Using Feasibility Property to rank scores

Trevor Maynard *

February 14, 2017

The Feasibility property highlights that certain scores tend to assign good values to forecasts that assign high probability to events that are near impossible. If Feasibility is accepted as a desirable property then scores which do not posses this property should be passed over in favour of those that do. Specifically the Mean Squared Error and CRPS scores are not 'Feasible' and would therefore have a low rank according to this criteria.

*LSE, Lloyd's of London

Feasibility Feasibility (see Maynard in [2]) for a negatively oriented score is defined as follows:

A score S is **'Feasible'** if it assigns bad scores to forecasts that give material probability to highly improbable events. Specifically, let $\lambda = \inf\{p(z)|z \in \text{supp}(p)\}$, this is the probability density of the least likely outcome, the infimum (where supp(p) denotes the support of the random variable with pdf p). For any $\epsilon > 0$ define a set $M_\epsilon := \{z|p(z) < \lambda + \epsilon\}$; when ϵ is small these are the set of observations that the forecast ascribes small probability density to. Let $\mu = \inf\{S(z,p)|z \in M_\epsilon\}$, the best score amongst the minimal probability events. Then a score is Feasible if $S(z,p) \leq \mu \ \forall z \notin M_\epsilon$, that is, for any observation that is not in M_ϵ the skill score ascribes a better or equal score than μ to the forecast.

1 Feasibility of various score types

The following subsections show whether the various scores described above are Feasible. In summary, the following scores are Feasible: Ignorance, Naive and Proper Linear, Power Rule, Spherical and Brier. The CRPS and Mean Squared Error scores are not Feasible.

1.1 CRPS - not Feasible

This subsection demonstrates that the CRPS is not Feasible by providing two counterexamples.

Counterexample 1: Gaussian mixture The following sequence of graphics illustrate a major shortcoming of the CRPS score. Consider a bimodal forecast with PDF ($f(v)$) and Cumulative Density Function (CDF) ($F(v)$), for a given observation v. The PDF (figure (a)) and CDF (figure (b)) of one such distribution is shown in figure 1. Figure (b) also shows a Heaviside function (H(v) shown blue) centred at an observation value of $v = -1$. The difference, $\delta(z) = F(z) - H(z - v)$, between the CDF of the bimodal forecast and the Heaviside function is shaded green. It is this difference term which is squared and integrated over the whole support of the forecast which defines the CRPS. Note that the CDF has an inflection point at v=0 which corresponds to the trough between the two peaks.

Figure 2 illustrates various observations $v_1, ...v_9$ increasing from negative, through zero, to positive, these are shown by a blue vertical line plotted at the observation value. The graphic also shows the forecast f with a dotted red line and the integrand of the CRPS (i.e. $\delta(v)^2$) as a shaded green region. The CRPS score is the area of the green region. This is clearly least when $v = 0$ which is at the at the median of the forecast distribution. The fact that the score is minimised at the median is easily shown. Differentiate S to get $\frac{dS}{dv} = 2.\int_{-\infty}^{v} p(t)dt - 1$ (see appendix **??** for a proof), this is zero when the integral is equal to $\frac{1}{2}$, i.e. at the median of the forecast. In the above example the median occurs in the middle of the two peaks when the density is close to zero (the pdf of the forecast, in red, has been superimposed on the graphic to illustrate this). Suppose that an outcome of x=0 arose (which would be likely if the process generating the observations (the 'truth') was unimodal for example). Then the

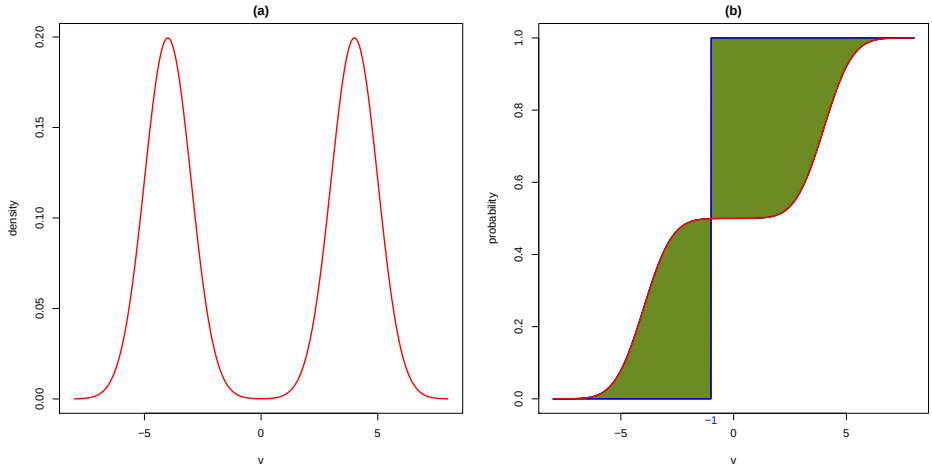

Figure 1: Figure (a) bimodal Gaussian distribution. Figure (b), the CDF of the bimodal distribution is shown by the red line, the Heaviside function is drawn at an observation of -1. This illustrates the region that is taken into the CRPS integral when the observation is -1.

CRPS score for the forecast, observation pair would be the best possible value. This is despite the fact that the forecast ascribed close to zero probability to the event that occurred. It can therefore be seen that CRPS ascribes a good score (in fact the best) to a highly improbable event, which is opposite of the behaviour required for a skill score to be 'Feasible', clearly CRPS does not have this desirable property.

Counterexample 2: Sawtooth The following extends the discussion presented in Smith et al [3]. Suppose the support of a forecast PDF (f) is comprised of 17 intervals from 0 to 17 each of length 1. Also suppose that the forecaster believes that values are highly likely to arise from only 8 of these intervals ('high density blocks') and that the likelihood is close to uniform within each block. In between these are 'low density blocks' with density approximately $f(x) \approx \frac{1}{1000}$. Each block is not quite uniform with a small positive or negative adjustment, each an order of magnitude lower than the average density within the block. This example is illustrated in figure 3(a),(b) and (c). The adjustments are made to make figure (c) clearer. The PDF f is illustrated in figure (a) with a red line - its median is shown with a black vertical line. The score value for each of the potential observations from 0 to 17 is shown in the figure (b): the Ignorance score is shown in green and the CRPS in black. Note that the Ignorance score is inversely related to the density of the forecast. The CRPS, however, behaves very strangely. First, as shown above, the best score is at the median of the forecast. This is in a region the forecaster believes is highly unlikely to occur. Also the score smoothly decreases from values to the left of the median and then smoothly increases after this. It passes through several other intervals of near-zero forecast probability, but there is no indication from CRPS the graph that this has occurred. CRPS is indifferent between intervals that are forecast to be highly probable and near impossible.

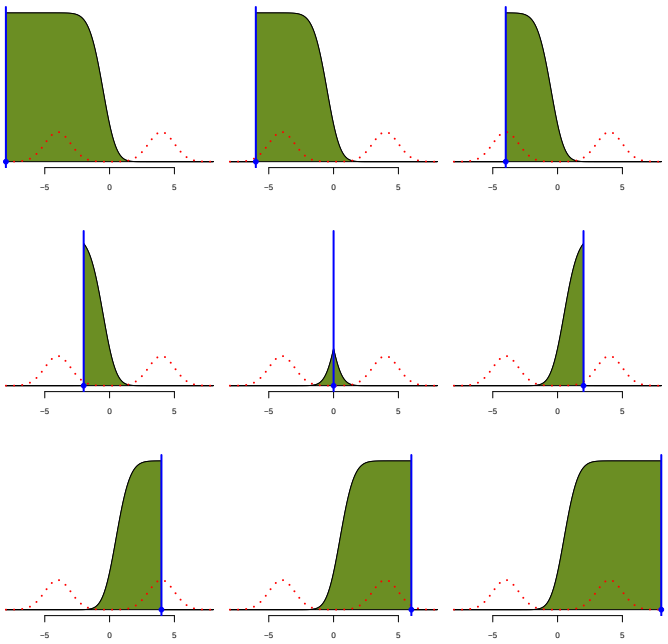

Figure 2: Integrand of the CRPS integral (shown green) for various observation values with respect to a Bimodal Gaussian forecast (shown red). This highlights visually that the CRPS score is best (i.e. a minimum) at the median of the distribution; in this example this is at a point of very low forecast density.

Assume the the true distribution is shifted one unit to right of the forecast. So that for every outcome that can occur the forecast assigned a near-zero probability. The *average* CRPS score (i.e. integrated over all possible outcomes) will be close to that of a perfect forecast; the average Ignorance score will be close to the worst possible score recognising that the forecast was completely wrong. Figure (c) illustrates this point a different way; the CRPS and Ignorance score values are plotted as $x(v)$ and $y(v)$ coordinates as functions of increasing observation values v. As the observation increases from 0 to 17 the Ignorance value fluctuates whilst the CRPS gradually descends to its minimum at the median, then the process reverses and the path is approximately traversed in the opposite direction. The small positive and negative adjustments from the multi-modal uniform distribution described above were made so that these two paths are discernible.

The sawtooth example can be used to prove that the CRPS is not Feasible as follows. Using the notation in the definition of Feasibility note that $\lambda \approx \frac{1}{1000}$ and let $\epsilon = \frac{2}{1000}$. In this case $M_\epsilon = \bigcup_{i=0}^{8}(2i, 2i+1)$ and from Figure 3(b) it is clear that $\mu < 2$, since all values on the black line, which represents the CRPS score for the low probability region $(8, 9) \subset M_\epsilon$, are less than 2. Choose $t = 1.5$ then $t \notin M_\epsilon$, then $S(p, t) > 4$. In conclusion, there exists a probability density p such that

$$\exists t \notin M_\epsilon \text{ s.t. } S(p, t) > 2 > \mu$$

and so CRPS is not Feasible.

Gneiting and Raftery [1] see this behaviour of CRPS as desirable. They note that with Local

scores 'no credit is given for assigning high probabilities to values near but not identical to the one materialising'. In the above example the sawtooth forecast looks very similar to the truth, it is just located in the wrong place. So in a sense the forecast *is* close, and this is what CRPS rewards. In the bimodal case, above, (with a unimodal truth) CRPS is not close, however. In decisions that require accuracy ('is outcome X likely at location Y or not?') this definition of 'close' is not helpful. It may therefore be arguable that the CRPS score could be useful in the context of model development where a forecast that resembles reality should be highlighted for further improvement. Using CRPS in a production setting (i.e. where real decisions are to be made) is inadvisable due to its lack of Feasibility.

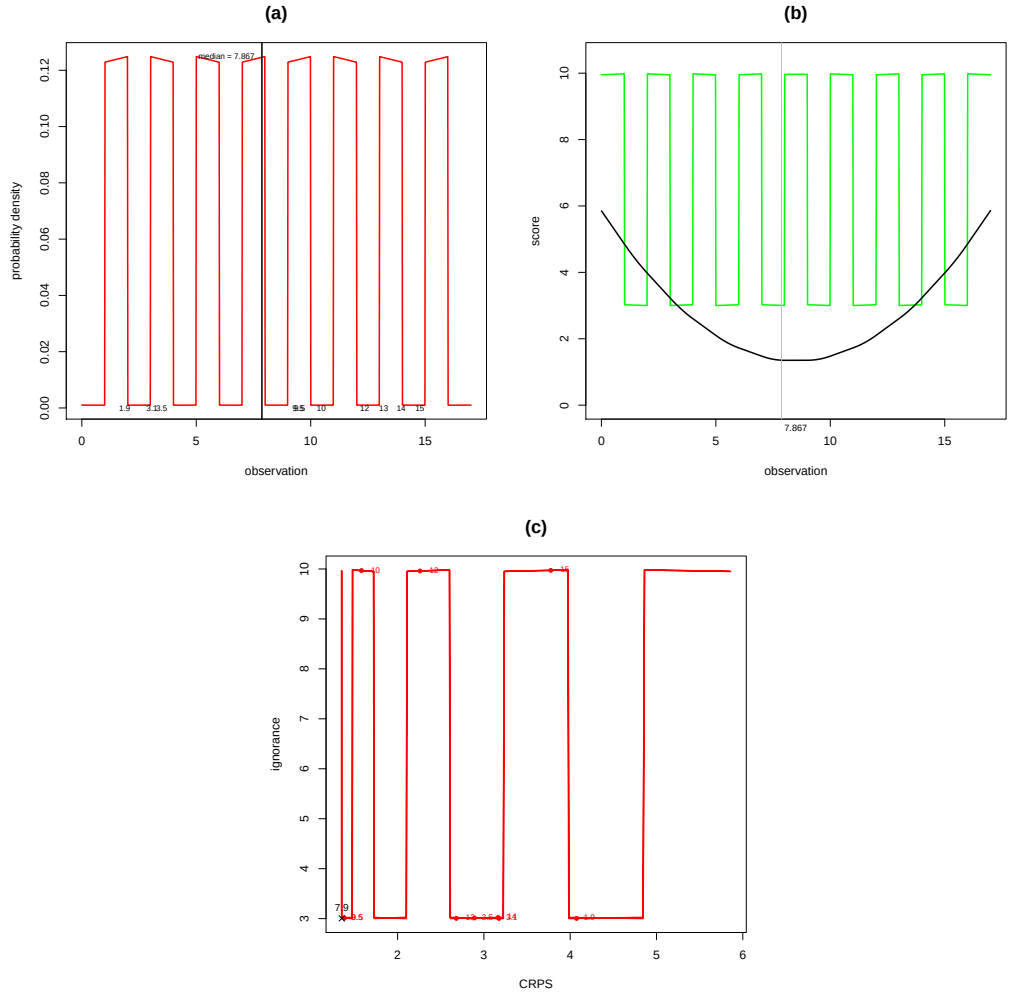

Figure 3: Implications of lack of 'Feasibility' - CRPS (not Feasible) vs Ignorance (Feasible). Top left: Probability density function of 8-uniform forecast - with median shown as vertical black line. Top right: Score value for various observed values. Ignorance shown in green and CRPS shown in black. Note that the Ignorance score reacts sensitively to the probability density of observations whereas the CRPS gives similar scores to observations that vary from highly likely (the peaks in probability density) and highly unlikely (the troughs) Bottom: CRPS score vs Ignorance score as the observation moves from lowest to highest value.

1.2 Mean Squared Error - not Feasible

The Mean Squared Error skill score S_{MSE} is not Feasible and this is shown by way of the following counterexample. This section uses the notation used in the definition of Feasibility. Recall, to show that a skill score is not Feasible we have to find a value outside of the minimal probability events (i.e. $z \notin M_\epsilon$) for which the score is worse than the best score μ for all points within M_ϵ. In short we must find a probable event that scores worse than an improbable one. Where 'Improbable events' are defined for a given value of ϵ.

Let $\epsilon = \frac{1}{4}$ and consider a bimodal uniform distribution p defined for $\delta < \frac{\epsilon}{8}$, as:

$$p(z) = \begin{cases} \frac{1}{2} - \frac{\delta}{2} & z \in [-2, -1] \cup [1, 2] \\ \frac{\delta}{2} & z \in (-1, 1) \end{cases} \tag{1}$$

Consider two possible outcomes 0 and 2. Note that, the probability density of the least likely outcomes $\lambda = \frac{\delta}{2}$ which is the density for all outcomes in the open interval $(-1, 1)$ so that $M_\epsilon = (-1, 1)$. In particular note that $0 \in M_\epsilon$. Note, however, that $p(2) = \frac{1}{2} - \frac{\delta}{2}$ and so $2 \notin M_\epsilon$. By integration, $S_{MSE}(p, 0) = \frac{7}{3} - 2\delta$, and $S_{MSE}(p, 2) = \frac{19}{3} - 2\delta$. The outcome with the best score in M_ϵ must have score μ that is lower than or equal to the score for the observation 0, by definition of the infemum. Therefore $\mu \leq \frac{7}{3} - 2\delta = S(p, 0) < \frac{19}{3} - 2\delta = S(p, 2)$.

The above has shown that $\exists z \notin M_\epsilon$ (i.e. $z = 2$) such that $S(p, z) > \mu$. Therefore MSE is not Feasible.

1.3 Ignorance - Feasible

By defintion, $\forall t \notin M_\epsilon$ we have $p(t) \geq \lambda + \epsilon$, and also $\forall z \in M_\epsilon$ we have $p(z) < \lambda + \epsilon$. Since Ignorance is defined as $S(p, v) = -log_2(p(v))$ and by the continuity and monotonicity of log it is always the case that $S(p, t) \leq -log(\lambda + \epsilon) < S(p, z)$ where t and z are defined as above. Hence $S(p, t) \leq \inf(S(p, z)) = \mu$. Therefore the Ignorance score is Feasible, since the above inequality is true for any $t \notin M_\epsilon$.

1.4 Linear scores - Feasible

'Linear' scores are taken to include: Naive Linear, Proper Linear, Spherical and Power Rule scores. In each case the score $S(p, v) = A(p) + \frac{-p(v)^\alpha}{B(p)}$ where α is a positive real number and $A(p)$ and $B(p)$ are non-negative terms that depend only on p (and for a given p are therefore constant terms). In particular, with the spherical score A(p)=0 and for the other scores B(p)=1. Note that $p(v)^\alpha$ is monotonic increasing and continuous. Since both A and B are non-negative and constant for all observations (z or t), dividing by B and adding A does not change the inequality so, with the same definition of t and z in section 1.3:

$$S(p, t) \leq A + \frac{-p(\lambda + \epsilon)^\alpha}{B} < S(p, z) \tag{2}$$

Hence, $S(p,t) \leq \inf(S(p,z)) = \mu$ and since the above inequality is true for any $t \notin M_\epsilon$ the Linear scores are all Feasible.

1.5 Brier - Feasible

In the case of the Brier score there are only two possible observations - the event occurs or it does not. These are denoted $e_1 = (1,0)$ and $e_2 = (0,1)$ respectively. The assigned probability that the event occurs is denoted p. There are then three cases to consider: $p < 0.5$, $p = 0.5$ and $p > 0.5$. **Case 1:** If $p = 0.5$ then $\lambda = 0.5$ and $M_\epsilon = \{e_1, e_2\}$ for all $\epsilon > 0$, there are therefore no points $z \notin M_\epsilon$ and the condition for Feasibility is met trivially. **Case 2:** consider the case when $p < 0.5$, then $\lambda = p$. If $\epsilon > 1 - 2p$ then $M_\epsilon = \{e_1, e_2\}$ and again the condition is met trivially. Consider the other situation when $\epsilon \leq 1 - 2p$ then $M_\epsilon = \{e_1\}$ and $\mu = |(p, 1-p) - (1,0)|^2 = 2(p-1)^2$. In this situation $e_2 \notin M_\epsilon$ and $S(e_2) = 2p^2 < 2(p-1)^2 = \mu$ since $p < 0.5 < 1 - p$. **Case 3:** the case for $p > 0.5$ is similar, here $\lambda = 1 - p$. If $\epsilon > 2p - 1$ then $M_\epsilon = \{e_1, e_2\}$ and, as before, the Feasibility criteria is trivially met. If $\epsilon \leq 2p - 1$ then $M_\epsilon = \{e_2\}$ and, by definition, $\mu = |(p, 1-p) - (0,1)|^2 = 2p^2$. Under these conditions $e_1 \notin M_\epsilon$ and
$S(e_1) = 2(1-p)^2 < 2p^2 = \mu$. Therefore in all possible cases the Feasibility criteria is met and the Briar score is Feasible.

2 Conclusions

The following scores have the Feasibility property:

- Ignorance
- Naive linear
- Proper linear
- Power rule
- Spherical
- Brier

The following scores do not have the Feasibility property:

- CRPS
- MSE

On the presumption that the Feasibility property is considered desirable the CRPS and MSE are therefore of low rank on this basis.

References

[1] Gneiting and Raftery. Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association, 102(477):359– 376, March 2007.

[2] T. Maynard. Extreme Insurance and the Dynamics of Risk. PhD thesis, London School of Econmics and Political Science, 2016.

[3] L. A. Smith, E. B. Suckling, E. L. Thompson, T. Maynard, and H. Du. Towards improving the framework for probabilistic forecast evaluation. Springer Climatic Change, 2015.