Forecast evaluation II

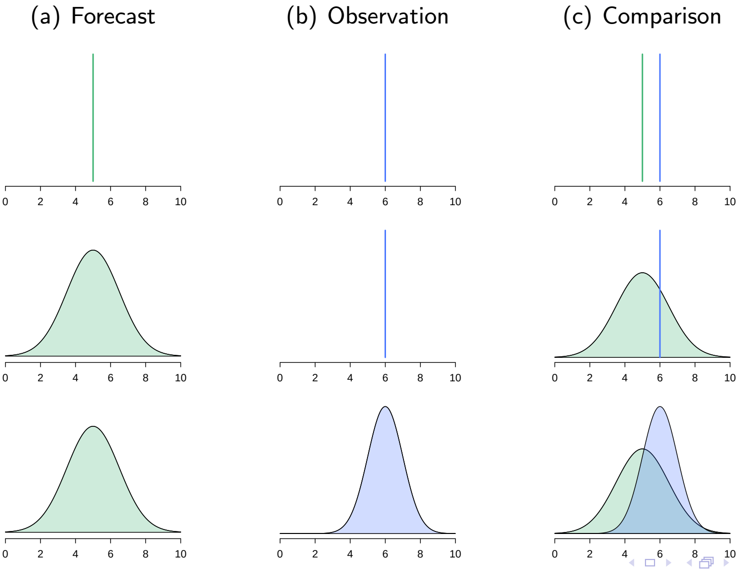
Thordis L. Thorarinsdottir

Norwegian Computing Center, Oslo, Norway

`www.nr.no/~thordis`

CUSO winter school 2021

Forecast and observation classes



What is a good probabilistic forecast?

> *There should be consistency between the forecaster's judgement and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user.*
>
> Murphy (WAF, 1993)

> *We propose a diagnostic approach to the evaluation of predictive performance that is based on the paradigm of maximizing the sharpness of the predictive distribution subject to calibration.*
>
> Gneiting, Balabdaoui and Raftery (JRSSB, 2007)

Outline for this lecture

Assume we have a prediction $p \in \mathcal{P}$ and an observation $o \in \mathcal{O}$ where we wish to measure the skill of the prediction by applying a function

$$s : \mathcal{P} \times \mathcal{O} \longrightarrow \mathbb{R}$$

with a lower function value indicating a better skill.

What are good theoretical properties for s?

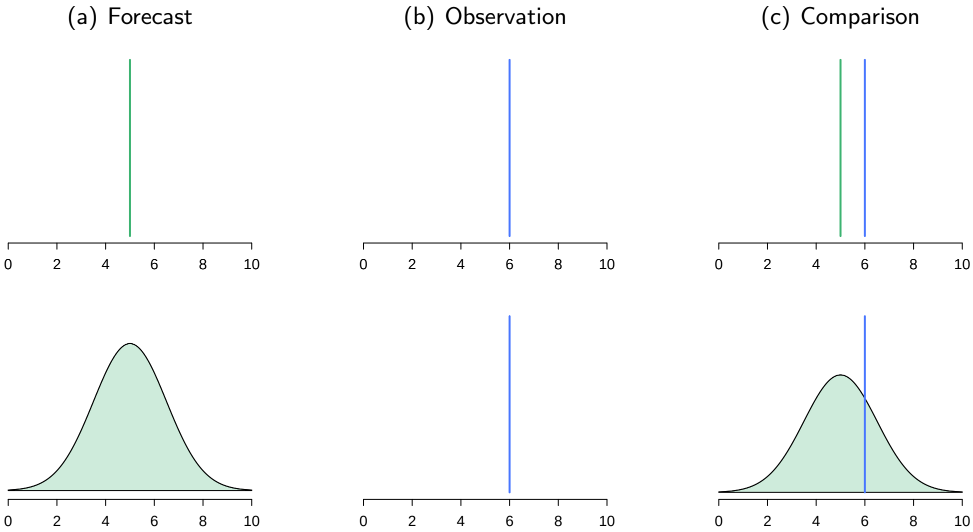
General framework without any formulas...

- Assume G is Nature's distribution of some event y and denote our forecast for y by F.
- For forecast evaluation, we should use performance metrics that follow the principle

  in the long run, we will obtain the optimal performance for $F = G$

  where "in the long run" means "over very many pairs (y_i, F)".
- Note that this is an abstract quality which is checked theoretically for general classes of distributions F and G.
- If we agree that this is a sensible framework, we can then, in many cases, just pick a (few) such metric(s) and perform our forecast evaluation using those.

Deterministic vs. probabilistic forecasts



Which deterministic value to choose?

In the absence of explicit guideance, forecasters may report different distributional features as their point predictions.

Engelberg, Manski and Williams (JBES, 2009)

A **decision-theoretic approach** provides a unifying framework for the evaluation of both probabilistic and deterministic forecasts.

Scoring functions apply to deterministic forecasts

The forecast x is evaluated against the observation y using **scoring functions** such as

Squared Error (SE)	$\mathrm{S}(x, y) = (x - y)^2$
Absolute Error (AE)	$\mathrm{S}(x, y) = \lvert x - y \rvert$

Generally, we assume that

$$\mathrm{S} : \mathbb{R} \times \mathbb{R} \to [0, \infty) \quad \text{or} \quad \mathrm{S} : (0, \infty) \times (0, \infty) \to [0, \infty),$$

with the regularity conditions

(S0) $\mathrm{S}(x, y) \geq 0$ with equality if $x = y$
(S1) $\mathrm{S}(x, y)$ is continuous in x
(S2) The partial derivative $\partial_x \mathrm{S}(x, y)$ exists and is continuous if $y \neq x$

Average scores facilitate comparison across methods

Assume various forecasting methods $m = 1, \ldots, M$ compete

They issue point forecasts x_{mn} with observed values y_n, at a finite set of times, locations or instances $n = 1, \ldots, N$

The methods are assessed and ranked by the mean score

$$\bar{\mathrm{S}}_N^m = \frac{1}{N} \sum_{n=1}^{N} \mathrm{S}(x_{mn}, y_n) \qquad \text{for } m = 1, \ldots, M.$$

Testing equal predictive performance: Diebold-Mariano test

If the forecast cases are indpendent, a test of equal predictive performance can be based on the statistic

$$t_N = \sqrt{N}\frac{\bar{\mathrm{S}}_N^{m_1} - \bar{\mathrm{S}}_N^{m_2}}{\hat{\sigma}_N},$$

where

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\mathrm{S}(x_{m_1 n}, y_n) - \mathrm{S}(x_{m_2 n}, y_n)\right)^2.$$

For correlated forecast errors, the variance estimate needs to be adjusted (Diebold and Mariano, JBES, 1995).

Testing equal predictive performance: Permutation test

Alternatively, m_1 and m_2 can be compared using the statistic

$$s_N = \frac{1}{N}\sum_{n=1}^{N}\left(\mathrm{S}(x_{m_1 n}, y_n) - \mathrm{S}(x_{m_2 n}, y_n)\right).$$

The permuation test is based on resampling copies of s_N with random number of labels swapped. Under the null hypothesis, m_1 and m_2 perform equally well and the permutations have the same limiting distributions as s_N for $N \to \infty$. An asymptotic test is obtained by considering the rank of s_N within the permutations (Good, 2013).

Bayes predictors should be used for probilistic forecasts

For a probabilistic forecast F, **decision theory** tells us that if the scoring function S is given, we should issue the **Bayes predictor**,

$$\hat{x} = \arg\min_x \mathbb{E}_F \left[\mathrm{S}(x, Y)\right]$$

as the point forecast, where the expectation is with respect to F.

Squared Error (SE)	$\mathrm{S}(x, y) = (x - y)^2$	$\hat{x} = \mathrm{mean}(F)$
Absolute Error (AE)	$\mathrm{S}(x, y) = \lvert x - y \rvert$	$\hat{x} = \mathrm{median}(F)$

Consistency and elicitability

Conversly, assume we only have one functional T of F which we know to be, say, the mean value.

Here, we may apply any scoring function that is **consistent** for the functional T, in the sense that

$$\mathbb{E}_F\left[\mathrm{S}(\mathrm{T}(F), Y)\right] \leq \mathbb{E}_F\left[\mathrm{S}(x, Y)\right]$$

for all x.

A functional is **elicitable** if there exists a scoring function that is **strictly consistent** for it, in the sense that equality holds if, and only if, $x = \mathrm{T}(F)$.

The variance and the mode are not elicitable (Gneiting, JASA, 2011; Heinrich, B, 2014).

Probabilistic forecasts should generally be evaluated using proper scoring rules

A consistent scoring function is a special case of a proper scoring rule for probabilistic forecasts

Definition

If $\mathcal{F}$ denotes a class of probabilistic forecasts on $\mathbb{R}$, a **proper scoring rule** is any function

$$\mathrm{R} : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$$

such that

$$\mathrm{R}(G, G) := \mathbb{E}_G \, \mathrm{R}(G, Y) \leq \mathbb{E}_G \, \mathrm{R}(F, Y) =: \mathrm{R}(F, G)$$

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

Proper scoring rules prevent hedging

Is it possible to hedge the following scoring rule?

$$R^*(F, y) = \frac{(\text{mean}(F) - y)^2}{\text{var}(F)}$$

Proper scoring rules prevent hedging

The proper **Dawid-Sebastiani score** is given by

$$R(F, y) = \log(\mathrm{var}(F)) + \frac{(\mathrm{mean}(F) - y)^2}{\mathrm{var}(F)}$$

Consistent scoring functions are proper scoring rules

Any consistent scoring function induces a proper scoring rule: if the scoring function

$$\mathrm{S} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$$

is consistent for the functional T, the relationship

$$\mathrm{R} : \mathcal{F} \times \mathbb{R} \longrightarrow [0, \infty), \quad (F, y) \longmapsto \mathrm{R}(F, y) = \mathrm{S}(\mathrm{T}(F), y)$$

defines a proper scoring rule.

Squared Error (SE)	$\mathrm{R}(F, y) = (\mathrm{mean}(F) - y)^2$
Absolute Error (AE)	$\mathrm{R}(F, y) = \lvert\mathrm{median}(F) - y\rvert$

The class of proper scoring rules is large

A commonly used score is the logarithmic or ignorance score,

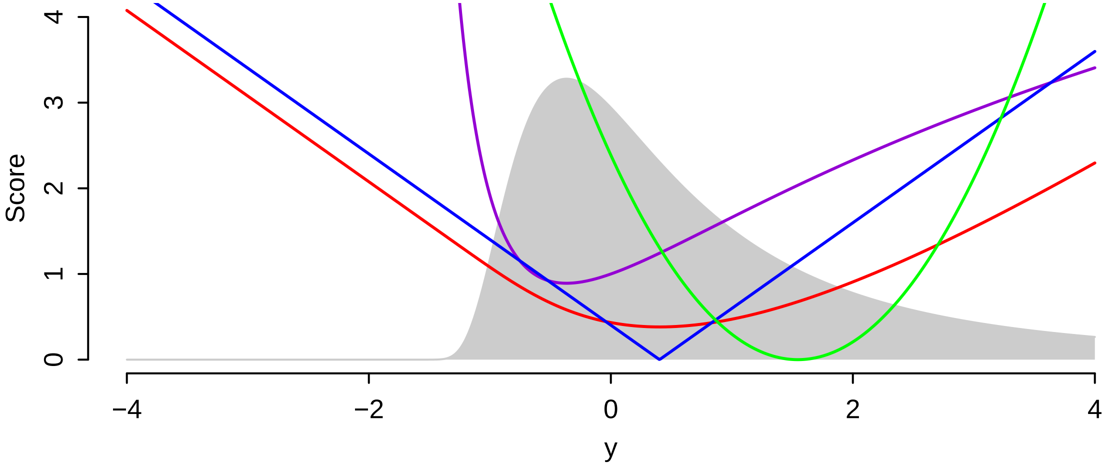
$$\mathrm{R}(F, y) = -\log(f(y)),$$

The continuous ranked probability score (CRPS) is given by

$$\begin{aligned}
\mathrm{R}(F, y) &= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F\mathbb{E}_F|X - X'| \\
&= \int [F(x) - \mathbb{1}\{x \geq y\}]^2 dx \\
&= \int_0^1 \left(F^{-1}(\tau) - y\right)\left(\mathbb{1}\{y \leq F^{-1}(\tau)\} - \tau\right)\mathrm{d}\tau,
\end{aligned}$$

where the integrands are the Brier score and the quantile score, respectively (Gneiting and Raftery, JASA, 2007).

The different scores behave somewhat differently



SE AE CRPS IGN

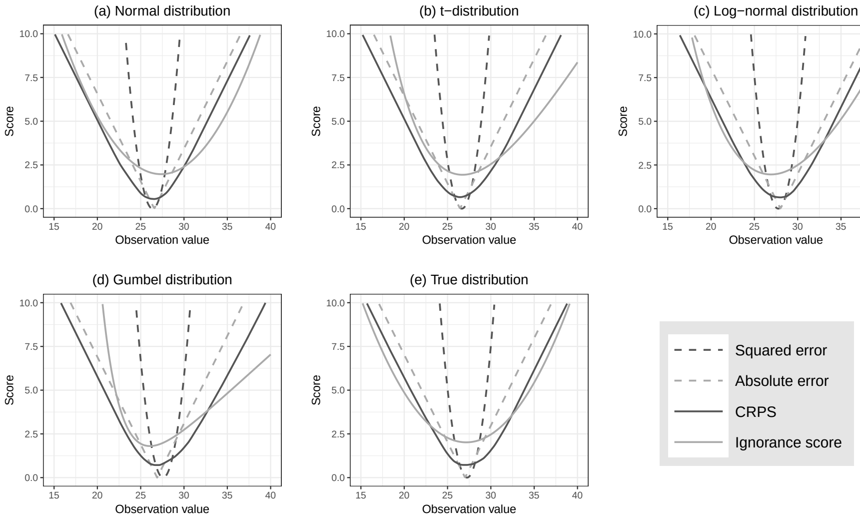
Back to our example from yesterday

Distribution	$F(Y)$	$\mathbb{E}(Y)$	$\mathrm{Var}(Y)$
Normal	$\mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(25, 1)$	$\sigma^2 = 9$
Gumbel	$\mathrm{G}(\mu, \sigma)$	$\mu + \sigma \cdot \gamma \sim \mathcal{N}(25, 1)$	$\frac{\pi^2}{6}\sigma^2 = \frac{3\pi^2}{2}$

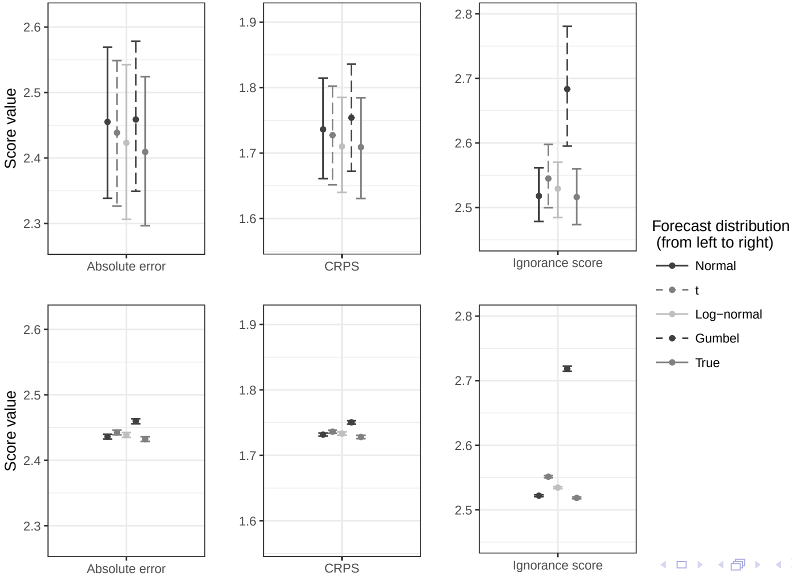
- Competing forecasts: Normal, non-central t, log-normal, Gumbel
- Each forecast is estimated based on 300 i.i.d. observations using methods of moments
- Case 1: 1 000 forecast-observation pairs
- Case 2: 1 000 000 forecast-observation pairs

(Thorarinsdottir and Schuhen, 2018)

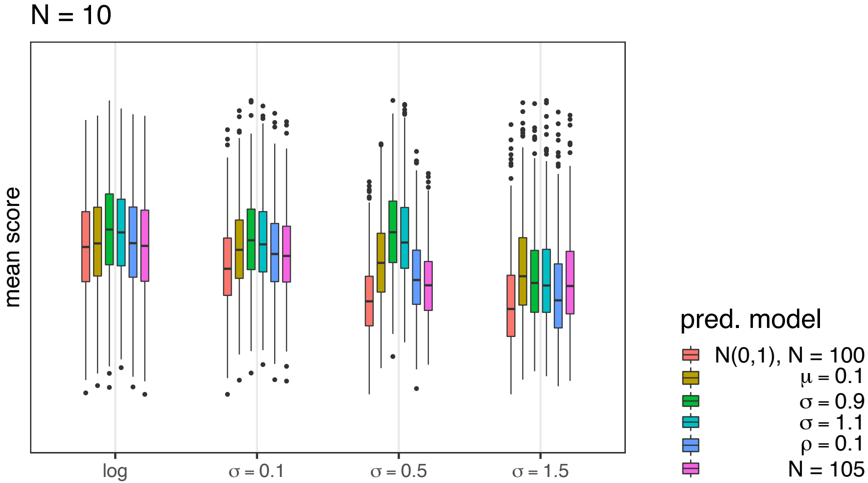
Score behavior for normal truth



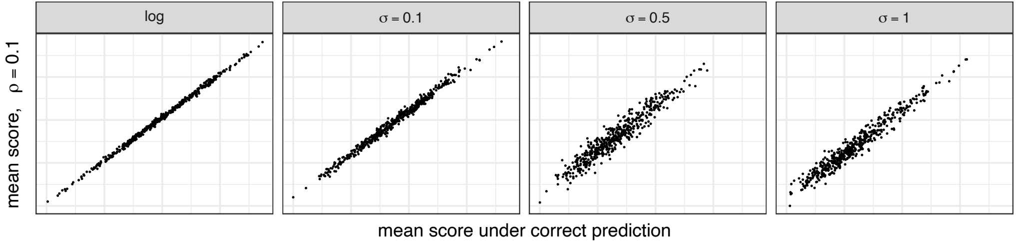
Scores for normal truth



Uncertainty in scores vs. distribution of scores



Uncertainty in scores vs. distribution of scores



The CRPS is appealing but not convenient to calculate: `scoringRules` to the rescue!

Dist. on	Dist. on $_{>0}$	Dist. on intervals	Discrete dist.
Gaussian	Exponential	Generalized extreme value	Poisson
t	Gamma	Generalized Pareto	Neg. binomial
Logistic	Log-Gaussian	Trunc. Gaussian	
Laplace	Log-logistic	Trunc. t	
Two-piece Gaussian	Log-Laplace	Trunc. logistic	
Two-piece exponential		Trunc. exponential	
Mixture of Gaussians		Uniform	
		Beta	

Truncated families can be defined with or without a point mass at the support boundaries.

What to do if the predictive distribution is not available in closed form?

Assume our predictive distribution is the posterior predictive distribution of a Bayesian forecasting model,

$$F(y) = \int F_c(y|\theta)\mathrm{d}P_{\text{post}}(\theta).$$

We then have various options to estimate F:

- **Mixture-of-parameters**: $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} F_c(y|\theta_i)$ for posterior sample $\{\theta_i\}_{i=1}^{n}$
- **Empirical CDF**: $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y \geq Y_i\}$ for $Y_i \sim F_c(\cdot|\theta_i)$
- **Kernel density estimator**: $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{y-Y_i}{h_n}\right)$ with bandwidth h_n
- **Gaussian approximation**: $\hat{F}(y) = \Phi\left(\frac{y-\hat{\mu}}{\hat{\sigma}}\right)$ for posterior mean $\hat{\mu}$ and sd $\hat{\sigma}$

(Krüger et al., ISR, 2020)

How do these approximations compare?

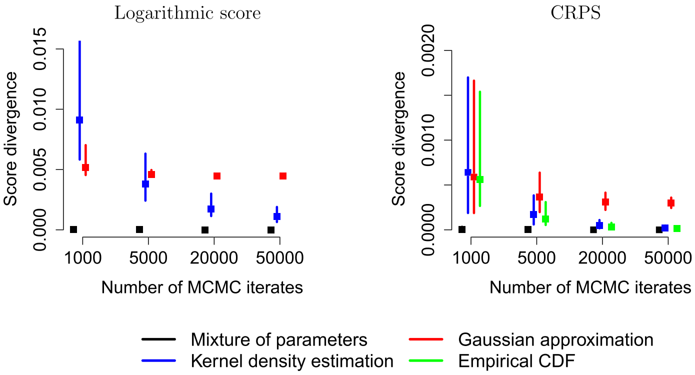
Simulation study with

$$F_c(y|\theta) = \Phi\left(\frac{y}{\theta}\right)$$

and

$$F(y) = \mathrm{T}(y|0, \alpha_1, \alpha_2).$$

That is, F is the CDF of a variable Z where $Z/\sqrt{\alpha_1}$ follows a t distribution with α_2 degrees of freedom.



Conclusions

- The performance measure used in forecast evaluation may influence the results of a comparative study and should be selected with care.
- Different verification measures focus on different aspects of the model output; it is thus useful to apply multiple complementary measures.

References

1. A Murphy (1993): What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8, 281-293.
2. T Gneiting, F Balabdaoui and A Raftery (2007): Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, 69, 243-268.
3. J Engelberg, C F Manski and J Williams (2009): Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, 27, 30-41.
4. F Diebold and R Mariano (1995): Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253-263.
5. P Good (2013): *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, New York: Springer-Verlag.
6. T Gneiting (2011): Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746-762.
7. C Heinrich (2014): The mode functional is not elicitable. *Biometrika*, 101, 245-251.
8. T Gneiting and A Raftery (2007): Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359-378.
9. T Thorarinsdottir and N Schuhen (2018): Verification: Assessment of calibration and accuracy. In *Statistical Postprocessing of Ensemble Forecasts*, pp. 155-186.
10. F Krüger, S Lerch, T Thorarinsdottir and T Gneiting (2020): Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review*, in press.