Elicitability and backtesting

Johanna F. Ziegel

University of Bern

joint work with
Natalia Nolde, UBC

17 November 2017
Research Seminar at the
Institute for Statistics and Mathematics, WU Vienna

Introduction

Let X be the single-period return of some financial asset.

- A *risk measure* ρ assigns a real number $\rho(X)$ to X (interpreted as the *risk* of the asset).

Risk measures are used for

- external regulatory capital calculation
- management, optimization and decision making
- performance analysis
- capital allocation

Introduction

Let X be the single-period return of some financial asset.

- A *risk measure* ρ assigns a real number $\rho(X)$ to X (interpreted as the *risk* of the asset).

Risk measures are used for

- external regulatory capital calculation
- management, optimization and decision making
- performance analysis
- capital allocation

Risk measures

We assume that the risk $\rho(X)$ only depends on the distribution of X, that is,

$$\rho : \mathcal{P} \to \mathbb{R},$$

where $\mathcal{P}$ is a suitable class of probability measures on $\mathbb{R}$.
(We write both: $\rho(X) = \rho(P)$ for $X \sim P$.)

- Impose theoretical requirements on ρ motivated by economic principles: monetary risk measures, coherent risk measures,. . . .
- Consider statistical aspects of the resulting functionals.

Sign convention

- Losses are positive, profits are negative.
- Risky positions yield positive values of ρ.

Risk measures

We assume that the risk $\rho(X)$ only depends on the distribution of X, that is,

$$\rho : \mathcal{P} \to \mathbb{R},$$

where $\mathcal{P}$ is a suitable class of probability measures on $\mathbb{R}$.
(We write both: $\rho(X) = \rho(P)$ for $X \sim P$.)

- Impose theoretical requirements on ρ motivated by economic principles: monetary risk measures, coherent risk measures,. . . .
- Consider statistical aspects of the resulting functionals.

Sign convention

- Losses are positive, profits are negative.
- Risky positions yield positive values of ρ.

Risk measures

We assume that the risk $\rho(X)$ only depends on the distribution of X, that is,

$$\rho : \mathcal{P} \to \mathbb{R},$$

where $\mathcal{P}$ is a suitable class of probability measures on $\mathbb{R}$.
(We write both: $\rho(X) = \rho(P)$ for $X \sim P$.)

- Impose theoretical requirements on ρ motivated by economic principles: monetary risk measures, coherent risk measures,. . . .
- Consider statistical aspects of the resulting functionals.

Sign convention

- Losses are positive, profits are negative.
- Risky positions yield positive values of ρ.

Examples

Example (Value-at-Risk (VaR))

For $\alpha \in (0,1)$ and a random variable X with distribution function F, we define

$$\mathrm{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\} = F^{-1}(\alpha).$$

VaR is a co-monotonic additive, monetary risk measure.

Example (Expected shortfall (ES))

If X has finite mean, we define

$$\mathrm{ES}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(X)\,\mathrm{d}u \quad \Big(= \mathbb{E}[X \mid X \geq \mathrm{VaR}_\alpha(X)]\Big).$$

ES is a co-monotonic additive, coherent risk measure.

Examples

Example (Value-at-Risk (VaR))

For $\alpha \in (0,1)$ and a random variable X with distribution function F, we define

$$\mathrm{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\} = F^{-1}(\alpha).$$

VaR is a co-monotonic additive, monetary risk measure.

Example (Expected shortfall (ES))

If X has finite mean, we define

$$\mathrm{ES}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(X)\,\mathrm{d}u \quad \Big(= \mathbb{E}[X|X \geq \mathrm{VaR}_\alpha(X)]\Big).$$

ES is a co-monotonic additive, coherent risk measure.

Examples

Example (Value-at-Risk (VaR))

For $\alpha \in (0,1)$ and a random variable X with distribution function F, we define

$$\mathrm{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\} = F^{-1}(\alpha).$$

VaR is a co-monotonic additive, monetary risk measure.

Example (Expected shortfall (ES))

If X has finite mean, we define

$$\mathrm{ES}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(X)\,\mathrm{d}u \quad \Big(= \mathbb{E}[X|X \geq \mathrm{VaR}_\alpha(X)]\Big).$$

ES is a co-monotonic additive, coherent risk measure.

Statistical aspects

Returns $\{X_t\}_{t\in\mathbb{N}}$ (covariates $\{Z_t\}_{t\in\mathbb{N}}$) adapted to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$
Historical data: returns $x_1, \dots, x_n$, (covariates $z_1, \dots, z_n$)

Estimation

Approximate $\rho(X_{n+1}|\mathcal{F}_n)$ by $r_{n+1} = \hat{\rho}(x_1, \dots, x_n, z_1, \dots, z_n)$. ($\rightarrow$ elicitability/identifiability)

Forecast evaluation

- Given $r_1, \dots, r_n$, evaluate the quality of the predictions (traditional backtesting). ($\rightarrow$ identifiability)
- Given $r_1, \dots, r_n$, $r_1^*, \dots, r_n^*$, say which method has better predictive power (comparative backtesting). ($\rightarrow$ elicitability)

Statistical aspects

Returns $\{X_t\}_{t\in\mathbb{N}}$ (covariates $\{Z_t\}_{t\in\mathbb{N}}$) adapted to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$
Historical data: returns $x_1,\dots,x_n$, (covariates $z_1,\dots,z_n$)

Estimation

Approximate $\rho(X_{n+1}|\mathcal{F}_n)$ by $r_{n+1} = \hat{\rho}(x_1,\dots,x_n,z_1,\dots,z_n)$. ($\to$ elicitability/identifiability)

Forecast evaluation

- Given $r_1,\dots,r_n$, evaluate the quality of the predictions (traditional backtesting). ($\to$ identifiability)
- Given $r_1,\dots,r_n$, $r_1^*,\dots,r_n^*$, say which method has better predictive power (comparative backtesting). ($\to$ elicitability)

Statistical aspects

Returns $\{X_t\}_{t\in\mathbb{N}}$ (covariates $\{Z_t\}_{t\in\mathbb{N}}$) adapted to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$
Historical data: returns $x_1,\dots,x_n$, (covariates $z_1,\dots,z_n$)

Estimation

Approximate $\rho(X_{n+1}|\mathcal{F}_n)$ by $r_{n+1}=\hat{\rho}(x_1,\dots,x_n,z_1,\dots,z_n)$. ($\rightarrow$ elicitability/identifiability)

Forecast evaluation

- Given $r_1,\dots,r_n$, evaluate the quality of the predictions (traditional backtesting). ($\rightarrow$ identifiability)
- Given $r_1,\dots,r_n$, $r_1^*,\dots,r_n^*$, say which method has better predictive power (comparative backtesting). ($\rightarrow$ elicitability)

Statistical aspects

Returns $\{X_t\}_{t\in\mathbb{N}}$ (covariates $\{Z_t\}_{t\in\mathbb{N}}$) adapted to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$
Historical data: returns $x_1, \dots, x_n$, (covariates $z_1, \dots, z_n$)

Estimation

Approximate $\rho(X_{n+1}|\mathcal{F}_n)$ by $r_{n+1} = \hat{\rho}(x_1, \dots, x_n, z_1, \dots, z_n)$. ($\rightarrow$ elicitability/identifiability)

Forecast evaluation

- Given $r_1, \dots, r_n$, evaluate the quality of the predictions (traditional backtesting). ($\rightarrow$ identifiability)
- Given $r_1, \dots, r_n$, $r_1^*, \dots, r_n^*$, say which method has better predictive power (comparative backtesting). ($\rightarrow$ elicitability)

Statistical aspects

Returns $\{X_t\}_{t\in\mathbb{N}}$ (covariates $\{Z_t\}_{t\in\mathbb{N}}$) adapted to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$
Historical data: returns $x_1, \ldots, x_n$, (covariates $z_1, \ldots, z_n$)

Estimation

Approximate $\rho(X_{n+1}|\mathcal{F}_n)$ by $r_{n+1} = \hat{\rho}(x_1, \ldots, x_n, z_1, \ldots, z_n)$. ($\rightarrow$ elicitability/identifiability)

Forecast evaluation

- Given $r_1, \ldots, r_n$, evaluate the quality of the predictions (traditional backtesting). ($\rightarrow$ identifiability)
- Given $r_1, \ldots, r_n$, $r_1^*, \ldots, r_n^*$, say which method has better predictive power (comparative backtesting). ($\rightarrow$ elicitability)

Elicitability and identifiability

Let $\mathsf{A} \subset \mathbb{R}$ be such that

$$\rho\colon \mathcal{P} \to \mathsf{A}, \quad P \mapsto \rho(P)$$

Definition

A loss function $L : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ is *consistent for* ρ *(relative to* $\mathcal{P}$*)*, if

$$\mathbb{E}[L(\rho(P), X)] \leq \mathbb{E}[L(r, X)], \quad P \in \mathcal{P},\ X \sim P,\ r \in \mathsf{A}.$$

It is *strictly consistent (relative to* $\mathcal{P}$*)* if "=" implies $r = \rho(P)$.

The risk measure ρ is called *elicitable (relative to* $\mathcal{P}$*)* if there exists a loss function L that is strictly consistent for it.

In other words

$$\rho(P) = \arg\min_{r \in \mathsf{A}} \mathbb{E}L(r, X).$$

Elicitability and identifiability

Let $\mathsf{A} \subset \mathbb{R}$ be such that

$$\rho \colon \mathcal{P} \to \mathsf{A}, \quad P \mapsto \rho(P)$$

Definition

A loss function $L : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ is *consistent for* ρ *(relative to* $\mathcal{P}$*)*, if

$$\mathbb{E}[L(\rho(P), X)] \leq \mathbb{E}[L(r, X)], \quad P \in \mathcal{P},\ X \sim P,\ r \in \mathsf{A}.$$

It is *strictly consistent (relative to* $\mathcal{P}$*)* if "$=$" implies $r = \rho(P)$.

The risk measure ρ is called *elicitable (relative to* $\mathcal{P}$*)* if there exists a loss function L that is strictly consistent for it.

In other words

$$\rho(P) = \arg\min_{r \in \mathsf{A}} \mathbb{E} L(r, X).$$

Elicitability and identifiability

Let $\mathsf{A} \subset \mathbb{R}$ be such that

$$\rho\colon \mathcal{P} \to \mathsf{A}, \quad P \mapsto \rho(P)$$

Definition

ρ is *identifiable (relative to $\mathcal{P}$)*, if there is a function $V : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{E}(V(r, X)) = 0 \quad \Longleftrightarrow \quad r = \rho(X),$$

for all $P \in \mathcal{P}$, $X \sim P$, $r \in \mathsf{A}$.

Some examples

- Mean

  $L(r,x) = (x-r)^2 \qquad V(r,x) = x - r$

  - Least squares regression
  - Comparison of models/forecast performance in terms of MSE

- α-Quantiles/VaR$_\alpha$ (Median)

  $L(r,x) = (\mathbb{1}\{x \le r\} - \alpha)(r-x) \qquad V(r,x) = \mathbb{1}\{x \le r\} - \alpha$

  - Quantile/Median regression

- α-Expectiles (Newey and Powell, 1987)

  $L(r,x) = |\mathbb{1}\{x \le r\} - \alpha|(r-x)^2$
  $V(r,x) = |\mathbb{1}\{x \le r\} - \alpha|(r-x)$

  - Expectile regression

Elicitable and non-elicitable functionals

Elicitable

- Mean, moments
- Median, quantiles/Value-at-Risk
- Expectiles (Newey and Powell, 1987)

Not elicitable

- Variance
- Expected shortfall (Weber, 2006, Gneiting, 2011)
- Spectral risk measures (Weber, 2006, Z, 2014)

k-Elicitability and k-identifiability

Let $\mathsf{A} \subset \mathbb{R}^k$ be such that

$$\underline{\rho}\colon \mathcal{P} \to \mathsf{A}, \quad P \mapsto \underline{\rho}(P)$$

Definition

A loss function $L : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ is *consistent for* $\underline{\rho}$ *(relative to* $\mathcal{P}$*)*, if

$$\mathbb{E}[L(\underline{\rho}(P), X)] \leq \mathbb{E}[L(r, X)], \quad P \in \mathcal{P},\ X \sim P,\ r = (r_1, \ldots, r_k) \in \mathsf{A}.$$

It is *strictly consistent (relative to* $\mathcal{P}$*)* if "$=$" implies $r = \underline{\rho}(P)$.

The vector of risk measures $\underline{\rho}$ is called *k-elicitable (relative to* $\mathcal{P}$*)* if there exists a loss function L that is strictly consistent for it.

In other words

$$\underline{\rho}(P) = \arg\min_{r \in \mathsf{A}} \mathbb{E}L(r, X).$$

k-Elicitability and identifiability

Let $\mathsf{A} \subset \mathbb{R}^k$ be such that

$$\underline{\rho}\colon \mathcal{P} \to \mathsf{A}, \quad P \mapsto \underline{\rho}(P)$$

Definition

$\underline{\rho}$ is *k-identifiable (relative to $\mathcal{P}$)*, if there is a function $V : \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ such that

$$\mathbb{E}(V(r, X)) = 0 \quad \Longleftrightarrow \quad r = \underline{\rho}(X),$$

for all $P \in \mathcal{P}$, $X \sim P$, $r = (r_1, \dots, r_k) \in \mathsf{A}$.

Elicitable and identifiable functionals

1-Elicitable and 1-identifiable

- Mean, moments
- Median, quantiles/VaR
- Expectiles (Newey and Powell, 1987)

2-Elicitable and 2-identifiable

- Mean and variance
- Second moment and variance
- VaR and ES (Acerbi and Szekely, 2014, Fissler and Z, 2016)

***k*-Elicitable and *k*-identifiable**

- Some spectral risk measures together with several VaRs at certain levels
  (Fissler and Z, 2016)

$\underline{\rho} = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

Theorem (Fissler and Z, 2016)

Let $\alpha \in (0,1)$, and $\mathsf{A}_0 := \{r = (q,e) \in \mathbb{R}^2 : q \le e\}$. Let $\mathcal{P}$ be a class of probability measures on $\mathbb{R}$ with finite first moments and unique α-quantiles. Any loss function $L\colon \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ of the form

$$L(q,e,x) = \big(1 - \alpha - \mathbb{1}\{x > q\}\big) g(q) + \mathbb{1}\{x > q\} g(x) + \phi'(e)\left(\big(1 - \alpha - \mathbb{1}\{x > q\}\big)\frac{q}{1-\alpha} + \mathbb{1}\{x > q\}\frac{x}{1-\alpha} - e\right) + \phi(e)$$

is consistent for $\underline{\rho} = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ if $\mathbb{1}_{[q,\infty)} g$ is $\mathcal{P}$-integrable and

- *g is increasing and ϕ is increasing and concave.*

It is strictly consistent if, additionally,

- *ϕ is strictly increasing and strictly concave.*

Evaluating forecasts of expected shortfall

Filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$

Prediction-observation triples

$$(Q_t, E_t, X_t)_{t\in\mathbb{N}}$$

Q_t: VaR_α prediction for time point t, $\mathcal{F}_{t-1}$-measurable
E_t: ES_α prediction for time point t, $\mathcal{F}_{t-1}$-measurable
X_t: Realization at time point t, $\mathcal{F}_t$-measurable

Absolute evaluation

Let V be an identification function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, i.e.

$$V(q,e,x) = A(q,e) \begin{pmatrix} 1-\alpha - \mathbb{1}\{x > q\} \\ q - e - \frac{1}{1-\alpha}\mathbb{1}\{x > q\}(q-x) \end{pmatrix},$$

where $A(q,e) \in \mathbb{R}^{2\times 2}$ with $\det(A(q,e)) \neq 0$. (We take $A = I_2$.)

Definition (Calibration)

The sequence of predictions $\{(Q_t, E_t)\}_{t\in\mathbb{N}}$ is *conditionally calibrated* for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ if

$$\mathbb{E}\left(V(Q_t, E_t, X_t)|\mathcal{F}_{t-1}\right) = 0 \quad \text{for all } t \in \mathbb{N}.$$

Compare Davis (2016).

Calibration and optimal prediction

- Given information $\mathcal{F}_{t-1}$ at time point $t-1$, the best prediction for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ of X_t is

$$\big(\mathrm{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right), \mathrm{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right)\big).$$

This is the only $\mathcal{F}_{t-1}$-measurable prediction which is conditionally calibrated.

- If $\mathcal{F}^*_{t-1} \supset \mathcal{F}_{t-1}$, then

$$\big(\mathrm{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right), \mathrm{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right)\big).$$

is also conditionally calibrated (with respect to $\mathcal{F}_{t-1}$).

Calibration and optimal prediction

- Given information $\mathcal{F}_{t-1}$ at time point $t-1$, the best prediction for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ of X_t is

$$\big(\mathrm{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right), \mathrm{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right)\big).$$

  This is the only $\mathcal{F}_{t-1}$-measurable prediction which is conditionally calibrated.

- If $\mathcal{F}^*_{t-1} \supset \mathcal{F}_{t-1}$, then

$$\big(\mathrm{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right), \mathrm{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right)\big).$$

  is also conditionally calibrated (with respect to $\mathcal{F}_{t-1}$).

Calibration and optimal prediction

- Given information $\mathcal{F}_{t-1}$ at time point $t-1$, the best prediction for $(\text{VaR}_\alpha, \text{ES}_\alpha)$ of X_t is

$$\big(\text{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right), \text{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}_{t-1})\right)\big).$$

  This is the only $\mathcal{F}_{t-1}$-measurable prediction which is conditionally calibrated.

- If $\mathcal{F}^*_{t-1} \supset \mathcal{F}_{t-1}$, then

$$\big(\text{VaR}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right), \text{ES}_\alpha\left(\mathcal{L}(X_t|\mathcal{F}^*_{t-1})\right)\big).$$

  is also conditionally calibrated (with respect to $\mathcal{F}_{t-1}$).

Traditional backtesting

H_0^C: The sequence of predictions $\{(Q_t, E_t)\}_{t\in\mathbb{N}}$ is conditionally calibrated.

- Backtesting decision: If we do not reject H_0^C, the risk measure estimates are adequate.
- Many existing backtests can be described as a test for conditional calibration (with different choices for the identification function, different model assumptions). (McNeil and Frey, 2000, Acerbi and Szekely 2014)
- Does not give guidance for decision between methods.
- Does not respect increasing information sets.

Traditional backtesting

H_0^C: The sequence of predictions $\{(Q_t, E_t)\}_{t\in\mathbb{N}}$ is conditionally calibrated.

- Backtesting decision: If we do not reject H_0^C, the risk measure estimates are adequate.
- Many existing backtests can be described as a test for conditional calibration (with different choices for the identification function, different model assumptions). (McNeil and Frey, 2000, Acerbi and Szekely 2014)
- Does not give guidance for decision between methods.
- Does not respect increasing information sets.

Constructing conditional calibration tests

- $\mathbb{E}(V(Q_t, E_t, X_t)|\mathcal{F}_{t-1}) = 0$ is equivalent to $\mathbb{E}(h_t^\top V(Q_t, E_t, X_t)) = 0$ for all $\mathcal{F}_{t-1}$-measurable $\mathbb{R}^2$-valued functions h_t.
- Choose $\mathcal{F}$-predictable sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$ of $q \times 2$-matrices $\mathbf{h}_t$ of *test functions*: Ideally, the rows of $\mathbf{h}_t$ generate $\mathcal{F}_{t-1}$.
- Construct test statistic: For example,

$$T_1 = n\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big)^\top \widehat{\Omega}_n^{-1}\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big),$$

  where

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{h}_t V(Q_t, E_t, X_t))(\mathbf{h}_t V(Q_t, E_t, X_t))^\top.$$

  Giacomini and White (2006) give general conditions under which T_1 is asymptotically χ^2_q.
- For these traditional ES_α backtests, VaR_α is also needed but they are robust with respect to model misspecification.

Constructing conditional calibration tests

- $\mathbb{E}(V(Q_t, E_t, X_t)|\mathcal{F}_{t-1}) = 0$ is equivalent to $\mathbb{E}(h_t^\top V(Q_t, E_t, X_t)) = 0$ for all $\mathcal{F}_{t-1}$-measurable $\mathbb{R}^2$-valued functions h_t.
- Choose $\mathcal{F}$-predictable sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$ of $q \times 2$-matrices $\mathbf{h}_t$ of *test functions*: Ideally, the rows of $\mathbf{h}_t$ generate $\mathcal{F}_{t-1}$.
- Construct test statistic: For example,

$$T_1 = n\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big)^\top \widehat{\Omega}_n^{-1}\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big),$$

  where

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{h}_t V(Q_t, E_t, X_t))(\mathbf{h}_t V(Q_t, E_t, X_t))^\top.$$

  Giacomini and White (2006) give general conditions under which T_1 is asymptotically χ^2_q.
- For these traditional ES_α backtests, VaR_α is also needed but they are robust with respect to model misspecification.

Constructing conditional calibration tests

- $\mathbb{E}(V(Q_t, E_t, X_t)|\mathcal{F}_{t-1}) = 0$ is equivalent to $\mathbb{E}(h_t^\top V(Q_t, E_t, X_t)) = 0$ for all $\mathcal{F}_{t-1}$-measurable $\mathbb{R}^2$-valued functions h_t.
- Choose $\mathcal{F}$-predictable sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$ of $q \times 2$-matrices $\mathbf{h}_t$ of *test functions*: Ideally, the rows of $\mathbf{h}_t$ generate $\mathcal{F}_{t-1}$.
- Construct test statistic: For example,

$$T_1 = n\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big)^\top \widehat{\Omega}_n^{-1}\Big(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V(Q_t, E_t, X_t)\Big),$$

  where

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{h}_t V(Q_t, E_t, X_t))(\mathbf{h}_t V(Q_t, E_t, X_t))^\top.$$

  Giacomini and White (2006) give general conditions under which T_1 is asymptotically χ^2_q.
- For these traditional ES_α backtests, VaR_α is also needed but they are robust with respect to model misspecification.

Examples of a traditional ES_α backtests

- Mean of exceedance residuals (McNeil and Frey, 2000):

$$\mathbf{h}_t = \frac{1}{\hat{\sigma}_t}\left(\frac{E_t - V_t}{1-\alpha}, 1\right),$$

  where $\hat{\sigma}_t$ is an $\mathcal{F}_{t-1}$-measurable estimator of the volatility of X_t.
  They assume a model of the form

$$X_t = \mu_t + \sigma_t Z_t$$

  to calculate the distribution of $(1/n)\sum_{t=1}^{n} \mathbf{h}_t V(Q_t, E_t, X_t)$ under H_0^C.
- $\mathbf{h}_t = (0, 1)$.
- Other options: Acerbi and Szekely (2014)

A simulation study

AR(1)-GARCH(1,1)-model:

$$X_t = \mu_t + \varepsilon_t, \quad \mu_t = -0.05 + 0.3X_{t-1},$$

$$\varepsilon_t = \sigma_t Z_t, \quad \sigma_t^2 = 0.01 + 0.1\varepsilon_{t-1}^2 + 0.85\sigma_{t-1}^2,$$

(Z_t) iid with skewed t distribution with shape $= 5$ and skewness $= 1.5$.

Estimation procedures:

- Fully parametric (n-FP, t-FP, st-FP)
- Filtered historical simulation (n-FHS, t-FHS, st-FHS)
- EVT based semi-parametric estimation (n-EVT, t-EVT, st-EVT)

Moving window of size 500 for estimation
5000 out-of-sample verifying observations

P-values of traditional backtests for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

	$\alpha = 0.754$		$\alpha = 0.975$	
	simple	general	simple	general
n-FP	**0.000**	**0.000**	**0.000**	**0.000**
n-FHS	0.881	0.184	0.653	0.231
n-EVT	0.754	0.672	0.886	0.226
t-FP	0.086	**0.006**	**0.000**	**0.000**
t-FHS	0.936	0.512	0.697	0.717
t-EVT	0.880	0.475	0.995	0.498
st-FP	0.569	0.824	0.695	0.419
st-FHS	0.909	0.796	0.843	0.758
st-EVT	0.935	0.706	0.962	0.564
opt	0.401	0.337	0.131	0.571

Comparative evaluation

Filtrations $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ and $\mathcal{F}^* = \{\mathcal{F}_t^*\}_{t\in\mathbb{N}}$

Q_t, Q_t^*: VaR_α predictions for time point t
E_t, E_t^*: ES_α predictions for time point t

Q_t, E_t: internal model, $\mathcal{F}_{t-1}$-measurable
Q_t^*, E_t^*: standard model, $\mathcal{F}_{t-1}^*$-measurable

X_t: Realization at time point t, $\mathcal{F}_t$-measurable and $\mathcal{F}_t^*$-measurable

Comparative evaluation

Filtrations $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ and $\mathcal{F}^* = \{\mathcal{F}_t^*\}_{t\in\mathbb{N}}$

Q_t, Q_t^*: VaR_α predictions for time point t
E_t, E_t^*: ES_α predictions for time point t

Q_t, E_t: internal model, $\mathcal{F}_{t-1}$-measurable
Q_t^*, E_t^*: standard model, $\mathcal{F}_{t-1}^*$-measurable

X_t: Realization at time point t, $\mathcal{F}_t$-measurable and $\mathcal{F}_t^*$-measurable

Comparative evaluation

Filtrations $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ and $\mathcal{F}^* = \{\mathcal{F}_t^*\}_{t\in\mathbb{N}}$

Q_t, Q_t^*: VaR_α predictions for time point t
E_t, E_t^*: ES_α predictions for time point t

Q_t, E_t: internal model, $\mathcal{F}_{t-1}$-measurable
Q_t^*, E_t^*: standard model, $\mathcal{F}_{t-1}^*$-measurable

X_t: Realization at time point t, $\mathcal{F}_t$-measurable and $\mathcal{F}_t^*$-measurable

Forecast dominance

Let L be a consistent loss function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$.

Definition (S-Dominance)

The sequence of predictions $\{(Q_t, E_t)\}_{t\in\mathbb{N}}$ *(L-)dominates* $\{(Q_t^*, E_t^*)\}_{t\in\mathbb{N}}$ if

$$\mathbb{E}(L(Q_t, E_t, X_t) - L(Q_t^*, E_t^*, X_t)) \leq 0, \quad \text{for all } t \in \mathbb{N}.$$

- Diebold-Mariano tests for difference in predictive performance (Diebold and Mariano, 1995).
- Test statistic:

$$\frac{\sqrt{n}}{\hat{\Sigma}_n}\frac{1}{n}\sum_{t=1}^{n}(L(Q_t, E_t, X_t) - L(Q_t^*, E_t^*, X_t)).$$

  Asymptotically normal under suitable conditions (Giacomini and White, 2006).

Forecast dominance

Let L be a consistent loss function for $(\text{VaR}_\alpha, \text{ES}_\alpha)$.

Definition (S-Dominance)

The sequence of predictions $\{(Q_t, E_t)\}_{t\in\mathbb{N}}$ *(L-)dominates* $\{(Q_t^*, E_t^*)\}_{t\in\mathbb{N}}$ if

$$\mathbb{E}(L(Q_t, E_t, X_t) - L(Q_t^*, E_t^*, X_t)) \leq 0, \quad \text{for all } t \in \mathbb{N}.$$

- Diebold-Mariano tests for difference in predictive performance (Diebold and Mariano, 1995).
- Test statistic:

$$\frac{\sqrt{n}}{\hat{\Sigma}_n} \frac{1}{n} \sum_{t=1}^{n} (L(Q_t, E_t, X_t) - L(Q_t^*, E_t^*, X_t)).$$

Asymptotically normal under suitable conditions (Giacomini and White, 2006).

Comparative backtesting

H_0^-: Internal model dominates the standard model.

H_0^+: Internal model is dominated by the standard model.

- Backtesting decision using H_0^-: If we do not reject H_0^-, the risk measure estimates are acceptable (compared to the standard).
- Backtesting decision using H_0^+: If we reject H_0^+, the risk measure estimates are acceptable (compared to the standard).
- Elicitability is crucial for robust comparative backtests.
- Allows for comparison between methods.
- Necessitates a standard reference model.
- Respects increasing information sets (Holzmann and Eulert, 2014).

Comparative backtesting

H_0^-: Internal model dominates the standard model.

H_0^+: Internal model is dominated by the standard model.

- Backtesting decision using H_0^-: If we do not reject H_0^-, the risk measure estimates are acceptable (compared to the standard).
- Backtesting decision using H_0^+: If we reject H_0^+, the risk measure estimates are acceptable (compared to the standard).
- Elicitability is crucial for robust comparative backtests.
- Allows for comparison between methods.
- Necessitates a standard reference model.
- Respects increasing information sets (Holzmann and Eulert, 2014).

Comparative backtesting

H_0^-: Internal model dominates the standard model.

H_0^+: Internal model is dominated by the standard model.

- Backtesting decision using H_0^-: If we do not reject H_0^-, the risk measure estimates are acceptable (compared to the standard).
- Backtesting decision using H_0^+: If we reject H_0^+, the risk measure estimates are acceptable (compared to the standard).
- Elicitability is crucial for robust comparative backtests.
- Allows for comparison between methods.
- Necessitates a standard reference model.
- Respects increasing information sets (Holzmann and Eulert, 2014).

Choice of a loss function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

A loss function L is called *positively homogeneous of degree b* if

$$L(cr, cx) = c^b L(r, x), \quad \text{for all } c > 0.$$

- Important in regression, forecast ranking; implies "unit consistency" (Efron, 1991, Patton, 2011, Acerbi and Szekely, 2014).

Let $\mathsf{A} = (\mathbb{R} \times (0, \infty)) \cap \mathsf{A}_0$.

- There are positively homogeneous strictly consistent loss functions of degree b if and only if $b \in (-\infty, 1) \backslash \{0\}$.
- There are strictly consistent loss functions such that the loss differences are positively homogeneous of degree $b = 0$:

$$L_0(q, e, x) = \mathbb{1}\{x > q\} \frac{x - q}{e} + (1 - \alpha) \left(\frac{q}{e} - 1 + \log(e) \right)$$

Choice of a loss function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

A loss function L is called *positively homogeneous of degree b* if

$$L(cr, cx) = c^b L(r, x), \quad \text{for all } c > 0.$$

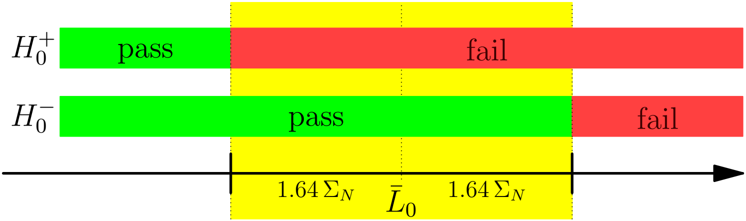
- Important in regression, forecast ranking; implies "unit consistency" (Efron, 1991, Patton, 2011, Acerbi and Szekely, 2014).

Let $\mathsf{A} = (\mathbb{R} \times (0, \infty)) \cap \mathsf{A}_0$.

- There are positively homogeneous strictly consistent loss functions of degree b if and only if $b \in (-\infty, 1) \setminus \{0\}$.
- There are strictly consistent loss functions such that the loss differences are positively homogeneous of degree $b = 0$:

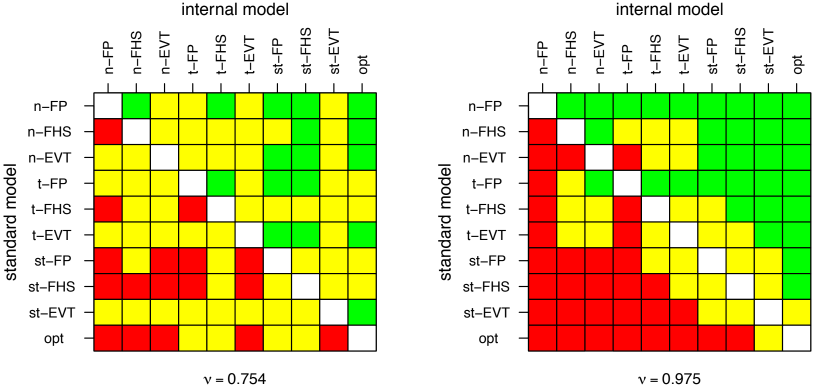
$$L_0(q, e, x) = \mathbb{1}\{x > q\} \frac{x - q}{e} + (1 - \alpha) \left(\frac{q}{e} - 1 + \log(e) \right)$$

Three zone approach for comparative backtesting



P-values of traditional backtests and L_0-ranking for $(\text{VaR}_\alpha, \text{ES}_\alpha)$

	$\alpha = 0.754$			$\alpha = 0.975$		
	simple	general	$\bar{L}_0$	simple	general	$\bar{L}_0$
n-FP	**0.000**	**0.000**	9	**0.000**	**0.000**	10
n-FHS	0.881	0.184	4	0.653	0.231	8
n-EVT	0.754	0.672	8	0.886	0.226	7
t-FP	0.086	**0.006**	10	**0.000**	**0.000**	9
t-FHS	0.936	0.512	5	0.697	0.717	6
t-EVT	0.880	0.475	7	0.995	0.498	5
st-FP	0.569	0.824	3	0.695	0.419	2
st-FHS	0.909	0.796	2	0.843	0.758	4
st-EVT	0.935	0.706	6	0.962	0.564	3
opt	0.401	0.337	1	0.131	0.571	1


internal model
n–FP
n–FHS
n–EVT
t–FP
t–FHS
t–EVT
st–FP
st–FHS
st–EVT
opt
standard model
ν = 0.754
ν = 0.975

Backtesting with small sample size

$(\mathrm{VaR}_{0.975}, \mathrm{ES}_{0.975})$										
	n-FP	n-FHS	n-EVT	t-FP	t-FHS	t-EVT	st-FP	st-FHS	st-EVT	opt
n-FP	0	84	86	84	84	86	86	84	86	89
n-FHS	16	0	58	23	54	58	61	57	59	74
n-EVT	14	42	0	22	45	53	55	49	58	72
t-FP	16	77	78	0	79	80	81	77	80	84
t-FHS	16	46	55	21	0	60	58	51	63	72
t-EVT	14	42	47	20	40	0	52	43	53	73
st-FP	14	39	45	19	42	48	0	40	52	71
st-FHS	16	43	51	23	49	57	60	0	60	72
st-EVT	14	41	42	20	37	47	48	40	0	73
opt	11	26	28	16	28	27	29	28	27	0

Conclusion

- Traditional backtests that are robust with respect to model misspecification necessitate an identifiable risk measure.
- Robust comparative backtests necessitate an elicitable risk measure.
- Robust traditional and comparative backtests for the whole tail of the distribution are also available. (Holzmann & Klar, 2017, Gordi, Lok & McNeil, 2017)
- k-Elicitability allows to find loss functions for functionals that are not elicitable individually.
- A relevant example in banking and insurance is the non-elicitable risk measure ES_α which is 2-elicitable with VaR_α.

Conclusion

- Traditional backtests that are robust with respect to model misspecification necessitate an identifiable risk measure.
- Robust comparative backtests necessitate an elicitable risk measure.
- Robust traditional and comparative backtests for the whole tail of the distribution are also available. (Holzmann & Klar, 2017, Gordi, Lok & McNeil, 2017)
- *k*-Elicitability allows to find loss functions for functionals that are not elicitable individually.
- A relevant example in banking and insurance is the non-elicitable risk measure ES_α which is 2-elicitable with VaR_α.

Outlook

- Characterization result for loss functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ allows for Murphy diagrams: Forecast comparison without the choice of a specific loss function (Z, Krüger, Jordan, Fasciati, 2017).
- The loss functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ allow for M-estimation (Zwingmann & Holzmann, 2016), generalized regression (Bayer & Dimitriadis, 2017, Barendse, 2017), semi-parametric time series models (Patton, Z, Chen, 2017)

Some references

T. Fissler and J. F. Ziegel (2016). *Higher order elicitability and Osband's principle*. Annals of Statistics, 44:1680–1707.

T. Fissler, J. F. Ziegel and T. Gneiting (2016). *Expected Shortfall is jointly elicitable with Value at Risk – Implications for backtesting*. Risk Magazine, January 2016.

N. Nolde and J. F. Ziegel (2016). *Elicitability and backtesting*. Annals of Applied Statistics, to appear with discussion.

Thank you for your attention!