Evaluation of interval forecasts – a brief overview

Johannes Bracher

March 9, 2021

PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

Evaluating epidemic forecasts in an interval format

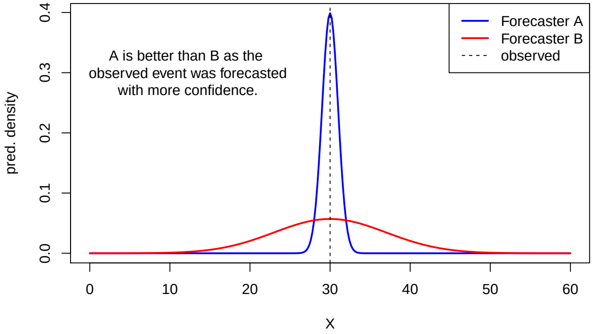
Johannes Bracher[1,2], **Evan L. Ray**[3], **Tilmann Gneiting**[2,4], **Nicholas G. Reich**[3]*

1 Chair of Statistics and Econometrics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, **2** Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, **3** School of Public Health and Health Sciences, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, United States of America, **4** Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008618

Why take into account uncertainty in forecast evaluation?

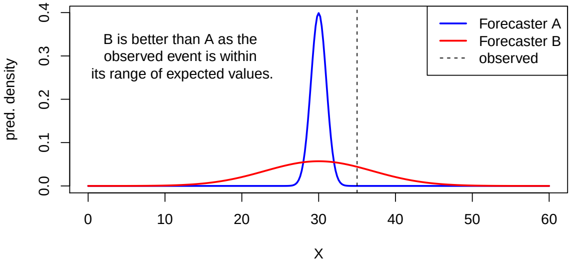
- Forecast quality cannot be fully described considering only the central tendency:



- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

Why take into account uncertainty in forecast evaluation?

- Forecast quality cannot be fully described considering only the central tendency:



- Good forecasts "maximize sharpness subject to calibration"
- Proper scoring rules (Gneiting and Raftery 2007) allow us to compare probabilistic forecasts

Proper scoring rules

Gneiting and Ratery 2007, https://doi.org/10.1198/016214506000001437

- **Proper scoring rules** encourage honest forecasting
  - Forecasters maximize the (subjective) expected score by reporting their actual predictive distribution
  - No way to "cheat the score"
  - Good forecasts "maximize sharpness subject to calibration"

Proper scoring rules (continued)

- Popular choices:
  - **logarithmic score** / predictive log-likelihood:

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

    ie the predictive density at the observed value y.
  - **continuous ranked probability score** (CRPS):

    $$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \{F(x) - 1(x \geq y)\}^2 dx,$$

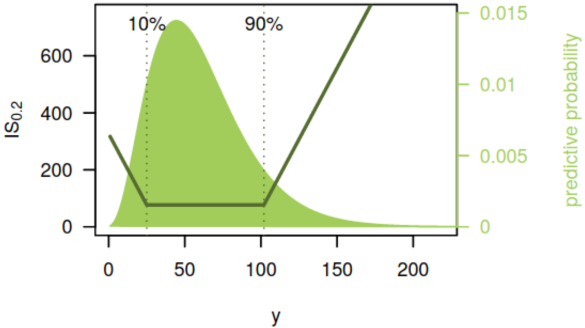
    ie the integrated squared distance between predictive and observed CDF.
- **Typically require full predictive distribution!**

The interval score

Consider a central $(1-\alpha) \times 100\%$ prediction interval $[l, u]$ and observation y. The **interval score** is given by

$$\mathrm{IS}_\alpha(F, y) = \underbrace{(u-l)}_{\text{spread}} + \underbrace{\frac{2}{\alpha}(l-y)\mathbb{1}(y<l)}_{\text{penalty for underprediction}} + \underbrace{\frac{2}{\alpha}(y-u)\mathbb{1}(y>u)}_{\text{penalty for overprediction}},$$

where $\mathbb{1}$ is the indicator function.



The weighted interval score

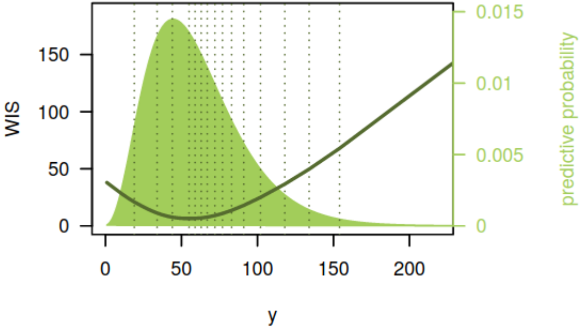
Bracher, Ray, Gneiting, Reich (2021)

To assess prediction intervals at levels $(1-\alpha_0, \ldots, 1-\alpha_K)$ simultaneously we can use the **weighted interval score**:

$$\mathrm{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K+1/2} \times \left\{ \frac{1}{2}|y-m| + \sum_{k=0}^{K} \frac{\alpha}{2} \times \mathrm{IS}_{\alpha_k}(F, y) \right\},$$

where m is the predictive median.

This approximates the CRPS and generalizes the AE.



The weighted interval score

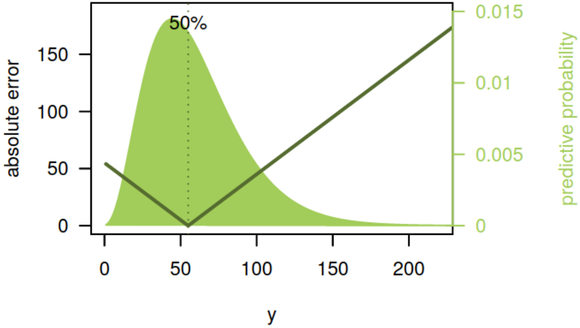
Bracher, Ray, Gneiting, Reich (2021)

To assess prediction intervals at levels $(1-\alpha_0, \ldots, 1-\alpha_K)$ simultaneously we can use the **weighted interval score**:

$$\mathrm{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K+1/2} \times \left\{ \frac{1}{2}|y-m| \; + \; \sum_{k=0}^{K} \frac{\alpha}{2} \times \mathrm{IS}_{\alpha_k}(F, y) \right\},$$

where m is the predictive median.

This approximates the CRPS and generalizes the AE.



The weighted interval score

Bracher, Ray, Gneiting, Reich (2021)

To assess prediction intervals at levels $(1 - \alpha_0, \dots, 1 - \alpha_K)$ simultaneously we can use the **weighted interval score**:

$$\text{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{K + 1/2} \times \left\{ \frac{1}{2}|y - m| \; + \; \sum_{k=0}^{K} \frac{\alpha}{2} \times \text{IS}_{\alpha_k}(F, y) \right\},$$

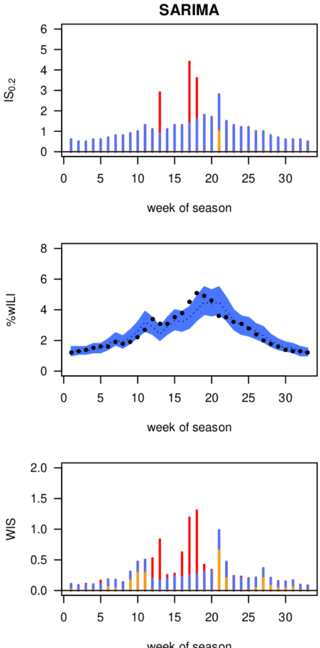
where m is the predictive median.

Equivalent to "pinball loss" known eg from quantile regression:

$$\text{WIS}_{\alpha_{0:K}}(F, y) = \frac{1}{2K + 1} \times \sum_{i=1}^{2K+1} 2 \times \{\mathbb{1}(y \leq q_{\tau_i}) - \tau_i\} \times (q_{\tau_i} - y),$$

where $q_{\tau_i}, i = 1, \dots, 2K$ are the $2K + 2$ available quantiles and τ_i are the respective levels.

Example (using FluSight data)



Application in practice

- Proper scores can be averaged across weeks/locations/targets.
- Typically complemented with measures of quality of point forecasts (note: WIS can be compared to absolute errors of deterministic forecasts.)
- Calibration can be assessed separately via coverage probabilities and PIT histograms.