Robust Forecast Evaluation

Identify questionable claims of forecast superiority

b
UNIVERSITÄT
BERN

When comparing the predictive performance of point forecasts as measured by a scoring function chosen from a usually large class, the ranking of forecasts often depends on the choice of scoring function. Simultaneous evaluation with multiple scoring functions gives an idea about the robustness of the forecast ranking.

Consistency in point forecasting

Point forecasts are popular due to various reasons, e.g. organizational requirements, ease of reporting, tradition.

A consensus among forecasters and evaluators needs to be reached about what **type of forecast** should be issued, where the type is defined in terms of a **(statistical) functional**, such as:

- Mean
- Value-at-Risk (VaR)
- Expected Shortfall (ES)

It is common to evaluate point forecasts with a **scoring function** that assigns a real-valued penalty depending on the forecast and the realization. This scoring function needs to be consistent for the functional.

Definition (Consistency)

The scoring function $S\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is consistent for the functional T relative to the class $\mathcal{F}$ of probability distributions if

$$\mathbb{E}_F\, S(T(F), Y) \leq \mathbb{E}_F S(x, Y)$$

for all $F \in \mathcal{F}$, and all $x \in \mathbb{R}^d$. That is,

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

Characterizations of scoring functions

While the requirement of consistency restricts the choice of scoring function for the evaluator, the remaining class is usually still large.

A scoring function $S\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is consistent for the **mean** functional if and only if it is of the form

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

where ϕ is **convex** with subgradient ϕ'.
Special case: $\phi(t) = t^2$ leads to the squared error

A scoring function $S\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is consistent for the **quantile functional (VaR)** at level $\alpha \in (0,1)$ if and only if it is of the form

$$S(q, y) = (\mathbb{1}\{y < q\} - \alpha)\big(g(q) - g(y)\big),$$

where g is **non-decreasing**.
Special case: $g(t) = t$ leads to the tick loss

A scoring function $S\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ is consistent for the functional **(VaR, ES)** at level $\alpha \in (0,1)$ if and only if it is of the form

$$\begin{aligned} S(v, e, y) = {} & (\mathbb{1}\{y < v\} - \alpha)\big(g(v) - g(y)\big) \\ & + \frac{\phi'(e)}{\alpha}(\mathbb{1}\{y < q\} - \alpha)(v - y) \\ & + \phi(y) - \phi(e) - \phi'(e)(y - e) \end{aligned}$$

where g is **non-decreasing** and ϕ is **convex**.

All three of these results depend on certain regularity conditions.

Forecast dominance

It has been demonstrated that **the choice of scoring function can influence the ranking** of two competing forecasts in the presence of model misspecification and non-nested information sets.

Definition (Dominance)

Let $\mathcal{S}$ be a class of consistent scoring functions for the functional T. For two forecasts X^A and X^B made by methods A and B, respectively, we say that method A weakly dominates method B with respect to $\mathcal{S}$ if

$$\mathbb{E}_F\, S(X^A, Y) \leq \mathbb{E}_F\, S(X^B, Y)$$

for all $S \in \mathcal{S}$, where F denotes the joint distribution of (X^A, X^B, Y).

Once **dominance** has been established for a given class $\mathcal{S}$, it **translates to** the extension including **all mixtures**, e.g., dominance with respect to $\{S_1, S_2\}$ implies dominance with respect to $\{aS_1 + bS_2 \colon a, b \geq 0\}$.

This simple observation is the basis for so-called Murphy diagrams which are **graphical tools** to check for forecast dominance empirically with respect to all consistent scoring functions.

Elementary scoring functions

We identify **linearly parameterized** classes of elementary scores that generate the previously identified classes of scoring functions.

We define the elementary scores that are consistent for the **mean** functional as

$$S_\eta(m, y) = (\mathbb{1}\{\eta \leq y\} - \mathbb{1}\{\eta \leq m\})(y - \eta),$$

where $\eta \in \mathbb{R}$.

We define the elementary scores that are consistent for the **quantile functional (VaR)** at level $\alpha \in (0,1)$ as

$$S_\eta(q, y) = (\mathbb{1}\{y \leq \eta\} - \alpha)(\mathbb{1}\{\eta \leq q\} - \mathbb{1}\{\eta \leq y\}).$$

where $\eta \in \mathbb{R}$.

We define the elementary scores that are consistent for the functional **(VaR, ES)** at level $\alpha \in (0,1)$ as

$$S_{\eta,1}(v, y) = (\mathbb{1}\{y \leq \eta\} - \alpha)(\mathbb{1}\{\eta \leq v\} - \mathbb{1}\{\eta \leq y\})$$

$$\begin{aligned} S_{\tilde{\eta},2}(v, e, y) = {} & \frac{\mathbb{1}\{\tilde{\eta} \leq e\}}{\alpha}(\mathbb{1}\{y \leq \tilde{\eta}\} - \alpha)(v - y) \\ & + (\mathbb{1}\{\tilde{\eta} \leq y\} - \mathbb{1}\{\tilde{\eta} \leq e\})(y - \tilde{\eta}), \end{aligned}$$

where $\eta, \tilde{\eta} \in \mathbb{R}$.

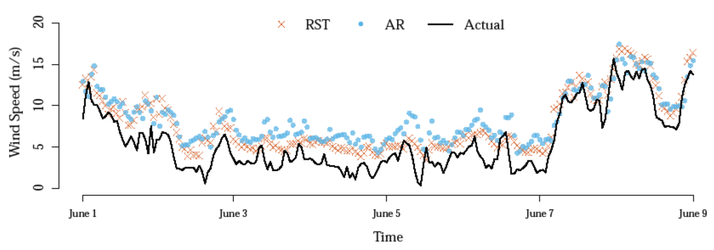
These elementary scores then lead to mixture representations

$$S(x, y) = \int S_\eta(x, y)\; dH(\eta),$$

where H is a nonnegative measure.

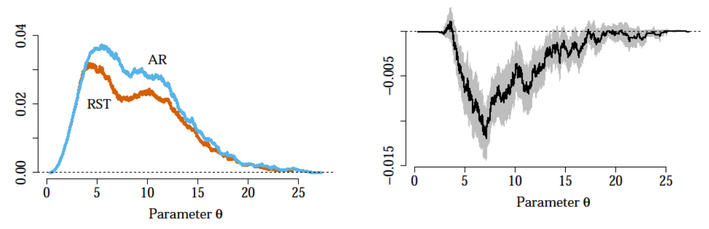
Case study

Data set from Gneiting et al. (2006)
90% quantile forecasts of wind speed, $n = 5136$.



Murphy diagram
(empirical expected elementary scores)

Score | Score difference
(< 0 means RST is preferable)

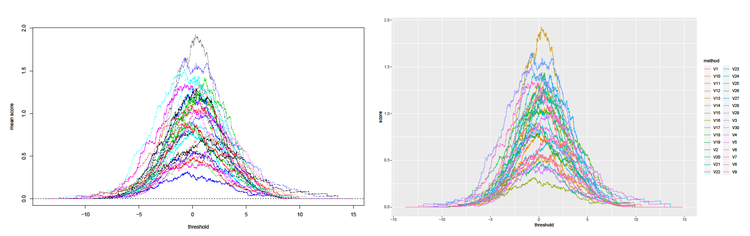


R-package: murphydiagram2

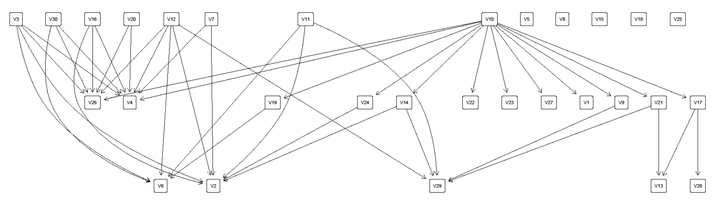
```
>  library(murphydiagram2)
>  library(ggplot2)

>  ex <- drawExample(n = 500, m = 30)
>  m <- murphydiag(ex$x, ex$y, "mean")
>  df <- as.data.frame(m)

>  plot(m)
>  ggplot(df, aes(x = threshold, y = score,
+     col = method)) + geom_line()
```



```
>  d <- dominance(m)
>  plot(d)
```



References

Ehm, W., Gneiting, T., Jordan, A., and Krüger F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings (with discussion and rejoinder). *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 78, 505 – 562.

Ziegel, J.F., Krüger, F., Jordan, A., Fasciati, F. (2018). Robust forecast evaluation of expected shortfall. arXiv:1705.04537.

Institute for Mathematical Statistics and Actuarial Science | Alexander I. Jordan | alexander.jordan@stat.unibe.ch