Huber loss as a scoring function

Rob Taggart, Bureau of Meteorology, Sydney

Harry Jack, Deryn Griffiths, Michael Foley, Nicholas Loveday

19 November 2020

Huber loss as a scoring function: Outline

1. Motivation

Applications of Huber loss

2. Verification with contaminated observations
3. Huber loss targets the *Huber mean*, which is a point summary of the centre of a distribution that has appealing properties
4. Huber mean arises naturally in optimal decision rules

See [Taggart 2020] for details and generalisations.

Contact: robert.taggart@bom.gov.au

Motivation

- Asked to assess quality of competing point forecasts for temperature
- Service not clearly defined (no directive in terms of a scoring function to minimise, or specific functional to target; diverse user group)

Two initial candidates:

- Squared error scoring function

$$S(x, y) = (x - y)^2$$

- Absolute error scoring function

$$S(x, y) = |x - y|$$

Here x is a point forecast and y is the verifying observation.

Subjective assessment of the cost of errors

- Which error sequence is better?

Error sequence	RMSE	MAE
$A = (1, 1, 1, 1, 1)$	1.0	1.0
$B = (0, 0, 0, 0, 5)$	2.2	1.0

  We prefer A and hence RMSE in this example

- Which error sequence is better?

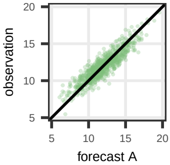
Error sequence	RMSE	MAE
$C = (22, 0)$	15.6	11.0
$D = (21, 5)$	15.3	13.0

  We prefer C and hence MAE in this example

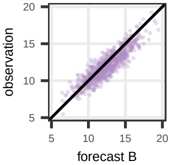
- What about sensitivity to contaminated observations?

Robust verification: example

System A: true errors $\sim \mathcal{N}(0,1)$ (no bias)



System B: true errors $\sim \mathcal{N}(0.5,1)$ (over forecast bias)



- Take a random sample of two years of daily forecast cases
- Null hypothesis: **"System A is no better than System B"**
  Alternative hypothesis: **"System A is better than System B"**

Likelihood that null hypothesis is rejected (at 5% significance level):

Scoring function	Likelihood based on true errors
Squared error scoring function	92%
Absolute error scoring function	90%

Robust verification: example

But now suppose that some observations are contaminated:

- 3% chance of a $+5$ measurement spike

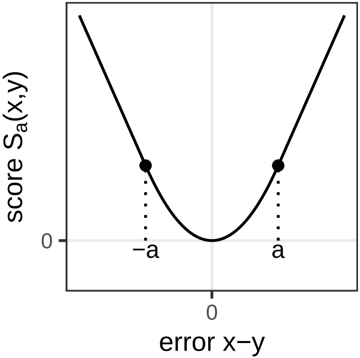
Likelihood that null hypothesis is rejected (at 5% significance level):

Scoring function	True errors	Contaminated errors
Squared error scoring function	92%	22%
Absolute error scoring function	90%	75%

A third candidate: Huber loss

Huber loss scoring function with tuning parameter a:

$$S_a(x, y) = \begin{cases} \frac{1}{2}(x-y)^2\,, & |x-y| \le a \\ a|x-y| - \frac{1}{2}a^2\,, & |x-y| > a \end{cases}$$



- Quadratic penalty for small errors
- Linear penalty for large errors

Introduced by Peter Huber (1964) because it gives rise to the most robust (in a certain sense) estimator of the location parameter for contaminated normal distributions.

Robust verification: example continued

Likelihood that null hypothesis is rejected (at 5% significance level):

Scoring function	True errors	Contaminated errors
Squared error scoring function	92%	22%
Absolute error scoring function	90%	75%
Huber loss scoring function ($a = 1.5$)	92%	70%

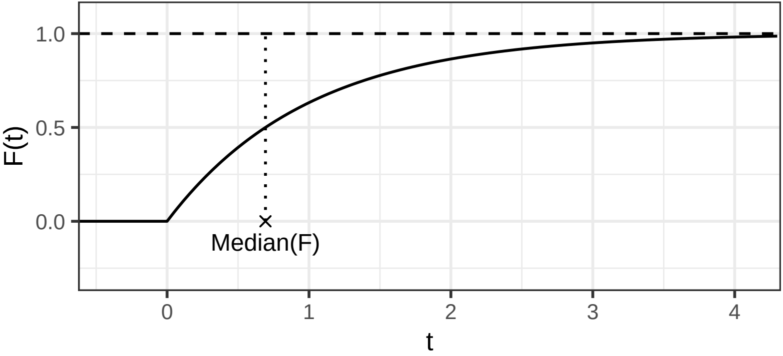
Tuning parameter $a = 1.5$ is suitable because

- most errors are between ± 1.5
- 1.5 is substantially less than the contaminating contribution $+5$

$S(x, y) = |x - y|$ **targets** $\mathrm{Median}(F)$

i.e., the optimal point forecast (for minimising expected score) is a median of one's predictive distribution F

Example: $F(t) = 1 - \exp(-t)$, $t \geq 0$

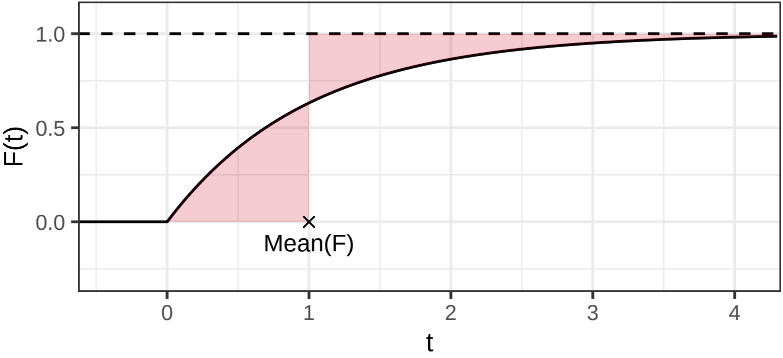


- two dotted vertical line segments have equal length

$S(x, y) = (x - y)^2$ **targets** $\mathrm{Mean}(F)$

i.e., the optimal point forecast (for minimising expected score) is the mean of one's predictive distribution F

Example: $F(t) = 1 - \exp(-t)$, $t \geq 0$

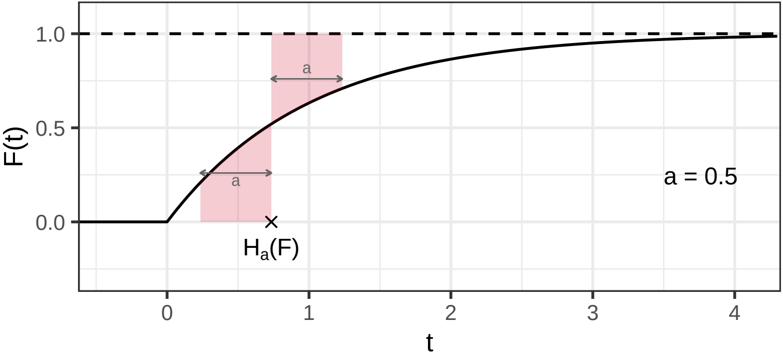


- two shaded regions have equal area

Huber loss $S_a(x, y)$ targets the Huber mean $\mathrm{H}_a(F)$

i.e., the optimal point forecast (for minimising expected score) is a *Huber mean* $\mathrm{H}_a(F)$ of one's predictive distribution F

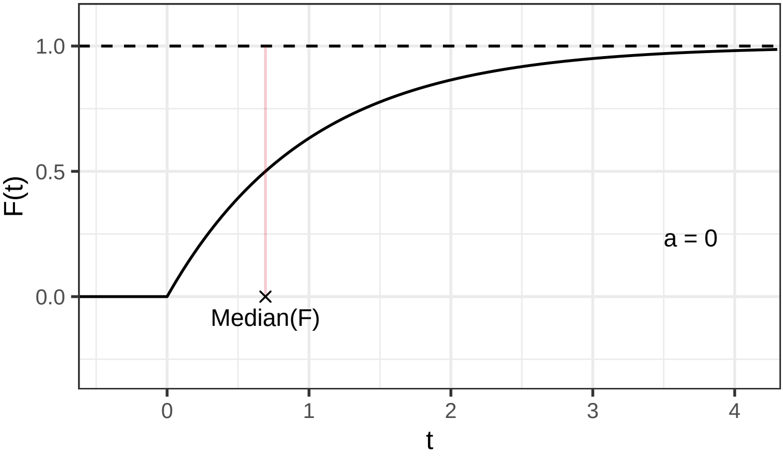
Example: $F(t) = 1 - \exp(-t)$, $t \geq 0$



- two shaded regions have equal area

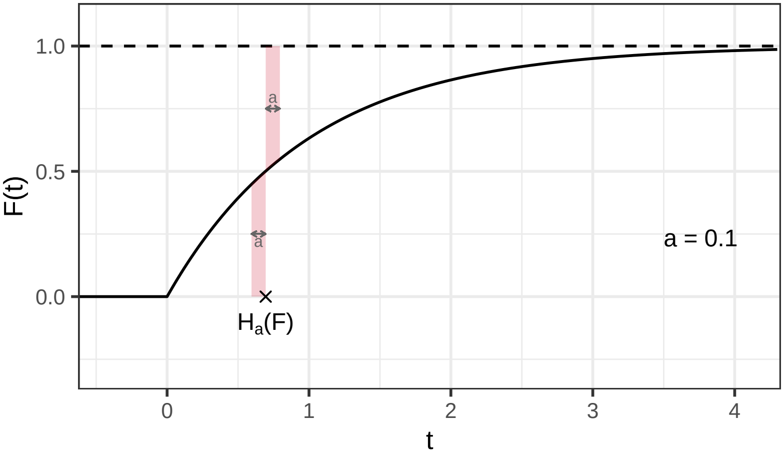
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



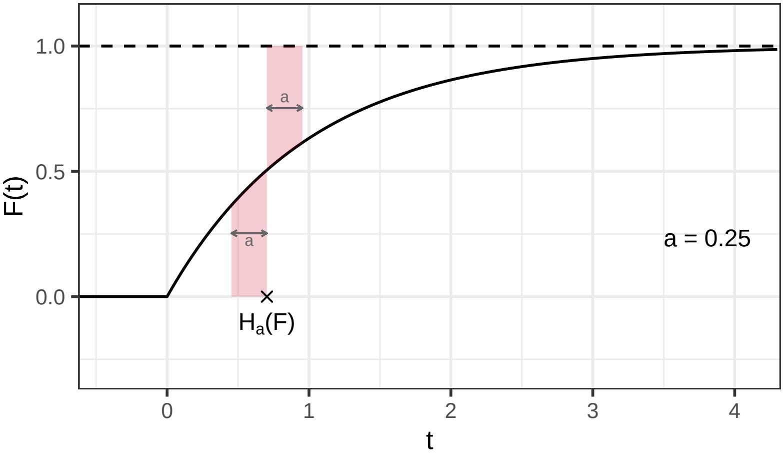
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



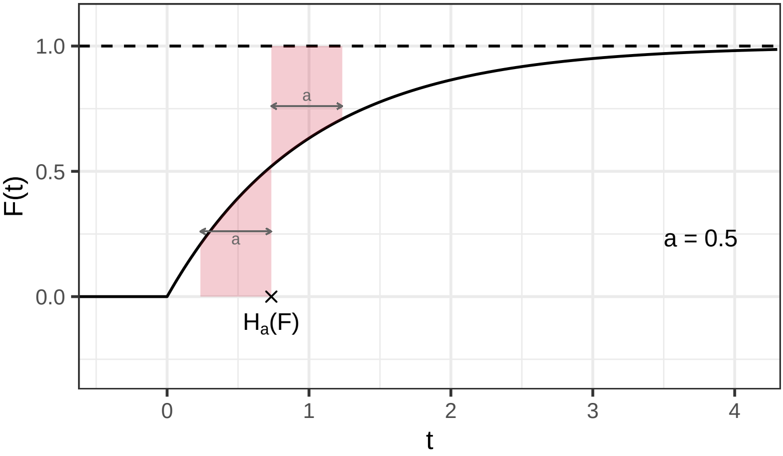
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



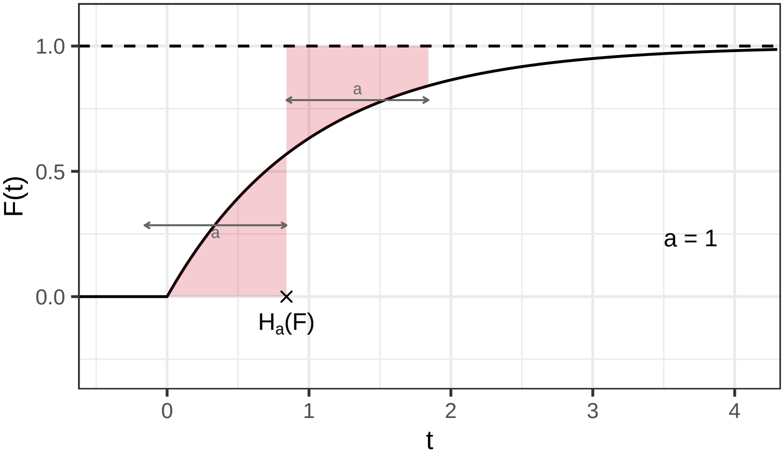
Median ... Huber mean ... Mean

Example: $F(t) = 1 - \exp(-t)$, $t \geq 0$



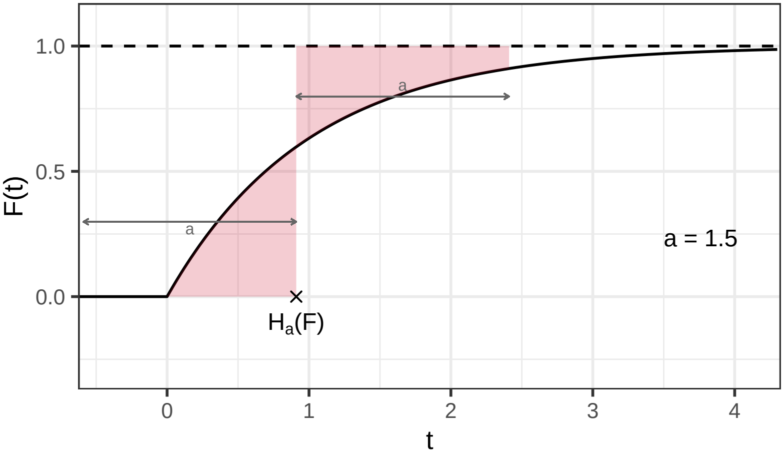
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



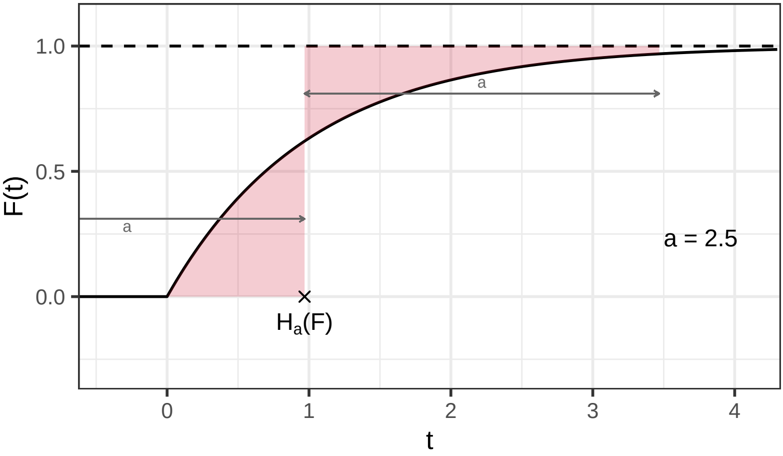
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



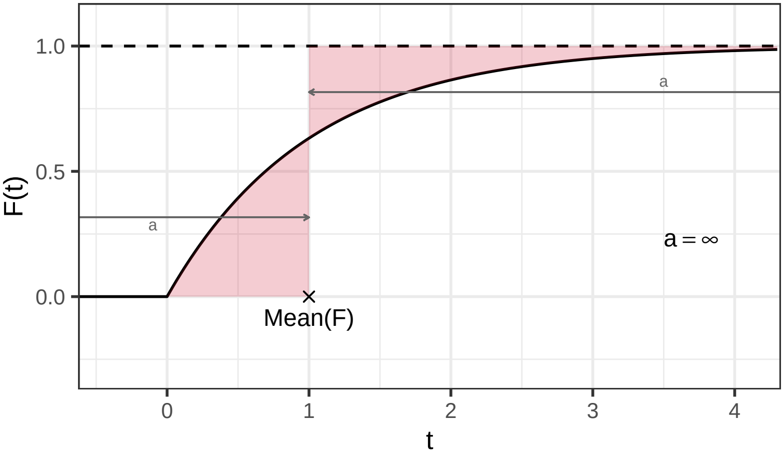
Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



Median . . . Huber mean . . . Mean

Example: $F(t) = 1 - \exp(-t),\ t \geq 0$



Some basic properties of the Huber mean $\mathrm{H}_a(F)$

1. $\mathrm{H}_a(F)$ is the midpoint of the 'central interval' of F with length $2a$
2. $\mathrm{H}_a(F) \to \mathrm{Median}(F)$ as $a \downarrow 0$
3. $\mathrm{H}_a(F) \to \mathrm{Mean}(F)$ as $a \to \infty$

In summary:

- The Huber mean is an intermediary between the median and mean.
- The Huber mean incorporates more information about the centre of a distribution than the median.
- The Huber mean is not sensitive to behaviour at the tails of a distribution, unlike the mean.
- The Huber loss scoring function is consistent (or proper) for the Huber mean.

See [Taggart 2020] for details and further properties; also [Huber 1964] for the case of finite discrete distributions.

Theoretical properties [Taggart 2020]

1. **Consistency:** S is consistent for the Huber mean H_a if and only if

$$S(x,y)=\begin{cases}\phi(y)-\phi(x)+\phi'(x)(x-y)\,, & |x-y|\le a\\ \phi(y)-\phi(y+a)+a\phi'(x)\,, & x-y>a\\ \phi(y)-\phi(y-a)-a\phi'(x)\,, & x-y<-a\end{cases}$$

   where ϕ is convex.

2. **Elicitability:** The Huber mean is elicitable.

3. **Mixture representation:** Every consistent scoring function S for the Huber mean H_a can be expressed as an integral

$$S(x,y)=\int_{-\infty}^{\infty} S_{\theta,a}(x,y)\,\mathrm{d}M(\theta)$$

   of *elementary scoring functions*

$$S_{\theta,a}(x,y)=\begin{cases}(1-\alpha)\min(\theta-y,a) & \text{if } y\le\theta<x\\ \alpha\min(y-\theta,a) & \text{if } x\le\theta<y\\ 0 & \text{otherwise}\,,\end{cases}$$

   where M is a nonnegative measure satisfying $\mathrm{d}M(\theta)=\mathrm{d}\phi(\theta)$.

Elementary scoring functions for the Huber mean

Elementary scoring functions for the Huber mean measure the economic regret, relative to actions based on a perfect forecast, of investment decisions with fixed up-front costs and where both profits and losses are capped.

Example. Each Friday, Joe decides whether to sell ice creams at a sports stadium the following afternoon.

- Up-front cost if he sells ice creams: \$120 (includes stadium fee)
- Expected profit p from sales depends on daily maximum temperature y:
$$p = 40y - 680, \quad y \geq 17.$$
- p capped by cart storage capacity: $0 \leq p \leq 240$

Joe makes a profit if and only if $y > 20^{\circ}\mathrm{C}$.

Elementary scoring functions for the Huber mean

Decision rule: sell ice creams if and only if point forecast x for maximum temperature exceeds 20°C.

Which point forecast x?

Do the maths... boils down to minimising the elementary score $S_{\theta,a}(x, y)$, where

- x is forecast maximum temperature
- y is the observed maximum temperature
- $\theta = 20°\mathrm{C}$
- $a = 3$

Optimal decision rule: Sell ice creams if and only if $x > 20°\mathrm{C}$, where

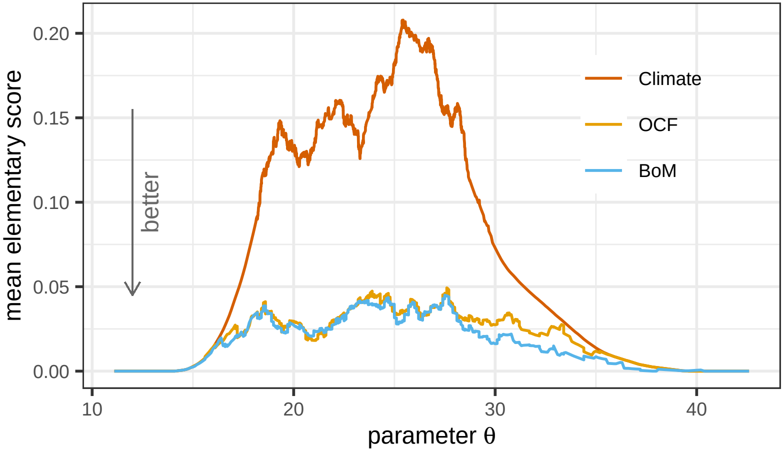
$$x = \mathrm{H}_3(F)$$

and F is Joe's predictive distribution for daily maximum temperature.

Murphy diagrams

A Murphy diagram is a graph of the mean elementary score $\bar{S}_{\theta,a}$ versus θ. See [Ehm et. al., 2016] for the cases of the mean, median and quantiles.

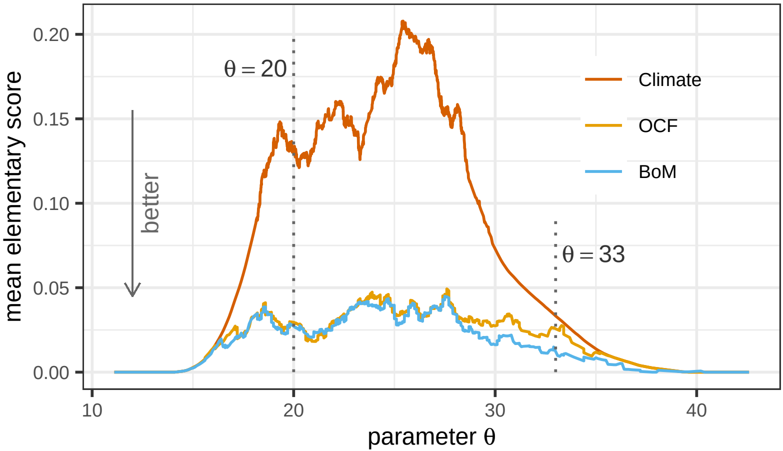
Three forecast systems targeting the Huber mean H_3 for maximim temperature at Sydney Observatory Hill (July 2018 to June 2020).



Murphy diagrams

Joe's decision rule: act if and only if H_3-forecast exceeds $\theta = 20°$C.
Joe should use BoM or OCF.

Wendy's decision rule: act if and only if H_3-forecast exceeds $\theta = 33°$C.
Wendy should use BoM.



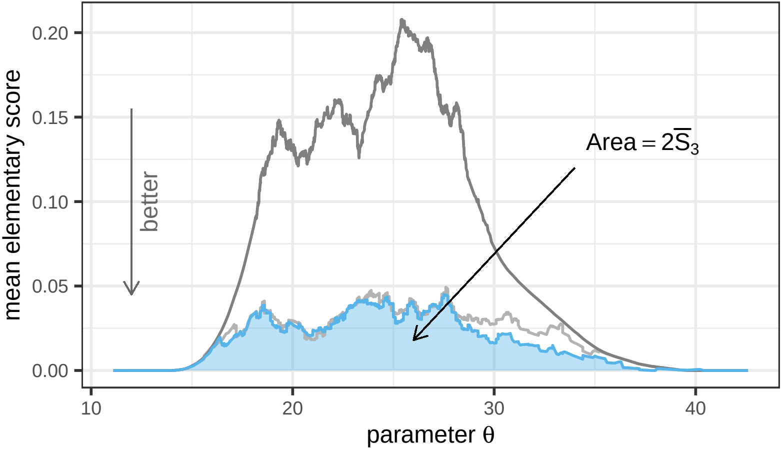
Murphy diagrams

Which forecast is best, on average, across all decision thresholds θ?

Mean Huber loss score $\bar{S}_3$ is twice the area under the Murphy Diagram.

BoM: $\bar{S}_3 = 1.001$

OCF: $\bar{S}_3 = 1.182$



Summary and references

1. Huber loss can be used as a robust scoring function.
2. The Huber mean is a good candidate statistic for summarising the centre of a distribution. It is an intermediary between the mean and median.
3. The Huber mean (more generally, Huber functional) arises naturally in optimal decision making for investment problems with fixed up-front costs and a cap on profits and losses.

Selected references

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). *Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings*, J. R. Statist. Soc. B, 78, 505–562.

Gneiting, T. (2011). *Making and evaluating point forecasts*, Journal of the American Statistical Association, 106, 746–762.

Huber, P. (1964). *Robust estimation of a location parameter*, Annals of Mathematical Statistics, 35, 73–101.

Taggart, R. (2020). *Point forecasting and forecast evaluation with generalised Huber loss*, Bureau of Meteorology Research Series (to appear).