Thorpex 2004

Scoring Probabilistic Forecasts

Jochen Bröcker

Devin Kilminster, Liam Clarke and Leonard Smith

Centre for the Analysis of Time Series
London School of Economics
Houghton Street
London WC2A 2AE

United Kingdom

`cats.lse.ac.uk`

Outline

- Why are probabilistic forecasts important?
- What is a skill score?
- What is a *proper* skill score?
- The importance of being proper
- Examples

The Forecast Problem

Problem: We want to forecast an observable T_n (e.g. temperature), where n is the time.

- We issue probabilistic forecasts: $p_n(T) = \{$Pobability of $T_n = T\}$

- Usually p_n is built upon some related side information (past observations, weather model simulations)

- Does *not* mean $p_n(t)$ the probability of T_n *given* that side information

Why Using Probabilistic Forecasts

- End users don't want to *know* T_n, they want to base *decisions* on T_n (and none of them care about models' 500mB height)

- To take reasonable action, the *risk* of taking that action must be factored into that decision

- To do that, information about the *uncertainty* of T_n must be known

How Do We Evaluate Probabilistic Forecasts?

We need a *general skill score*, that takes into account the probabilistic character of the forecasts and that is relevant to many different users (incl. model developers, meteorologists).

Skill Scores

- A skill score is a function $\mathcal{S}(p, t)$
- The empirical skill is a sample mean:

$$S = \frac{1}{N}\sum_n \mathcal{S}(\underbrace{p_n}_{\text{Our forecast}}, \underbrace{T_n}_{\text{Reality}})$$

What should skill scores actually measure?

Properties A Probabilistic Forecast Should Have

A good probabilistic forecast should have:

- Reliability – Looking at those days where a probability $p_n = r$ of rain is forecasted, a fraction r of them should have rain
- Sharpness – High probability is issued to events that acctually happen to occur

Skill scores should take this into account, since we *believe* that a probabilistic forecast having these properties is good for a *multitude* of specific problems.

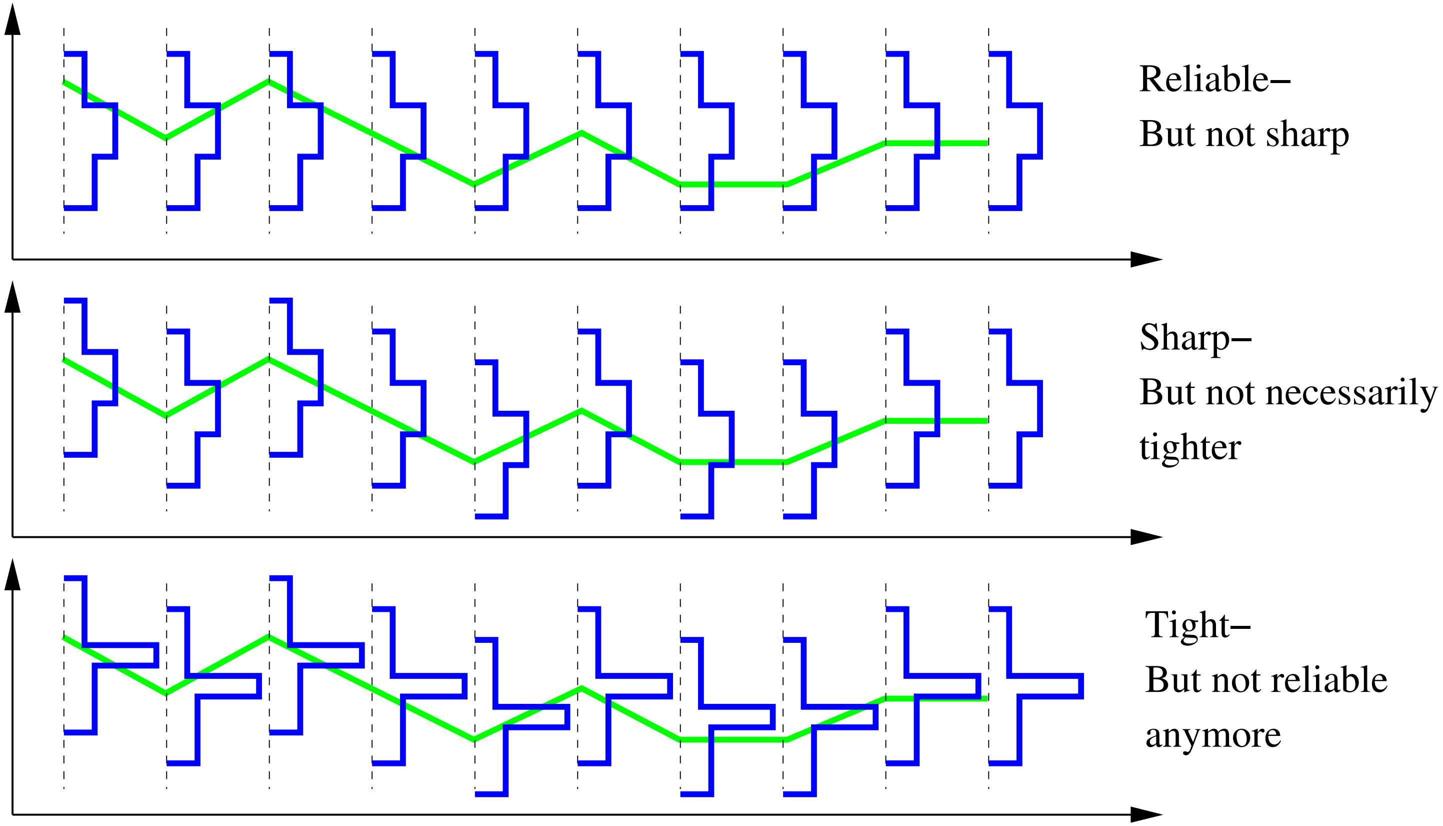
Reliability

- Take a fixed number r
- Count the instances where the probability of an event is forecasted to be π, i.e. $p_n = r$
- The event should actually occur at a fraction r of these instances.

Other equivalent formulations:

- $P(T_n = T|p_n) = p_n(T)$
- $p_n(T)$ can be written as a *conditional probability density*

Sharpness



This is an issue only when p_n actually depends on n

Proper Scores

Propriety is the key property for a skill score

- Assume p_n is our "best knowledge" probabilistic forecast

- Then, to the best of our knowledge, *our* forecast p_n has the skill

$$\mathcal{S}_p = \int \mathcal{S}(p_n, t) p_n(t) \mathrm{d}t$$

- To the best of our knowledge, *another* forecast q_n has the skill

$$\mathcal{S}_q = \int \mathcal{S}(q_n, t) p_n(t) \mathrm{d}t$$

- Believing p_n is right, we want p_n to have a better skill than q_n, otherwise we would not issue p_n

Proper Scores

Propriety is a property of the Skill Score alone, what the actual truth is doesn't matter

Proper Scores and Sharpness/Reliability

Two statements for proper scores:

If, given the available information, a reliable forecasts exists, it would yield a maximum score

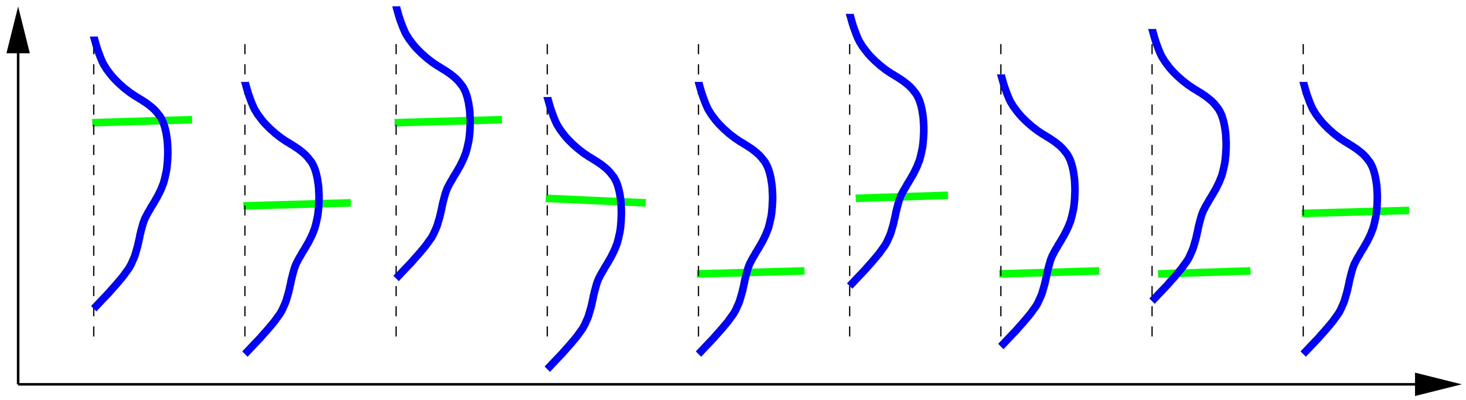
Of two equally reliable forecasts, the sharper one would score higher

Local Scores

Local Skill Scores are concerned only with verifications, that means:

The forecast $p_n(T)$ is scored only on what happened – the *verification*. How the forecast looks like at other points does not matter.

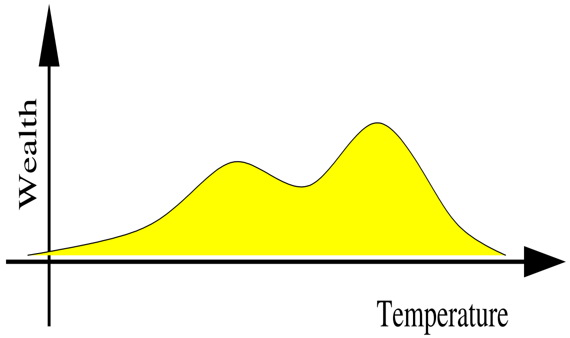
In other words, $\mathcal{S}(p_n, T) = \mathcal{S}(p_n(T))$.



An Example: Weather Roulette

Bet on temperature at London Heathrow. Objective: Maximize the expected return rate

- Strategy: Distribute your wealth



Distribution of wealth as a function $\alpha(t)$ of temperature

- Reasonable strategy (if odds are fair and you assume $p(t)$ is right): $\alpha_p(t) = p(t)$

An Example: Weather Roulette

- Actual wealth grow rate: $S = \frac{1}{N} \sum_n \log p(T_n) +$ something that depends on the odds only

Can this be used as a skill score?

The Ignorance

- The *Ignorance Skill Score* is

$$I(p) = -\log(p_n(T_n))$$

- The Ignorance is proper, local and smooth

- The Ignorance is the *only* proper, local and smooth score for continuous forecasts (Good 1952, Gneiting & Raftery 2004)

The Brier Skill Score

- The *Brier Score* considers binary events
- $$\mathcal{S} = (T_n - p_n(1))^2$$
- The Brier score is proper for binary events
- Taking any other function than p^2 here is *improper*

About Other Scores

- The *Linear Score* or p *Score* $\mathcal{S}(p_n, t) = p_n(t)$ is *improper*
- The RMS error depends only on some moments and therefore is *not* strictly proper
- Many proper *nonlocal* Scores have been suggested and used (see talk by Zoltan Toth about CRPS)

Take Home Points (And Questions)

- End users need probabilistic forecasts to make better decisions
- We need skill scores that measure desirable properties of probabilistic forecasts
- We need to use proper scores, since improper scores give misleading answers – we would reject even the optimal forecast
- There are only a handful of essentially different proper skill scores (see www.dime.lse.ac.uk)

Take Home Points (And Questions)

Questions

- Are there good reasons to use nonlocal scores?
- What is the actual connection between general skill scores and end users specific cost functions?
- Is weather really a stochastic process? If not, there will never be fully reliable forecasts
- Do probabili*stic* forecasts need to be probabili*ty* forecasts, and if not, what are the neccessary amendments to the concept of skill?