Skill score efficacy given sparse data

Trevor Maynard *

February 14, 2017

This experiment tests how well skill scores perform when the observations form a sparse data set. The (hypothetical) underlying is a unit normal distribution; three forecasts are produced named: Narrow, Perfect and Wide - each also being gaussian but with varying standard deviation. A given forecast is chosen if it has the best score over the sparse hypothetical observations. The experiment is repeated multiple times and the proportion of times the perfect forecast is chosen gives an objective measure of how well the skill score has peformed (i.e. if a skill score regularly leads to an incorrect forecast being chosen, then it has not performed well). The experiment is carried out for multiple skill scores to produce a ranking in this context. The Ignorance score performs best in the majority of situations.

*LSE, Lloyd's of London

1 Experiment design

In many real world problems the number of observations can be very low. The following experiment illustrates how well different scores perform when data is sparse and produces a ranking in that context.

Create observations Let $x_{k,j}$ be sampled from a unit normal distribution ($N(0,1)$). Define 'Observation Set k' as $O_k = \{x_{k,1}, ...x_{k,2^N}\}$, where N is an integer and $k \in \{1, 2, ...M\}$

Define forecasts Three forecasts distributions $P = \{p_{Narrow}, p_{Perfect}, p_{Wide}\}$ are tested:

1. $p_{Narrow} \sim N(0, \frac{1}{\sqrt{2}})$;
2. $p_{Perfect} \sim N(0, 1)$; and
3. $p_{Wide} \sim N(0, \sqrt{2})$.

Experiment algorithm The following experiment algorithm is used:

- For a given skill score $S(p, x)$
- For $k \in \{1, ...M\}$
  - Calculate the average score $\bar{S}_{p,k}$ for each of the three forecasts over the observations O_k. $\bar{S}_{p,k} = \frac{1}{2^N} \sum_{j=1}^{2^N} S(p, x_{k,j})$, where $p \in P$;
  - For $p \in P$ define $C_{p,k} = \begin{cases} 0 & \bar{S}_{p,k} \leq \bar{S}_{q,k} \text{ for some } q \neq p \\ 1 & \bar{S}_{p,k} > \bar{S}_{q,k} \ \forall \ q \neq p \end{cases}$
- Define $F_{p,S} = \frac{\sum_{k=1}^{M} C_{p,k}}{M}$, where S denotes the skill score.

In words, for each of the three forecasts the average score is calculated for a given observation set; the forecast which has the best average score is deemed to be 'chosen' by the score. This is repeated for M observation sets and the frequency F_p with which the score chooses each forecast p is calculated. Forecast $p_{Perfect}$ is correct and so scores with a high value of $F_{Perfect,S}$ have done well. F defines a ranking between scores: i.e. if $F_{Perfect,S_1} > F_{Perfect,S_2}$ then we say that score S_1 is better than S_2. The above description now allows experiment 2.3 to be defined:

Experiment C2.3 Sparse data

Parameters for observations: $M = 2^{10}$, $N \in \{0, 1, ..7\}$

Skill scores: The following skill scores are tested: Ignorance, Powerrule ($\alpha \in 1.5, 2, 2.5$), Spherical, CRPS, MSE, Naive Linear

Results for experiment C2.3 Figure 1 shows the results for the Ignorance score. When the number of observations is small (up to 2^2) the Narrow forecast is chosen more often. This is not surprising because there is a significant chance with so few observations that all of them are close to the mean, in which case the narrow distribution will give high probability to the observed values. If an event occurs far from the mean the narrow distribution would be penalised more than the wider ones, but such events are rare. Hence the narrow distribution will often avoid being penalised and will therefore be preferred by the Ignorance score. This behaviour is only observed when the number of observations is low. Indeed once the number of observations are equal to or greater than 2^5 the correct distribution is chosen by the Ignorance score over 80% of the time.

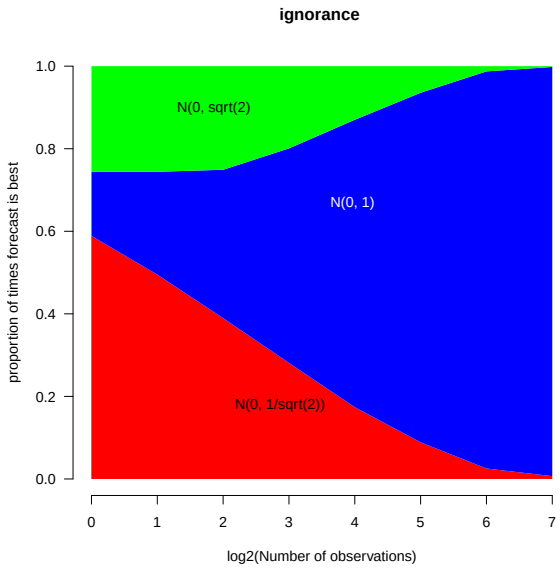

Figure 1: Sparse Data Example: Winning proportions for the various forecasts using the Ignorance score. Blue represents $F_{Perfect}$ the relative frequency of choosing $N(0,1)$, Green represents F_{Wide}, $N(0,\sqrt{2})$ and Red represents F_{Narrow}, $N(0,\frac{1}{\sqrt{2}})$. The proportions are shown on the y-axis for a given sample size of observations (N), the x-axis shows $\log_2(N)$. Observations are drawn from $N(0,1)$ distribution.

Figure 2 shows, for multiple skill scores, the proportion ($F_{Perfect}$) that the correct distribution is chosen (for comparison this is the width of the blue segment in figure 1). As is typical, the non-Proper scores perform poorly: MSE never picks the correct forecast (in fact it always picks the narrow distribution - not shown) and the Naive Linear score does little better. Apart from situations with very few observations (i.e. two or less) Ignorance performs best out of all the score types. CRPS has similar performance to the power rule score with $\alpha = 1.5$. As the α parameter increases the success rate for the power rule decreases - a similar result to experiment C2.1. Amongst Proper scores the Spherical score performs worse for larger sample sizes. The proportion appears to tend to 1 for all the Proper scores.

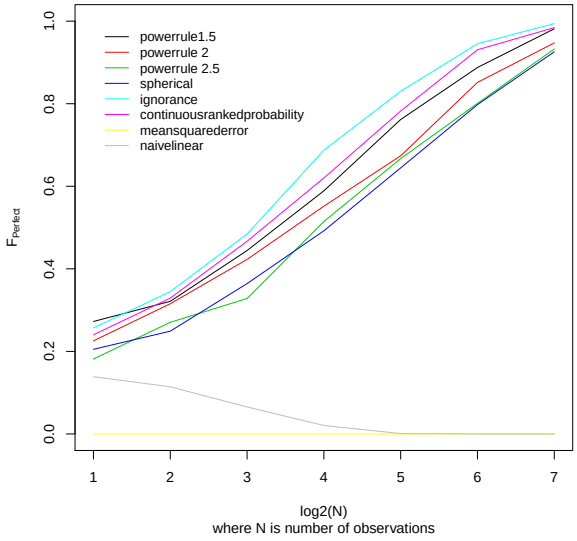

Figure 2: Sparse Data Example: Proportion of realisations in which the perfect forecast is correctly identified by different skill scores. Each line corresponds to a different skill score. When more than 2^1 observations are available the Ignorance score has the highest success rate.

2 Conclusions

The following conclusions are drawn from this experiment:

- The non-Proper scores perform poorly: MSE never picks the correct forecast (in fact it always picks the narrow distribution - not shown) and the Naive Linear score does little better.
- Apart from situations with very few observations (i.e. two or less) Ignorance performs best out of all the score types.
- CRPS has similar performance to the power rule score with $\alpha = 1.5$.
- As the α parameter increases the success rate for the power rule decreases.
- Amongst Proper scores the Spherical score performs worse for larger sample sizes.

References

[1] R. Selten. Axiomatic characterization of the quadratic scoring rule. Experimental Economics, 1:43–62, 1998.