School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

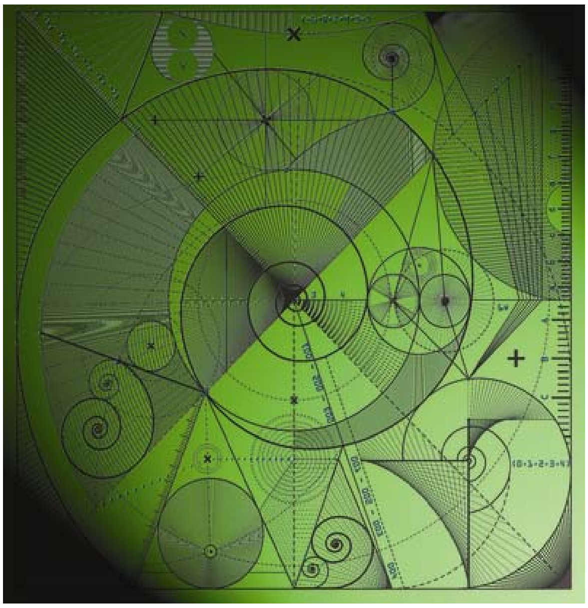
Preprint MPS-2011-15

01 November 2011

Contrasting Probabilistic Scoring Rules

by

Reason L. Machete



Contrasting Probabilistic Scoring Rules

Reason L. Machete*

Dept. of Mathematics and Statistics, P. O. Box 220, Reading, RG6 6AX, UK

Dated: November 01, 2011

Abstract

There are several scoring rules that one can choose from in order to score probabilistic forecasting models or estimate model parameters. Whilst it is generally agreed that proper scoring rules are preferable, there is no clear criterion for preferring one proper scoring rule above another. This manuscript contrasts properties of some commonly used proper scoring rules and provides incremental guidance on scoring rule selection. In particular, it is shown that the *logarithmic scoring rule* prefers erring on the side of caution, but the *Continuous Ranked Probability Score* tends to prefer over-confident forecasts.

1 Introduction

Issuing probabilistic forecasts is meant to express uncertainty about the future evolution of a dynamical system. The quality of probabilistic forecasts may be undermined by model mis-specification. This makes it necessary to assess forecast-quality. Forecast-quality can be assessed using either probability integral transforms (Diebold et al., 1998) or probabilistic scoring rules (Gneiting & Raftery, 2007). A probability integral transform (PIT) is obtained by evaluating a cumulative distribution function at an observation whilst a scoring rule assigns a numerical value based on the value that materialises. If predictive distributions coincide with ideal forecasts, then PITs will be uniform and identically distributed (Diebold et al., 1998; Dawid, 1984); it is then said that the forecasting model is *calibrated* (Corradi & Swanson, 2006).

It is often the case that a scientist has competing probabilistic forecasts that need to be ranked. For instance, one may want to know how forecasting performance changes with lead time; forecasting performance then needs to be ranked according to lead time. Alternatively, a scientist may want to know which of two competing mis-specified forecasting models is better than the other. In both of the aforementioned situations, the PITs are of no value. Instead of PITs, one can employ probabilistic scoring rules because they provide summary measures of forecasting performance.

We shall take scoring rules to be loss functions that a forecaster wishes to minimise. Scoring rules that are minimised if and only if the issued forecasts coincide with ideal forecasts are said to be *strictly proper* (Gneiting & Raftery, 2007; Brocker & Smith, 2007). Since there is no incentive not to issue ideal forecasts, we shall restrict our attention to strictly proper scoring rules. Nonetheless using scoring rules to rank competing forecasting models poses a problem; as noted by Diebold et al. (1998), scoring rules do not provide a universally acceptable ranking of performance, i.e. the ranking of probabilistic forecasts can vary with the scoring rule used. There is, therefore, a need to understand how different scoring rules would rank competing forecasts.

*Corresponding author: `r.l.machete@reading.ac.uk`, tel: +44(0)118 378 6378

This paper presents a novel theoretical analysis of strictly proper scoring rules. It focuses upon those scoring rules that are commonly used in the forecasting literature, including econometrics and meteorology. It unravels and contrasts how the different scoring rules would rank competing forecasts of specified departures from ideal forecasts and provides incremental guidance on scoring rule selection. This work may be viewed as an extension of the empirical studies of Boero et al. (2011) who compared how three scoring rules would rank competing forecasts; but our study considers both categorical forecasts and density forecasts. More specifically, we contrast the relative information content of forecasts preferred by different scoring rules.

In section 2, we consider the case of scoring categorical forecasts by the Brier score (Brier, 1950) and the logarithmic scoring rule (Good, 1952). For simplicity, special attention is focused on binary forecasts. This section then inspires our study of density forecasts in section 3, where we consider three scoring rules: the Quadratic Score (Gneiting & Raftery, 2007), Logarithmic Score (Good, 1952) and Continuous Ranked Probability Score (Epstein, 1969). We conclude with a discussion of the results in section 4.

2 Categorical Forecasts

In this section, we consider the scoring of categorical forecasts. We compare and contrast how different scoring rules would rank two given forecasts. The scoring rules considered are Brier score (Brier, 1950) and the logarithmic scoring rule. In order to aid intuition in the next section, here we focus on the binary case. Another commonly used scoring rule for categorical forecasts is the Ranked Probability Score (RPS) (Epstein, 1969). In the binary case, the RPS score reduces to the Brier score.

2.1 The Brier score

Consider a probabilistic forecast $\{f_i\}_{i=1}^m$ of m categorical events. Suppose the true distribution is $\{p_i\}_{i=1}^m$. If the actual outcome is the jth category, the Brier score is given by (Brier, 1950)

$$BS(f,j) = \frac{1}{m}\sum_{i=1}^{m}(f_i - \delta_{ij})^2,$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. If follows that if we expand out the bracket we get

$$BS(f,j) = \frac{1}{m}\sum_{i=1}^{m}(f_i^2 - 2f_j + 1).$$

The expected Brier score is then given by

$$\mathbb{E}[BS] = \sum_{j=1}^{m} p_j BS(f,j) = \frac{1}{m}\sum_{i=1}^{m}(f_i^2 - 2p_i f_i + 1).$$

When $m = 2$, we have can put $f_1 = p + \gamma$, $p_1 = p$ and $p_2 = q$ and obtain

$$\mathbb{E}[BS] = \gamma^2 - p^2 + p.$$

It follows that $\pm\gamma$ will yield the same Brier score. This means the Brier score does not discriminate between over-estimating and under-estimating the probabilities with the same amount. Further more, for any two forecasts $\boldsymbol{f}_i = (p + \gamma_i, q - \gamma_i)$, $i = 1, 2$, with $|\gamma_1| < |\gamma_2|$, the Brier score would prefer the forecast corresponding to γ_1.

2.2 Logarithmic scoring rule

The logarithmic scoring rule was proposed by Good (1952). It was later termed Ignorance by Roulston & Smith (2002) when they introduced it to the meteorological community. Given a probabilistic forecast $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)$, the Ignorance score is given by $IGN(\boldsymbol{f}, j) = -\log f_j$, where j denotes the category that materialises. Let us consider the expected logarithmic score (Good, 1952) of the forecasting scheme $\boldsymbol{f} = (p + \gamma, q - \gamma)$:

$$\mathbb{E}[IGN] = -p \log(p + \gamma) - q \log(q - \gamma). \tag{1}$$

The above expectation is also referred to as the Kullback-Leibler Information Crieterion (Corradi & Swanson, 2006). Without loss of generality, we assume that $\gamma > 0$, so that

$$\mathbb{E}[IGN]_{\pm} = p \log\left(\frac{p-\gamma}{p+\gamma}\right) + q \log\left(\frac{q+\gamma}{q-\gamma}\right) \tag{2}$$

Note that when $p = q = 0.5$, then $\mathbb{E}[IGN]_{\pm} = 0$, otherwise $\mathbb{E}[IGN]_{\pm} \neq 0$. Differentiating (2) with respect to γ yields

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}[IGN]_{\pm} = \frac{2\gamma^2(p^2 - q^2)}{(p^2 - \gamma^2)(q^2 - \gamma^2)} \tag{3}$$

Expressions (2) and (3) are well defined provided $\gamma < \min(p, q)$.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}[IGN]_{\pm} &> 0, \quad \text{if } p > q \\ \frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}[IGN]_{\pm} &< 0, \quad \text{if } p < q \end{aligned}$$

It follows that $\mathbb{E}[IGN]_{\pm} > 0$ if $p > q$ and $\mathbb{E}[IGN]_{\pm} < 0$ if $p < q$. In other words, the logarithmic score penalises over confidence on the likely outcome and rewards erring on the side of caution. Given forecasting schemes that are equally calibrated, the logarithmic score will prefer the one with a higher entropy. To explain this further, let us denote the entropy of the forecast corresponding to γ by $h(\gamma)$, i.e.

$$h(\gamma) = -(p + \gamma) \log(p + \gamma) - (q - \gamma) \log(q - \gamma). \tag{4}$$

We now define the function $G(\gamma) = h(\gamma) - h(-\gamma)$ and claim that $G(\gamma) < 0$ for $0 < \gamma < q < p$. To prove this claim, we first note that $G(0) = 0$. It then suffices to show that $G'(0) < 0$. Note that

$$\begin{aligned} G'(\gamma) &= -\log(p + \gamma) + \log(q - \gamma) - \log(p - \gamma) + \log(q + \gamma) \\ &= -\log\left(\frac{p+\gamma}{q+\gamma}\right) + \log\left(\frac{q-\gamma}{p-\gamma}\right). \end{aligned}$$

The condition $p > q$ implies that $G'(\gamma) < 0$ for all $\gamma \in (0, q)$. Therefore, it is evident that, of the two forecasts, the logarithmic score prefers the one with a higher entropy. We have thus proved the following proposition:

Proposition 1 *Given two forecasts,* $\boldsymbol{f}_+ = (p + \gamma, q - \gamma)$ *and* $\boldsymbol{f}_- = (p - \gamma, q + \gamma)$, *where* $0 < \gamma < q < p$, *the logarithmic scoring rule prefers* $\boldsymbol{f}_-$. *Moreover,* $\boldsymbol{f}_-$ *has a higher entropy than* $\boldsymbol{f}_+$.

What about when there are two forecasts $\boldsymbol{f}_i = (p + \gamma_i, q - \gamma_i)$, $i = 1, 2$ with $0 < \gamma_1 < \gamma_2 < q$ and $p > q$? It is obvious that the Brier score will prefer $\boldsymbol{f}_1$ over $\boldsymbol{f}_2$. The question is, which of the two forecasts will the logarithmic scoring rule prefer? We answer this question by stating the following proposition:

Proposition 2 *Given two forecasts* $\boldsymbol{f}_i = (p + \gamma_i, q - \gamma_i)$, $i = 1, 2$ *with* $0 < \gamma_1 < \gamma_2 < q$ *and* $p > q$, *the logarithmic scoring rule prefers* $\boldsymbol{f}_1$ *over* $\boldsymbol{f}_2$, *in agreement with the Brier score.*

Proof: In order to prove this proposition, it is sufficient to consider the expected logarithmic score of the forecast $\boldsymbol{f} = (p + \gamma, q - \gamma)$, which is given by equation (1). Differentiating the equation with respect to γ yields

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}[IGN] = \frac{\gamma}{(p + \gamma)(q - \gamma)} \tag{5}$$

Equation (5) implies that, if $q > \gamma > 0$, $\mathbb{E}[IGN]$ is an increasing function of γ. Hence, the logarithmic scoring rule prefers the forecast $\boldsymbol{f}_1$, in agreement with the Brier score.

On the other hand, if $\gamma < 0$ with $|\gamma| < p$, then equation (5) implies that $\mathbb{E}[IGN]$ is a decreasing function of γ. It then follow that, given $\gamma_2 < \gamma_1 < 0$ with $|\gamma_2| < p$, the logarithmic scoring rule will prefer the forecast $\boldsymbol{f}_1$, again in agreement with the Brier score.

Finally, let us consider the case of two forecasts $\boldsymbol{f}_1 = (p + \gamma_1, q - \gamma_1)$ and $\boldsymbol{f}_2 = (p - \gamma_2, q + \gamma_2)$, where $0 < \gamma_1 < \gamma_2 < q < p$. Again, it is clear that the Brier score will prefer the forecast $\boldsymbol{f}_1$ over $\boldsymbol{f}_2$. It remains to be seen which forecast the logarithmic scoring rule will prefer. This may be determined by considering the function $H(\gamma_1, \gamma_2)$, where

$$H(\gamma_1, \gamma_2) = p \log\left(\frac{p - \gamma_2}{p + \gamma_1}\right) + q \log\left(\frac{q + \gamma_2}{q - \gamma_1}\right) \tag{6}$$

Note that $H(\gamma_1, \gamma_2) = \mathbb{E}[IGN]_1 - \mathbb{E}[IGN]_2$. The forecast $\boldsymbol{f}_1$ is preferred if $H(\gamma_1, \gamma_2) < 0$. The following proposition gives insights of relative forecast performance in the parameter space.

Proposition 3 *Given that* $0 < \gamma_2 < q < p$, *there exists* $\gamma^* \in (0, \gamma_2)$ *such that (a)* $H(\gamma*, \gamma_2) = 0$, *(b)* $H(\gamma_1, \gamma_2) > 0$ *for* $\gamma_1 \in (\gamma^*, \gamma_2)$ *and (c)* $H(\gamma_1, \gamma_2) < 0$ *for* $\gamma_1 \in (0, \gamma^*)$.

Before proving the above proposition, we remark that $H(\gamma_1, \gamma_2) < 0$ if and only if the logarithmic scoring rule prefers the forecast $\boldsymbol{f}_1$. This proposition implies that the logarithmic scoring rule and the Brier score prefer different forecasts when $\gamma_1 \in (\gamma^*, \gamma_2)$. Let us now consider the proof of this proposition.

Proof: In proving this proposition, it is useful to bear in mind that $H(\gamma_2, \gamma_2) > 0$. The partial derivatives of equation (6) are given by

$$\frac{\partial H}{\partial \gamma_1} = \frac{\gamma_1}{(p + \gamma_1)(q - \gamma_1)} \quad \text{and} \quad \frac{\partial H}{\partial \gamma_2} = \frac{-\gamma_2}{(p + \gamma_2)(q - \gamma_2)}. \tag{7}$$

Further more, we can differentiate equations (7) to obtain

$$\frac{\partial^2 H}{\partial \gamma_1^2} = \frac{pq + \gamma_1^2}{(p + \gamma_1)^2(q - \gamma_1)^2} \quad \text{and} \quad \frac{\partial^2 H}{\partial \gamma_2^2} = \frac{-(pq + \gamma_2^2)}{(p - \gamma_2)^2(q + \gamma_2)^2}. \tag{8}$$

It follows from equations (7) that $\partial H/\partial\gamma_1 = 0$ at $\gamma_1 = 0$ and $\partial H/\partial\gamma_2 = 0$ at $\gamma_2 = 0$. Since $\partial^2 H/\partial\gamma_1^2 > 0$ for all γ_1, $H(\gamma_1, \cdot)$ has a global minimum at $\gamma_1 = 0$. Similarly, $H(\cdot, \gamma_2)$ has a global maximum at $\gamma_2 = 0$ since $\partial^2 H/\partial\gamma_2^2 < 0$ for all γ_2 and the first partial derivative with respect to γ_2 vanishes there. In particular, $H(0, \gamma_2) \leq H(0, 0) = 0$, i.e. $H(0, \gamma_2) \leq 0$. For $\gamma_2 > 0$, we have the strict inequality, $H(0, \gamma_2) < 0$. But we also have $H(\gamma_2, \gamma_2) > 0$ from Proposition 1. It, therefore, follows from the intermediate value theorem that $H(\gamma_1, \gamma_2) = 0$ for some $\gamma_1 = \gamma^* \in (0, \gamma_2)$, which completes the proof.

As we conclude the subsection, we state the following proposition:

Proposition 4 *For positive γ_1 and γ_2 such that $\gamma_1 < q < p$ and $\gamma_2 < p$, the entropy of the forecast $\boldsymbol{f}_1 = (p+\gamma_1, q-\gamma_1)$ is lower than that of the forecast $\boldsymbol{f}_2 = (p-\gamma_2, q+\gamma_2)$ whenever $\gamma_2 \leq (p-q)/2$.*

A consequence of this proposition is that the forecast corresponding to $\gamma_1 = \gamma^*$ is more informative than $\boldsymbol{f}_2$ provided $\gamma_2 \leq (p-q)/2$. Otherwise, either forecast could be more informative than the other. We now give the proof of this proposition.

Proof: To prove the above proposition, we consider the derivative of equation (4):

$$\frac{\mathrm{d}h}{\mathrm{d}\gamma} = -\log\left(\frac{p+\gamma}{q-\gamma}\right).$$

We then note that $\mathrm{d}h/\mathrm{d}\gamma < 0$ provided that $(p-q) > -2\gamma$. If $\gamma > 0$, this inequality is trivially satisfied. On the other hand, if $\gamma < 0$, then the inequality is satisfied provided $|\gamma| < (p-q)/2$. If $\gamma_2 < (p-q)/2$, then $h(\gamma)$ is a strictly decreasing function for all $\gamma \in [-\gamma_2, \gamma_2]$, which implies that $h(\gamma_1) > h(\gamma_2)$. If $\gamma_2 > (p-q)/2$, then $h(\gamma)$ is an increasing function for all $\gamma \in (-\gamma_2, -(p-q)/2)$ (provided $p > 3q$) and strictly decreasing function in $(-(p-q)/2, \gamma_1)$, which implies that $h(-(p-q)/2) > \max\{h(\gamma_1), h(-\gamma_2)\}$. Hence, in this case, we cannot determine which of $h(\gamma_1)$ and $h(-\gamma_2)$ is lower.

3 Density Forecasts

This section considers scoring rules for for forecasts of continuous variables. It is in some sense a generalisation of the previous section. As before, we consider how each scoring rule would rank two competing predictive distributions of fairly good quality. In the case of the logarithmic scoring rule and the Continuous Ranked Probability Score, we consider errors of each predictive distribution, $f(x)$, from the target distribution, $p(x)$, that are odd functions, i.e. $\gamma(x) = f(x) - p(x)$ with $\gamma(-x) = -\gamma(x)$.

3.1 The Quadratic scoring rule

A continuous counterpart of the Brier score is the quadratic score (Gneiting & Raftery, 2007), given by

$$QS(f, X) = ||f||_2^2 - 2f(X),$$

where X is a random variable. Taking the expectation yields

$$\mathbb{E}[QS(f, X)] = ||f - p||_2^2 - ||p||_2^2. \tag{9}$$

We can now write $f(x) = p(x) + \gamma(x)$, where $\int \gamma(x)\mathrm{d}x = 0$, and substitute it into (9) to obtain

$$\mathbb{E}[QS(f, X)] = ||\gamma||_2^2 - ||p||_2^2 \tag{10}$$

As was the case with the Brier score, the functions $\pm\gamma(x)$ yield the same quadratic score. For any two forecasts, $f_i(x) = p(x) + \gamma_i(x)$, $i = 1, 2$ with $||\gamma_1||_2 < ||\gamma_2||_2$, the quadratic scoring rule would prefer $f_1(x)$.

3.2 The Logarithmic scoring rule

The expectation of the logarithmic (or Ignorance) score for this forecast is

$$\mathbb{E}[IGN(f, X))] = -\int p(x)\log(p(x)+\gamma(x))\mathrm{d}x.$$

Further more,

$$\mathbb{E}[IGN]_{\pm} = \int p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x. \tag{11}$$

It is necessary that $|\gamma(x)| \leq p(x)$ for (11) to be well defined. Consider the case when $p(x) = p(-x)$. If, in addition, $\gamma(x)$ is an odd function, i.e. $\gamma(-x) = -\gamma(x)$, then equation (11) yields $\mathbb{E}[IGN]_{\pm} = 0$.

We now turn to the general case where $p(x)$ is not necessarily even and $\gamma(x)$ is not necessarily odd. If we let $\varphi(x)$ be a test function, then the functional derivative of $\mathbb{E}[IGN]_{\pm}$, denoted by $\delta\mathbb{E}[IGN]_{\pm}/\delta\gamma$, satisfies the relation

$$\left\langle \frac{\delta}{\delta\gamma}\mathbb{E}[IGN]_{\pm}, \varphi \right\rangle = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int p(x)\log\left(\frac{p(x)-\gamma(x)-\varepsilon\varphi(x)}{p(x)+\gamma(x)+\varepsilon\varphi(x)}\right)\mathrm{d}x\bigg|_{\varepsilon=0},$$

from which it follows that

$$\frac{\delta}{\delta\gamma}\mathbb{E}[IGN]_{\pm} = \frac{-2p^2(x)}{p^2(x)-\gamma^2(x)}. \tag{12}$$

Hence, the functional derivative is negative for all $\gamma(x)$ such that $|\gamma(x)| < p(x)$, which implies that $\mathbb{E}[IGN]_{\pm}$ is a decreasing functional of $\gamma(x)$. This means that, starting at $\gamma(x) = 0$, making $\gamma(x)$ negative will make $\mathbb{E}[IGN]_{\pm}$ positive while increasing $\gamma(x)$ will make $\mathbb{E}[IGN]_{\pm}$ negative.

Considering the case when $\gamma(-x) = -\gamma(x)$ and $\int_{\infty}^{0} p(x)\mathrm{d}x > 0.5$, we state the following proposition:

Proposition 5 *Given that $\gamma(-x) = -\gamma(x)$ with $\gamma(|x|) < 0$ and $p(|x|) \leq p(x)$, then $\mathbb{E}[IGN]_{\pm} \geq 0$.*

The above proposition gives conditions under which the forecast f_+ is preferred by the logarithmic scoring rule over f_-.

Proof: The proof proceeds as follows:

$$\begin{aligned}
\mathbb{E}[IGN]_{\pm} &= \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x \\
&= \int_{-\infty}^{0} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x + \int_{0}^{\infty} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x
\end{aligned}$$

If we now perform a change of variable $u = -x$ and in the right hand integral and then replace u by x, we obtain

$$\begin{aligned}
\mathbb{E}[IGN]_{\pm} &= \int_{-\infty}^{0} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x - \int_{0}^{-\infty} p(-x)\log\left(\frac{p(-x)-\gamma(-x)}{p(-x)+\gamma(-x)}\right)\mathrm{d}x \\
&= \int_{-\infty}^{0} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x + \int_{-\infty}^{0} p(-x)\log\left(\frac{p(-x)+\gamma(x)}{p(-x)-\gamma(x)}\right)\mathrm{d}x \\
&\geq \int_{-\infty}^{0} p(x)\log\left(\frac{p(x)-\gamma(x)}{p(x)+\gamma(x)}\right)\mathrm{d}x + \int_{-\infty}^{0} p(x)\log\left(\frac{p(x)+\gamma(x)}{p(x)-\gamma(x)}\right)\mathrm{d}x = 0,
\end{aligned}$$

where we used $p(|x|) \leq p(x)$ to obtain the last inequality. To justify the use of this inequality, we need to show that the function

$$\Phi(p) = p \log\left(\frac{p+\gamma}{p-\gamma}\right)$$

is a decreasing function for $\gamma \in (0, p)$. Differentiating Φ with respect to p yields

$$\Phi'(p) = \log\left(\frac{p+\gamma}{p-\gamma}\right) - \frac{2p\gamma}{p^2-\gamma^2}.$$

It now suffices to show that $\Phi'(p) < 0$ for all p. Let us introduce the notation $W(p) = \log[(p+\gamma)/(p-\gamma)]$ and $Y(p) = 2p\gamma/(p^2-\gamma^2)$ so that $\Phi'(p) = W(p) - Y(p)$. Note that $W(2\gamma) = \log 2$ and $Y(2\gamma) = 4/3 = \log(e^{4/3})$. Hence $W(2\gamma) < Y(2\gamma)$, which implies that $\Phi'(2\gamma) < 0$. Differentiating $W(p)$ and $Y(p)$ with respect to p yields

$$W'(p) = \frac{-2\gamma}{p^2-\gamma^2} \quad \text{and} \quad Y'(p) = \frac{-2\gamma(p^2+\gamma^2)}{(p^2-\gamma^2)^2}.$$

It is now clear that $W'(p) < 0$ and $Y'(p) < 0$ for all p. Further more, $Y'(p) < W'(p)$. Hence $W(p) < Y(p)$ for all $p \in (\gamma, 2\gamma]$, which implies that $\Phi'(p) < 0$ for all $p \in (\gamma, 2\gamma]$. It now remains to be shown that $\Phi'(p) < 0$ for all $p \in (2\gamma, \infty)$. It suffices to consider the asymptotic behaviour as $p \to \infty$. Applying L'Hopital's rule, we obtain

$$\lim_{p\to\infty} \frac{|W(p)|}{|Y(p)|} = \lim_{p\to\infty} \frac{|W'(p)|}{|Y'(p)|} = \lim_{p\to\infty} \frac{p^2-\gamma^2}{p^2+\gamma^2} = 1.$$

Hence, $\lim_{p\to\infty} \Phi'(p) = 0$ and the proof is complete. The condition $p(|x|) < p(x)$ implies that $\int_{-\infty}^{0} p(x)\mathrm{d}x \geq 0.5$. It corresponds to the case $p > q$ in the previous discrete case.

We now want to compare the entropies of the forecasts $f(x) = p(x) \pm \gamma(x)$ when $\gamma(-x) = -\gamma(x)$ and $\gamma(|x|) \leq \gamma(x)$. The entropy of the function $f(x) = p(x) + \gamma(x)$ is then given by

$$h(\gamma) = -\int (p(x)+\gamma(x)) \log(p(x)+\gamma(x))\mathrm{d}x. \tag{13}$$

The functional derivative of $h(\gamma)$ with respect to γ is then given by

$$\begin{aligned} \frac{\delta h(\gamma)}{\delta\gamma(x)} &= -\frac{\partial}{\partial\gamma(x)}\{(p(x)+\gamma(x))\log(p(x)+\gamma(x))\} \\ &= -[\log(p+\gamma)+1]. \end{aligned} \tag{14}$$

The order $O(\varepsilon)$ part of $h(\gamma+\varepsilon\delta\gamma) - h(\gamma)$ is given by (see Stone & Goldbart (2008) for further insights)

$$\delta h(\gamma) = \int_{-\infty}^{\infty} \frac{\delta h(\gamma)}{\delta\gamma(x)}\delta\gamma(x)\mathrm{d}x. \tag{15}$$

Plugging (15) into (15) yields

$$\begin{aligned}
\delta h(\gamma) &= -\int_{-\infty}^{\infty} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x \\
&= -\int_{-\infty}^{0} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x - \int_{0}^{\infty} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x \\
&= -\int_{-\infty}^{0} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x + \int_{0}^{-\infty} [\log(p(-x)+\gamma(-x))+1]\delta\gamma(-x)\mathrm{d}x \\
&= -\int_{-\infty}^{0} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x - \int_{-\infty}^{0} [\log(p(-x)+\gamma(-x))+1]\delta\gamma(-x)\mathrm{d}x \\
&= -\int_{-\infty}^{0} [\log(p(x)+\gamma(x))+1]\delta\gamma(x)\mathrm{d}x + \int_{-\infty}^{0} [\log(p(-x)-\gamma(x))+1]\delta\gamma(x)\mathrm{d}x \\
&= -\int_{-\infty}^{0} \log\left(\frac{p(x)+\gamma(x)}{p(-x)-\gamma(x)}\right)\delta\gamma(x)\mathrm{d}x,
\end{aligned}$$

where we have applied a change of variable $x \to -x$ in the second integral of the third line and assumed $\delta\gamma(-x) = -\delta\gamma(x)$ in the fifth line. In particular,

$$\delta h(\gamma)|_{\gamma=0} = -\int_{-\infty}^{0} \log\left(\frac{p(x)}{p(-x)}\right)\delta\gamma(x)\mathrm{d}x.$$

Using the assumption that $p(x) \geq p(-x)$ whenever $x < 0$, we consequently obtain

$$\delta h(\gamma)|_{\gamma=0} \leq 0, \tag{16}$$

if $\delta\gamma(x) > 0$ for all $x < 0$. In effect, we have just proved the following proposition:

Proposition 6 *Given that $\gamma(-x) = -\gamma(x)$, $\int \gamma(x)\mathrm{d}x = 0$, $\gamma(|x|) \leq 0$, $p(|x|) \leq p(x)$ and $|\gamma(x)| < p(x)$, then the entropy of the forecast density $f_+(x) = p(x)+\gamma(x)$ is lower than that of the forecast density $f_-(x) = p(x) - \gamma(x)$.*

Propositions 5 and 6 imply that the logarithmic scoring rule prefers the forecast density that is less informative, which is in agreement with the categorical case considered in the previous section.

Proposition 7 *Given two forecasts $f_i(x) = p(x) + \gamma_i(x)$, $i = 1, 2$, with (i) $|\gamma_1(x)| < |\gamma_2(x)|$, (ii) $\gamma_i(|x|) \leq 0$, (iii) $\gamma_i(-x) = -\gamma_i(x)$, (iv) $|\gamma_i(x)| \leq p(x)$ and (v) $p(|x|) \leq p(x)$, then the logarithmic scoring rule prefers forecast $f_1(x)$ over forecast $f_2(x)$.*

Proof: To prove the above proposition, we consider the functional derivative of the expected logarithmic scoring, $\mathbb{E}[IGN] = \int_{-\infty}^{\infty} p(x)\log(p(x)+\gamma(x))\mathrm{d}x$. The functional derivative with respect to $\gamma(x)$ is

$$\frac{\delta}{\delta\gamma}\mathbb{E}[IGN] = \frac{-p(x)}{p(x)+\gamma(x)}.$$

Using this result, we obtain the first variation of $\mathbb{E}[IGN]$ as

$$\begin{aligned}
\delta\mathbb{E}[IGN] &= \int_{-\infty}^{\infty} \frac{\delta\mathbb{E}[IGN]}{\delta\gamma(x)}\delta\gamma(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{-p(x)}{p(x)+\gamma(x)}\delta\gamma(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)+\gamma(x)}\delta\gamma(x)\mathrm{d}x + \int_{0}^{\infty} \frac{-p(x)}{p(x)+\gamma(x)}\delta\gamma(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)+\gamma(x)}\delta\gamma(x)\mathrm{d}x + \int_{0}^{-\infty} \frac{p(-x)}{p(-x)+\gamma(-x)}\delta\gamma(-x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)+\gamma(x)}\delta\gamma(x)\mathrm{d}x + \int_{-\infty}^{0} \frac{p(-x)}{p(-x)-\gamma(x)}\delta\gamma(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \left[\frac{p(-x)}{p(-x)-\gamma(x)} - \frac{p(x)}{p(x)+\gamma(x)}\right]\delta\gamma(x)\mathrm{d}x \\
&\geq \int_{-\infty}^{0} \left[\frac{p(x)}{p(x)-\gamma(x)} - \frac{p(x)}{p(x)+\gamma(x)}\right]\delta\gamma(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{2p(x)\gamma(x)}{p^2(x)-\gamma^2(x)}\delta\gamma(x)\mathrm{d}x \\
&\geq 0,
\end{aligned}$$

provided $\delta\gamma(x) > 0$ for $x < 0$ and $\delta\gamma(-x) = -\delta\gamma(x)$, and this completes the proof.

We shall now consider two forecasts, $f_1(x) = p(x) + \gamma_1(x)$ and $f_2(x) = p(x) - \gamma_2(x)$ with $|\gamma_1(x)| \leq |\gamma_2(x)| \leq p(x)$. In this case, the quadratics scoring rule would prefer $f_1(x)$ over $f_2(x)$. In order to determine which forecast the logarithmic scoring would prefer, we consider the functional

$$\mathcal{H}(\gamma_1, \gamma_2) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)-\gamma_2(x)}{p(x)+\gamma_1(x)}\right) \mathrm{d}x. \tag{17}$$

Then the following proposition holds

Proposition 8 *Given that $|\gamma_1(x)| \leq |\gamma_2(x)| \leq p(x)$, there exists $\gamma*(x)$ satisfying the inequality $|\gamma^*(x)| \leq |\gamma_2(x)|$ such that (a) $\mathcal{H}(\gamma^*, \gamma_2) = 0$, (b) $\mathcal{H}(\gamma_1, \gamma_2) > 0$ for $|\gamma^*| < |\gamma_1|$ and (c) $\mathcal{H}(\gamma_1, \gamma_2) < 0$ for $|\gamma^*| < |\gamma_1|$.*

Proof: It is helpful to first note that Proposition 5 implies that $\mathcal{H}(\gamma_2, \gamma_2) > 0$ when $\gamma_2 \neq 0$.

Thinking of $\gamma_1(x)$ as fixed, the first variation of $\mathcal{H}(\cdot,\gamma_2)$ with respect to $\gamma_2(x)$ is given by

$$\begin{aligned}
\delta\mathcal{H}(\cdot,\gamma_2) &= \int_{-\infty}^{\infty} \frac{\delta\mathcal{H}(\cdot,\gamma_2)}{\delta\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{-p(x)}{p(x)-\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)-\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x + \int_{0}^{\infty} \frac{-p(x)}{p(x)-\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)-\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x - \int_{0}^{-\infty} \frac{p(-x)}{p(-x)+\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-p(x)}{p(x)-\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x + \int_{-\infty}^{0} \frac{p(-x)}{p(-x)+\gamma_2(x)}\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \left[\frac{p(-x)}{p(-x)+\gamma_2(x)} - \frac{p(x)}{p(x)-\gamma_2(x)}\right]\delta\gamma_2(x)\mathrm{d}x \\
&\leq \int_{-\infty}^{0} \left[\frac{p(x)}{p(x)+\gamma_2(x)} - \frac{p(x)}{p(x)-\gamma_2(x)}\right]\delta\gamma_2(x)\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{-2p(x)\gamma_2(x)}{p^2(x)-\gamma_2^2(x)}\delta\gamma_2(x)\mathrm{d}x.
\end{aligned}$$

Hence, $\delta\mathcal{H}(\cdot,\gamma_2) \leq 0$ provided $\delta\gamma_2 > 0$. It follows that $\mathcal{H}(\cdot,\gamma_2)$ has a maximum when $\gamma_2 = 0$, i.e. $\mathcal{H}(\cdot,\gamma_2) \leq \mathcal{H}(\cdot,0)$. In particular, $\mathcal{H}(0,\gamma_2) \leq \mathcal{H}(0,0) = 0$. For $\gamma_2 \neq 0$, we have the strict inequality, $\mathcal{H}(0,\gamma_2) < 0$. But, $\mathcal{H}(\gamma_2,\gamma_2) > 0$. Therefore, continuity implies that $\mathcal{H}(\gamma_1,\gamma_2) = 0$ for some $\gamma_1(x) = \gamma^*(x)$ such that $|\gamma^*| < |\gamma_2|$, and this completes the proof.

3.3 Continuous Ranked Probability Score

Finally, we consider the *Continuous Ranked Probability Score* (CRPS) of the density forecast $f(x)$ whose cumulative distribution is $F(x)$. The CRPS is a function of F and the verification X and is defined by (Gneiting & Raftery, 2007)

$$\mathrm{CRPS}(F,X) = \int_{-\infty}^{\infty} (F(y) - \mathbb{I}\{y \geq X\})^2 \mathrm{d}y. \tag{18}$$

Its associated entropy function is (Gneiting & Raftery, 2007)

$$G(F) = \int_{-\infty}^{\infty} F(y)(1-F(y))\mathrm{d}y \tag{19}$$

and its divergence function is

$$d(P,F) = -\int_{-\infty}^{\infty} (P(y)-F(y))^2 \mathrm{d}y, \tag{20}$$

where P is the true cumulative distribution function, i.e. $P(x) = \int_{-\infty}^{x} p(\tau)\mathrm{d}\tau$. Again, we consider the case when $p(|x|) \leq p(x)$. Since $\mathbb{E}[\mathrm{CRPS}(F,X)] = G(F) - d(P,F)$, it follows from (19) and (20) that

$$\mathbb{E}[\mathrm{CRPS}(F,X)] = \int_{-\infty}^{\infty} F(y)(1-F(y))\mathrm{d}y + \int_{-\infty}^{\infty} (P(y)-F(y))^2 \mathrm{d}y. \tag{21}$$

Define $F_{\pm}(x) = P(x) \pm \Gamma(x)$, where $\Gamma(x) = \int_{-\infty}^{x} \gamma(\tau)\mathrm{d}\tau$. It can be shown that $\gamma(-x) = -\gamma(x)$ implies that $\Gamma(-x) = \Gamma(x)$. If we now define $\Delta G(F_{\pm}) = G(F_+) - G(F_-)$, then the following proposition holds:

Proposition 9 *If $\delta\gamma(x) \geq 0$ for every $x < 0$ and $\delta\gamma(-x) = -\delta\gamma(x)$, then the first variation of $\Delta G(F_\pm)$ is non-negative, i.e.*

$$\delta\{\Delta(F_\pm)\} \leq 0.$$

Proof: The functional derivative of $\Delta G(F_\pm)$ with respect to $\Gamma(x)$ is

$$\frac{\delta\{\Delta G(F_\pm)\}}{\delta\Gamma(x)} = 2 - 4P(x).$$

It follows that the first variation of $\Delta G(F_\pm)$ is given by

$$\begin{aligned}
\delta\{\Delta G(F_\pm)\} &= 2\int_{-\infty}^{\infty}\{1-2P(x)\}\delta\Gamma(x)\mathrm{d}x \\
&= 2\int_{-\infty}^{0}\{1-2P(x)\}\delta\Gamma(x)\mathrm{d}x + 2\int_{0}^{\infty}\{1-2P(x)\}\delta\Gamma(x)\mathrm{d}x \\
&= 2\int_{-\infty}^{0}\{1-2P(x)\}\delta\Gamma(x)\mathrm{d}x - 2\int_{0}^{-\infty}\{1-2P(-x)\}\delta\Gamma(-x)\mathrm{d}x \\
&= 2\int_{-\infty}^{0}\{1-2P(x)\}\delta\Gamma(x)\mathrm{d}x + 2\int_{-\infty}^{0}\{1-2P(-x)\}\delta\Gamma(x)\mathrm{d}x \\
&= 4\int_{-\infty}^{0}\{1-P(x)-P(-x)\}\delta\Gamma(x)\mathrm{d}x
\end{aligned}$$

Since $p(|x|) \leq p(x)$ implies that $P(-x) \geq 1 - P(x)$ whenever $x < 0$, it follows that $\delta\{\Delta(F_\pm)\} \leq 0$.

In consequence of the foregoing proposition, the following lemma holds:

Lemma 1 *The continuous rank probability score prefers the forecast distribution F_+ over F_-, i.e.* $\mathbb{E}[\mathrm{CRPS}(F_+, X)] \leq \mathbb{E}[\mathrm{CRPS}(F_-, X)]$.

Proof: Introduce the notation $\mathbb{E}[\mathrm{CRPS}]_\pm = \mathbb{E}[\mathrm{CRPS}(F_+, X)] - \mathbb{E}[\mathrm{CRPS}(F_-, X)]$. We then have

$$\mathbb{E}[\mathrm{CRPS}]_\pm = \Delta G(F_\pm) + ||\Gamma||_2^2.$$

Hence $\mathbb{E}[\mathrm{CRPS}]_\pm$ decreases (resp. increases) if and only if $\Delta G(F_\pm)$ decreases (resp. increases). In particular, using Proposition 9, $\Gamma = 0$, $\delta\gamma(x) \geq 0$ for $x < 0$ and $\delta\gamma(-x) = -\delta\gamma(x)$ results in decrease of $\mathbb{E}[\mathrm{CRPS}]_\pm$, which proves the Lemma.

According to Proposition 6, the entropy of F_+ is lower than that of F_-. Hence we see that, unlike the logarithmic scoring rule, the CRPS prefers the forecast distribution that is more informative (or more confident).

As a final remark, we note that the second term in the expectation of CRPS somewhat resembles the *mean squared error criterion* discussed in Corradi & Swanson (2006). The mean squared error of the forecast $F(x)$ is given by $\int p(x)\Gamma^2(x)\mathrm{d}x$. Evidently, just like the quadratic scoring rule, the mean squared error does not distinguish between the two forecasts, $f_\pm(x)$.

4 Discussion

This manuscript contrasted how certain popular scoring rules would rank competing forecasts of specified departures from the target distribution. In the categorical case, we considered the Brier Score and the logarithmic scoring rule, focusing on the binary case. Given two forecasts whose errors from the target distribution differ only by the sign, we found that the logarithmic scoring rule prefers the higher entropy forecast. Preferring the higher entropy forecast may be thought of as taking a more cautious stance of less confidence. The logarithmic scoring rule selects a lower entropy forecast only if it is nearer to the target distribution in the sense of the L^2 norm.

We extended the investigation from the binary forecasts to the continuous case, where we considered the Quadratic score, Logarithmic score and the Continuous Ranked Probability Score (CRPS). Just like the Brier score in the binary case, the Quadratic Score does not distinguish between forecasts with equal L^2 norms of their error from the target distribution. Given two density forecasts whose errors from the target forecast differ by a sign, the logarithmic scoring rule prefers the distribution with higher entropy. On the other hand, the CRPS prefers the forecast distribution with lower entropy; bear in mind that lower entropy corresponds to more confidence (Shannon, 1948).

Our findings indicate that the logarithmic scoring rule encourages a more cautious decision when forecasts depart from the ideal forecast. This is in agreement with the idiom that we should "err on the side of caution." We consider this to be an advantage over the CRPS which ecourages erring on the side of risk. In an investment scenario, erring on the side of risk can result in substantial losses. Some have critised the logarithmic scoring rule for placing a heavy penalty on assigning zero probability to events that materialise (e.g. Boero et al., 2011; Gneiting & Raftery, 2007); but assigning zero probability to events that are possible is also discouraged by *Laplace's rule of succession* (Jaynes, 2003). The logarithmic scoring rule is good at highlighting misplaced confidence of forecasts. Such forecasts may have to be dealt with appropriately. One way of dealing with over-confident forecasts is to apply shrinkage estimators discussed in Casella (1985); Efron & Morris (1977).

Acknowledgements

This work was supported by the RCUK Digital Economy Programme via EPSRC grant EP/G065802/1 The Horizon Digital Economy Hub.

References

Boero, G., Smith, J. & Wallis, K. F. (2011). Scoring rules and survey density forecasts. *International Journal of Forecasting* **27**, 379–393.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Brocker, J. & Smith, L. A. (2007). Scoring Probabilistic Forecasts: The importance of Being proper. *Weather and Forecasting* **22**, 382–388.

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *The American Statistician* **39**, 83–87.

Corradi, V. & Swanson, N. R. (2006). Predictive density evaluation. In *Handbook for Econometric Forecasting*, G. Elliott, C. W. J. Granger & A. Timmermann, eds., vol. 1. North-Holland, pp. 197–284.

Dawid, A. P. (1984). Present position and potential developments: Some Personal Views: Statistical Theory: The Prequencial Approach. *J. R. Statist. Soc. A* **147**, 278–292.

Diebold, F. X., Gunther, T. A. & Tay, A. S. (1998). Evaluating density forecasts with application to financial risk management. *International Economics Review* **39**, 863–883.

Efron, B. & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American* **236**, 119–127.

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology* **8**, 985–987.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *J. Amer. Math. Soc.* **102**, 359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**, 107–114.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Roulston, M. S. & Smith, L. A. (2002). Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review* **130**, 1653–1660.

Shannon, C. E. (1948). A Mathematical theory of communication. *Bell Systems Technology Journal* **27**, 379–423,623–656.

Stone, M. & Goldbart, P. (2008). *Mathematics for Physics I*. PIMANDER-CASAUBON.