Higher order elicitability

Johanna F. Ziegel

University of Bern

joint work with
Fernando Fasciati, Tobias Fissler, Tilmann Gneiting, Alexander Jordan, Fabian Krüger and Natalia Nolde

2 June 2017
Van Dantzig Seminar

Outline

1. Elicitability
    - Definition and a simple example
    - Risk measures
    - *k*-Elicitability
    - Osband's principle
2. Evaluating forecasts of expected shortfall
    - Absolute forecast evaluation
    - Classical comparative forecast evaluation
    - Comparative forecast evaluation with Murphy diagrams
3. Summary

Outline

Elicitability

Let $\mathcal{P}$ be a class of probability measures on $\mathsf{O} \subseteq \mathbb{R}^d$. Let

$$T : \mathcal{P} \to \mathsf{A}, \quad F \mapsto T(F)$$

be a functional where $\mathsf{A} \subseteq \mathbb{R}$.

Definition

A scoring (or loss) function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *consistent for T relative to* $\mathcal{P}$, if

$$\mathbb{E}_F S(T(F), Y) \le \mathbb{E}_F S(x, Y), \quad F \in \mathcal{P}, x \in \mathsf{A}.$$

It is *strictly consistent* if "$=$" implies $x = T(F)$.
The functional T is called *elicitable relative to* $\mathcal{P}$ if there exists a scoring function S that is strictly consistent for it.

In other words

$$T(F) = \arg\min_{x \in \mathsf{A}} \mathbb{E}_F S(x, Y).$$

A simple example – the mean

Let Y be a random variable with distribution function F. Suppose that $\mathbb{E}_F Y^2 < \infty$. Then,

$$\mathbb{E}_F Y = \arg\min_{x \in \mathbb{R}} \mathbb{E}_F (Y - x)^2.$$

- The mean is elicitable with respect to the class of all probability measures with finite second moment.

A simple example – the mean

Theorem (Savage, 1971)

Let $\mathcal{P}$ be a class of probability measures with finite first moments. Let ϕ be a (strictly) convex function such that $\mathbb{E}_F \phi(Y)$ exists and is finite for all $F \in \mathcal{P}$. Then,

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x)$$

is (strictly) consistent for the mean.

- Under suitable assumptions on $\mathcal{P}$, the *Bregman functions* are the only consistent scoring functions for the mean.
- Choosing $\phi(y) = y^2/(1 + |y|)$ shows that the mean is elicitable with respect to the class of all probability measures with finite first moment.

A simple example – the mean

Theorem (Savage, 1971)

Let $\mathcal{P}$ be a class of probability measures with finite first moments. Let ϕ be a (strictly) convex function such that $\mathbb{E}_F\phi(Y)$ exists and is finite for all $F \in \mathcal{P}$. Then,

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x)$$

is (strictly) consistent for the mean.

- Under suitable assumptions on $\mathcal{P}$, the *Bregman functions* are the only consistent scoring functions for the mean.
- Choosing $\phi(y) = y^2/(1 + |y|)$ shows that the mean is elicitable with respect to the class of all probability measures with finite first moment.

A simple example – the mean

Theorem (Savage, 1971)

Let $\mathcal{P}$ be a class of probability measures with finite first moments. Let ϕ be a (strictly) convex function such that $\mathbb{E}_F \phi(Y)$ exists and is finite for all $F \in \mathcal{P}$. Then,

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x)$$

is (strictly) consistent for the mean.

- Under suitable assumptions on $\mathcal{P}$, the *Bregman functions* are the only consistent scoring functions for the mean.
- Choosing $\phi(y) = y^2/(1 + |y|)$ shows that the mean is elicitable with respect to the class of all probability measures with finite first moment.

Why is elicitability interesting?

Generalized regression/M-estimation

Assume the following model

$$T(\mathcal{L}(Y|Z)) = m(Z, \beta)$$

parametrized by $\beta \in \Theta$ and let S be a strictly consistent scoring function for T. Suppose we have iid observations (z_i, y_i), $i = 1, \ldots, n$ from (Z, Y). Then, we can estimate β by

$$\hat{\beta} = \arg\min_{\beta' \in \Theta} \frac{1}{n} \sum_{i=1}^{n} S(y_i, m(z_i, \beta')).$$

- Least squares regression
- Quantile regression
- Logistic regression

Why is elicitability interesting?

Generalized regression/M-estimation

Assume the following model

$$T(\mathcal{L}(Y|Z)) = m(Z, \beta)$$

parametrized by $\beta \in \Theta$ and let S be a strictly consistent scoring function for T. Suppose we have iid observations (z_i, y_i), $i = 1, \dots, n$ from (Z, Y). Then, we can estimate β by

$$\hat{\beta} = \arg\min_{\beta' \in \Theta} \frac{1}{n} \sum_{i=1}^{n} S(y_i, m(z_i, \beta')).$$

- Least squares regression
- Quantile regression
- Logistic regression

Why is elicitability interesting?

Forecast comparison/Model selection
Suppose we have sequences of competing forecasts $x_1^A, \ldots, x_n^A$, $x_1^B, \ldots, x_n^B$ for T and observations $y_1, \ldots, y_n$. Let S be a strictly consistent scoring function for T.
Then it is natural to prefer method A over method B if

$$\frac{1}{n}\sum_{i=1}^{n} S(x_i^A, y_i) < \frac{1}{n}\sum_{i=1}^{n} S(x_i^B, y_i).$$

Why is elicitability interesting?

Forecast comparison/Model selection

Suppose we have sequences of competing forecasts $x_1^A, \dots, x_n^A$, $x_1^B, \dots, x_n^B$ for T and observations $y_1, \dots, y_n$. Let S be a strictly consistent scoring function for T.

Then it is natural to prefer method A over method B if

$$\frac{1}{n}\sum_{i=1}^{n} S(x_i^A, y_i) < \frac{1}{n}\sum_{i=1}^{n} S(x_i^B, y_i).$$

Risk measures

Let $Y \sim F$ be the single-period return of some financial asset.

- A *risk measure* assigns a real number to Y (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

Let $Y \sim F$ be the single-period return of some financial asset.

- A *risk measure* assigns a real number to Y (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

Value at Risk and expected shortfall

Let $Y \sim F$, $\alpha \in (0,1)$.

Value at Risk (VaR)

$$\mathrm{VaR}_\alpha(Y) = q_\alpha(F) = \inf\{x \in \mathbb{R} : \mathbb{P}(Y \leq x) \geq \alpha\},$$

Expected shortfall (ES)

$$\mathrm{ES}_\alpha(Y) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(Y)\,\mathrm{d}u.$$

- Profits are positive.
- We consider α close to zero ($\alpha = 0.01$, $\alpha = 0.025$).
- Risky positions yield large negative values of VaR_α and ES_α.

Criticism of VaR as a risk measure

Lack of super-additivity

- Usually there are several $Y^{(1)}, Y^{(2)}, \ldots$ to be considered with limited knowledge of their dependence.
- Goal: Bound on risk of the total $\sum_i Y^{(i)}$.
- VaR is not super-additive: There are $Y^{(1)}, Y^{(2)}$ such that

$$\mathrm{VaR}_\alpha(Y^{(1)} + Y^{(2)}) < \mathrm{VaR}_\alpha(Y^{(1)}) + \mathrm{VaR}_\alpha(Y^{(2)}).$$

- Problematic for risk aggregation.
- Counterintuitive to diversification.

It is just a quantile...

- VaR_α does not take sizes of losses beyond the threshold α into account.

Criticism of VaR as a risk measure

Lack of super-additivity

- Usually there are several $Y^{(1)}, Y^{(2)}, \ldots$ to be considered with limited knowledge of their dependence.
- Goal: Bound on risk of the total $\sum_i Y^{(i)}$.
- VaR is not super-additive: There are $Y^{(1)}, Y^{(2)}$ such that

$$\mathrm{VaR}_\alpha(Y^{(1)} + Y^{(2)}) < \mathrm{VaR}_\alpha(Y^{(1)}) + \mathrm{VaR}_\alpha(Y^{(2)}).$$

- Problematic for risk aggregation.
- Counterintuitive to diversification.

It is just a quantile...

- VaR_α does not take sizes of losses beyond the threshold α into account.

Expected shortfall

For continuous distributions, we have

$$\mathrm{ES}_\alpha(Y) = \mathbb{E}(Y \mid Y \leq \mathrm{VaR}_\alpha(Y)).$$

- ES_α is a coherent risk measure, so in particular super-additive.
- It takes the entire tail of the distribution into account.
- Largest coherent risk measure is dominated by VaR_α.
- It has a natural interpretation.

Elicitable and non-elicitable functionals

Elicitable

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

Not elicitable

- Variance
- Expected Shortfall (Weber, 2006, Gneiting, 2011)

Elicitable...

- **coherent risk measures:** Expectiles
- **convex risk measures:** Shortfall risk measures
- **distortion risk measures:** VaR and mean

(Weber, 2006, Z 2014, Bellini and Bignozzi, 2014, Delbaen et al. 2015, Kou and Peng 2014, Wang and Z 2015)

k-Elicitability

Let $\mathcal{P}$ be a class of probability measures on $\mathsf{O} \subseteq \mathbb{R}^d$. Let

$$T : \mathcal{P} \to \mathsf{A}, \quad F \mapsto T(F)$$

be a functional where $\mathsf{A} \subseteq \mathbb{R}^k$.

Definition

A scoring (or loss) function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *$\mathcal{P}$-consistent for T*, if

$$\mathbb{E}_F S(T(F), Y) \leq \mathbb{E}_F S(x, Y), \quad F \in \mathcal{P}, x \in \mathsf{A}.$$

It is *strictly $\mathcal{P}$-consistent* if "=" implies $x = T(F)$.
The functional T is called *k-elicitable relative to $\mathcal{P}$* if there exists a scoring function S that is strictly consistent for it.

In other words

$$T(F) = \arg\min_{x \in \mathsf{A}} \mathbb{E}_F S(x, Y).$$

Elicitable functionals

1-Elicitable

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

2-Elicitable

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

k-**Elicitable**

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

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

Theorem (Fissler and Z, 2016)

Let $\alpha \in (0,1)$, and $\mathsf{A}_0 := \{x \in \mathbb{R}^2 : x_1 \geq x_2\}$. Let $\mathcal{P}$ be a class of probability measures on $\mathbb{R}$ with finite first moments and unique α-quantiles. Any scoring function $S\colon \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ of the form

$$S(x_1, x_2, y) = \big(\mathbb{1}\{y \leq x_1\} - \alpha\big) G_1(x_1) - \mathbb{1}\{y \leq x_1\} G_1(y) + G_2(x_2)\left(x_2 - x_1 + \frac{1}{\alpha}\mathbb{1}\{y \leq x_1\}(x_1 - y)\right) - \mathcal{G}_2(x_2)$$

with $\mathcal{G}_2' = G_2$, is $\mathcal{P}$-consistent for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ if $\mathbb{1}_{(-\infty, x_1]} G_1$ is $\mathcal{P}$-integrable and

- *G_1 is increasing and $\mathcal{G}_2$ is increasing and convex.*

It is strictly $\mathcal{P}$-consistent if, additionally,

- *$\mathcal{G}_2$ is strictly increasing and strictly convex.*

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

Theorem (Fissler and Z, 2016, Part 2)

If $T(\mathcal{P}) = \mathsf{A}_0$, the class $\mathcal{P}$ is rich enough and S fulfils some smoothness conditions, all strictly $\mathcal{P}$-consistent scoring functions for T are of the above form (up to equivalence).

Corollary

If the elements of $\mathcal{P}$ have finite first moment and unique α-quantiles, then the pair $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha) \colon \mathcal{P} \to \mathsf{A}_0$ is 2-elicitable.

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

Theorem (Fissler and Z, 2016, Part 2)

If $T(\mathcal{P}) = \mathsf{A}_0$, the class $\mathcal{P}$ is rich enough and S fulfils some smoothness conditions, all strictly $\mathcal{P}$-consistent scoring functions for T are of the above form (up to equivalence).

Corollary

If the elements of $\mathcal{P}$ have finite first moment and unique α-quantiles, then the pair $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha) \colon \mathcal{P} \to \mathsf{A}_0$ is 2-elicitable.

Osband's Principle

- Osband's principle originates from Osband (1985) and gives necessary condition for strictly consistent scoring functions.
- It gives a connection of partial derivatives of the expected score and an expected identification function.

Definition

An *$\mathcal{P}$-identification function* for a functional T is a function $V : \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ such that

$$\mathbb{E}_F V(x, Y) = 0 \iff x = T(F)$$

for all $F \in \mathcal{P}$ and for all $x \in \mathsf{A}$.

Examples:

- Mean: $V(x, y) = x - y$
- α-quantile: $V(x, y) = \mathbb{1}\{y \leq x\} - \alpha$.

Osband's Principle

- Osband's principle originates from Osband (1985) and gives necessary condition for strictly consistent scoring functions.
- It gives a connection of partial derivatives of the expected score and an expected identification function.

Definition

An *$\mathcal{P}$-identification function* for a functional T is a function $V : \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ such that

$$\mathbb{E}_F V(x, Y) = 0 \quad \Longleftrightarrow \quad x = T(F)$$

for all $F \in \mathcal{P}$ and for all $x \in \mathsf{A}$.

Examples:

- Mean: $V(x, y) = x - y$
- α-quantile: $V(x, y) = \mathbb{1}\{y \le x\} - \alpha$.

Osband's Principle

Theorem (Osband's Principle; Fissler and Z (2016))

Let $T\colon \mathcal{P} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a surjective, elicitable and identifiable functional with $\mathcal{P}$-identification function $V\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ and a strictly $\mathcal{P}$-consistent scoring function $S\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$. Under some assumptions, there exists a matrix-valued function $h\colon \operatorname{int}(\mathsf{A}) \to \mathbb{R}^{k\times k}$ such that

$$\nabla_x \mathbb{E}_F S(x, Y) = h(x)\, \mathbb{E}_F V(x, Y)$$

for all $x \in \operatorname{int}(\mathsf{A})$ and $F \in \mathcal{P}$.

Key idea: Exploit the first order condition of the minimization problem:

$$\rightsquigarrow \nabla_x \mathbb{E}_F S(x, Y) = 0 \qquad \text{for } x = T(F) \text{ for all } F \in \mathcal{P}.$$

"The gradient ∇S is an identification function."

Osband's Principle

Theorem (Osband's Principle; Fissler and Z (2016))

Let $T\colon \mathcal{P} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a surjective, elicitable and identifiable functional with $\mathcal{P}$-identification function $V\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ and a strictly $\mathcal{P}$-consistent scoring function $S\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$. Under some assumptions, there exists a matrix-valued function $h\colon \operatorname{int}(\mathsf{A}) \to \mathbb{R}^{k\times k}$ such that

$$\nabla_x \mathbb{E}_F S(x, Y) = h(x)\, \mathbb{E}_F V(x, Y)$$

for all $x \in \operatorname{int}(\mathsf{A})$ and $F \in \mathcal{P}$.

Key idea: Exploit the first order condition of the minimization problem:

$$\rightsquigarrow \nabla_x \mathbb{E}_F S(x, Y) = 0 \qquad \text{for } x = T(F) \text{ for all } F \in \mathcal{P}.$$

"The gradient ∇S is an identification function."

Osband's Principle

(Under some smoothness conditions)
Second order conditions for the minimization problem: The Hessian

$$\nabla_x^2 \left[\mathbb{E}_F S(x, Y)\right] \in \mathbb{R}^{k \times k}$$

must be symmetric for all $x \in \mathsf{A}, F \in \mathcal{P}$, and positive semi-definite at $x = T(F)$.

- For $k = 1$ the necessary conditions of Osband's principle directly lead to sufficient conditions: For an oriented identification function, choose some $h > 0$ and integrate.
- Harder for $k > 1$:
  - Symmetry/positive semi-definiteness of the Hessian imposes (complicated) restrictions on the function h.
  - Even if $x \mapsto \mathbb{E}_F S(x, Y)$ has only one critical point and the Hessian is positive definite there, we can only guarantee a local minimum!
- $\rightsquigarrow$ Generally, we must verify sufficient conditions on a case by case basis.

Osband's Principle

(Under some smoothness conditions)
Second order conditions for the minimization problem: The Hessian

$$\nabla_x^2 \left[\mathbb{E}_F S(x, Y)\right] \in \mathbb{R}^{k \times k}$$

must be symmetric for all $x \in \mathsf{A}, F \in \mathcal{P}$, and positive semi-definite at $x = T(F)$.

- For $k = 1$ the necessary conditions of Osband's principle directly lead to sufficient conditions: For an oriented identification function, choose some $h > 0$ and integrate.
- Harder for $k > 1$:
  - Symmetry/positive semi-definiteness of the Hessian imposes (complicated) restrictions on the function h.
  - Even if $x \mapsto \mathbb{E}_F S(x, Y)$ has only one critical point and the Hessian is positive definite there, we can only guarantee a local minimum!
- $\rightsquigarrow$ Generally, we must verify sufficient conditions on a case by case basis.

Osband's Principle

(Under some smoothness conditions)
Second order conditions for the minimization problem: The Hessian

$$\nabla_x^2 \left[\mathbb{E}_F S(x, Y)\right] \in \mathbb{R}^{k \times k}$$

must be symmetric for all $x \in \mathsf{A}, F \in \mathcal{P}$, and positive semi-definite at $x = T(F)$.

- For $k = 1$ the necessary conditions of Osband's principle directly lead to sufficient conditions: For an oriented identification function, choose some $h > 0$ and integrate.
- Harder for $k > 1$:
  - Symmetry/positive semi-definiteness of the Hessian imposes (complicated) restrictions on the function h.
  - Even if $x \mapsto \mathbb{E}_F S(x, Y)$ has only one critical point and the Hessian is positive definite there, we can only guarantee a local minimum!
- $\rightsquigarrow$ Generally, we must verify sufficient conditions on a case by case basis.

Application examples

In Fissler and Z (2016), we considered:

- Functionals with elicitable components (vectors of quantiles, expectiles, ratios of expectations,. . .)
- Spectral risk measures with finitely supported spectral measure
- In particular: $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

Functionals with elicitable components – two examples

Vectors of quantiles

- Let $T(F) = (q_{\alpha_1}(F), \ldots, q_{\alpha_k}(F))$ with pairwise different $\alpha_1, \ldots, \alpha_k$.
- Strictly consistent scoring functions are of the form

$$S(x_1, \ldots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y).$$

Vectors of expectations

- Let $T(F) = \mathbb{E}_F(p(Y))$ for some $p : \mathbb{R}^d \to \mathbb{R}^k$.
- Strictly consistent scoring functions are of the form

$$S(x, y) = \phi(y) - \phi(x) - \langle \nabla\phi(x), p(y) - x \rangle.$$

where ϕ is strictly convex.

Functionals with elicitable components – two examples

Vectors of quantiles

- Let $T(F) = (q_{\alpha_1}(F), \ldots, q_{\alpha_k}(F))$ with pairwise different $\alpha_1, \ldots, \alpha_k$.
- Strictly consistent scoring functions are of the form

$$S(x_1, \ldots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y).$$

Vectors of expectations

- Let $T(F) = \mathbb{E}_F(p(Y))$ for some $p : \mathbb{R}^d \to \mathbb{R}^k$.
- Strictly consistent scoring functions are of the form

$$S(x, y) = \phi(y) - \phi(x) - \langle \nabla\phi(x), p(y) - x \rangle.$$

where ϕ is strictly convex.

Outline

1. Elicitability
   - Definition and a simple example
   - Risk measures in banking
   - *k*-Elicitability
2. Evaluating forecasts of expected shortfall
   - Absolute forecast evaluation
   - Classical comparative forecast evaluation
   - Comparative forecast evaluation with Murphy diagrams
3. Summary

Evaluating forecasts of expected shortfall

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

Prediction-observation triples

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

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
Y_t: Realization at time point t, $\mathcal{F}_t$-measurable

Absolute evaluation: Model verification

Let V be an identification function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, that is

$$\mathbb{E}(V(q, v, Y)) = 0 \quad \Leftrightarrow \quad (q, v) = (\mathrm{VaR}_\alpha(Y), \mathrm{ES}_\alpha(Y)).$$

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

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

Compare Davis (2016).

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.
- Most existing backtests can be described as a test for conditional calibration. (McNeil and Frey, 2000, Acerbi and Szekely 2014)
- Elicitability is not relevant.
- 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.
- Most existing backtests can be described as a test for conditional calibration. (McNeil and Frey, 2000, Acerbi and Szekely 2014)
- Elicitability is not relevant.
- Does not give guidance for decision between methods.
- Does not respect increasing information sets.

Comparative evaluation: Model selection

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

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

Comparative evaluation: Model selection

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

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

Comparative evaluation: Model selection

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

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

Forecast dominance

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

Definition (S-Dominance)

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

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

Comparative backtesting

$$\lambda^* := \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(S(Q_t, E_t, X_t) - S(Q_t^*, E_t^*, X_t)),$$

$$\lambda_* := \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(S(Q_t, E_{t,t}\ ) - S(Q_t^*, E_t^*, X_t)).$$

- S-dominance implies $\lambda_* \leq \lambda^* \leq 0$.
- $\lambda^* \leq 0$: Internal model is *at least as good* as the standard model.
- $\lambda_* \geq 0$: Internal model *predicts at most as well* as the standard model.

Comparative backtesting

$$H_0^- : \lambda^* \leq 0, \qquad H_0^+ : \lambda_* \geq 0.$$

- $\Delta_n \bar{S} := \frac{1}{n} \sum_{t=1}^{n} (S(Q_t, E_t, Y_t) - S(Q_t^*, E_t^*, Y_t^*))$.
- Under suitable assumptions on the process of score differences: Asymptotically normal test statistic

$$T_2 = \frac{\Delta_n \bar{S}}{\hat{\sigma}_n / \sqrt{n}},$$

where $\hat{\sigma}_n{}^2$ is an HAC estimator of $\sigma_n^2 = \mathrm{var}(\sqrt{n}\Delta_n \bar{S})$.
- Reject H_0^- if T_2 is "too much" ≥ 0.
- Reject H_0^+ if T_2 is "too much" ≤ 0.

(Diebold and Mariano, 1995, Giacomini and White, 2006)

Comparative backtesting

- 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.
- Allows for sensible comparison between methods.
- Necessitates a standard reference model.
- Respects increasing information sets (Holzmann and Eulert, 2014).

Comparative backtesting

- 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.
- Allows for sensible comparison between methods.
- Necessitates a standard reference model.
- Respects increasing information sets (Holzmann and Eulert, 2014).

Three zone approaches

BIS three zone approach for VaR_α

- Traditional backtest: One-sided binomial test.
- Backtesting decision:

	Red	Yellow	Green
p-value	very small	moderately small	sufficiently big

- Generalization of three zone approach for ES_α by Costanzino and Curran (2015).

Three zone approaches

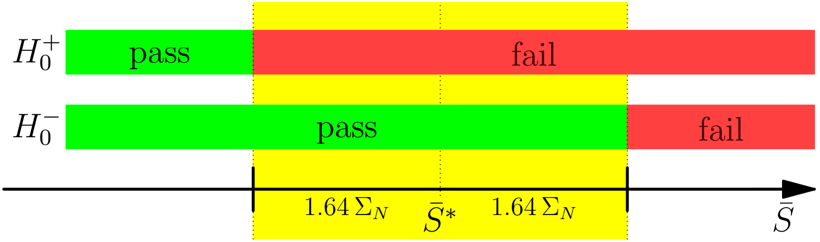
BIS three zone approach for VaR_α

- Traditional backtest: One-sided binomial test.
- Backtesting decision:

	Red	Yellow	Green
p-value	very small	moderately small	sufficiently big

- Generalization of three zone approach for ES_α by Costanzino and Curran (2015).

Three zone approach for comparative backtesting



A numerical illustration on nested information sets

$(\mu_t)_{t=1,\dots,N}$ iid standard normal,

$$Y_t \sim \mathcal{N}(\mu_t, 1), \quad \text{conditional on } \mu_t.$$

Scenario A	
(v_t, e_t)	$= (\mathrm{VaR}_\alpha(\mathcal{N}(\mu_t, 1)), \mathrm{ES}_\alpha(\mathcal{N}(\mu_t, 1)))$
(v_t^*, e_t^*)	$= (\mathrm{VaR}_\alpha(\mathcal{N}(0, 2)), \mathrm{ES}_\alpha(\mathcal{N}(0, 2)))$

Scenario B	
(v_t, e_t)	$= (\mathrm{VaR}_\alpha(\mathcal{N}(0, 2)), \mathrm{ES}_\alpha(\mathcal{N}(0, 2)))$
(v_t^*, e_t^*)	$= (\mathrm{VaR}_\alpha(\mathcal{N}(\mu_t, 1)), \mathrm{ES}_\alpha(\mathcal{N}(\mu_t, 1)))$

A numerical illustration – cont'd

Scenario A		**Green**	Yellow	Red
Traditional	$VaR_{0.01}$	89.35	10.65	0.00
Traditional	$ES_{0.025}$	93.62	6.36	0.02
Comparative	$VaR_{0.01}$	88.23	11.77	0.00
Comparative	$ES_{0.025}$	87.22	12.78	0.00

Scenario B		Green	Yellow	**Red**
Traditional	$VaR_{0.01}$	89.33	10.67	0.00
Traditional	$ES_{0.025}$	93.80	6.18	0.02
Comparative	$VaR_{0.01}$	0.00	11.77	88.23
Comparative	$ES_{0.025}$	0.00	12.78	87.22

$N = 250$; 10'000 simulations

Choice of a scoring function

- Fissler and Z (2016): $G_1(x_1) = x_1$, $G_2(x_2) = e^{x_2}$
- Fissler, Z and Gneiting (2016): $G_1(x_1) = x_1$, $G_2(x_2) = e^{x_2}/(1 + e^{x_2})$

A scoring function S is called *positively homogeneous of degree b* if

$$S(cx, cy) = c^b S(x, y), \quad \text{for all } c > 0.$$

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

Choice of a scoring function

- Fissler and Z (2016): $G_1(x_1) = x_1$, $G_2(x_2) = e^{x_2}$
- Fissler, Z and Gneiting (2016): $G_1(x_1) = x_1$, $G_2(x_2) = e^{x_2}/(1 + e^{x_2})$

A scoring function S is called *positively homogeneous of degree b* if

$$S(cx, cy) = c^b S(x, y), \quad \text{for all } c > 0.$$

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

Homogeneous scores for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

- For the action domain $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$, there are positively homogeneous strictly consistent scoring functions of degree $b \in (-\infty, 1)\backslash\{0\}$.
- There are strictly consistent scoring functions on $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$ such that the score differences are positively homogeneous of degree $b = 0$.
- For $b \geq 1$ positively homogeneous strictly consistent scoring functions can only be defined on smaller action domains $\mathsf{A} = \{Wx_1 < x_2\}$, compare the proposal by Acerbi and Szekely (2014).

Details can be found in Nolde and Z (2016).

Homogeneous scores for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

- For the action domain $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$, there are positively homogeneous strictly consistent scoring functions of degree $b \in (-\infty, 1)\backslash\{0\}$.
- There are strictly consistent scoring functions on $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$ such that the score differences are positively homogeneous of degree $b = 0$.
- For $b \geq 1$ positively homogeneous strictly consistent scoring functions can only be defined on smaller action domains $\mathsf{A} = \{Wx_1 < x_2\}$, compare the proposal by Acerbi and Szekely (2014).

Details can be found in Nolde and Z (2016).

Homogeneous scores for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$

- For the action domain $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$, there are positively homogeneous strictly consistent scoring functions of degree $b \in (-\infty, 1)\backslash\{0\}$.
- There are strictly consistent scoring functions on $\mathsf{A} = \mathbb{R} \times (-\infty, 0)$ such that the score differences are positively homogeneous of degree $b = 0$.
- For $b \geq 1$ positively homogeneous strictly consistent scoring functions can only be defined on smaller action domains $\mathsf{A} = \{Wx_1 < x_2\}$, compare the proposal by Acerbi and Szekely (2014).

Details can be found in Nolde and Z (2016).

A larger simulation study on comparative backtesting

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

$$Y_t = \mu_t + \varepsilon_t, \quad \mu_t = -0.05 + 0.3 Y_{t-1},$$

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

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

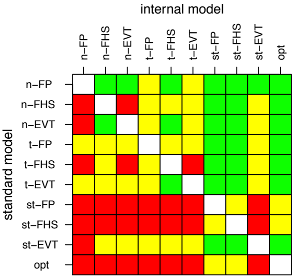
Estimation procedures:

- 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)

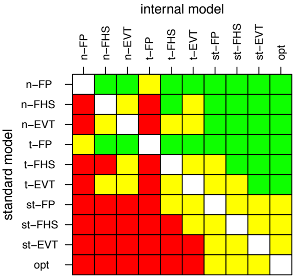
Moving window of size 500
5000 out-of-sample verifying observations

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

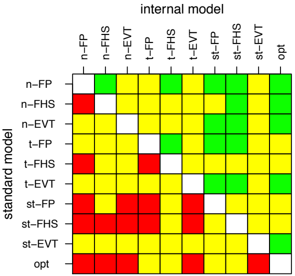
	$\alpha = 0.246$			$\alpha = 0.025$	
	simple	general		simple	general
n-FP	**0.000**	**0.000**	n-FP	**0.000**	**0.000**
n-FHS	0.881	0.184	n-FHS	0.653	0.231
n-EVT	0.754	0.672	n-EVT	0.886	0.226
t-FP	0.086	**0.006**	t-FP	**0.000**	**0.000**
t-FHS	0.936	0.512	t-FHS	0.697	0.717
t-EVT	0.880	0.475	t-EVT	0.995	0.498
st-FP	0.569	0.824	st-FP	0.695	0.419
st-FHS	0.909	0.796	st-FHS	0.843	0.758
st-EVT	0.935	0.706	st-EVT	0.962	0.564
opt	0.401	0.337	opt	0.131	0.571



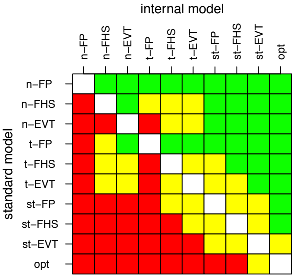
ν = 0.754



ν = 0.975



ν = 0.754



ν = 0.975

Can we avoid the choice of a specific scoring function for forecast comparison?

Forecast dominance

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

Definition (S-Dominance)

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

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

Forecast dominance

Definition (Dominance)

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

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

and for all consistent scoring functions S for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$.

Mixture representation

Proposition

Let $\alpha \in (0,1)$*. For* $v_1, v_2, y \in \mathbb{R}$*,* $(x_1, x_2) \in \mathrm{A}$*, we define*

$$S_{v_1}(x_1, y) = (\mathbb{1}\{y \leq x_1\} - \alpha)\big(\mathbb{1}\{v_1 \leq x_1\} - \mathbb{1}\{v_1 \leq y\}\big)$$

$$S_{v_2}(x_1, x_2, y) = \mathbb{1}\{v_2 \leq x_2\}\left(\frac{1}{\alpha}\mathbb{1}\{y \leq x_1\}(x_1 - y) - (x_1 - v_2)\right) + \mathbb{1}\{v_2 \leq y\}(y - v_2).$$

All scoring functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ *can be written as*

$$S(x_1, x_2, y) = \int S_{v_1}(x_1, y)\,\mathrm{d}H_1(v_1) + \int S_{v_2}(x_1, x_2, y)\,\mathrm{d}H_2(v_2),$$

where H_1 *is a locally finite measure and* H_2 *is a measure that is finite on all intervals of the form* $(-\infty, x]$*,* $x \in \mathbb{R}$*.*

Assessing forecast dominance

Corollary

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

$$\mathbb{E}(S_{v_1}(Q_t, Y_t) - S_{v_1}(Q_t^*, Y_t)) \leq 0, \quad \textit{for all } t \in \mathbb{N},$$

and

$$\mathbb{E}(S_{v_2}(Q_t, E_t, Y_t) - S_{v_2}(Q_t^*, E_t^*, Y_t)) \leq 0, \quad \textit{for all } t \in \mathbb{N},$$

and for all $v_1, v_2 \in \mathbb{R}$*.*

- Forecast dominance can be assessed by considering a two-parameter family of consistent scoring functions, only.
- We are (primarily) interested in the ES forecast. Consider S_{v_2} only.

Murphy diagrams

Simplifying assumption

Assume that $(Q_t, E_t, Y_t)_{t\in\mathbb{N}}$, $(Q_t^*, E_t^*, Y_t)_{t\in\mathbb{N}}$ are stationary and ergodic.

Murphy diagram

Plot

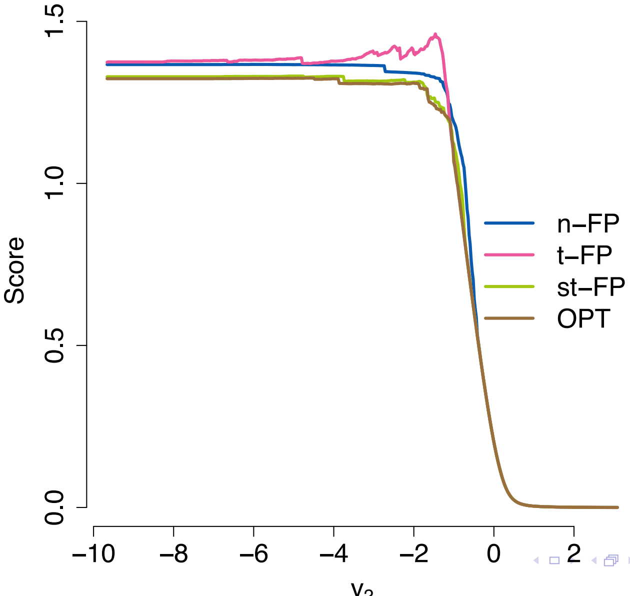
$$v_2 \mapsto \frac{1}{n}\sum_{t=1}^{n}(S_{v_2}(Q_t, E_t, Y_t) - S_{v_2}(Q_t^*, E_t^*, Y_t))$$

as an estimate of

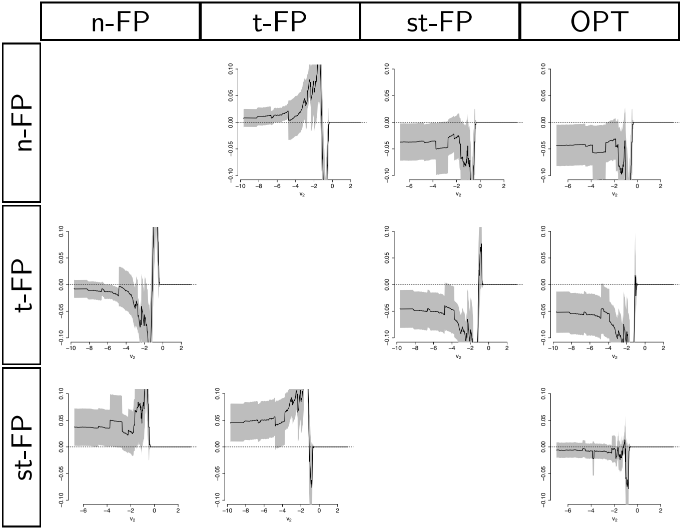
$$v_2 \mapsto \mathbb{E}(S_{v_2}(Q_t, E_t, Y_t) - S_{v_2}(Q_t^*, E_t^*, Y_t))$$

Idea of Murphy diagrams: Ehm et al. (2016, JRSSB).

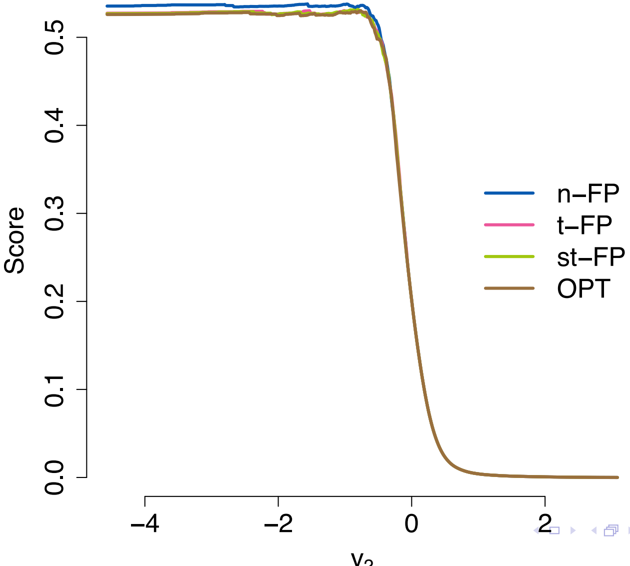
Comparison of parametric models, $\alpha = 0.025$



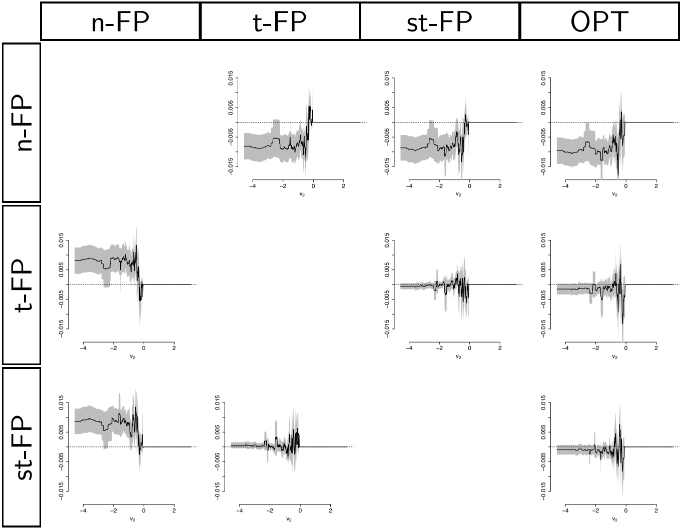
Comparison of parametric models, $\alpha = 0.025$



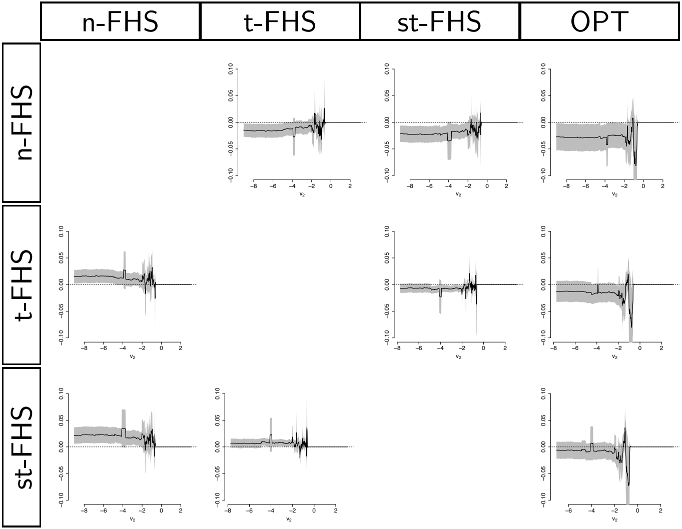
Comparison of parametric models, $\alpha = 0.246$



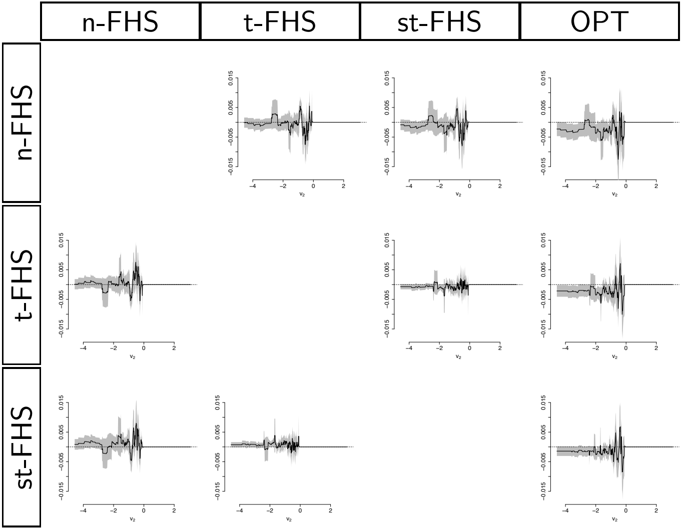
Comparison of parametric models, $\alpha = 0.246$



Influence of the filtering distribution, $\alpha = 0.025$



Influence of the filtering distribution, $\alpha = 0.246$



Formal tests for forecast dominance

- Formal tests for forecast dominance are possible.
- We have suggested the following procedure:
  - Diebold-Mariano tests for each grid point v_2.
  - Adjust p-values for multiple testing by the Westfall-Young procedure.
- Test works well in simulation examples, theoretical properties not fully understood, yet.

Outline

1. Elicitability
   - Definition and a simple example
   - Risk measures in banking
   - *k*-Elicitability
2. Evaluating forecasts of expected shortfall
   - Absolute forecast evaluation
   - Classical comparative forecast evaluation
   - Comparative forecast evaluation with Murphy diagrams
3. Summary

Summary

- k-Elicitability allows to find scoring 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.
- Consistent scoring functions can be used for forecast comparison.
- Characterization results for consistent scoring functions may allow for Murphy diagrams. These can be used for forecast comparison without the choice of a specific scoring function.
- The scoring functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ allow for M-estimation (Zwingmann & Holzmann, 2016), generalized regression (Bayer & Dimitriadis, 2017, Barendse, 2017).

References

(Almost) all mentioned references can be found in:

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.

J. F. Ziegel, F. Krüger, A. Jordan and F. Fasciati (2017). *Murphy Diagrams: Forecast Evaluation of Expected Shortfall*. Preprint, `arXiv:1705.04537`.

Thank you for your attention!