediction will generate a strictly worse (i.e. more distant from zero) expected backtest function

$$\begin{cases} \mathbb{E}_F\left[Z_{\mathbf{y}}(y_1, X)\right] < \mathbb{E}_F\left[Z_{\mathbf{y}}(y_2, X)\right] < 0 & \text{if} \quad y_1 < y_2 < \mathbf{y}(F) \\ \mathbb{E}_F\left[Z_{\mathbf{y}}(y_1, X)\right] > \mathbb{E}_F\left[Z_{\mathbf{y}}(y_2, X)\right] > 0 & \text{if} \quad y_1 > y_2 > \mathbf{y}(F) \end{cases} \tag{13}$$

Backtestability is slightly more restrictive than identifiability in that it requires the expected value of the identification function to be strictly increasing in the prediction variable y. For this, it is necessary that the identification function be nondecreasing in y for all x and it is sufficient that it be strictly increasing.

The canonical identification functions of $\boldsymbol{\mu}$ and $\mathbf{e}_\alpha$ are strictly increasing in y, and therefore constitute valid backtest functions on the same distribution class $\mathcal{F}$ of identifiability, which is maximal.

The identification function (8) of the quantile $\mathbf{q}_\alpha$ is nondecreasing for any c, but not strictly increasing. We therefore need to check under what distribution functions we obtain strict monotonicity of the expected identification function under y. From (9), remembering that $F(y^-) \leq F_d(y) \leq F(y^+)$, we can immediately conclude that for the expectation to be strictly increasing in y we need to restrict to distribution functions that are overall strictly increasing[4]. We can conclude that

[4]We call strictly increasing a distribution function when it is strictly increasing except where it takes values 0 or 1.

$\mathbf{y}$	$Z_{\mathbf{y}}(y,x)$	$\mathcal{F}_Z$
$\boldsymbol{\mu}$	$y-x$	maximal
$\mathbf{q}_{1/2}$	$(y>x)-(y<x)+c(x=y)$	$F(x)$ cont. in $\mathbf{q}_{1/2}$ and str. incr.
$\mathbf{q}_\alpha$	$(1-\alpha)(y>x)-\alpha(y<x)+c(x=y)$	$F(x)$ cont. in $\mathbf{q}_\alpha$ and str. incr.
$\mathbf{e}_\alpha$	$(1-\alpha)(x-y)_- - \alpha(x-y)_+$	maximal

Table 3: Common examples of canonical backtest functions. In the case of the mean and of the quantile, these are also the only possible backtest functions (see corollary 3.7). We conjecture that also the backtest function of the expectile is unique.

Proposition 3.2. *The quantile $\mathbf{q}_\alpha$ is backtestable only on the class*

$$\mathcal{F} = \{F|\ \textit{overall strictly monotonic; continuous in}\ \mathbf{q}_\alpha\} \tag{14}$$

The scope of a quantile backtest will typically be the smaller class of distributions which are overall both continuous and strictly monotonic.

Remark 3.3. Given that the quantile is backtestable only on strictly increasing distributions on which the quantile is single–valued, we restrict throughout this section to statistics $\mathbf{y}$ which are single–valued. We don't have other interesting examples of interval–valued backtestable statistics.

In all the examples shown in table 2, the identification functions are non-decreasing in y and as we have shown represent also valid backtest functions although in the case of the quantile on a smaller class of distributions. We summarize the results in table 3. In fact we don't know a single example of identifiable statistic which is not also backtestable on some class of distributions, although we can't exclude this possibility.

There exist, on the other hand, identification functions which are not backtest functions as the below proposition shows.

Proposition 3.4. *In the family of identification functions of the quantile*[5]

$$I_{\mathbf{q}_\alpha}(y,x) = g'(y)((y \geq x) - \alpha) \qquad\qquad g\ \textit{increasing} \tag{15}$$

derived from the most general scoring function (3), the only one which is non-decreasing in y for all x, is the canonical one obtained with $g(x) = x$, up to a positive constant.

Proof: We need to impose that

$$\partial_y[g'(y)((y \geq x) - \alpha)] = g''(y)((y \geq x) - \alpha) + g'(y)\delta(x-y) \tag{16}$$

be ≥ 0 for all y and x. But whatever the sign of $g''(y)$, we know that the term $((y \geq x) - \alpha)$ is both positive and negative depending on x and y. The only

[5]For simplicity we fix $c = 1 - \alpha$ in the general expression (8). The choice is irrelevant for the result.

possibility for the above expression for being always positive is therefore that $g''(y) = 0$, which imposes $g(x) = x$. □

An identical result holds for the scoring function of the mean.

Proposition 3.5. *In the family of identification functions of the mean*

$$I_{\boldsymbol{\mu}}(y, x) = \phi''(y)(y - x) \qquad \phi \text{ convex} \tag{17}$$

derived from the most general scoring function (4), the only one which is non-decreasing in y for all x, is the canonical one obtained with $\phi(x) = x^2$, up to a positive constant.

Proof: We need to impose that

$$\partial_y[\phi''(y)(y - x)] = \phi'''(y)(y - x) + \phi''(y) \tag{18}$$

be ≥ 0. We know that $\phi'' \geq 0$. But if $\phi'''(y) \neq 0$ at some y, there exists some x for which the above expression goes negative. Therefore $\phi'''(y) = 0$. But given that $\phi(x) = x$ makes (17) vanish altogether, the only possible solution is $\phi(x) = x^2$. □

Corollary 3.6. *Up to inessential transformations, the only scoring functions for the mean and the quantile that are y–convex are the ones listed in table 1.*

Corollary 3.7. *Up to a positive constant, the only backtest functions of the mean and the quantile are the canonical functions listed in table 3.*

It is immediate to show that backtestability implies elicitability.

Proposition 3.8. *If* $\mathbf{y}$ *is $\mathcal{F}$–backtestable with backtest function $Z_{\mathbf{y}}(y, x)$, then it is also $\mathcal{F}'$–elicitable with y–convex scoring function*

$$S_{\mathbf{y}}(y, x) = \int^{y} Z_{\mathbf{y}}(t, x)\, dt \tag{19}$$

on the class $\mathcal{F}'$ of distributions F for which $S_{\mathbf{y}}(y, X)$ is integrable.

Proof: $y \mapsto \mathbb{E}_F[Z_{\mathbf{y}}(y, X)]$ strictly increasing ensures that $y \mapsto \mathbb{E}_F[S_{\mathbf{y}}(y, X)]$ is convex and has a global minimum in $y \in \mathbf{y}(F)$. □

The above proposition tells us that convexity of the scoring function is a necessary condition for backtestability. And that the backtest function is the y–subdifferential of the scoring function[6].

Elicitability with a strictly y–convex and smooth scoring function, implies backtestability. If the smoothness condition is not met, like in the case of the quantile, restrictions on $\mathcal{F}$ may be necessary to ensure backtestablity.

[6]It is interesting to notice that the condition on c in remark 2.4, although derived from different requirements, corresponds exactly to the subdifferential range in the non differentiable point of the scoring function of the quantile.

Proposition 3.9. *If* $\mathbf{y}$ *is* $\mathcal{F}$*–elicitable with scoring function* $S_{\mathbf{y}}(y,x)$ *which is strictly convex and continuously differentiable in* y*, then it is also* $\mathcal{F}'$*–backtestable with* $\mathcal{F}' \supseteq \mathcal{F}$ *and with backtest function*

$$Z_{\mathbf{y}}(y,x) = \partial_y S_{\mathbf{y}}(y,x) \tag{20}$$

Proof: straightforward. □

3.1.1 Model selection via backtest function

In analogy with section 2.2.1, one can use a realized mean backtest function for performing a selection among models that compete on a sequence of predictions of $\mathbf{y}$, based on realizations $X_t = x_t$

$$\bar{z}_{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T} Z_{\mathbf{y}}(y_t, x_t) \tag{21}$$

Also this test, like the one based on (5) is relative and not absolute. However, the test is directional. A positive (negative) realized value is a sign of over-estimation (resp. underestimation) of the true statistic. The test provides a natural preference criterion (the closer to zero the better) between models that both over– (or under–) estimate. On the other hand, a further criterion has to be supplemented for choosing between models that display opposite tendencies.

3.1.2 Model validation: proper backtesting procedure

To obtain a proper model validation of absolute type, we need[7] to consider the case in which the statistic prediction $y_t = \mathbf{y}(P_t)$ is formulated via an entire predictive distribution function P_t. We can then exploit (12) which becomes

$$\mathbb{E}_F\left[Z_{\mathbf{y}}(\mathbf{y}(P), X)\right] \lesseqgtr \mathbb{E}_P\left[Z_{\mathbf{y}}(\mathbf{y}(P), X)\right] = 0 \quad \text{if} \quad \mathbf{y}(P) \lesseqgtr \mathbf{y}(F) \tag{22}$$

to test whether the prediction $\mathbf{y}(P)$ over/underestimates the real value $\mathbf{y}(F)$. A standard hypothesis test can be set up, checking the compatibility between the model distribution $P_{\bar{Z}}$ of the mean backtest function

$$\bar{Z}_{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T} Z_{\mathbf{y}}(y_t, X_t), \qquad X_t \sim P_t \tag{23}$$

and its realization $\bar{z}_{\mathbf{y}}$, eq. (21). As usual, a preliminary significance level $\eta \lesssim 1$ (such as $\eta = 95\%$), will have to be set, which determines in turn a rejection region of P–measure $1-\eta$ for $\bar{z}_{\mathbf{y}}$. This will span one or both tails of the distribution, depending on whether the test is one– or two–sided.

[7] Quantile will represent an exception, as it does not require resampling of the predictive backtest distribution.

Let's consider for instance a one–sided test on a risk measure $\mathbf{y}$ intended to check if the model is prudential enough. The null hypothesis is that the model issues predictions that do not underestimate the risk measure:

$$H_0 : \mathbf{y}(P_t) \geq \mathbf{y}(F_t) \qquad \forall t = 1, \dots, T$$

The rejection region will span the left tail of the distribution: $\bar{z}_{\mathbf{y}} \in (-\infty, c_\eta]$, where $c_\eta = P_{\bar{Z}}^{-1}(1-\eta)$.

In the absence of analytical results, one needs to simulate the model distribution $P_{\bar{Z}}$ of the random variable $\bar{Z}_{\mathbf{y}}$, generating a suitably large number M of scenarios

$$\begin{aligned} &\text{simulate independent} && X_t^i \sim P_t && \forall t = 1, \dots, T,\ \forall i = 1, \dots, M \\ &\text{compute test scenarios} && \bar{z}_{\mathbf{y}}^i = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} Z_{\mathbf{y}}(\mathbf{y}(P_t), X_t^i) && \forall i = 1, \dots, M \end{aligned}$$

The critical threshold will be estimated by $c_\eta = \bar{z}_{\mathbf{y}}^{M(1-\eta):M}$, where $\bar{z}_{\mathbf{y}}^{i:M}$ denotes as usual the sample's order statistics. Finally, the null hypothesis will be rejected if $\bar{z}_{\mathbf{y}} \leq c_\eta$.

Remark 3.10. The case of quantile backtesting (like the standard **VaR** backtest in Basel regulation) is exceptional in the sense that the distribution $P_{\bar{Z}}$ is binomial, independently of the model, and for this reason requires no montecarlo resampling. Implementation of the backtest is therefore much more straightforward, in particular for what concerns data storage. To backtest a quantile it is sufficient to record, for every t, the numerical prediction y_t and the realization x_t, whereas for other statistics, it is necessary to record also the entire probability distribution P_t, for resampling purposes.

However, from a conceptual point of view, the standard quantile backtest based on the counting of "exceedances" (i.e. on $Z_{\mathbf{q}_\alpha}(y, x) = (x \leq y) - \alpha$), follows exactly the same logic as the above test of hypothesis.

Remark 3.11. We may be tempted to use a scoring function instead of a backtest function for a similar bootstrap procedure, to create some type of two–sided test. By simulating a model distribution $P_{\bar{S}_{\mathbf{y}}}$ for the realized mean score, we may believe that the realization under a different real distribution F should have a larger expected value, unless the prediction is correct. This argument is however flawed, because it assumes the following property to hold, in analogy with (22).

$$\mathbb{E}_F[S_{\mathbf{y}}(\mathbf{y}(P), X)] \geq \mathbb{E}_P[S_{\mathbf{y}}(\mathbf{y}(P), X)], \qquad \forall P, F \in \mathcal{F}_{\mathcal{S}}, \qquad \text{false in general} \quad (24)$$

The problem is that this property does not hold for a (and probably for any) scoring function.

As a counterexample, let's consider the scoring function $S_{\boldsymbol{\mu}}$ in table 1. We can easily compute

$$\mathbb{E}_F[S_{\boldsymbol{\mu}}(\boldsymbol{\mu}(P), X)] = \boldsymbol{\sigma}^2(F) + (\boldsymbol{\mu}(F) - \boldsymbol{\mu}(P))^2$$

$$\mathbb{E}_P[S_{\boldsymbol{\mu}}(\boldsymbol{\mu}(P), X)] = \boldsymbol{\sigma}^2(P)$$

and realize that the difference between these two expressions has a positive contribution coming from a wrong prediction of $\boldsymbol{\mu}$ but also a contribution given by the difference between the two variances. The difference can therefore be negative as soon as $\boldsymbol{\sigma}^2(F) < \boldsymbol{\sigma}^2(P) - (\boldsymbol{\mu}(F) - \boldsymbol{\mu}(P))^2$ when this last expression is larger than zero.

3.2 Sharpness

Backtestability requires the expectation of the backtest function to be strictly monotonic with respect to the prediction variable: worse predictions rank worse, given a real distribution. We can however reverse the logic and ask ourselves: is the opposite true? Is a given prediction ranking worse if the real value of the statistic is closer than if it is farther? The question is legitimate: we would like our validation system to send louder warnings for larger discrepancies. If that weren't the case, for instance, a traffic light indicator for ongoing risk model validation could turn from yellow to red over time when in fact the prediction power is improving and not worsening. Or viceversa, from red to yellow when it is worsening and not improving.

And yet, natural as it sounds, the requirement,

$$\mathbb{E}_{F_1}[Z_{\mathbf{y}}(y, X)] > \mathbb{E}_{F_2}[Z_{\mathbf{y}}(y, X)] \quad \text{iff} \;\; \mathbf{y}(F_1) < \mathbf{y}(F_2), \;\; \forall y \tag{25}$$

namely that the expected backtest function be strictly decreasing with respect to the true value of the statistics, is not implied by the definition of backtest function, and in fact fails to hold in the notable cases of quantiles and expectiles. We therefore add to backtestability the above requirement and pose the following

Definition 3.12. We define the backtest $Z_{\mathbf{y}}(y, x)$ of a $\mathcal{F}$–backtestable statistic $\mathbf{y}$ to be *sharp* if it is strictly decreasing in the value $\mathbf{y}(F)$ of the statistic, in the sense that

$$\mathbb{E}_{F_1}\left[Z_{\mathbf{y}}(y, X)\right] > \mathbb{E}_{F_2}\left[Z_{\mathbf{y}}(y, X)\right] \quad \text{iff} \;\; \mathbf{y}(F_1) < \mathbf{y}(F_2); \;\; \forall F_1, F_2 \in \mathcal{F}, \forall y \tag{26}$$

Condition (26) implies in particular that $\mathbb{E}_{F_1}(Z_y(y, X)) = \mathbb{E}_{F_2}(Z_y(y, X))$ if $\mathbf{y}(F_1) = \mathbf{y}(F_2)$ which means that $\mathbb{E}_F(Z_y(y, X))$ is a function of $\mathbf{y}(F)$ only, and not of all F. An equivalent definition is therefore

Definition 3.13. We define the backtest $Z_{\mathbf{y}}(y, x)$ of $\mathcal{F}$–backtestable statistic $\mathbf{y}$ to be *sharp*, if for all $F \in \mathcal{F}$, the expectation

$$\mathbb{E}_F\left[Z_{\mathbf{y}}(y, X)\right] = \psi(y, \mathbf{y}(F)) \tag{27}$$

is a function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of the prediction y and of the statistic value $\mathbf{y}(F)$, which is strictly increasing in the former and strictly decreasing in the latter.

The importance of sharp backtests (and the reason why we call them sharp) is understood by noting that for any value z of the expected value (27) there's one and only one compatible value

$$\mathbf{y}(F) = \psi^{-1}(y, z) \tag{28}$$

of the real statistic, where by ψ^{-1} we denote the inverse function of the second argument of ψ. This is a fundamental point: a sharp backtest will quantify the discrepancy of a prediction and not only test whether or not the real value is compatible with the model as a generic backtest does.

The point has been overlooked for long time in the case of **VaR** backtesting for banking regulation. A bank model is withdrawn the regulator's approval if it repeatedly ranks red in a traffic light system even if, as we will see, this could be caused by very small discrepancies between prediction and real value. And more worryingly, the traffic light could be yellow or even green even when the underestimation is huge. As a matter of fact, the color of the traffic light, in the case of **VaR** tells absolutely nothing about the real **VaR** value, as proposition 3.15 will show. Sharp backtests, on the contrary, measure both the likelihood and the extent of a wrong prediction.

It is immediate to check that the backtest of the mean μ is sharp. As a matter of fact, it is obvious that all statistics that are explicitly defined as an expectation or strictly increasing functions thereof, have a sharp backtest.

Proposition 3.14. *Let $g, h : \mathbb{R} \to \mathbb{R}$, g strictly monotonic and let*

$$\mathbf{y}(F) \equiv g(\mathbb{E}_F[h(X)])$$

. Then $\mathbf{y}$ is $\mathcal{F}$–backtestable on maximal $\mathcal{F}$, with sharp backtest function

$$Z_{\mathbf{y}}(y, x) = \nu\, g^{-1}(y) - \nu\, h(x) \tag{29}$$

where $\nu \in \{\pm 1\}$ is the sign of g'.

Proof: obvious. □

3.2.1 Information content of the Quantile and Expectile backtests

We will show that both the backtests of the quantile and the expectile are not sharp. It is therefore interesting to study how informative a backtest of these statistic is in the localization of the real value.

Let's begin by analyzing the quantile, recalling that its backtest function is unique, up to a positive multiplicative constant and the choice of c in (8).

Proposition 3.15. *The backtest of the quantile $\mathbf{q}_\alpha$ is not sharp on its $\mathcal{F}$–backtestability class (14).*

For any prediction y and expected value $z = \mathbb{E}_F[Z_{\mathbf{q}_\alpha}(y, X)]$ of the backtest function, the possible range of $\mathbf{q}_\alpha(F)$ is given by

$$\begin{cases} \mathbf{q}_\alpha(F) \in (-\infty, y) & \text{if} \quad z > 0 \\ \mathbf{q}_\alpha(F) \in (y, +\infty) & \text{if} \quad z < 0 \end{cases} \tag{30}$$

Proof: For the proof we fix $c = 1 - \alpha$ in (8). It is easy to see that the choice is completely inessential. If $F \in \mathcal{F}$ then $z = F(y) - \alpha$. If $z > 0$, $\exists F \in \mathcal{F}$ such that $F^{-1}(\alpha) = y - \epsilon$ as well as $F^{-1}(\alpha) = -M$, for arbitrarily small $\epsilon > 0$ and large $M > 0$. If $z < 0$, $\exists F \in \mathcal{F}$ such that $F^{-1}(\alpha) = y + \epsilon$ as well as $F^{-1}(\alpha) = +M$. This proves the range conditions on $\mathbf{q}_\alpha = F^{-1}(\alpha)$.

Then $Z_{\mathbf{q}_\alpha}$ is not sharp because the above would contradict (28). □

The result of the above proposition is quite dramatic. In plain English it means that you have no clue whatsoever on the location of the real quantile based on the result of a backtest, even if you knew the exact expected backtest function. It can be anywhere. Also notice that to enforce the missing property (26), we should restrict to a class $\mathcal{F}'$ such that

$$\mathbb{E}_{F_1}\left[Z_{\mathbf{q}_\alpha}(y, X)\right] - \mathbb{E}_{F_2}\left[Z_{\mathbf{q}_\alpha}(y, X)\right] = F_1(y) - F_2(y) > 0$$

when $\mathbf{q}_\alpha(F_1) < \mathbf{q}_\alpha(F_2)$, for all $F_1, F_2 \in \mathcal{F}'$ and for all y. But this means $F_1 <_{1^{st}} F_2$ in the sense of first stochastic dominance. Therefore the only classes on which the backtest of the quantile can be made sharp are totally ordered by first stochastic dominance. No class of this type makes any sense for practical application purposes, so that in practical terms we can affirm that the quantile backtest is not sharp at all.

The only information content of the backtest of the quantile, is restricted to saying whether the real value is above or below the prediction.

Remark 3.16. One may confuse the result of proposition 3.15 with the elementary fact that a quantile is blind to the magnitude of the risks in the tail, in the sense that the position of the quantile doesn't tell you anything on how severe the further tail events are. The above result has nothing to with this. The backtest of the quantile tells nothing on the position of the quantile which in turn would tell nothing on the position of the risks of the tail. We are speaking in other words of a (new) problem on top of another (well known) one.

We now consider the expectile. We pose the following

Conjecture 3.17. *Also for the expectile, as for the quantile, the backtest function is unique (table 3) up to a positive multiplicative constant*

Proposition 3.18. *The backtest of the expectile $\mathbf{e}_\alpha$ is not sharp on its maximal $\mathcal{F}$–backtestability class unless $\alpha = 1/2$.*

For any prediction y and expected value $z = \mathbb{E}_F[Z_{\mathbf{e}_\alpha}(y, X)]$ of the backtest

function, the possible range of $\mathbf{e}_\alpha(F)$ *is given by*

$$\begin{cases} \mathbf{e}_\alpha(F) \in (y - z/\alpha, y - z/(1-\alpha)] & \text{if} \quad z > 0 \quad \text{and} \quad \alpha < 1/2 \\ \mathbf{e}_\alpha(F) \in (y - z/(1-\alpha), y - z/\alpha] & \text{if} \quad z < 0 \quad \text{and} \quad \alpha < 1/2 \\ \mathbf{e}_\alpha(F) \in [y - z/(1-\alpha), y - z/\alpha) & \text{if} \quad z > 0 \quad \text{and} \quad \alpha > 1/2 \\ \mathbf{e}_\alpha(F) \in [y - z/\alpha, y - z/(1-\alpha)) & \text{if} \quad z < 0 \quad \text{and} \quad \alpha > 1/2 \\ \mathbf{e}_\alpha(F) = y - 2z & \text{if} \quad \alpha = 1/2 \end{cases} \tag{31}$$

Proof: If $F \in \mathcal{F}$ then the map $t \mapsto h(t) = \mathbb{E}_F[Z_{e_\alpha}(t, X)]$ can be shown to be strictly increasing and convex (concave) for $\alpha < 1/2$ (resp. $\alpha > 1/2$) with first derivative $h'(t) = \alpha + (1 - 2\alpha)F(t)$ and asymptotes

$$\begin{array}{ll} h(t) \sim \alpha(t - \mathbb{E}_F[X]) & \text{if} \quad t \to -\infty \\ h(t) \sim (1-\alpha)(t - \mathbb{E}_F[X]) & \text{if} \quad t \to +\infty \end{array}$$

Let us prove (31) for $z > 0$ and $\alpha < 1/2$. It is convenient to plot the function on the plane $(t, h(t))$. The expectile is the value t that solves $h(t) = 0$. From $h(y) = z > 0$ and the fact that the minimum and maximum slopes of the function h are α and $1 - \alpha$, we conclude that $\mathbf{e}_\alpha \in [y - z/\alpha, y - z/(1-\alpha)]$. But given that $z > 0$, the point $(t, h(t)) = (\mathbf{e}_\alpha, 0)$ can be on the right asymptote, but not on the left. Therefore, the right extreme of the interval can be attained, but the left one can not.

A distribution family that spans all the range is given by

$$F_\chi(t) = \chi \left(t \geq y - \frac{z}{\chi(1-\alpha)} \right) + (1 - \chi)(t \geq y) \qquad \chi \in (0, 1]$$

One can easily prove that $\mathbb{E}_{F_\chi}[Z_{\mathbf{e}_\alpha}(y, X)] = z$ and

$$\mathbf{e}_\alpha(F) = y - \frac{z}{\chi + \alpha - 2\chi\alpha} \in (y - z/\alpha, y - z/(1-\alpha)]$$

The three other cases in (31) can be proven in analogous way.

Then $Z_{\mathbf{e}_\alpha}$ is not sharp for $\alpha \neq 1/2$ because the above would contradict (28). □

Notice that the discrete nature of the distributions used in the proof is inessential, apart from the attained extreme of the interval. Smoothing the jumps of the distribution with arbitrarily peaked gaussians would still span $\mathbf{e}_\alpha(F) \in (y - z/\alpha, y - z/(1-\alpha))$.

The result is clear. The expectile has a sharp backtest on $\mathcal{F}$ if and only if $\alpha = 1/2$, where it coincides with the mean of the distribution. For $\alpha \neq 1/2$, the real expectile is bounded in a range which is smaller and smaller as α approaches $1/2$. The information content of the backtest in other words decreases when the statistic is used as an extreme tail statistic. In the limits $\alpha \to 0, 1$ the range size diverges.

Interestingly, from the proof we learn that $h'(t) \propto F(t)$ so that $h(t) \propto \int^t F$. So, similarly to the quantile case, to impose the expectile backtest to be sharp

Figure 3: Risk management on a sharp ridge

we should restrict to distributions F_1, F_2 such that $\forall y$

$$\mathbb{E}_{F_1}\left[Z_{\mathbf{e}_\alpha}(y, X)\right] - \mathbb{E}_{F_2}\left[Z_{\mathbf{e}_\alpha}(y, X)\right] \propto \int^y (F_1 - F_2) > 0$$

whenever $e_\alpha(F_1) < e_\alpha(F_2)$, which means that we should restrict to classes of distributions which are totally ordered by second order stochastic dominance $F_1 <_{2^{nd}} F_2$; as in the case of the quantile, this is again an inconceivable condition for any practical application. We can say with no hesitation that the expectile admits no sharp backtest altogether, unless $\alpha = 1/2$.

We conclude the section noticing that the backtest of the expectile is sharper than the one of the quantile, but less and less so for the values of α which are most interesting for risk management purposes. It is important to recall that this uncertainty in the position of the real statistic is there also in the absence of statistical errors, namely assuming the perfect knowledge of the expected value of the backtest function. Adding to this the inevitable statistical errors that the estimation of the expected value brings along, raises serious questions as to whether the non–sharp backtest of a statistic should be relied upon for risk management purposes at all.

4 Ridge backtests

We have learned that the first thing to look at, in order to see whether a statistic $\mathbf{y}$ is backtestable or not is the existence of an expression of the type $\mathbb{E}[I_{\mathbf{y}}(\mathbf{y}(X), X)] = 0$, namely a null expectation which involves only the statistic itself and the random variable X. This is the most intuitive way for understanding whether or not a statistic is in turn elicitable, identifiable, backtestable. If

the definition of the statistic doesn't immediately lend itself to a representation of this type, the statistic is probably not elicitable to begin with, although a formal proof may be nontrivial.

Variables like the variance $\boldsymbol{\sigma}^2$ and the tail mean **TM** (or the expected shortfall **ES** $= -$**TM**) have been proven to be not elicitable [11, 8] and therefore are not backtestable. This should not sound surprising. A null expectation involving the variance, for instance, exists, following directly from its definition

$$\mathbb{E}_F[\boldsymbol{\sigma}^2(F) - (X - \boldsymbol{\mu}(F))^2] = 0 \tag{32}$$

but it irreducibly involves also another statistic, the mean $\boldsymbol{\mu}$. Similarly, the tail mean satisfies (53) from which it is immediate to obtain a null expectation,

$$\mathbb{E}_F \left[\mathbf{TM}_\alpha(F) - \mathbf{q}_\alpha(F) + \frac{1}{\alpha}(X - \mathbf{q}_\alpha(F))_- \right] = 0 \tag{33}$$

Also this expression, however, involves another statistic, the quantile $\mathbf{q}$.

In general, when there exists a condition like

$$\mathbb{E}[I(\mathbf{y}_2(X), \mathbf{y}_1(X), X)] = 0 \tag{34}$$

involving two statistics $\mathbf{y}_1$ and $\mathbf{y}_2$, a single backtest may tell very little on the prediction quality of either statistic, because deviations from zero expectation may be driven by discrepancies in both predictions. One may try to perform a preliminary backtest on say $\mathbf{y}_1$, if it is separately backtestable, and then use the joint relationship to backtest $\mathbf{y}_2$, conditionally to $\mathbf{y}_1$ being correct. This is essentially the strategy we adopted in [1], with the backtests Z_1 and Z_2 for $\mathbf{y}_2 = \mathbf{ES}$, in the assumption that the prediction for $\mathbf{y}_1 = \mathbf{VaR}$ had previously been tested. This strategy, however, poses two serious problems. First of all, what if the prediction of $\mathbf{y}_1$ turned out to be not correct? What is the sensitivity of the backtest for $\mathbf{y}_2$ to this discrepancy? But secondly, and more importantly: how can one ascertain at all that the prediction for $\mathbf{y}_1$ is spot on, given that via hypothesis testing we can at most rule out large discrepancies, and only up to some significance level?

It is clear from these arguments that a useful backtest methodology for $\mathbf{y}_2$ based on a condition of type (34) can be obtained only if some model–independent mechanism ensures small sensitivity on predictions to $\mathbf{y}_1$. Incidentally, this is exactly the case of the variance and the tail mean. These two statistics are very special in that they are the attained minimum (up to a sign for **TM**) of the scoring function of a partner variable, $\boldsymbol{\mu}$ and $\mathbf{q}$ respectively.

We pose the following

Definition 4.1. We say that a statistic $\mathbf{y}_2$ admits a *ridge $\mathcal{F}$–backtest*

$$Z_{\mathbf{y}_2}(y_2, y_1, x) = h(y_2) - \nu\, S_{\mathbf{y}_1}(y_1, x) \tag{35}$$

if it can be expressed (up to a strictly monotonic function $g : \mathbb{R} \to \mathbb{R}$) as the minimum of the expected $\mathcal{F}$–scoring function $S_{\mathbf{y}_1}$ of an elicitable *auxiliary*

statistic $\mathbf{y}_1$

$$\begin{cases} \mathbf{y}_2(F) = g\left(\min_y E_F[S_{\mathbf{y}_1}(y, X)]\right) \\ \mathbf{y}_1(F) = \arg\min_y E_F[S_{\mathbf{y}_1}(y, X)] \end{cases}, \qquad F \in \mathcal{F} \tag{36}$$

In (35), $\nu \in \{\pm 1\}$ is the sign of g', and we denote $h(x) \equiv \nu\, g^{-1}(y)$.

The utility of this definition is explained by the following

Proposition 4.2. *Let $Z_{\mathbf{y}_2}$ be a ridge $\mathcal{F}$–backtest for $\mathbf{y}_2$ with auxiliary statistic $\mathbf{y}_1$ as in definition 4.1. Then*

- *the expected backtest is zero in the correct predictions for $\mathbf{y}_2$ and $\mathbf{y}_1$*

$$\mathbb{E}_F[Z_{\mathbf{y}_2}(\mathbf{y}_2(F), \mathbf{y}_1(F), X)] = 0 \qquad \forall F \in \mathcal{F} \tag{37}$$

- *$Z_{\mathbf{y}_2}$ acts as a backtest for $\mathbf{y}_2$ with a one–sided bias depending only on y_1 in the sense that*

$$\mathbb{E}_F[Z_{\mathbf{y}_2}(y_2, y_1, X)] = (h(y_2) - h(\mathbf{y}_2(F))) - \nu\, \mathbb{E}_F[S_{\mathbf{y}_1}(y_1, X) - S_{\mathbf{y}_1}(\mathbf{y}_1(F), X)] \tag{38}$$

so that

$$\begin{cases} \mathbb{E}_F[Z_{\mathbf{y}_2}(y_2, y_1, X)] \leq (h(y_2) - h(\mathbf{y}_2(F))) & \text{if} \quad \nu > 0 \\ \mathbb{E}_F[Z_{\mathbf{y}_2}(y_2, y_1, X)] \geq (h(y_2) - h(\mathbf{y}_2(F))) & \text{if} \quad \nu < 0 \end{cases} \qquad \forall y_1 \tag{39}$$

- *If $\mathbb{E}_F[S_{\mathbf{y}_1}(y_1, X)]$ is continuously differentiable in $y_1 = \mathbf{y}_1(F)$ for $F \in \mathcal{F}' \subseteq \mathcal{F}$, then*

$$\mathbb{E}_F[Z_{\mathbf{y}_2}(y_2, y_1, X)] = (h(y_2) - h(\mathbf{y}_2(F))) + \mathcal{O}(y_1 - \mathbf{y}_1(F))^2 \tag{40}$$

- *$Z_{\mathbf{y}_2}$ is sharp for $\mathbf{y}_2$ up to terms $\mathcal{O}(y_1 - \mathbf{y}_1(F))^2$*

Proof: Eq. (38) follows directly from the definition (35), noting that (36) implies $\mathbf{y}_2(F) = g(\mathbb{E}_F[S_{\mathbf{y}_1}(\mathbf{y}_1(F), X)])$. Eq. (39) follows from (38) and the second equality in (36). Eq. (37) follows from (38).

If $\mathbb{E}_F[S_{\mathbf{y}_1}(y_1, X)]$ is continuously differentiable in the minimum $y_1 = \mathbf{y}_1(F)$, it is at least quadratic. This shows (40) which in turn proves the last assertion. □

The above proposition contains some important facts. When you climb the ridge of a mountain, if you lose your way on either side of the edge, you can be sure of one thing: that you'll find yourself below where you should be. Similarly, a ridge backtest for a statistic $\mathbf{y}_2$ has a one–sided dependence only on the prediction y_1 of the auxiliary statistic $\mathbf{y}_1$. This allows to draw conclusions at least in one direction.

Suppose to fix the ideas that $\nu > 0$. And imagine that we want to test against underestimations of $\mathbf{y}_2$. These will imply a negative expected backtest

$$\mathbb{E}_F[Z_{\mathbf{y}_2}(y_2, y_1, X)] < 0 \quad \text{if} \quad y_2 < \mathbf{y}_2(F), \tag{41}$$

irrespective of the prediction y_1. An imperfect prediction y_1 will make the result even more negative, or in other words, will make the test a bit more prudential. Very importantly, the magnitude of this bias will be small around a correct prediction $y_1 = \mathbf{y}_1(F)$ where it attains the minimum. Under smooth conditions, the sensitivity to y_1 will be zero at first order.

With these observations in mind, a ridge backtest can be adopted both for model selection (section 3.1.1) and for model validation (section 3.1.2).

Examples of statistics admitting a ridge backtest can be obtained given any scoring function $S_{\mathbf{y}_1}$ (even non y–convex) and any strictly monotonic function g. Convexity of $S_{\mathbf{y}_1}$ plays no role in the ridge backtest mechanism.

4.1 Variance

The variance admits a ridge backtest. It is well known that

$$\boldsymbol{\sigma}^2(F) = \min_y \mathbb{E}_F[(X-y)^2] = \min_y \mathbb{E}_F[S_{\boldsymbol{\mu}}(y, X)] \tag{42}$$

where $S_{\boldsymbol{\mu}}$ is the canonical scoring function of the mean in table 1. The variance therefore admits a ridge $\mathcal{F}$–backtest on the maximal class where it is integrable, with $g(x) = x$, $\nu = +1$. Its backtest function is given by

$$Z_{\boldsymbol{\sigma}^2}(v, m, x) = v - (x-m)^2 \tag{43}$$

The sensitivity to predictions m of the mean is quadratic around $m = \boldsymbol{\mu}(F)$ because $S_{\boldsymbol{\mu}}(y, x)$ is continuously differentiable in y.

The existence of a ridge backtest explains the apparent paradox that despite the variance is not elicitable (hence not backtestable), it's never been difficult to perform effective backtests, under even rudimental predictions for the mean, such as $m = 0$. Estimation errors in the mean affect the backtest very mildly and only in the direction of penalizing possible underestimations of the variance.

As a simple example of robustness of backtestability under strictly monotonic transformations of the statistic, we notice that the standard deviation $\boldsymbol{\sigma}$ also admits a ridge backtest

$$Z_{\boldsymbol{\sigma}}(s, m, x) = s^2 - (x-m)^2 \tag{44}$$

as per definition 4.1 with $g(x) = \sqrt{x}$.

4.2 Tail Mean / Expected Shortfall

Let's consider the case of $\mathbf{TM} = -\mathbf{ES}$. For simplicity let's consider the expected shortfall, which also has $g(x) = x$, $\nu = +1$. To show that $\mathbf{ES}$ admits a ridge backtest, we recall from (55) that

$$\begin{aligned} \mathbf{ES}_\alpha &= \min_q \left\{-q + \tfrac{1}{\alpha}\mathbb{E}[(X-q)_-]\right\} \\ \mathbf{q}_\alpha &= \arg\min_q \left\{-q + \tfrac{1}{\alpha}\mathbb{E}[(X-q)_-]\right\} \end{aligned} \tag{45}$$

To show that $S'_{\mathbf{q}_\alpha}(y,x) = -y + (1/\alpha)(x-y)_-$ is a scoring function for $\mathbf{q}_\alpha$ we notice that it coincides with the scoring function $S_{\mathbf{q}_\alpha}$ in table 1 up to the inessential transformation:

$$S'_{\mathbf{q}_\alpha}(y,x) = S_{\mathbf{q}_\alpha}(y,x)/\alpha - x \tag{46}$$

The corresponding backtest function (which is defined up to a positive constant) for predictions e of $\mathbf{ES}_\alpha$ can be written as

$$Z_{\mathbf{ES}_\alpha}(e,q,x) = \alpha(e+q) - (x-q)_- = \alpha(e+q) + (x-q)(x-q<0) \tag{47}$$

If we express this in terms of predictions v for $\mathbf{VaR}_\alpha = -\mathbf{q}_\alpha$ we have

$$Z_{\mathbf{ES}_\alpha}(e,v,x) = \alpha(e-v) + (x+v)(x+v<0) \tag{48}$$

where the name of the variables should prevent possible confusion between these last two expressions. This shows that **ES** admits a ridge $\mathcal{F}$–backtest on all distributions $F \in \mathcal{F}$ where it exists.

The scoring function of the quantile is however not continuously differentiable. The sensitivity to q (or v) of the backtest is zero at first order only requiring the same conditions under which the quantile is identifiable, namely that the distribution F be continuous in $\mathbf{q}_\alpha(F) = -\mathbf{VaR}(F)$.

We observe that the ridge backtest (48) is very similar to test Z_2 proposed in [1], which can be written as

$$Z_2(e,v,x) = \alpha e + x(x+v<0) \tag{49}$$

This expression, was derived from (54) in the "simplifying" assumption of continuous distributions under which it satisfies $\mathbb{E}[Z_2(\mathbf{ES},\mathbf{VaR},X)] = 0$. In reality, this test suffers from significant linear sensitivity on the **VaR** prediction v, as it fails to reflect the mechanism of the ridge backtest for **ES**. The sensitivity is there even for continuous distributions, which may appear surprising. Backtest (48) has to be preferred in any circumstance over test (49).

Example 4.3. To illustrate this point, we consider the following experiment: let $F = N(0,1)$ be a real distribution and let $P_v = N(\mu_v, \sigma_v^2)$ be a family of predictive distributions which forecast $\mathbf{ES}_\alpha = 5\%$ exactly ($e = \mathbf{ES}_\alpha(P_v) = \mathbf{ES}_\alpha(F)$) but yield varying predictions $v = \mathbf{VaR}_\alpha(P_v)$ around $\mathbf{VaR}_\alpha(F)$. Figure 4 plots the behavior of the mean over $T = 250$ days of the two tests in a backtesting procedure (as in section 3.1.2) meant to exclude underestimations up to significance level $\eta = 95\%$.

From the plot, we see that Z_2 displays significant linear sensitivity to v, possibly leading to model rejection (a type I error) also for relatively small underestimations $v < \mathbf{VaR}_\alpha(F)$. Test $Z_{\mathbf{ES}}$ on the other end displays much lower sensitivity in v, in fact zero at first order, and is affected only in the prudential direction. Rejection can still occur, but only for huge estimation errors of $\mathbf{VaR}_\alpha$.

Let now instead consider predictions chosen in such a way that they all underestimate the expected shortfall $\mathbf{ES}_\alpha(P_v) = \mathbf{ES}_\alpha(F) - 1$. Figure 5 shows

that the linear sensitivity of Z_2 this time can be responsible for the failure to reject a wrong model (type II error), something that will never happen with $Z_{\mathbf{ES}}$ given that its bias sign is prudential.

Remark 4.4 (Ridge backtest and banking regulation). The fact that the **ES** admits a ridge backtest is relevant for model validation purposes in the financial industry, notably the banking regulation debate discussed in section 1.2. The **ES** has long been considered non–backtestable, and rightly so, strictly speaking, as our definition of backtestability confirms. We now know, however, that it admits a backtest which is biased in a prudential way and negligibly so unless **VaR** is grossly misspecified. And up to this bias, the backtest is sharp, namely provides information on the magnitude and not only the likelihood of a prediction discrepancy, as opposed to a traditional **VaR** backtest. The adoption of a ridge backtest for **ES** makes a lot of sense: if **VaR** is not grossly misspecified by the model, the backtest is sharp. If **VaR** were to be completely misspecified, either way, even a correct prediction of **ES** would be suspicious, because it would come from a predicted tail of wrong shape. The bias, in other words, could be welcome as a penalty for models that predict **ES** correctly only by pure luck.

Speaking of luck, it's worth observing that in a ridge backtest, the direction of the bias (the sign of ν) is built in with the statistic, and cannot be chosen at will. Despite Murphy's law, in the case of **ES**, the toast did not land butter–side down. It's only for a fortunate coincidence that the bias turns out to be prudential for tests meant to exclude underestimations of **ES**. If for whatever reason someone wanted to test for overestimations of **ES** there would be no ridge backtest that serves the purpose with a prudential bias.

5 Conclusions

A clear definition of backtestability of a statistic was in demand, after decades of practices. If there's one feature that deserves this name, is the existence of an expectation involving only the statistic prediction and the random variable (the only two quantities that are observable in an experiment), which is strictly monotonic in the prediction and zero when it's correct (for detecting under– and overestimations and ranking prediction accuracy). This is what enables model selection based on the statistic point predictions and hypothesis backtesting based on entire predictive distributions.

Not all backtests, however, are also strictly increasing in the real value of the statistic, for a fixed prediction. We call this property sharpness because the expected value of such a backtest determines exactly the real value of the statistic, providing information also on the extent of a wrong prediction. Sharpness is a natural requirement if we want louder warnings for worse discrepancies. Non sharp backtests (quantile, expectile) provide limited if any information at all on the possible real values of the statistic.

Some statistics (variance, Expected Shortfall), despite being not backtestable, admit a null expectation involving another auxiliary statistics (resp. the mean, Value at Risk), which is extremal in the latter. The corresponding backtest

(ridge backtest) has limited sensitivity in the prediction of the auxiliary variable and known bias sign. This is the mechanism that has always allowed for effective tests of variance predictions without caring much if the predictions of the mean were accurate.

We show that the same mechanism allows for effective ways to backtest the Expected Shortfall. For a fortunate circumstance, the modest bias to Value at Risk predictions has always prudential direction when it comes to detecting underestimations of the Expected Shortfall, the natural use case for a risk measure. This backtest is also sharp, as opposed to the backtest of Value at Risk which is completely blind to the magnitude of prediction errors. This result is relevant for banking regulatory standards, which now adopt Expected Shortfall as a risk measure for capital adequacy, but are still based on backtests of Value at Risk.

A Definitions: quantile, tail mean, expectile

We collect the definitions adopted for the concepts of quantile, tail mean, and expectile. When changed sign, these statistics are used to define tail risk measures on a profit–loss distribution, respectively known as Value at Risk, Expected Shortfall and Expectile Risk.

In this section, X denotes a random variable with distribution function F and integrability properties specified case by case. We will denote by $\alpha \in (0,1)$ a confidence level, typically small.

A.1 Quantile and Value at Risk

Definition A.1 (**Quantile**)**.** Given a r.v. $X \in L_0$ we define the α–quantile as the interval–valued quantity

$$\mathbf{q}_\alpha(F) = [x_\alpha, x^\alpha] \tag{50}$$

where the lower and upper quantiles are respectively defined by

$$\begin{cases} x_\alpha = \inf\{x | F(x) \geq \alpha\} \\ x^\alpha = \sup\{x | F(x) \leq \alpha\} \end{cases} \tag{51}$$

$\mathbf{q}_\alpha$ is a single–valued statistic if and only if the inverse image $F^{-1}(\alpha)$ is single–valued or empty.

We define *Value at Risk* as the opposite of the quantile, $\mathbf{VaR}_\alpha = -\mathbf{q}_\alpha$.

A.2 Tail Mean and Expected Shortfall

Definition A.2 (**Tail Mean**)**.** Given a random variable[8] $X \in L_1$, we define the tail mean as [2]

$$\mathbf{TM}_\alpha(F) = \frac{1}{\alpha}\int_0^\alpha \mathbf{q}_t(F)\, dt \tag{52}$$

We define the *Expected Shortfall* as the opposite of the tail mean, $\mathbf{ES}_\alpha = -\mathbf{TM}_\alpha$.

An equivalent useful formulation [2] is given by[9]

$$\begin{aligned} \mathbf{TM}_\alpha(F) &= \frac{1}{\alpha}\mathbb{E}_F\{X(X - \mathbf{q}_\alpha(F) < 0) + \mathbf{q}_\alpha(F)[\alpha - (X - \mathbf{q}_\alpha(F) < 0)]\} \\ &= \mathbf{q}_\alpha(F) - \frac{1}{\alpha}\mathbb{E}_F[(X - \mathbf{q}_\alpha(F))_-] \end{aligned} \tag{53}$$

Notice that when $\mathbf{TM}_\alpha$ exists, it is always single–valued, even when $\mathbf{q}_p$ is interval–valued for some $p \in [0, \alpha]$. Formulae (52) and (53) do not depend

[8] It is sufficient to require that only the left tail be L_1.

[9] We use the standard notation for the positive part $(x)_+ = \max(x, 0) = x(x > 0)$ and negative part $(x)_- = -\min(x, 0) = -x(x < 0)$ of a number x.

on the choice of a single value $\mathbf{q}_p$ in what may possibly be an interval. Also, (53) remains valid if strict inequalities $<$ are replaced with $\leq$. See [2].

We report here also two widely used alternative expressions that hold for continuous distributions[10].

$$\mathbf{TM}_\alpha(F) \overset{F\,\text{cont.}}{=} \frac{1}{\alpha}\mathbb{E}_F\{X(X-\mathbf{q}_\alpha(F)<0)\} \overset{F\,\text{cont.}}{=} \mathbb{E}_F\{X|X<\mathbf{q}_\alpha(F)\} \tag{54}$$

More generally, the equalities hold for all distributions for which $Prob\{X < \mathbf{q}_\alpha(F)\} = \alpha$. For non–continuous distributions, on the contrary, these three expressions may all be different.

Finally, we recall the classical result of Uryasev and Rockafellar [13, 2] that shows that eq. (53) can be expressed as an optimization, in which $\mathbf{TM}_\alpha$ plays the role of the attained maximum and $\mathbf{q}_\alpha$ of the minimizer.

$$\begin{aligned}\mathbf{TM}_\alpha(F) &= \max_q \left\{q - \tfrac{1}{\alpha}\mathbb{E}_F[(X-q)_-]\right\}\\ \mathbf{q}_\alpha(F) &= \arg\max_q \left\{q - \tfrac{1}{\alpha}\mathbb{E}_F[(X-q)_-]\right\}\end{aligned} \tag{55}$$

A.3 Expectile

Definition A.3 (**Expectile**)**.** For $X \in L_1$, the expectile $\mathbf{e}_\alpha(F)$ is defined as the solution of [12, 7]

$$\alpha\mathbb{E}_F[(X-\mathbf{e}_\alpha(F))_+] = (1-\alpha)\mathbb{E}_F[(X-\mathbf{e}_\alpha(F))_-] \tag{56}$$

In the special case $\alpha = 1/2$, the expectile coincides with the mean of the distribution.

The corresponding tail risk measure, differing by an overall sign is also called expectile in most literature, without ambiguity.

[10] More generally, the equalities hold for all distributions for which $Prob\{X < \mathbf{q}_\alpha(F)\} = \alpha$.

References

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

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

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

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

[5] Artzner, P.; Delbaen, F.; Eber, J. M.; Heath, D. (1999) Coherent Measures of Risk, Mathematical Finance 9 (3)

[6] Basel Committee on Banking Supervision (2013) Fundamental review of the trading book: A revised market risk framework, Consultative Paper.

[7] Bellini, F., Klar, B., Müller, A. and Rosazza Gianin, E. (2014) Generalized quantiles as risk measures. Insurance: Mathematics and Economics **54**, (41–48)

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

[9] International Association of Insurance Supervisors (2014) Risk-based Global Insurance Capital Standard, public consultation document.

[10] International Association of Insurance Supervisors (2016) Risk-based Global Insurance Capital Standard, Version 1.0, public consultation document.

[11] Lambert, N., Pennock, D. M., Shoham, Y. (2008) Eliciting Properties of Probability Distributions, Proceedings of the 9^{th} ACM Conference on Electronic Commerce, EC 08

[12] Newey, W. K. and Powell, J. L. (1987) Asymmetric least squares estimation and testing. Econometrica, **55** (819-847)

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

[14] Ziegel, J. F. Coherence and elicitability, to appear on Mathematical Finance, preprint available at `http://arxiv.org/abs/1303.1690v2`

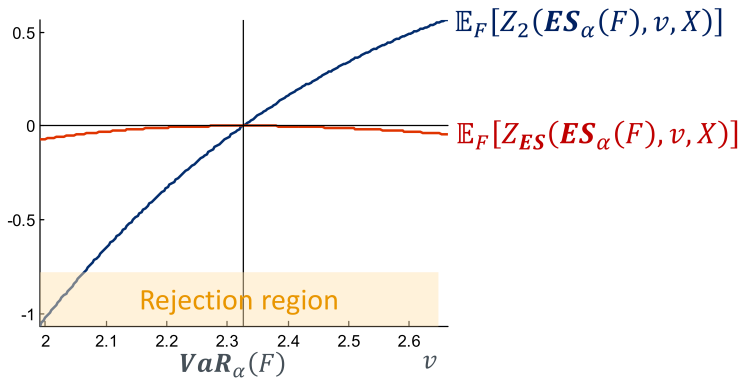

Figure 4: Dependence on v of tests $Z_{\mathbf{ES}}$ and Z_2 in the case of correct predictions for **ES**. Notice the linear sensitivity of the latter and the muted, quadratic sensitivity of the former. We can see that Z_2 can easily generate a type I error.

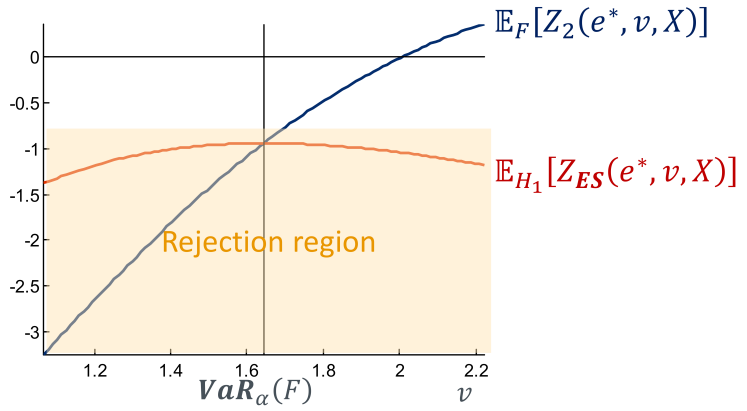

Figure 5: Similar example in the case of an underestimation $e^* = \mathbf{ES}_\alpha(F) - 1$. Z_2 can generate a type II error, while $Z_{\mathbf{ES}}$ can not.