Forecaster's dilemma: Extreme events and forecast evaluation

Sebastian Lerch

Karlsruhe Institute of Technology

Heidelberg Institute for Theoretical Studies

7th International Verification Methods Workshop
Berlin, May 8, 2017

joint work with Thordis Thorarinsdottir, Francesco Ravazzolo and Tilmann Gneiting

Motivation

THE SPECTATOR

HOME | COFFEE HOUSE | ELECTION 2015 | MAGAZINE | COLUMNISTS | CULTURE HOUSE | PODCAST

The Week | Features | Columnists | Books | Arts | Life | Cartoons | Classified

Forecast failure: how the Met Office lost touch with reality

Ideology has corrupted a valuable British institution

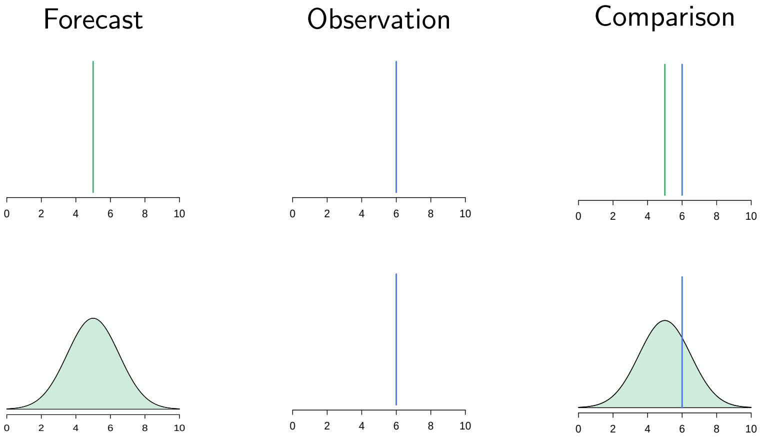
Rupert Darwall 13 July 2013 118 Comments

http://www.spectator.co.uk/features/8959941/whats-wrong-with-the-met-office/

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events

Probabilistic vs. point forecasts



Evaluation of probabilistic forecasts: Proper scoring rules

A proper scoring rule is any function

$$S(F, y)$$

such that

$$\mathbb{E}_{Y\sim G} S(G, Y) \leq \mathbb{E}_{Y\sim G} S(F, Y)$$

for all $F, G \in \mathcal{F}$.

We consider scores to be negatively oriented penalties that forecasters aim to minimize.

Gneiting, T. and Raftery, A. E. (2007) **Strictly proper scoring rules, prediction, and estimation**. *Journal of the American Statistical Association*, 102, 359–378.

Examples

Popular examples of proper scoring rules include

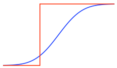
- the logarithmic score

$$\mathrm{LogS}(F, y) = -\log(f(y)),$$

  where f is the density of F,

- the continuous ranked probability score

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 dz,$$



  where the probabilistic forecast F is represented as a CDF.

Advertisement

R package scoringRules (joint work with Alexander Jordan and Fabian Krüger)

- implementations of popular proper scoring rules for ensemble forecasts and (many previously unavailable) parametric distributions
- implementations of multivariate scoring rules
- computationally efficient, statistically principled default choices

Available on CRAN, development version at `https://github.com/FK83/scoringRules`.

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events

Media attention often exclusively falls on prediction performance in the case of extreme events

He told us so

They called him Dr Doom. He was the economist who three years ago predicted in detail a collapse of the housing market and worldwide recession - and was roundly ridiculed for it. Emma Brockes asks Nouriel Roubini what he foresees now

http://www.theguardian.com/business/2009/jan/24/nouriel-roubini-credit-crunch

Toy example

We compare Alice's and Bob's forecasts for $Y \sim \mathcal{N}(0,1)$,

$$F_{\text{Alice}} = \mathcal{N}(0,1), \qquad F_{\text{Bob}} = \mathcal{N}(4,1)$$

Based on all 10 000 replicates,

Forecaster	CRPS	LogS
Alice	**0.56**	**1.42**
Bob	3.53	9.36

When the evaluation is restricted to the largest ten observations,

Forecaster	R-CRPS	R-LogS
Alice	2.70	6.29
Bob	**0.46**	**1.21**

Verifying only the extremes erases propriety

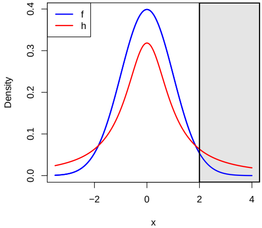
Some econometric papers use the restricted logarithmic score

$$\text{R-LogS}_{\geq r}(F, y) = -\mathbb{1}\{y \geq r\} \log f(y).$$

However, if $h(x) > f(x)$ for all $x \geq r$, then

$$\mathbb{E}\,\text{R-LogS}_{\geq r}(H, Y) < \mathbb{E}\,\text{R-LogS}_{\geq r}(F, Y)$$

independently of the true density.



In fact, if the forecaster's belief is F, her best prediction under $\text{R-LogS}_{\geq r}$ is

$$f^*(z) = \frac{\mathbb{1}(z \geq r) f(z)}{\int_r^\infty f(x)dx}.$$

The forecaster's dilemma

Given any (non-trivial) proper scoring rule S and any non-constant weight function w, any scoring rule of the form

$$S^*(F, y) = w(y)S(F, y)$$

is improper.

Forecaster's dilemma: Forecast evaluation based on a subset of extreme observations only corresponds to the use of an improper scoring rule and is bound to discredit skillful forecasters.

Outline

1. Probabilistic forecasting and forecast evaluation
2. The forecaster's dilemma
3. Proper forecast evaluation for extreme events

Proper weighted scoring rules I

Proper weighted scoring rules provide suitable alternatives.

Gneiting and Ranjan (2011) propose the threshold-weighted CRPS

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z)\, \mathrm{d}z$$

$w(z)$ is a weight function on the real line.

Weighted versions can also be constructed for the logarithmic score (Diks, Panchenko, and van Dijk, 2011).

Gneiting, T. and Ranjan, R. (2011) **Comparing density forecasts using threshold- and quantile-weighted scoring rules**. *Journal of Business and Economic Statistics*, 29, 411–422.

Role of the weight function

The weight function w can be tailored to the situation of interest.

For example, if interest focuses on the predictive performance in the right tail,

$$w_{\text{indicator}}(z) = \mathbb{1}\{z \geq r\}, \text{or}$$
$$w_{\text{Gaussian}}(z) = \Phi(z|\mu_r, \sigma_r^2)$$

Choices for the parameters r, μ_r, σ_r can be motivated and justified by applications at hand.

Toy example revisited

Recall Alice's and Bob's forecasts for $Y \sim \mathcal{N}(0,1)$,

$$F_{\text{Alice}} = \mathcal{N}(0,1), \qquad F_{\text{Bob}} = \mathcal{N}(4,1)$$

based on all 10 000 replicates

Forecaster	CRPS	LogS
Alice	**0.56**	**1.42**
Bob	3.53	9.36

based the largest 10 observations

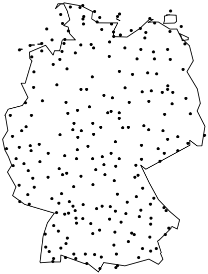
Forecaster	R-CRPS	R-LogS
Alice	2.70	6.29
Bob	**0.46**	**1.21**

threshold-weighted CRPS, with indicator weight $w(z) = \mathbb{1}\{z \geq 2\}$ and Gaussian weight $w(z) = \Phi(z|\mu_r = 2, \sigma = 1)$

Forecaster	$w_{\text{indicator}}$	w_{Gaussian}
Alice	**0.076**	**0.129**
Bob	2.355	2.255

Case study: Probabilistic wind speed forecasting

- Forecasts and observations of daily maximum wind speed
- Prediction horizon of 1-day ahead
- 228 observation stations over Germany
- Evaluation period: May 2010 – April 2011
- 90% of observations $\in [2.7\frac{m}{s}, 11.7\frac{m}{s}]$



Probabilistic forecasts:

- ECMWF ensemble (maximum over forecast period)
- Bob: for every forecast case,

$$F = \mathcal{N}(15, 1)$$

Case study: Results

based on all observations

Forecaster	CRPS
ECMWF	**1.26**
Bob	8.49

based on observations > 14

Forecaster	R-CRPS
ECMWF	6.87
Bob	**1.80**

threshold-weighted CRPS, with indicator weight $w(z) = \mathbb{1}\{z \geq 14\}$ and Gaussian weight $w(z) = \Phi(z|\mu_r = 14, \sigma = 1)$

Forecaster	$w_{\text{indicator}}$	w_{Gaussian}
ECMWF	**0.059**	**0.063**
Bob	0.653	0.761

Post-processing models and improvements for high wind speeds:

Lerch, S. and Thorarinsdottir, T.L. (2013) **Comparison of non-homogeneous regression models for probabilistic wind speed forecasting**. *Tellus A*, 65: 21206.

Summary and conclusions

- Forecaster's dilemma: Verification on extreme events only is bound to discredit skillful forecasters.
- The only remedy is to consider all available cases when evaluating predictive performance.
- Proper weighted scoring rules emphasize specific regions of interest, such as tails, and facilitate interpretation, while avoiding the forecaster's dilemma.
- In particular, the weighted versions of the CRPS share (almost all of) the desirable properties of the unweighted CRPS.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017)
Forecaster's dilemma: Extreme events and forecast evaluation.
Statistical Science, 32, 106–127.

Thank you for your attention!