Using the Skill Gap to create a skill scores ranking

Trevor Maynard *

February 14, 2017

The Skill Gap is a generalisation of the Information Deficit which compares how close a score is to its expected value. For a given series of forecasts the distribution of the skill gap can be estimated and this allows an erroneous forecast to be rejected after a certain time if the observed value falls outside of a pre-selected level of significance. A family of underlying distributions indexed within a triangle of parameters is created, each member of this family can also act as a forecast for any member of the family. For each pair forecast/underlying the rejection time is calculated (having specified a level of confidence and significance). This allows an objective measure of skill score performance (i.e. a skill score that takes a long time to reject an erroneous forecast has performed "badly") - which allows comparison of different skill scores. The key finding in this experiment is that the tested scores do well in some cases and poorly in others, suggesting the use of multiple skill scores would be useful.

*LSE, Lloyd's of London

1 Experiment Design

This section explores the speed with which forecasts from a structurally incorrect model are **'Rejected'**[1] by various scores. A sequence of forecasts from a given model will be referred to as a **'forecast system'** in this chapter. The running average Skill Gap (defined in equation 1 below) is calculated. The forecast system can be rejected if the Skill Gap falls outside of a chosen confidence interval. A family of distributions are defined from which the underlying truth and forecasts are chosen. The concept of 'Rejection Time' is introduced as the number of observations after which the forecast system can be rejected with a chosen probability. Initially the Rejection Time is explored using the Ignorance score and then one case is illustrated for the Naive Linear, Proper Linear and Spherical Scores. The Ignorance score is shown to perform well in some circumstances, but other scores sometimes do better.

Definition of Skill Gap For a forecast system $\underline{p} = \{p_t\}_{t=1}^{\tau}$, skill score S, and observations $\underline{X} = X_1, ...X_\tau$ the Skill Gap(G) is defined as:

$$G_S(\tau, \underline{X}, \underline{p}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left(S(p_t, X_t) - \int_{-\infty}^{\infty} p_t(x) S(p_t, x) dx \right) \tag{1}$$

The integral term is the expected score assuming the observation is drawn from the forecast distribution. Where this converges, define

$$G_\infty(\underline{X}, \underline{p}) = \lim_{t \to \infty} G_S(t, \underline{X}, \underline{p}) \tag{2}$$

The score $S(p_t, X_t)$ is a random variable; therefore G_S is also a random variable. If the Ignorance score is used then the Skill Gap is the same as the Information Deficit [1, 2]; the name 'Skill Gap' is used rather than Information Deficit to emphasise that, while the Ignorance score is naturally interpreted as information (in bits) the other skill scores considered below are not.

Let $P(A)$ denote the probability of A and let $\inf(S)$ denote the infimum of the set S. Define the quantile (Q_{G_S}), at time t, of the Skill Gap for a forecast system $\underline{p}$ as:

$$Q_{G_S}(t, \underline{X}, \underline{p}, \lambda) = \inf\{x | P(G_S(t, \underline{X}, \underline{p}) < x) = \lambda\} \tag{3}$$

Definition of Rejection If the forecasts are correct in the sense that observations (Y_i say) are drawn from the forecast distributions, so that $Y_i \sim p_i$, then it is possible to calculate $Q_{G_S}(\tau, \underline{Y}, \underline{p}, \lambda)$ (either analytically, or estimated through simulation). Given actual observations $\underline{X}$ let $g = G_S(\tau, \underline{X}, \underline{p})$, $Q_1 = Q_{G_S}(\tau, \underline{Y}, \underline{p}, \lambda)$ and $Q_2 = Q_{G_S}(\tau, \underline{Y}, \underline{p}, 1 - \lambda)$. Then the forecast system can be **'Rejected'** if either $g > Q_1$ or $g < Q_2$.

[1] Here 'Rejected' means that the observed outcomes are inconsistent with the probability distributions from the forecast system. The forecasts may still be useful and they may be informative, but it is not appropriate to use them as probability forecasts.

Note that the definitions of the Skill Gap and Rejection make no assumptions about the process generating the observations which may be unknown and even potentially unknowable. The following definition, however, considers a situation where observations are generated from known underlying distributions $\underline{q}$.

Definition of Rejection Time For a forecast system $\underline{p}$, underlying distributions $\underline{q}$ and observations $X_t \sim q_t$, for chosen confidence level λ and probability γ and a given skill score S: the **'Rejection Time'** $RT_S(\underline{q}, \underline{p}, \lambda, \gamma)$ is defined as follows:

$$RT_S(\underline{q}, \underline{p}, \lambda, \gamma) = \begin{cases} \inf\{t | Q_{G_S}(t, \underline{X}, q, 1-\gamma) = Q_{G_S}(t, \underline{X}, p, \lambda)\} & \text{if } G_\infty(q) > 0 \\ \infty & \text{if } G_\infty(q) = 0 \\ \inf\{t | Q_{G_S}(t, \underline{X}, q, \gamma) = Q_{G_S}(t, \underline{X}, p, 1-\lambda)\} & \text{if } G_\infty(q) < 0 \\ \text{undefined} & G_\infty \text{ doesn't converge} \end{cases} \tag{4}$$

Note that, for given components $\underline{p}$, $\underline{q}$, λ and γ the Rejection Time is a property of the given skill score S; allowing different scores to be compared. Specifically if $RT_{S_1} < RT_{S_2}$ for scores S_1 and S_2 then score S_1 can be said to be 'better' than S_2 for those particular components. The Rejection Time has been defined in general; in the following examples, however, a forecast PDF (p_0) is chosen and then used for every time t (i.e. $p_t = p_0 \ \forall t$). In this case the integral term is constant for each t so that

$$G_S(\tau, \underline{X}, \underline{p}) = \frac{1}{\tau}\left(\sum_{t=1}^{\tau} S(p_0, X_t)\right) - \int_{-\infty}^{\infty} p_0(x) S(p_0, x) dx \tag{5}$$

If observations are drawn from the forecast distribution, for a proper score, the summation term will converge to the integral and hence G_∞ is defined in each case. The following family of distributions will be used to test the Rejection Time concept in a specific controlled environment.

Definition of controlled distribution families Define (f) as the weighted sum of Lognormal, Gamma and Pareto distributions each with the same mean (μ) and variance (σ^2). Specifically:

$$f_{(w_1,w_2,w_3)}(x) := w_1 f_{Lognormal}(x) + w_2 f_{Gamma}(x) + w_3 f_{Pareto}(x) \tag{6}$$

Where, $\sum_{i=1}^{3} w_i = 1$ and $w_i > 0 \ \forall i$. Once two weights are chosen the other is determined since they are constrained to sum to unity. The set of weights that meet the criteria above fall in a triangular region of the plane and distributions can therefore be defined uniquely by points (w_1, w_2) in the triangle. Each point represents the distribution $f_{w_1,w_2,1-w_1-w_2}$. The **'distance'** between two distributions f_1 and f_2 is defined to be the Euclidean distance between the points in the triangle defining them.

Figure 1 (called a 'Rejection Time diagram') illustrates the calculation of Rejection Time for two points in the triangular distribution space. The series of truth distributions $\underline{q}$ is defined by the red dot in the triangle (note these are all the same) and the forecasts $\underline{p}$ by the blue dot. The quantile lines

for various values of t are illustrated for quantiles $\lambda, \gamma \in \{5, 10, 25, 50, 75, 90, 95\}$, these are created by sampling and are therefore not smooth as they would be in theory. $Q_G(t, \underline{q}, \lambda)$ are illustrated by the grey quantile lines and those of $\underline{p}$ by the black lines. The chosen confidence level for rejection is chosen to be 90% (i.e. $\lambda = 90$), the desired probability of rejection is chosen to be 75% (i.e. $\gamma = 75$). The long run Skill Gap is positive in this example because the grey lines converge above the x-axis. Therefore we look for the intersection of the 90th quantile of the black lines with the 25th quantile of the grey lines, illustrated by a red cross - and vertical line in figure. The time at which the lines cross (i.e. the Rejection Time) occurs at the 102nd observation.

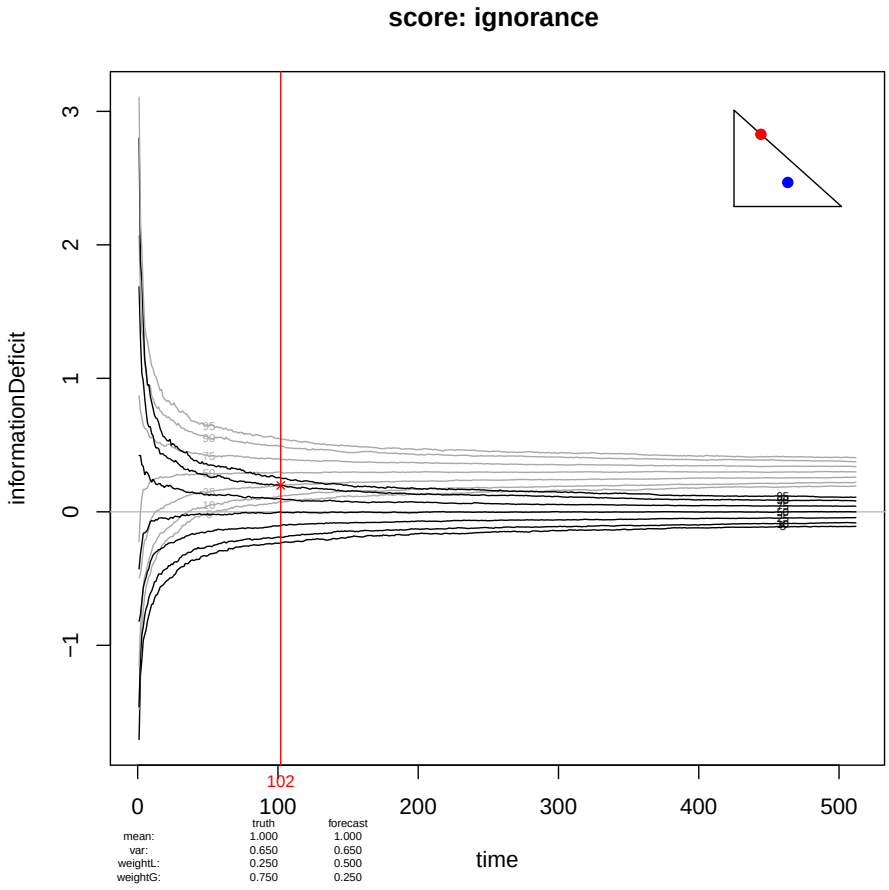

Figure 1: Rejection Time diagram: The Rejection Time is illustrated by red vertical line. Grey quantile lines show the observed Skill Gap, black lines show the expected Skill Gap if the forecast is correct. By time 102 we will have rejected the forecast system 75% of the time (at a 90% confidence level). The top right triangle graphically illustrates the chosen truth and forecast distributions.

The following algorithm (see also figure 2) uses an empirical approximation of $Q_G(t, q, p, \lambda)$ by sampling M values - sorting them and taking the 100λth largest value.

Experiment 2.4: Algorithm to calculate the Rejection Time

- **Choose confidence level λ and probability γ**
- **Choose forecast and truth** Let p' and q' be points in the triangle of allowable distributions
- For $i \in \{1...M\}$
  - **Sample observations** Let $X_{i,1}, ...X_{i,N_{obs}}$ be sampled from a distribution with PDF q' and let $Y_{i,1}, ...Y_{i,N_{obs}}$ a sample from PDF p'
  - **Calculate Skill Gap sequences** Let $S_X = \{G_i(j, q', p')\}_{j=1}^{N_{obs}}$ and let $S_Y = \{G_i(j, p', p')\}_{j=1}^{N_{obs}}$
  - If $i = M$ stop, otherwise continue for next value
- **Calculate proxy for** $G_{\infty(p',q')}$ Let the median $Q_G(M, p', q', 0.5)$ be a proxy for $G_{\infty(p',q')}$
- **If** $G_{\infty(p',q')} > 0$
  - **Estimate** Q_G For each $j \in \{1, ...N_{obs}\}$ estimate Q_G using the empirical approximation. Let $Q_X(j) = Q_G(j, q', p', 1-\gamma)$ and $Q_Y(j) = Q_G(j, p', p', \lambda)$
- **If** $G_{\infty(p',q')} < 0$
  - **Estimate** Q_G For each $j \in \{1, ...N_{obs}\}$ estimate Q_G using the empirical approximation. Let $Q_X(j) = Q_G(j, q', p', \gamma)$ and $Q_Y(j) = Q_G(j, p', p', 1-\lambda)$
- **Estimate Rejection Time** Let $RT(q', p', \lambda, \gamma)$ be the smallest value j' at which $Q_X(j') = Q_Y(j')$

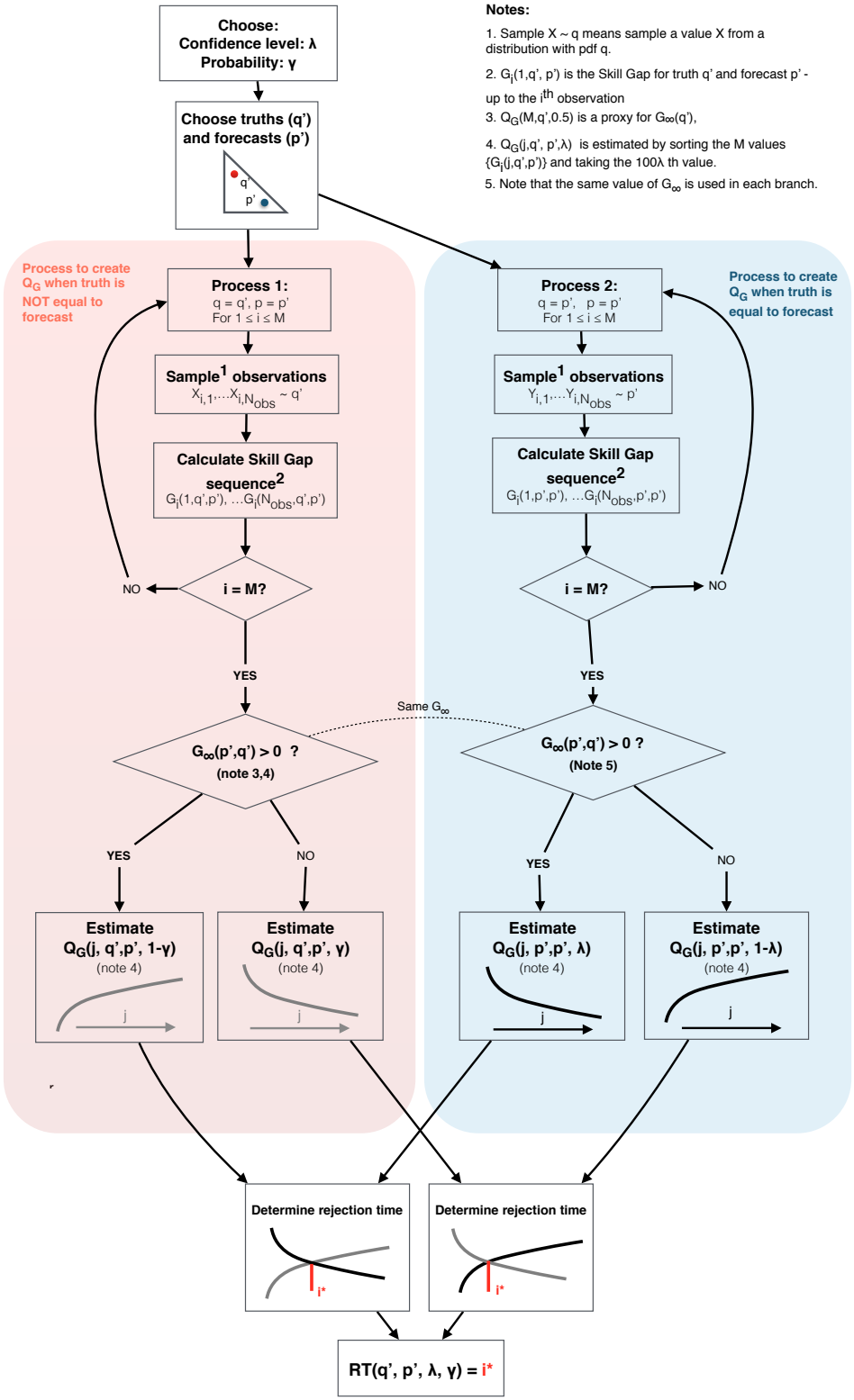

Figure 2: Experiment 2.4 Flowchart illustrating the algorithm to estimate the Rejection Time

Experiment C2.4.1 Find Rejection Times for specified forecast/truth pairs using Ignorance.

$\lambda = 0.75$, $\gamma = 0.75$ (these are chosen to keep run times lower)

$\mu = 1$, $\sigma^2 = 0.65$

The available distributions are constrained to (1) lie inside the triangle defined above and (2) be defined by grid points of the form $(\frac{a}{4}, \frac{b}{4})$, where $a, b \in \{0, 1, 2, 3, 4\}$. The exception to this is the point $(0, 0)$ which represents a Pareto distribution. The Pareto distribution requires that values less than its defining parameter be impossible; which leads to infinite Ignorance scores for some observations since the other distributions considered allow any value greater than or equal to zero. To avoid this the Pareto was not tested and a **'HybridPareto'** , defined as having the PDF $f_{(0.025,0.025,0.95)}$, is used in its place. Hence the point $(0, 0)$ is replaced by $(0.025, 0.025)$ in the triangle. The forecast is always denoted by a blue dot. All possible truth distributions within the available distribution space are tested (except where the forecast and truth are the same).

$N_{obs} = 2^9$, $M = 2^{10}$

Skill Scores = {Ignorance}

Experiment C.2.4.2 Find Rejection Times for multiple skill scores.

As for C2.4.1 except for the following:

$\underline{p}$ restricted to only consider a Gamma forecast (i.e. $f_{(0,1,0)}$).

Skill Scores = {Ignorance, Naive Linear, Proper Linear, Spherical}

2 Results

Results for experiment C.2.4.1 Figure 3 shows the Rejection Times plotted at the coordinates of the two dimensional weight combination that determines the true underlying distribution. For example, in the triangle in column two and row four of the graphic, the forecast is $p = f_{(0.25,0.75,0.00)}$. Consider the truth $q = f_{0.25,0.25,0.50}$, this has Rejection Time $RT(\underline{q}, \underline{p}, 0.75, 0.75) = 127$. The bottom graphic in figure 3 shows the results for this combination of truth and forecast for 11 different seeds this level of uncertainty would not alter the key conclusions.

Some key findings from the graphics are:

- The further the truth is from the forecast the shorter the time it takes to reject the forecast system.
- If the forecast is close to a HybridPareto and the outcome is drawn from a different distribution - then the method rejects it quickly whereas it takes longer to reject Gamma-like forecasts.
- Some truth distributions did not lead to a rejection at all within the timeframes considered (2048 time steps, shown as 9999 in the graphic).

Results for experiment C.2.4.2 Figure 4 shows Rejection Times. Some key observations are:

- An initially surprising result, all the scores apart from the Ignorance score have the same Rejection Times. This is because the extensions of the Naive Linear score all involve integral terms that are constant for a given forecast, these cause the Skill Gap of these related scores to be scalar multiples of one another and hence the same Rejection Times arise. This is proved on page 12 below.
- For all scores, the further truth is from the forecast the faster the forecast system will be rejected.
- Ignorance gives a shorter Rejection Time when the truth is HybridPareto - but the Naive Linear, ProperLinear and Spherical do better (in some cases much better) when the truth is Lognormal like. This leads to a key conclusion that **using multiple scores would be useful in some contexts**.

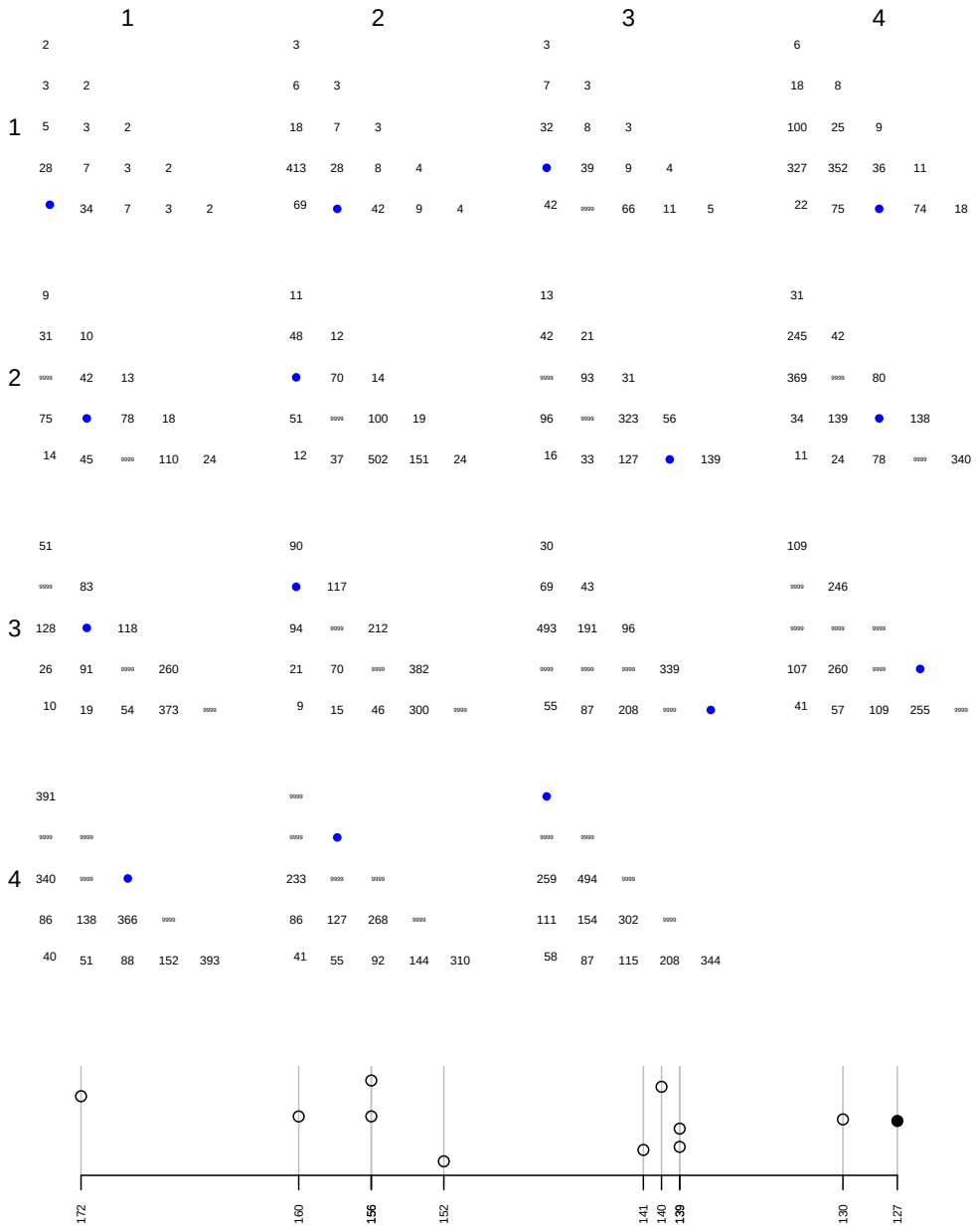

Figure 3: Rejection times as truth and forecast (blue dot) vary over the available weights - for fixed mean ($\mu = 1$) and variance ($\sigma^2 = 0.65$). 9999 denotes non-convergence within 2048 observations. The bottom plot illustrates the degree of sampling error by considering 10 different seeds when the forecast is (0.25,0.75,0) and truth is (0.25, 0.25, 0.5) (from the triangle in column 2 and row 4 of the top graphic), the black filled dot shows the results from the seed used in the top graphic, the hollow plot characters show 10 other seeds - the vertical height allows duplicate cases to be shown without overlap.

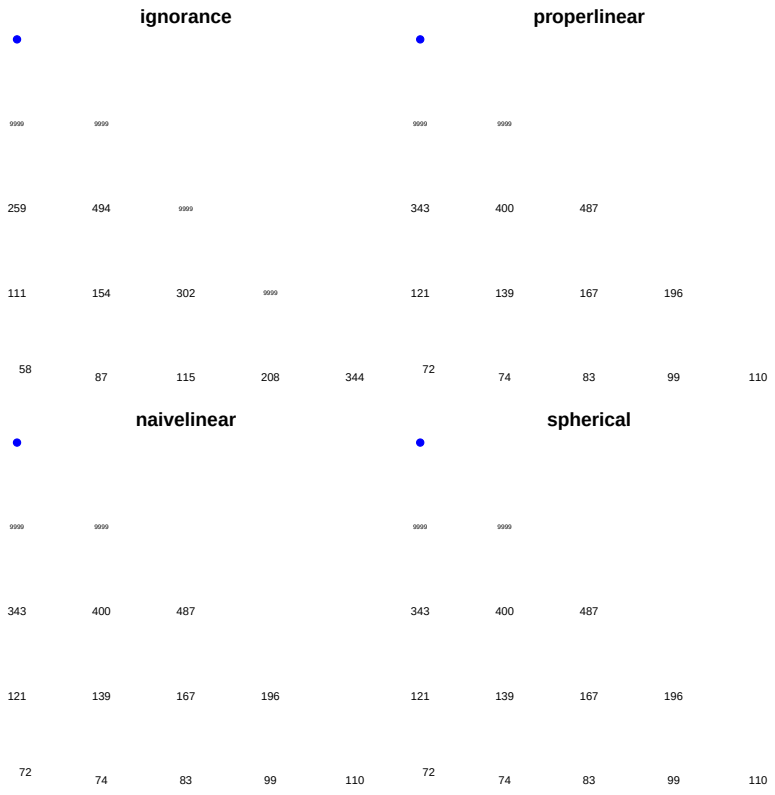

Figure 4: Rejection times for different score types, for forecast $f_{(0,1,0)}$ (Gamma distribution, denoted by a blue dot). Note that the Rejection Times for the Proper Linear, Naive Linear and Spherical scores are all the same. When observations are drawn from a Hybrid Pareto distribution (bottom left vertex of triangle) the Ignorance score rejects the forecast after 58 observations compared to (and faster than) 72 for the other scores. When the observations are drawn from a Lognormal (bottom right vertex) distribution, however, the Ignorance score required 344 observations to reject the forecast compared to 110 for the other scores. **This illustrates that there are situations where using multiple proper skill scores will be informative.** Values of 9999 (in small font) show cases where the forecast is not rejected within the maximum number (2048) of observations tested.

3 Conclusions

- An initially surprising result, all the scores apart from the Ignorance score have the same Rejection Times. This is because the extensions of the Naive Linear score all involve integral terms that are constant for a given forecast, these cause the Skill Gap of these related scores to be scalar multiples of one another and hence the same Rejection Times arise (see appendix)
- For all scores, the further truth is from the forecast the faster the forecast system will be rejected.
- Ignorance gives a shorter Rejection Time when the truth is HybridPareto - but the Naive Linear, ProperLinear and Spherical do better (in some cases much better) when the truth is Lognormal like.
- This leads to a key conclusion that using multiple scores would be useful in some contexts.

Appendix: Demonstration why the Rejection Times are the same for Naive Linear, Proper Linear and Spherical scores

In the following let G_S denote the Skill Gap for score type S. Then G_{NL} refers to the Naive Linear score, G_{PL} to Proper Linear and G_{SP} to the Spherical score. When the Naive Linear score (equation **??**) is substituted, equation 5 becomes:

$$G_{NL}(t) = \frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx \tag{7}$$

Lemma 1: If $G_S = \alpha G_R$ for two scores S and R then the Rejection Time from S will be the same as R. This is because the Rejection Time is calculated as the intersection point of two quantile lines for the Skill Gap. Since, for every sequence of observations, the Skill Gap for S is a scalar multiple of R its Rejection Time diagram will simply be a stretch in the y-axis direction - this doesn't affect where the lines cross and so Rejection Times will be the same.

Lemma 2: Proper linear $G_{PL} = 2G_{NL}$. When the Proper Linear score (equation **??**) is substituted into equation 5 the equation becomes:

$$G_{PL}(t) = \frac{1}{t}\sum_{i=1}^{t}(\int q^2(z)dz - 2q(X_i)) - \int q(x)\{\int q^2(z)dz - 2q(x)\}dx \tag{8}$$

Now $\int q^2(z)dz$ is just a constant so:

$$G_{PL}(t) = \frac{t}{t}\int q^2(z)dz + \frac{-2}{t}\sum_{i=1}^{t}(q(X_i)) - \int q^2(z)dz \int q(x)dx + 2\int q^2(x)dx \tag{9}$$

Because q is a PDF $\int q(x) = 1$, hence the two $\int q^2(z)dz$ terms cancel out. So the Skill Gap reduces to:

$$G_{PL}(t) = 2(\frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx) \tag{10}$$

This is exactly double the expression for the naive linear score stated in equation 7. So $G_{PL} = 2G_{NL}$.

Lemma 3: Spherical $G_{SP} = \frac{1}{\kappa}G_{NL}$, where $\kappa = (\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}$. When the spherical score (equation **??**) is substituted into the Skill Gap equation we see:

$$G_{SP}(t) = \frac{-1}{t}\sum_{i=1}^{t}\frac{q(X_i)}{(\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}} + \int q(x)\frac{q(x)}{(\int_{-\infty}^{\infty} q^2(z)dz)^{\frac{1}{2}}}dx \tag{11}$$

κ, defined above, is a constant so:

$$G_{SP}(t) = \frac{1}{\kappa}\{\frac{-1}{t}\sum_{i=1}^{t} q(X_i) + \int q(x)q(x)dx\} \tag{12}$$

This is a scalar multiple of the Naive score, so $G_{SP} = \frac{1}{\kappa}G_{NL}$.

Corollary: The Rejection times are always the same for Naive Linear, Proper Linear and Spherical scores Let $R = NL$ then lemma 2 shows that the condition of lemma 1 is met for $S = PL$ where $\alpha = 2$ and lemma 3 shows the condition of lemma 1 is met for $S = SP$ where $\alpha = \frac{1}{\kappa}$. So the Rejection Times for NL,SP and PL are the same.

References

[1] H. Du and L. Smith. Parameter estimation through ignorance. Physical Review E, 2012.

[2] M. Roulston and L. Smith. Evaluating probabilistic forecasts using information theory. Notes and Correspondence: American Meteorological Society, 130:1653–1660, 2002.