Discussion of "Forecasters's Dilemma: Extreme Events and Forecast Evaluation"

By Sebastian Lerch, Thordis L. Thorarinsdottir, Francesco Ravazzolo and Tilman Gneiting

Anthony Garratt

ECB Forecasting Workshop 3-4 June 2016

What Does the Paper Do?

- Describes, discusses and comments on the "forecasters dilemma" in the context of extreme events: using theoretical considerations, simulations and a case study.
- **Forecasters dilemma**: situation where either deliberately or unknowingly you evaluate forecasts (of extreme events) based on misguided inferential procedures.
- **Why?** Because you may be deemed as someone who gets it right when other don't and be "declared wise" as opposed to being regarded as "foolish" despite careful consideration which arguably would serve you well in more normal times - honesty of a forecaster (benchmark choices!).

What Does the Paper Do?

- Illustrated by simulation: data from a normal distribution, with a given mean and variance, construct (10,000) point and probability forecasts from:
  - a *perfect* model -with same mean and variance
  - an *unconditional* model with mean zero and variance of 1 and
  - an *extreme* model with biased mean and same variance
- For probabilistic forecasts, using *unweighted* (and proper) CRPS and LogS, the *perfect* model is preferred when using the whole sample but mistakenly the *extreme* model is preferred if the CRPS and LogS are conditioned on the observed extreme outcomes - illustrating the dilemma and emphasising the incompatibility with the theoretical assumptions of established forecast evaluation methods.
- For probabilistic forecasts, *proper weighted scoring rules* have been proposed as decision theoretically justifiable alternatives, with an emphasis on extreme events. If these are applied then the *perfect* model is again preferred.

What Does the Paper Do?

- A key concept in this context is whether the scoring rule is *proper*
  - i.e. has the feature that the highest expected reward is obtained by reporting the true probability distribution
  - which encourages the forecaster to be honest so as to maximise expected reward.(see Gneiting, 2007, JASA)
- Theoretical arguments suggest that conditioning on the observed outcomes - which gives rise to the forecasters dilemma - renders a *proper* scoring rule *improper*.
- Indicator weight functions used in the *proper weighted scoring functions* (Diks *et al* 2011, Gneiting and Ranjan, 2011) do not correspond to restricting the evaluation to observed observations which are deemed extreme– instead of excluding them it assigns a zero weight , so the ranking of competing forecasts is retained..

Simulations: Signal to Noise

- Forecasters dilemma is more acute in systems where the signal to noise ratios are low i.e. at high values of σ
  - perfect forecaster knowing μ is less valuable; increase in σ allows for a better match between the probabilistic forecasts of extreme event forecaster and true (more variable) distribution
  - The ranking of unweighted and restricted CRPS and LogS same except at high values of σ - suggesting forecasters dilemma not so pronounced
- **Comment 1:** Could this be linked to literature like Jerker Denrell (2013) Havard Business Review, "Experts that Beat the Odds Are Probably Just Lucky"?
- Conflate tasks that require refined skills (forecasting for all events) which ones that don't (forecasting extreme events stocks)
  - where there is inherent variability in the activity, like finance (a case study?) then unskilled people can strike it lucky, but where skill is involved it is much harder to succeed.

Revised Simulations using DM test

- Compares two forecast distributions, neither of which corresponds to the true sampling distribution (where $N = 100$).
- Three distributions: (i) standard normal Φ with density ϕ (ii) heavy right hand tailed distribution H, with a normal left tail and density $h(x)$ and (iii) an equally weighted mixture F of ϕ and $h(x)$
- Two scenarios are then examined: (i) *Scenario A:* data sampled from standard normal Φ then compare forecasts from F and H (ii) *Scenario B*: data sampled from H then compare forecasts from F and Φ
  - in both F is a weighted mixture of the true distribution and mis-specified competitor, which suggests we might expect F to be preferred (might mixtures be quite exotic?)
- Plots of rejection rates, of DM tests using scores, in favour of F and H, as a function of threshold r in the indicator function used for computing the proper unweighted/weighted scoring rules (CRPS and LogS).

Results of Revised Simulations

- *Scenario A*: frequency of desired rejections in favour of F increases with larger thresholds for *proper weighted scores*, suggesting improved discrimination properties at higher thresholds (note CL decreases)
- *Scenario B*: but for proper weighted scoring rules, the frequency of desired rejections in favour of F decays to zero with increasing thresholds values, whereas the frequency of undesired rejections in favour of Φ rises for larger threshold values.
- **Comment 2**: Does this suggest that the true underlying distribution therefore make a difference? Where for large thresholds a heavy tailed DGP's gives rise to low power on the weighted scoring rules, which presumably is not what you would want?
- Proper weighted scoring rules do not dominate proper unweighted scoring rules
- **Comment 3:** Score values favour the unweighted LogS in A, but unweighted CRPS in B - could the paper make more of the relative performance of the CRPS and LogS?

Explanation of results

- Tail behaviour, if a threshold exceeds the maximum of a given sample (or only very few observations exceed it) then the scores do not (or barely) depend on the observations and are solely determined by the respective tail probabilities with the lighter tails (i.e. Φ) receiving a better score
  - so when the emphasis lies on a low-probability event region with no or few observations, a forecaster assigning a low probability to this region will be preferred
  - traditional unweighted scoring rules do not depend on thresholds and thus do not suffer this deficiency
- Leads to a loss of finite sample discrimination, so **Comment(s) 4**:
  - should there be a greater focus on the size of the threshold and the effect on the use of weighted versus unweighted?
  - should the simulations try different sample sizes?

General comments on simulations and case study

- Revised simulation does not involve any estimation and hence parameter uncertainty - which is clean
- **Comment 5:** Should a DGP with dependence be examined? The case study looks at a macro example and the use of VARs which builds in dependence but there is a disconnect between the simulations and the case study
  - DGP which includes some form of serially dependence, estimate a series of VAR models on the generated data
- **Comment(s) 6**: Case study:
  - can models with and without TVP/SV be related related to the signal to noise ratio point made in the earlier simulations as the forecast densities with SV are narrower
  - Given reduced variance is there less of an issue in terms of the forecaster dilemma for these TVP/SV models?
  - finance high frequency example would be of interest

General comments on simulations and case study

- **Comment 7:** Is model averaging interesting here?
  - weights that change according to recent forecast performance so the density might vary in sharpness according to the model which receives the weight.
  - so you try and have an expert forecasters perform well for extreme events as well as more normal times.
  - how do these tests perform with a combined density relative to single model?
- **Comment 8:** Given the focus is on extreme events how does all this relate proper scoring rules for quantiles and intervals (as in Gneiting and Raferty, JASA 2007)? ROC/hit rates and false alarm rates/economic evaluation.

What are the general lessons?

- Do not use observations on extreme events to alter evaluation period
- High signal to noise ratios make forecasters dilemma more likely
- Use proper scoring rules but where it is not clear whether to use unweighted or weighted scoring rules depending on the circumstances/underlying DGP