Statistical evaluation of forecasts

Chris Ferro

University of Exeter, UK

Workshop on understanding, modeling and predicting weather and climate extremes (5–7 October 2015, Oslo)

1. Evaluating probability forecasts

Probability forecasts

A probability forecast is a probability distribution representing our uncertainty about a predictand.

The predictand can be multi-dimensional, so can be a spatial field, a climatological distribution, a time-series with a trend or cycle, a description of the evolution of some phenomenon, etc.

What makes a good forecast?

Sequence of cases indistinguishable at initialisation...

G = distribution of outcomes produced by Nature

F = forecast probability distribution

Act to minimize expected loss calculated from forecast

Long-run loss is minimized if F = G (Diebold et al. 1998)

Whatever our loss function, the best forecast matches the distribution of outcomes produced by Nature.

How do we measure performance?

We get only one outcome from G, but we can measure performance to reward optimal forecasts in long-run.

Evaluate a **score** s(F,y) for the forecast F and outcome y.

Only **proper** scores favour optimal forecasts in long-run.

The long-run score $E_{y\sim G}[s(F,y)]$ is optimized when F = G if and only if s is proper.

Proper scores can favour forecasts that match only some features of optimal forecasts, e.g. to evaluate only mean.

Example: proper/improper scores

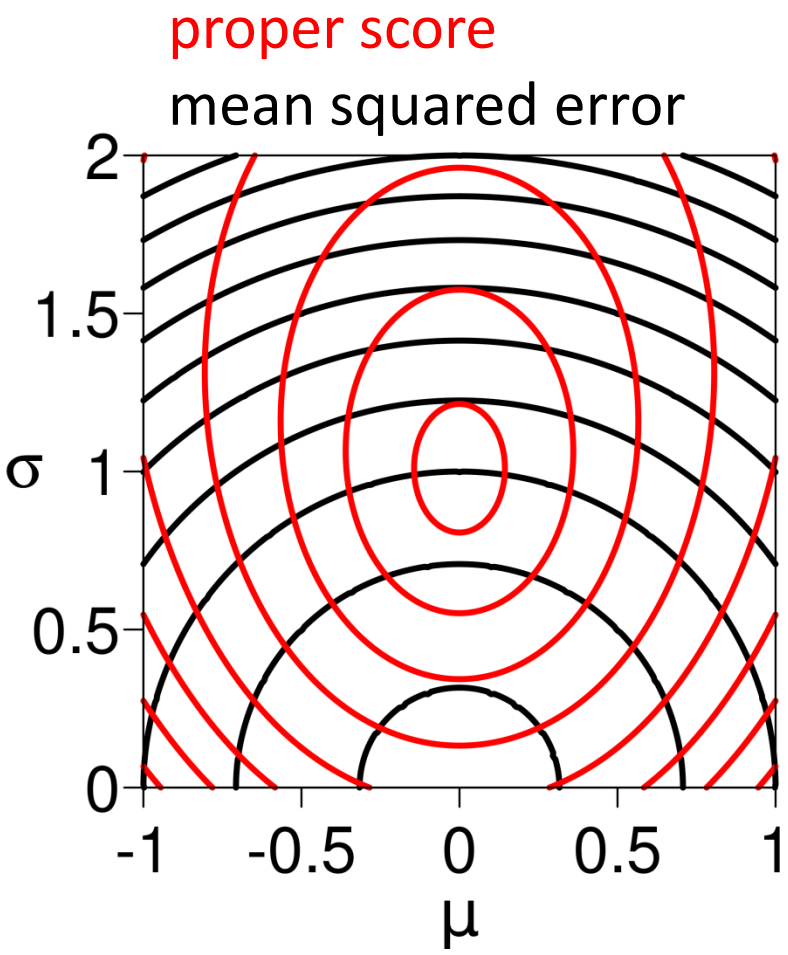
Nature: N(0,1)

Forecast: $N(\mu,\sigma^2)$

Contours of long-run score.

Proper score favours ideal forecast ($\mu = 0$, $\sigma = 1$).

Mean squared error favours unbiased ($\mu = 0$) but under-dispersed ($\sigma = 0$) forecast.



Example: proper/improper scores

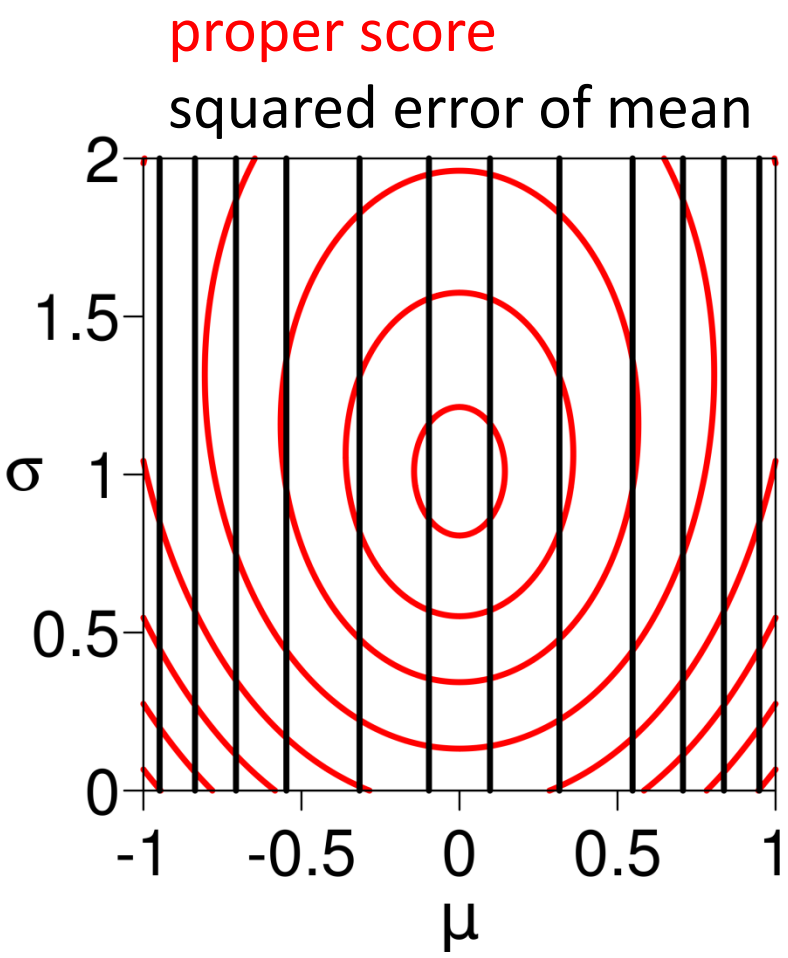
Nature: N(0,1)

Forecast: N(μ,σ^2)

Contours of long-run score.

Proper score favours ideal forecast ($\mu = 0$, $\sigma = 1$).

Squared error of the mean favours unbiased forecasts ($\mu = 0$) and ignores spread.



Probability forecasts: summary

Measures should favour optimal forecasts in long-run.

Only proper scores do this.

Proper scores can be chosen to evaluate specific features (e.g. mean, variance) of forecasts.

Multiple scores are needed to identify which features are in error.

2. Evaluating models

Evaluating models

Should we favour models that

a) better simulate Nature, or

b) support better forecasts?

We consider evaluating how well a model simulates Nature based on an initial-condition ensemble. ('Model' includes initialisation and ensemble generation.)

Ensemble needn't be a 'forecast' for forecast evaluation to be relevant. Just need a commensurate observation.

What makes a good model?

Sequence of cases indistinguishable at initialisation...

G = distribution of outcomes produced by Nature

F = distribution from which the ensembles are samples

If F ≠ G, the model is behaving differently to Nature.

The best model samples ensembles from a distribution that matches the distribution produced by Nature.

Evaluate the ensemble distribution

Ensemble 1: all members very close to outcome, unbiased on average, outcome often outside ensemble

Ensemble 2: some members quite far from outcome, biased on average, outcome usually inside ensemble

Evaluating the full ensemble distribution may favour #2.

We get only one outcome, so we don't know what range of behaviour is realistic. Being 'close' is not enough.

Evaluate only the ensemble mean?

Too demanding to require well dispersed ensembles?

Evaluating only the mean says little about the realism of ensemble members (especially in multiple dimensions?).

Poor ensemble spread etc. indicates not only under-/ over-confidence, but that the model differs from Nature.

How do we measure performance?

Evaluate a score s(x,y) for ensemble x and outcome y.

Only **fair** scores favour optimal ensembles in long-run.

The long-run score $E_{x\sim F,y\sim G}[s(x,y)]$ is optimized when F = G if and only if s is fair (Ferro 2014).

Fair scores effectively evaluate the (imperfectly known) distribution from which the ensembles are samples.

Fair scores can favour ensembles that match only some features of optimal ensembles, e.g. to evaluate mean.

Model ensembles: summary

Ensemble distribution, not only its mean, is important.

Measures should favour optimal ensembles in long-run.

Only fair scores do this.

Fair scores can be chosen to evaluate specific features (e.g. mean, variance) of ensembles.

Multiple scores are needed to identify which features are in error.

3. Extremes and understanding

Evaluating extremes

Evaluate fair scores only when the outcome is extreme?

This will summarize the performance in those cases, but will not favour optimal ensembles. (Forecaster's Dilemma – S. Lerch)

Alternative: use weighted scores to require ensemble distributions to match Nature better at certain possible values of the predictand (Gneiting and Ranjan 2011).

Relevance of scores

Only fair scores should be used to *rank* models (otherwise perfect models will be penalized).

But scores hide key information, e.g. direction of bias.

Other methods are needed to understand performance.

See Marion's talk next...

Summary and questions

Prefer models to simulate, or help forecast, Nature?

Entire ensemble distribution reflects ability to simulate.

Fair scores should be used if models are to be ranked.

Are existing scores sensitive to process differences?

What ensemble sizes needed to detect differences?

How handle observation error/lack of observations?

What other forecast evaluation methods are useful?

References

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39, 863–882

Ferro CAT (2014) Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140, 1917–1923

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

Characterization: binary case

Let y = 1 if an event occurs, and let y = 0 otherwise.

Let $s_{i,y}$ be the (finite) score when i of m ensemble members forecast the event and reality is y.

The (negatively oriented) score is fair if

$$(m - i)(s_{i+1,0} - s_{i,0}) = i(s_{i-1,1} - s_{i,1})$$

for i = 0, 1, ..., m and $s_{i+1,0} \geq s_{i,0}$ for i = 0, 1, ..., m – 1.

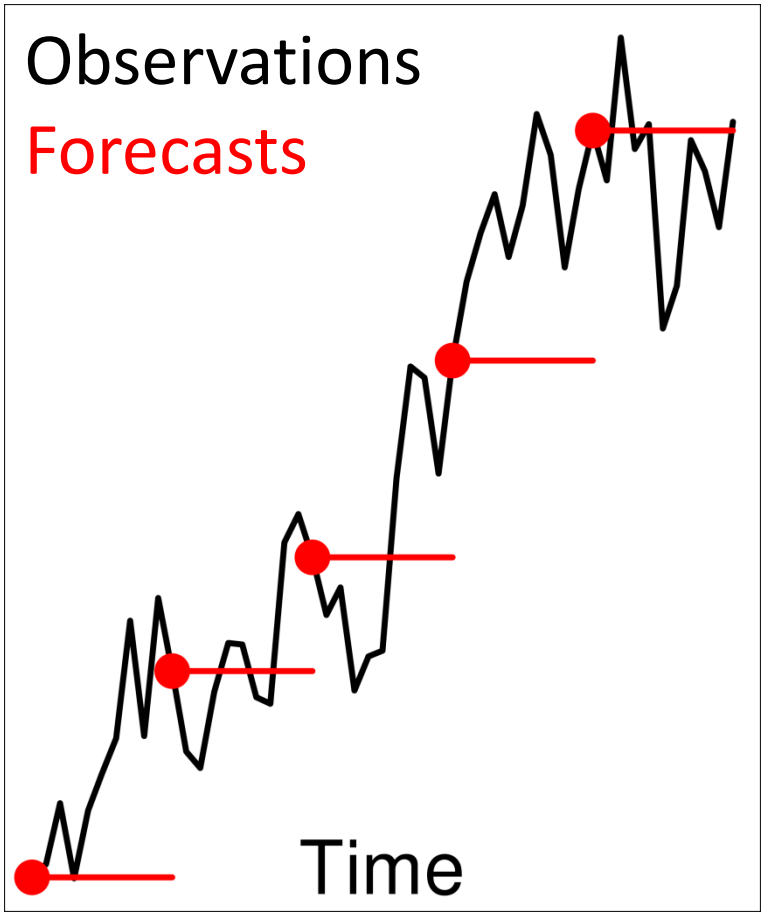
Ferro (2014, Q. J. Roy. Met. Soc.)

Measuring performance: scores

Calculate a score for each forecast and then average over the forecasts.

Other types of measure (e.g. correlation) are prone to spurious inflation due to trends in the data.

Example: naive forecasts achieve correlation 0.95.



Probability forecasts: attributes

Proper scoring rules reward **calibration** and **sharpness**.

Calibration: reality looks like random draws from the forecast distributions

Sharpness: forecast distributions are concentrated

The long-run (proper) score decomposes as

$$S(F,G) = S(G,G) + C(F,G)$$

where S(G,G) is concave and measures sharpness, and C(F,G) is non-negative and measures calibration.