Proper Scoring Rules for Interval Probabilistic Forecasts

K. Mitchell[a*] and C.A.T. Ferro[a]

[a]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK*

[*]Correspondence to: K. Mitchell, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK. E-mail: km411@exeter.ac.uk

Interval probabilistic forecasts for a binary event are forecasts issued as a range of probabilities for the occurrence of the event, for example, 'chance of rain: 10-20%'. To verify interval probabilistic forecasts, use can be made of a scoring rule that assigns a score to each forecast-outcome pair. An important requirement for scoring rules, if they are to provide a faithful assessment of a forecaster, is that they be proper, by which is meant that they direct forecasters to issue their true beliefs as their forecasts. Proper scoring rules for probabilistic forecasts issued as precise numbers have been studied extensively. But, applying such a proper scoring rule to, for example, the mid-point of an interval probabilistic forecast, does not, typically, produce a proper scoring rule for interval probabilistic forecasts. Complementing parallel work by other authors, we derive a general characterisation of scoring rules that are proper for interval probabilistic forecasts and from this characterisation we determine particular scoring rules for interval probabilistic forecasts that correspond to the familiar scoring rules used for probabilistic forecasts given as precise probabilities. All the scoring rules we derive apply immediately to rounded probabilistic forecasts, being a special case of interval probabilistic forecasts.

Key Words: interval probabilistic forecasts; rounded probabilistic forecasts; forecast verification; proper scoring rules

Received . . .

1. Introduction

Consider an event that can have one of two outcomes. When forecasting which outcome will occur, the word 'forecast' is often read as 'point forecast', a statement about what the outcome of the event will be. One may though, also speak of a 'probabilistic forecast', a statement about how likely it is that each outcome will occur. Probabilistic forecasts are not new (see the historical account by Murphy 1998) and, already familiar in meteorology, are of increasing interest in many other disciplines (for a broad map of applications, see Gneiting and Katzfuss 2014). Studies

of how well a probabilistic forecaster performs, which is the subject of probabilistic forecast verification, have up to now as far as we are aware, taken the forecast probability to be a precise number; we will refer to such probabilistic forecasts as *precise probabilistic forecasts* (a thorough overview of this type of probabilistic forecasting is given in Dawid 1986).

Yet a probabilistic forecast is often expressed as a range of probabilities (for example, "Chance of rain: 25-30%"). We assume that the forecaster can compute their forecast probability precisely but must issue a range of probabilities. For example, meteorological offices around the world communicate their forecasts for precipitation as ranges of probabilities. We call a probabilistic forecast issued as a range of probabilities, an *interval probabilistic forecast*.

Rounded probabilistic forecasts are a special case of interval probabilistic forecasts. Each rounded probability represents a range of probabilities, namely those probabilities that, when rounded, reduce to the forecast probability. For example, if probabilistic forecasts are rounded to the nearest 10%, a rounded probabilistic forecast of 20% can be represented as the interval of probabilities from 15% (inclusive) to 25% (exclusive).

To verify precise probabilistic forecasts, the standard formal approach is to use a scoring rule (see for example, Winkler 1996), a rule that assigns to each possible outcome of the event and each (precise) probabilistic forecast of the event, a score. A forecaster's accuracy is measured by their average score. There are many scoring rules from which to choose when calculating a forecaster's accuracy. There are no prescriptions about which rule should be chosen, but, the scoring rule used must satisfy the condition of being *proper*. A scoring rule is proper if a forecast matching the forecaster's actual judgment about the event's outcomes will optimise the score the forecaster expects to receive; a scoring rule is strictly proper *only if* a forecast reflecting the forecaster's actual judgment about the event's outcomes will optimise the forecaster's expected score (see Murphy and Epstein 1967).

Consider the following setting. Suppose that X is 1 if it rains tomorrow and 0 otherwise. Let the precise probability q be the forecaster's actual belief that it will rain tomorrow. The forecaster issues the probabilistic forecast p (which may or may not equal q). A scoring rule, S, assigns to each precise forecast probability p and each value x of X a score $S(p, x)$. Before we know the value of X, the forecaster can compute their expected score (with respect to their actual belief q) when they issue the precise forecast p. We denote this expected score by $S[p, q] = \mathbb{E}_q[S(p, X)]$. We assume that S is negatively oriented, that is, lower values of S are better (Winkler and Murphy 1968). With this assumption, the scoring rule, S, is said to be a proper scoring rule if $S[q, q] \leq S[p, q]$ for all p and q and strictly proper only if $S[q, q] < S[p, q]$ when $p \neq q$. For precise probabilistic forecasts there are many well-known proper scoring rules from which to choose (see for example, Gneiting and Raftery 2007). For ease of reference, we shall refer to proper scoring rules for precise probabilistic forecasts as *precise-proper scoring rules*.

Impropriety gives the forecaster the opportunity to hedge: obtain better accuracy by publishing forecasts that differ from their actual judgments, and, in allowing such dissemblance, impropriety undermines the credibility of the forecasts. Consider, for example, the apparently reasonable absolute error scoring rule (Murphy and Epstein 1967), $S(p, X) = |p - X|$. The forecaster will receive a score of $|p - 1|$ if it does rain tomorrow and a score of $|p|$ if it does not rain tomorrow; a lower score is a better score (p being closer to the outcome of X). The expected score of the forecaster is $S[p, q] = \mathbb{E}_q[|p - X|] = p + q - 2pq$. It is then evident that if the forecaster's true belief is $q = \frac{1}{2}$, $S[p, q] = \frac{1}{2}$, so the forecaster will receive the same expected score no matter what value they issue for p. Similarly, if $q < \frac{1}{2}$ the forecaster will receive the best (i.e. lowest) expected score by issuing $p = 0$. And if $q > \frac{1}{2}$, the forecaster will receive the best expected score by issuing $p = 1$. The published probabilistic forecasts will then always be either 0 or 1 (or, if $q = \frac{1}{2}$, an arbitrary value) and do not represent the forecaster's true views (unless the forecaster is always certain about whether there will be rain tomorrow i.e. $q = 0$ or $q = 1$).

Maintaining the above setting, suppose that the forecaster can articulate their precise true belief, q, that $X = 1$, but must issue an interval of probabilities that $X = 1$. Let $0 = a_0 < a_1 < \ldots < a_{n-1} < a_n = 1$ be a partition of the interval $[0, 1]$, with subintervals $I_1 = [a_0, a_1]$ and $I_i = (a_{i-1}, a_i]$ for $i = 2, \ldots, n$. An interval probabilistic forecast is the selection of I_i for some $0 < i \leq n$. A scoring rule, s, for such an interval probabilistic forecast, gives a value $s(I_i, x)$ when the value of X is x. Having issued the interval I_i, the forecaster's expected score, with respect to their actual (and precise) belief q that $X = 1$, is denoted $s[I_i, q] = \mathbb{E}_q[s(I_i, X)]$. A scoring rule for interval probabilistic forecasts is proper if the interval containing q optimises the expected score and is strictly proper if the only interval that optimises the forecaster's expected score is the interval that contains q. Assuming that lower values of s indicate better scores, we say, formally,

Definition 1.1 *(Propriety for Interval Probabilistic Forecasts) Let $X \in \{0, 1\}$, be a random variable, and let the probability q be the forecaster's actual belief that $X = 1$. The scoring rule, s, is defined to be proper if $s[I_i, q] \leq s\big[I_j, q\big]$ for all i, j and $q \in I_i$; s is strictly proper only if $s[I_i, q] < s\big[I_j, q\big]$ for all i, j and $q \in I_i$, $q \notin I_j$.*

We refer to scoring rules that are proper for interval probabilistic forecasts as *interval-proper scoring rules*.

Given a precise-proper scoring rule S, there are many possible ways of constructing an interval scoring rule, s, from S (e.g. maximum of S over an interval, average of S over an interval). However, an illustration in the next section shows that even when S is precise-proper and for each i, $s(I_i, X)$ is defined simply as the value of S at the mid-point of I_i, s need not be an interval-proper scoring rule. In response to this difficulty, we present in section 3 a general expression for any interval-proper scoring rule. This result is a special case of more general results that have been proved by Lambert *et al.* (2008); Lambert and Shoham (2009); Lambert (2013) and Frongillo and Kash (2014). But, their results, while powerful, are abstract and this has prompted us to offer a short new proof of the characterisation of interval-proper scoring rules for events with only two outcomes. From this general expression, we derive particular interval-proper scoring rules that are analogues of some familiar precise-proper scoring rules. In section 4 we demonstrate the effects of using improper scoring rules for interval probabilistic forecasts, with verification studies based on probability of precipitation (PoP) forecasts issued by the Australian Bureau of Meteorology and the United Kingdom Meteorological Office. Section 5 concludes. Proofs appear in the appendices.

2. An Illustration

We ask whether, at a particular time in the future, an event will occur (e.g. will it rain tomorrow?). Let X be a random variable that will take the value 0 if the event does not occur (e.g. no rain tomorrow) and 1 if the event does occur (e.g. rain tomorrow). A precise probabilistic forecast for X is a statement of the precise value for the probability that $X = 1$ (e.g. "chance of rain tomorrow, 0.2 (20%)"); such a value lies in the interval $[0, 1]$. An interval probabilistic forecast is a statement that the probability that $X = 1$ lies in a subinterval of $[0, 1]$ (e.g. "chance of rain tomorrow, 0.15-0.25 (15-25%)).

To evaluate a precise probabilistic forecast, choose the Brier scoring rule (Brier 1950), S, defined by $S(p, x) = (p - x)^2$ where x is the observed value of X and p is the precise probabilistic forecast that $X = 1$; S is negatively-oriented. It is known (Murphy and Epstein 1967) that the Brier scoring rule is proper, that is $S[q, q] \leq S[p, q]$ for all values of $p, q \in [0, 1]$, where $S[p, q] = \mathbb{E}_q[S(p, X)] = p^2 - 2pq + q$.

Suppose that the forecaster does not issue the precise probabilistic forecast p, but issues an interval I_i. A scoring rule, s, for an interval probabilistic forecast might be defined by

$$s(I_i, X) = S(\hat{p}_i, X)$$

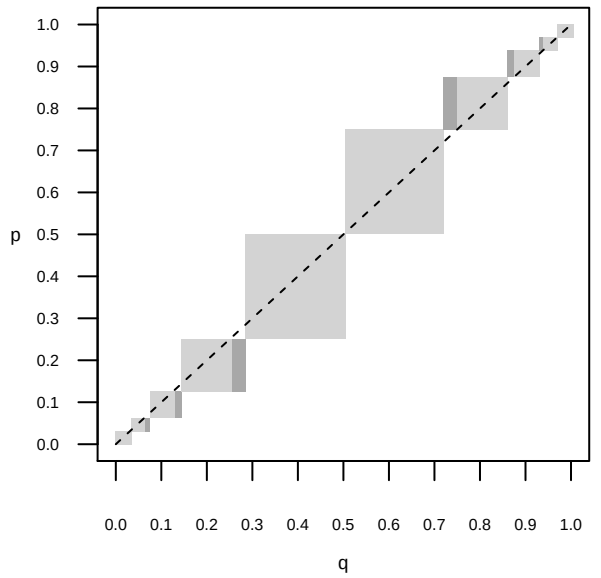
where $\hat{p}_i = \frac{1}{2}(a_{i-1} + a_i)$, is the mid-point of I_i. We shall refer to the resulting scoring rule, $S(\hat{p}, X)$, as the mid-point Brier scoring rule.

The following proposition gives conditions under which the mid-point Brier scoring rule is proper.

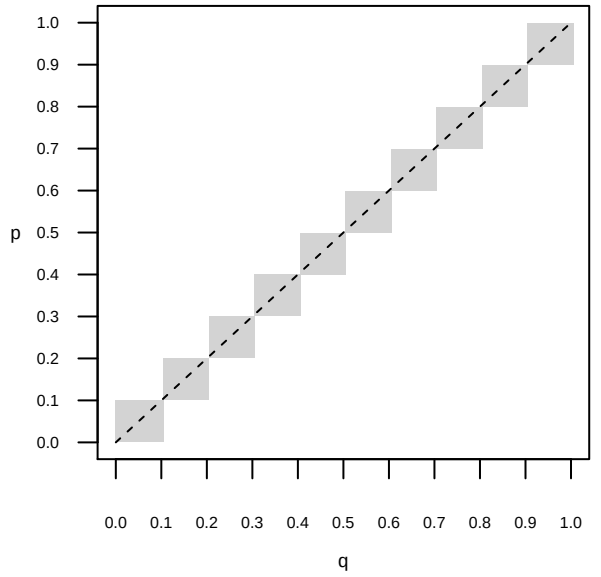
Proposition 2.1 *The mid-point Brier scoring rule is interval-proper if and only if the a_i are equally-spaced (i.e. $a_i = i/n$, for all $0 \leq i \leq n$).*

Proof. See Appendix A. ■

It is immediate that, for an unequally-spaced partition, the mid-point Brier scoring rule is not an interval-proper scoring rule. Under an equally-spaced partition, the interval probabilistic forecast will include the forecaster's true belief, q, that $X = 1$, but under an unequally-spaced partition the forecaster will find it advantageous, for some values of q, to hedge and issue an interval probabilistic forecast that does not contain q. For example, in figures 1a and 1b, the horizontal axis is the forecaster's true belief, q; the vertical axis shows the forecaster's precise probabilistic forecast, p. The forecaster must issue as their forecast an interval from the partition $0 = a_0 < a_1 < \ldots < a_n = 1$. The chosen interval is the interval I_i, at which the expected mid-point Brier score, $S[\hat{p}_i, q] = \hat{p}_i^2 - 2\hat{p}_i q + q$, is a minimum. In each figure, the forecast interval is displayed and coloured dark-grey if the interval does not contain q, or light-grey if the interval does contain q. The mid-point Brier scoring rule is proper if and only if there are no dark-grey intervals, as is the case in Figure 1b where the partition has equal spacing.



(a)



(b)

Figure 1. For each value of q, the interval probabilistic forecast that gives the lowest expected mid-point Brier score is shown. If q does not lie in this interval (indicating impropriety), the interval is coloured dark-grey, otherwise the interval is coloured light-grey. The mid-point Brier scoring rule is, therefore, proper if and only if there are no dark-grey intervals. (a) Unequally-spaced partition: $a_0 = 0 < \frac{1}{32} < \frac{1}{16} < \frac{1}{8} < \frac{1}{4} < \frac{1}{2} < \frac{3}{4} < \frac{7}{8} < \frac{15}{16} < \frac{31}{32} < 1 = a_{10}$ (b) Equally-spaced partition: $a_i = i/10$, $i = 0, \ldots, 10$.

3. A General Result

3.1. Characterisation Theorem

We would like to be able to write down the general form of those scoring rules that are proper for interval probabilistic forecasts. Interval probabilistic forecasts are a particular example of the wider class of statistical functionals (see for example, Gneiting 2011). Recently, Lambert *et al.* (2008); Lambert and Shoham (2009); Lambert (2013) and Frongillo and Kash (2014) have derived a general expression for scoring rules that are proper for statistical functionals (scoring rules that are proper for some particular statistical functionals are given in Gneiting (2011)). To arrive at a form for scoring rules that are proper for interval probabilistic forecasts, we can therefore, contextualise these general results, in particular those of Lambert (2013), to our setting. With this indirect approach, however, we risk being opaque. Moreover, for interval forecasts of a binary random variable, it is possible to give a straightforward derivation of the functional form that an interval-proper scoring rule must have,

and this we now do. The reader inclined more to application may move immediately to Theorem 3.1.

As in the previous section, X is a random variable taking only the values 0 and 1, for which the forecaster issues an interval probabilistic forecast, I_i, for some $0 < i \leq n$. What is the general expression for the strictly interval-proper scoring rule s?

Recalling that the expected value of $s(I_i, X)$ when the probability that $X = 1$ is q, is defined by

$$\begin{aligned} s[I_i, q] &= \mathbb{E}_q[s(I_i, X)] \\ &= s(I_i, 0)(1-q) + s(I_i, 1)q \end{aligned} \tag{1}$$

the propriety of s gives

$$s[I_k, q] \leq s\big[I_j, q\big] \qquad \text{for all } k, j \text{ and } q \in I_k. \tag{2}$$

The condition (2) must be satisfied for every $q \in I_k$ and, in particular, for $q = a_k$. Therefore, letting $j = k+1$ and $q = a_k$, we have

$$s[I_k, a_k] \leq s[I_{k+1}, a_k]\,. \tag{3}$$

From the strict propriety of s,

$$s[I_{k+1}, q] < s[I_k, q] \quad \forall q \in I_{k+1}.$$

By equation (1), for each i, $s[I_i, q]$ is a continuous function of q (a reasonable property: a forecaster who changes their true belief, q, by a small amount should not wish their expected score to change substantially). The continuity of $s[I_i, q]$ in q for all i gives, in particular, $\lim_{q \to a_j^+} s[I_i, q] = s\big[I_i, a_j\big]$, $\forall i, j$. This smoothness condition coupled with strict propriety gives

$$\begin{aligned} s[I_k, a_k] &= \lim_{q \to a_k^+} s[I_k, q] \\ &\geq \lim_{q \to a_k^+} s[I_{k+1}, q] \\ &= s[I_{k+1}, a_k]\,. \end{aligned} \tag{4}$$

Consequently, from (3) and (4), we have

$$s[I_k, a_k] = s[I_{k+1}, a_k]\,. \tag{5}$$

Using (1), equation (5) may be written as

$$\begin{aligned} &\big\{s(I_k, 0) - s(I_{k+1}, 0)\big\}(1 - a_k) \\ &\qquad + \big\{s(I_k, 1) - s(I_{k+1}, 1)\big\}a_k = 0 \end{aligned} \tag{6}$$

and this must hold for every $k = 1, \ldots, n-1$.

One possible solution to (6) is the trivial solution $s(I_k, X) = 0$ for all values of k and X. But such a solution violates the condition of strict propriety: suppose that $i < j$ and choose $q \in I_i$; by strict propriety we should have $s[I_i, q] < s\big[I_j, q\big]$, but because $s(I_k, X) = 0$ for all k and X we have $s[I_i, q] = s\big[I_j, q\big]$, a contradiction. So, the trivial solution is inadmissible.

Excluding the trivial solution, for each $k = 1, \ldots, n-1$, the solution must then have the form,

$$\begin{aligned} s(I_k, 0) - s(I_{k+1}, 0) &= -a_k \gamma_k \\ s(I_k, 1) - s(I_{k+1}, 1) &= (1 - a_k)\gamma_k \end{aligned} \tag{7}$$

where γ_k is a constant. We now show that γ_k is non-negative. For $k > 1$, from the propriety of s we have that $s[I_k, q] \leq s[I_{k+1}, q]$ for all $q \in I_k = (a_{k-1}, a_k]$, which together with the smoothness of s, gives

$$s[I_k, a_{k-1}] = \lim_{q \to a_{k-1}^+} s[I_k, q]$$
$$\leq \lim_{q \to a_{k-1}^+} s[I_{k+1}, q] = s[I_{k+1}, a_{k-1}]\,.$$

For $k = 1$, $I_1 = [a_0, a_1]$ and the propriety of s alone gives $s[I_1, a_0] \leq s[I_2, a_0]$. So, for all k,

$$s[I_k, a_{k-1}] \leq s[I_{k+1}, a_{k-1}]\,. \tag{8}$$

Applying (1) to $s[I_k, a_{k-1}]$ and $s[I_{k+1}, a_{k-1}]$ in (8) and rearranging the terms in the inequality,

$$\big\{s(I_k, 0) - s(I_{k+1}, 0)\big\}(1 - a_{k-1}) + \big\{s(I_k, 1) - s(I_{k+1}, 1)\big\}a_{k-1} \leq 0.$$

Substituting from (7) gives

$$\begin{aligned} & -a_k\gamma_k(1 - a_{k-1}) + (1 - a_k)\gamma_k a_{k-1} \leq 0 \\ \Leftrightarrow \quad & \gamma_k(a_{k-1} - a_k) \leq 0 \\ \Leftrightarrow \quad & \gamma_k \geq 0. \end{aligned}$$

We can, therefore, write

$$s(I_k, X) - s(I_{k+1}, X) = \gamma_k(X - a_k) \quad \text{for } k = 1, \ldots, n-1 \tag{9}$$

for non-negative constants γ_k. The difference equation (9) has a solution

$$s(I_k, X) = f(X) - \sum_{i=1}^{k-1} \gamma_i(X - a_i) \tag{10}$$

with f an arbitrary function of X. Defining the function g by $g(i) - g(i-1) = \gamma_i$, $i = 1, \ldots, n-1$, we have proved the following theorem

Theorem 3.1 *(Characterisation for Interval-Proper Scoring Rules) Let $X \in \{0, 1\}$ be a future binary observation. Given a partition $0 = a_0 < a_1 < \ldots < a_{n-1} < a_n = 1$, let s be a strictly interval-proper scoring rule for interval probabilistic forecasts $I_1 = [a_0, a_1]$ and $I_k = (a_{k-1}, a_k]$ for $k = 2, \ldots, n$ of the outcome $X = 1$. Then s has the form*

$$s(I_k, X) = f(X) - \sum_{i=1}^{k-1} \big(g(i) - g(i-1)\big)(X - a_i) \tag{11}$$

where f is an arbitrary function and g is a non-decreasing function.

Note that under s given by equation (11), interval probabilistic forecasts that are closer to the outcome for X receive a lower (that is, better) score than interval probabilistic forecasts that are further from the outcome for X. Suppose that $X = 0$. We have

$$s(I_k, 0) = f(0) + \sum_{i=1}^{k-1} \big(g(i) - g(i-1)\big)a_i$$

and the summation term increases as k increases (g being a non-decreasing function) so that as I_k moves further away from X (as k increases) $s(I_k, 0)$ increases. Similarly, if $X = 1$,

$$s(I_k, 1) = f(1) - \sum_{i=1}^{k-1} \big(g(i) - g(i-1)\big)(1 - a_i)$$

and the summation term is always positive and increases in size as k increases so that $s(I_k, 1)$ increases as I_k moves away from X (as k decreases).

3.2. Choosing f and g

In equation (11), each choice for the function f and for the non-decreasing function g, will give a new proper scoring rule for interval probabilistic forecasts. How should the functions f and g be chosen? While *any* real-valued function may be chosen for f and *any* non-decreasing real-valued function may be chosen for g, it is helpful to have some method to guide these choices. Here we suggest one such method.

To begin, choose $\xi_k \in I_{k+1}$ for $k = 0, \ldots, n-1$ and define the function h by $h(\xi_k) = g(k)$. Replacing g by h in (11),

$$s(I_k, X) = f(X) - \sum_{i=1}^{k-1} \big(h(\xi_i) - h(\xi_{i-1})\big)(X - a_i) \qquad (12)$$

from which

$$s(I_{k+1}, X) - s(I_k, X) = \\ - (h(\xi_k) - h(\xi_{k-1}))(X - a_k). \qquad (13)$$

Restrict attention to those s for which, as n increases and all subintervals of the partition are made steadily smaller, the value of s for the interval containing p tends to the value of some precise-proper scoring rule S at p. Then (see Appendix B), for suitably smooth functions S and h, letting $n \to \infty$ in (13), gives

$$\frac{\partial S(p, X)}{\partial p} = \frac{dh(p)}{dp}(p - X). \qquad (14)$$

So, if we have a scoring rule, S, that is proper for precise probabilistic forecasts, we substitute for this scoring rule into the left-hand side of (14) and solve for h as a function of p; having done so, we set $g(k) = h\left(\xi_k\right)$ (for some predetermined choice for the ξ_k).

To interpret h, integrate both sides of (14) with respect to p to obtain

$$S(p, X) + a(X) = h(p)(p - X) - \int h(p)\, dp$$

where $a(\cdot)$ is a function of X alone. Taking the expectation in X under p gives

$$S[p, p] + \mathbb{E}_p[a(X)] = -\int h(p)\, dp.$$

With $X \in \{0, 1\}$, we can write $a(X) = a(0)(1 - X) + a(1)X$ so that $\mathbb{E}_p[a(X)] = a(0)(1 - p) + a(1)p$. The function $\mathrm{e}_S(p) = -S[p, p]$ is known as the entropy of p associated with S (Gneiting and Raftery 2007; Bröcker 2009). We have

$$\int h(p)\, dp = \mathrm{e}_S(p) - a(0)(1 - p) - a(1)p$$

from which, differentiating both sides with respect to p,

$$h(p) = \frac{d\mathrm{e}_S(p)}{dp} - (a(1) - a(0)). \qquad (15)$$

Equation (15) states that $h(p)$ is (up to a constant), the derivative of the entropy of p associated with S (we thank an anonymous referee for bringing this property of h to our attention and for suggesting that this property of h promises an interesting form for equation (12) in the limit, a form which we resolve in the next paragraph).

Lead by this interpretation of h, from equation (12), we have (where the indicator function $\mathbb{1}(\cdot)$ has the value 1 if its argument is true, and 0 otherwise)

$$\begin{aligned} s(I_k, X) &= f(X) - \sum_{i=1}^{k-1} \big(h(\xi_i) - h(\xi_{i-1})\big)(X - a_i) \\ &= f(X) - \sum_{i=1}^{n} (X - a_i)\mathbb{1}(a_i < a_k)\big(h(\xi_i) - h(\xi_{i-1})\big). \end{aligned} \qquad (16)$$

Allowing $n \to \infty$ in equation (16), we obtain

$$S(p, X) = f(X) - \int (X - q)\mathbb{1}(q < p) dh(q) \qquad (17)$$

which (for our choice of f, see below) is the Schervish-representation of a proper scoring rule for a binary event (Schervish (1989),Theorem 4.2,page 1861; see also Gneiting and Raftery (2007), page 364).

What of the function f? From equation (11) we have that

$$s\left(I_1, X\right) = f(X)$$

We choose $f(X) = S(\xi_0, X)$. This choice ensures that $s\left(I_1, X\right) \to S(0, X)$ as $n \to \infty$.

As examples of this method we take some familiar precise-proper scoring rules and derive the corresponding analogues that are interval-proper. In all cases, we assume that X takes only the values 0 and 1, the precise probabilistic forecast that $X = 1$ is

p and that the interval $[0,1]$ has n subintervals with end-points $0 = a_0 < a_1 < \ldots < a_n = 1$.

EXAMPLE *(Brier scoring rule (Brier 1950)).* The Brier scoring rule is $S(p,X) = (p-X)^2$. Substituting for S in (14), we have

$$-2(X-p) = (p-X)\frac{dh(p)}{dp}$$

giving $h(p) = 2p$. Identify points $\xi_k \in I_{k+1}$ for all $k = 0,\ldots,n-1$. Then $g(k) = h(\xi_k) = 2\xi_k$. Choose $f(X) = (\xi_0 - X)^2$.

With these choices of f and g, equation (11) gives the following Brier scoring rule for interval probabilistic forecasts

$$s(I_k,X) = (\xi_0 - X)^2 - \sum_{i=1}^{k-1}(2\xi_i - 2\xi_{i-1})(X - a_i)$$

which may be rewritten as

$$s(I_k,X) = (\xi_{k-1} - X)^2 - \sum_{i=1}^{k-1}\left\{(\xi_i - a_i)^2 - (\xi_{i-1} - a_i)^2\right\} \quad (18)$$

and the expected interval Brier score is

$$s[I_k,q] = q - 2q\xi_{k-1} + \xi_{k-1}^2 - \sum_{i=1}^{k-1}\left\{(\xi_i - a_i)^2 - (\xi_{i-1} - a_i)^2\right\}.$$

If we choose $\xi_k = \frac{1}{2}(a_k + a_{k+1})$, the mid-point of each subinterval, then

$$s(I_k,X) = \left(\frac{1}{2}(a_{k-1}+a_k) - X\right)^2 - \frac{1}{4}(a_k - a_{k-1})^2 + \frac{1}{4}a_1^2 \quad (19)$$

$$= (X - a_{k-1})(X - a_k) + \frac{1}{4}a_1^2. \quad (20)$$

Since propriety is preserved under translation, we define the *adjusted* interval-proper Brier scoring rule by

$$s(I_k,X) = (X - a_{k-1})(X - a_k). \quad (21)$$

Equation (19) also shows that when ξ_k is the mid-point of the $(k+1)$st interval, then, under *equally*-spaced subintervals,

$$s(I_k,X) = \left(\frac{1}{2}(a_{k-1}+a_k) - X\right)^2$$

which, from Proposition 2.1, is known to be proper.

□

EXAMPLE *(Ignorance scoring rule (Good 1952)).* The Ignorance scoring rule is defined by

$$S(p,X) = -X\log(p) - (1-X)\log(1-p)$$

for $p \in (0,1)$. Substituting into (14) gives

$$\frac{p-X}{p(1-p)} = (p-X)\frac{dh(p)}{dp}.$$

We have, therefore, that for $p \in (0,1)$, $h(p) = \log\{p/(1-p)\}$, from which $g(k) = h(\xi_k) = \log\{\xi_k/(1-\xi_k)\}$, for $0 \le k < n$. Choose $f(X) = S(\xi_0, X)$.

The expression for $s(I_k,X)$ may be written

$$s(I_k,X) = S(\xi_{k-1},X) - \sum_{i=1}^{k-1}\{S(\xi_i,a_i) - S(\xi_{i-1},a_i)\}$$

which is of the same form as equation (18) for the Brier scoring rule, although, for the ignorance scoring rule there is no apparent simplification similar to that by which equation (18) reduces to equation (20) for the Brier scoring rule.

□

EXAMPLE *(Pseudo-spherical scoring rule (Roby 1964)).* Fix $\alpha > 1$. The α-pseudo-spherical scoring rule is

$$S(p,X) = \frac{\{-Xp + (X-1)(1-p)\}^{\alpha-1}}{\{p^\alpha + (1-p)^\alpha\}^{\frac{\alpha-1}{\alpha}}}.$$

Replacing S in (14) gives

$$\frac{dh(p)}{dp} = \frac{(\alpha-1)\{(p-1)p\}^{\alpha-2}}{\{p^\alpha + (1-p)^\alpha\}^{2-\frac{1}{\alpha}}}.$$

Solving for h, we have

$$h(p) = \frac{p^{\alpha-1} - (1-p)^{\alpha-1}}{\{p^\alpha + (1-p)^\alpha\}^{\frac{\alpha-1}{\alpha}}}.$$

Set $g(k) = h(\xi_k)$ and choose $f(X) = S(\xi_0, X)$. (If $\alpha = 2$, the pseudo-spherical scoring rule is referred to as the spherical scoring rule.)

□

4. Consequences of Impropriety

Equation (11) presents the characteristic form that an interval-proper scoring rule must have. Yet it is unclear what the practical implications are if an improper interval scoring rule is used. In this section, we use actual precise probabilistic forecasts provided by two separate meteorological offices to construct hypothetical interval probabilistic forecasts when an improper interval scoring rule is in place. From these synthetic, yet representative, interval probabilistic forecasts, we can establish empirical measures of the influence of impropriety.

4.1. Data

Two separate data sets, in both cases precipitation data, were used. The amount of precipitation per day (the 24-hour period beginning at midnight local time) is converted into a binary variable, X, by choosing a threshold rainfall level (in mm) and defining $X = 1$ if the recorded amount of precipitation is greater than or equal to the threshold level; otherwise $X = 0$.

The UK Meteorological Office (UKMO) provided data for 58 lead-times (from 6 to 348 hours at 6-hourly intervals) and 2 locations; for each lead-time and location pair approximately two-years of daily data was available. For each day of each lead-time and location pair, the observation was a precipitation level (in mm) and the forecast was given as a set of nodes $(z_j, F(z_j))$ $j = 1, \ldots, m$ of the cumulative distribution function (F) of the precipitation level in mm (z), from which the precise probability of the precipitation level exceeding a threshold of 1mm was calculated; if necessary, the nodes were linearly interpolated and the tails were linearly extrapolated, that is, the upper limit of the cumulative distribution function was determined by

$$z^* = \left(\frac{1 - F(z_m)}{F(z_m) - F(z_{m-1})}\right)(z_m - z_{m-1}) + z_m$$

and the lower limit of the cumulative distribution function was calculated as

$$z_* = \max\left\{0, z_1 - \left(\frac{F(z_1) - 0}{F(z_2) - F(z_1)}\right)(z_2 - z_1)\right\}.$$

UKMO precise probabilistic forecasts were translated into interval probabilistic forecasts (see below) using the following partition of the interval $[0, 1]$ used by the UKMO: $a_0 = 0$, 0.025, 0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.90, 0.95, $1 = a_{15}$.

The Australian Bureau of Meteorology (ABOM) computes precise probabilistic forecasts for X based on a threshold level of 0.2mm. For each day, a total of 7 different forecasts are computed: a forecast being calculated at 12, 36, 60, 84, 108, 132 and 156 hours before the start of the day to which the recorded precipitation amount refers. The data consisted of the 7 precise probabilistic forecasts for each of 290 consecutive days for 18 different locations around Australia; missing data (either observed precipitation or precise probabilistic forecast) was omitted not imputed. The ABOM precise probabilistic forecasts were converted to interval probabilistic forecasts (see below) using the following partition of the interval $[0, 1]$: $a_0 = 0$, 0.025, 0.075, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.925, 0.975, $1 = a_{13}$; this partition is used by the ABOM.

4.2. Calculating Interval Probabilistic forecasts

The data described in the previous subsection are precise probabilistic forecasts. We now describe how these data may be used to calculate interval probabilistic forecasts. We begin

by assuming that each precise probabilistic forecast, p, is determined under a precise-proper scoring rule and so represents the forecaster's true belief that $X = 1$. Next, suppose that the forecaster is made aware of both the interval scoring rule, s, by which they will be evaluated (see for example, Gneiting (2011) on the need for the forecaster to be made aware of the scoring rule) and the partition $0 = a_0 < a_1 < \ldots < a_{n-1} < a_n = 1$ from which they must choose an interval. The interval chosen by the forecaster, I_k, is that which optimises their expected score, $s[I_k, p]$. If s is an interval-proper scoring rule, the interval issued by the forecaster will be the interval containing p. Under an interval-improper scoring rule, the forecast interval will not necessarily contain the forecaster's true belief p.

In this manner, for each precise probabilistic forecast in the data two interval probabilistic forecasts are computed: one when s is an interval-improper scoring rule and one when s is an interval-proper scoring rule. We emphasise that all interval probabilistic forecasts so calculated are hypothetical and are not actual interval probabilistic forecasts provided by either the UKMO or the ABOM.

4.3. *Skill*

Let x_i be the ith recorded binary observation and I_{k_i} be the interval probabilistic forecast associated with x_i, $i = 1, \ldots, N$. The forecaster's *accuracy* is their average score $\bar{s}_N = \frac{1}{N}\sum_{i=1}^{N} s(I_{k_i}, x_i)$. In the limit, the forecaster's average score is the expected value $\mathbb{E}[s(I, X)]$, where the expected value is taken over the *joint* distribution of the intervals I and the observation X. For large N, $\bar{s}_N$ is approximately normally distributed with mean $\mathbb{E}[s(I, X)]$ and variance $\hat{\sigma}^2/N$ where

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(s(I_{k_i}, x_i) - \bar{s}_N\right)^2$$

A forecaster's skill is evaluated by comparing their accuracy to the accuracy of other forecasters; specifically, we make comparisons with the accuracy of the perfect forecaster and the accuracy of the climatological forecaster. The climatological forecaster computes their actual precise belief from the distribution of precipitation over some agreed historical period. (The UKMO historical period is 1983-2012, and the ABOM historical period is 1981-2010. Both the UKMO and the ABOM provide site-specific climatological probabilistic forecasts as a set of nodes $(z_j, F_{\text{clim}}(z_j))$ for $j = 1, \ldots, m_{\text{clim}}$ from which the precise climatological forecast is calculated as the probability of exceeding the applicable threshold; linear interpolation and extrapolation are used where necessary in the manner described above.)

We define (Wilks 2006, page 259), the forecaster's skill by

$$\frac{\mathbb{E}[s(I,X)] - \mathbb{E}_{\text{clim}}[s(I,X)]}{\mathbb{E}_{\text{perf}}[s(I,X)] - \mathbb{E}_{\text{clim}}[s(I,X)]} \tag{22}$$

where $\mu_{\text{clim}} = \mathbb{E}_{\text{clim}}[s(I,X)]$ is the accuracy of the climatological forecaster and $\mu_{\text{perf}} = \mathbb{E}_{\text{perf}}[s(I,X)]$ is the accuracy of the perfect forecaster, from which, taking μ_{clim} and μ_{perf} as constant, a forecaster's skill is approximately normally distributed with mean

$$\frac{\mathbb{E}[s(I,X)] - \mu_{\text{clim}}}{\mu_{\text{perf}} - \mu_{\text{clim}}}$$

and variance

$$\frac{\hat{\sigma}^2}{N\left(\mu_{\text{perf}} - \mu_{\text{clim}}\right)^2}$$

A skill of 1 for a forecaster demonstrates a perfect forecast record for the forecaster, while a skill of 0 indicates that the forecaster is no more skillful than a climatological forecaster.

EXAMPLE *(Brier scoring rule (Brier 1950)).* Let $0 = a_0 < a_1 < \ldots < a_n = 1$ be a partition of *unequally*-spaced intervals. Choose ξ_k to be the mid-point of I_{k+1} for each $k = 0, \ldots, n-1$. Let s be the adjusted interval-proper Brier scoring rule (equation (21)) and $\tilde{s}$ be the interval-improper adjusted mid-point Brier scoring rule

$$\begin{aligned}\tilde{s}(I_k, X) &= (\xi_{k-1} - X)^2 - \frac{1}{4}a_1^2 \\ &= (X - a_{k-1})(X - a_k) + \frac{1}{4}(a_k - a_{k-1})^2 - \frac{1}{4}a_1^2\end{aligned} \tag{23}$$

(For equally-spaced intervals s and $\tilde{s}$ are equivalent; but, here, unequally-spaced intervals are supposed.)

Figures 2a and 2b compare forecaster skill under s (proper) and $\tilde{s}$ (improper). In figure 2a, the skill of interval probabilistic forecasts at Heathrow Airport for different lead-times is shown. In figure 2b, the skill of the 12-hour lead-time forecast at each of 18 different locations around Australia is plotted.

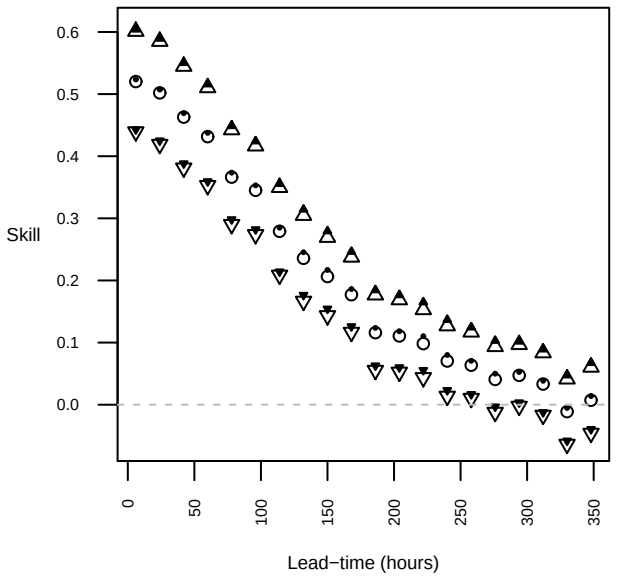
□

The immediate conclusion from the above example is that there appears to be no material difference in skill measured under the interval-proper and interval-improper (Brier) scoring rules.

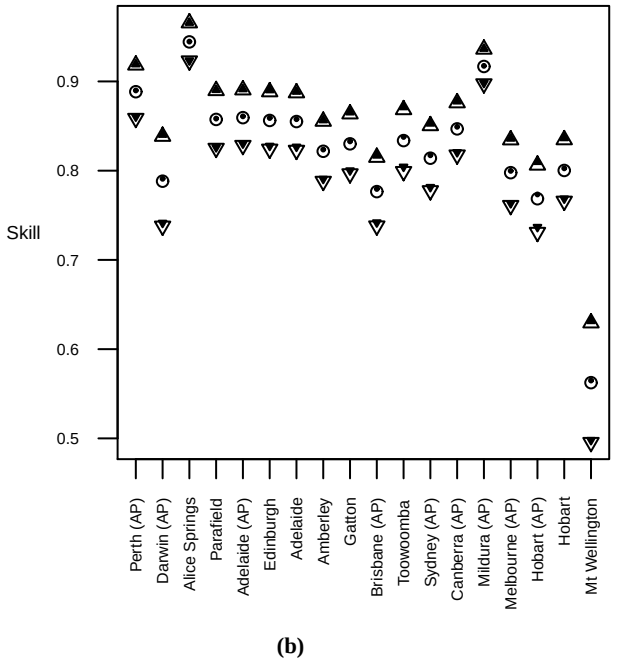
But, there is a more insidious danger from impropriety: impropriety permits hedging, wherein the forecaster chooses to publish an interval probabilistic forecast that differs from the interval they truly believe is appropriate. In such cases, a forecaster's accuracy (or skill) does not measure their true forecasts but measures their given forecasts, thereby misrepresenting their ability. In the presence of hedging, decisions based on the forecaster's ability, in particular whether one forecaster is better than another, are invalid.

For a given interval-improper scoring rule, $\tilde{s}$, and the forecaster's true (precise) belief that $X = 1$, q, whether a forecaster is induced to hedge depends on the values of $\tilde{s}[I, q]$ for different intervals I, and therefore, only on the partition from which the interval forecasts are selected. In the example that follows we demonstrate the effect of the choice of partition on a forecaster's hedging profile.

EXAMPLE *(Brier scoring rule (Brier 1950) cont.).* Assume unequally-spaced intervals and the interval-improper adjusted mid-point Brier scoring rule given by equation (23). We consider the interval probabilistic forecasts issued at Heathrow and Perth Airports.



(a)



(b)

Figure 2. In each figure, estimated forecaster skill (defined by equation (22)) and 95% confidence intervals are shown under the adjusted interval-proper Brier scoring rule (◀ • ▶) and the interval-improper adjusted mid-point Brier scoring rule (◁ ○ ▷). (a) Skill of interval probabilistic forecasts at Heathrow Airport for different forecast lead-times. (b) Skill of interval probabilistic forecasts for the 12-hour lead-time at different locations in Australia.

In the bar-graphs below, the height of each bar is the proportion of times the interval is issued as a forecast. For each bar, the white area (if any) is the proportion of times the interval is forecast and is a hedge that understates the forecaster's true belief; the dark-grey area (if any) is the proportion of times the interval is forecast and is a hedge that overstates the forecaster's true belief. (In all cases, an understated forecast is a forecast of the interval immediately below the true interval forecast and an overstated

forecast is a forecast of the interval immediately above the true interval forecast.) The points marked by •, are the proportion of times the interval is a hedge given the interval is forecast, that is, the propensity to hedge.

In figure 3, the distribution of the 12-hour lead-time forecasts at Heathrow Airport is shown. The relative frequency of hedging is 5% with hedging existing in both the lower and upper mid-ranges of the $[0, 1]$ interval. A hedge in the lower mid-ranges of the $[0, 1]$ interval may be either an understatement or an overstatement, as too a hedge in the upper mid-ranges may be. There is no simple trend in the propensity to hedge across the subintervals.

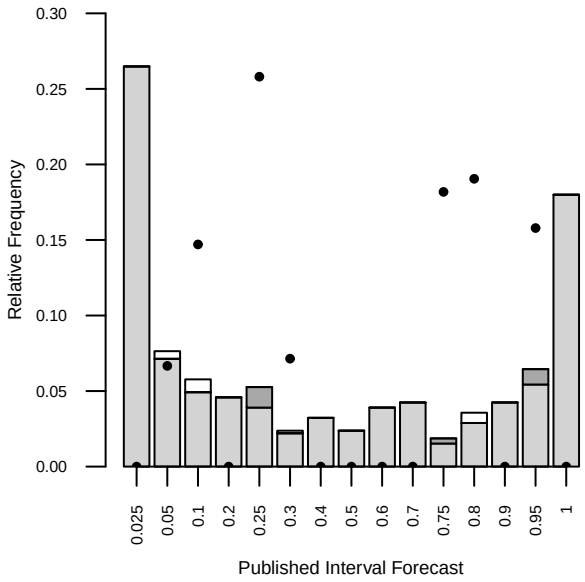


Figure 3. The relative frequency of 12-hour lead-time interval probabilistic forecasts issued for Heathrow Airport. The height of each *entire* bar is an estimate of the probability that the interval is forecast. The white portion of each bar is an estimate of the probability that the interval is the forecast published *and* is a hedge that understates the forecaster's true belief. The dark-grey portion of each bar is an estimate of the probability that the interval is the published forecast *and* is a hedge that overstates the forecaster's true belief. The • points are estimates of the conditional probability that when the interval is forecast, it is a hedge. (The tick-labels on the horizontal axis are the upper end-points of each subinterval.)

An altogether different set of features is displayed in figure 4, a bar-graph of the 12-hour lead-time interval probabilistic forecasts at Perth Airport. Here, the forecaster only tends to hedge when issuing forecasts in the extremities of the $[0, 1]$ interval. Further, the forecaster has a greater propensity to hedge the closer their published forecast lies to either extremity. Hedges in the lower ranges of the $[0, 1]$ interval will understate the forecaster's beliefs while hedges in the upper ranges of the $[0, 1]$ interval will overstate the forecaster's beliefs. The relative frequency of hedging is 11.5%.

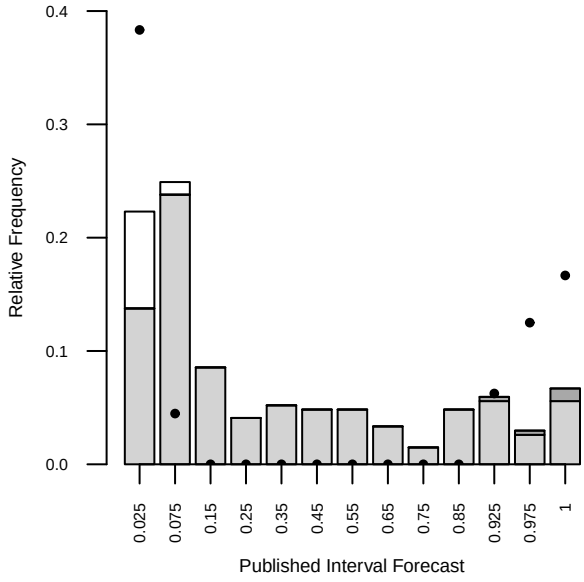


Figure 4. The relative frequency of 12-hour lead-time interval probabilistic forecasts issued for Perth Airport. For an interpretation of the bars see the caption to figure 3.

To examine the impact of the form of the partition on the hedging profile, figure 5 presents a bar-graph of the 12-hour lead-time interval probabilistic forecasts at Heathrow Airport assuming the same partition that was applied at Perth Airport. The relative frequency of hedging is 6% and hedging behaviour is now much more similar to the hedging behaviour seen for the Perth Airport forecasts.

□

An examination of single site and lead-time forecasts, for a predetermined partition and preselected interval-improper scoring rule, is helpful in assessing local properties of a forecaster's hedges. Of interest too, is aggregate hedging behaviour, the proportion of times the forecaster hedges (either understates or overstates their true beliefs) as the site and forecast lead-time changes. We investigate aggregate hedging behaviour by continuing the above example.

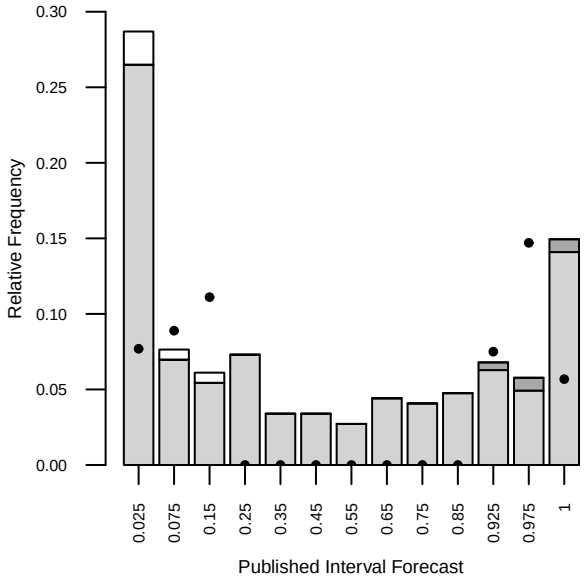


Figure 5. The relative frequency of 12-hour lead-time interval probabilistic forecasts issued for Heathrow Airport *under the same partition used for the forecasts issued for Perth Airport in figure 4*. For an interpretation of the bars see the caption to figure 3.

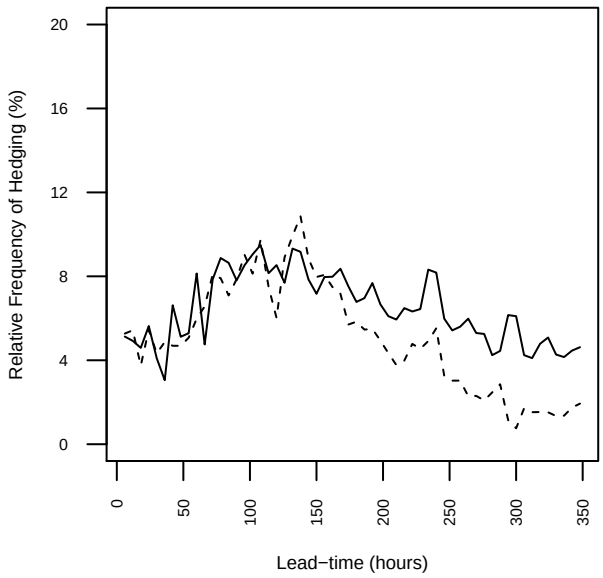


Figure 6. Relative frequency of hedging when issuing interval probabilistic forecasts for different lead-times (hours). Forecasts for two different locations, Heathrow Airport (solid line) and Eskdalemuir (dashed line), are compared.

EXAMPLE *(Brier scoring rule (Brier 1950) cont.).* In figure 6, the relative frequency of hedging is shown for different lead-times at two sites: Heathrow Airport and Eskdalemuir. Hedging is, on the whole, higher for Heathrow Airport than for Eskdalemuir, although the pattern of hedging is similar over the different lead-times: hedging occurs on no more than 12% or so of occasions, tending to peak shortly before the 150-hour lead-time forecast and is lowest for the longest lead-times.

In figure 7, the relative frequency of hedging when issuing interval probabilistic forecasts is compared for a number of lead-times across different locations in Australia. Here, hedging occurs on between approximately 0% and 20% of forecasts. Hedging levels are similar for sites that are geographically close. In general, hedging is higher for shorter duration lead-times with hedging decreasing as the lead-time increases.

□

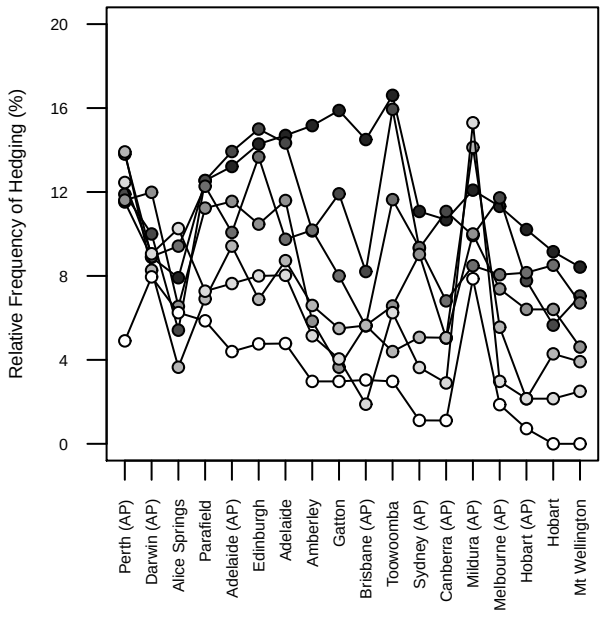


Figure 7. Relative frequency of hedging when publishing interval forecasts, for different locations. Each line represents forecasts for a different lead-time: 12-hour (●), 36-hour (●),60-hour (●),84-hour (●),108-hour (●),132-hour (○),156-hour (○).

5. Summary

We consider probabilistic forecasts for a future $0/1$ event. A precise probabilistic forecast is a statement of the exact value for the probability that 1 will occur. An interval probabilistic forecast is a range of values for the probability that 1 will occur. Interval probabilistic forecasts may be issued explicitly (e.g. 'chance of rain tomorrow: 10-20%') or implicitly as a rounded probabilistic forecast (the undeclared interval forecast being all those precise probabilities that round to the given rounded probabilistic forecast).

Probabilistic forecasts must be evaluated using proper scoring rules. Scoring rules that are proper when the forecast probability is a precise value are not, in general, proper when applied to a representative probability from the interval forecast. Analogous to the result of Lambert (2013), we present a general expression for scoring rules that are proper for interval probabilistic forecasts. Specific interval-proper scoring rules, corresponding to the more familiar precise-proper scoring rules (Brier scoring rule, Ignorance scoring rule and Pseudo-spherical scoring rule) are also given; of these, the interval-proper Brier scoring rule (equation (21)) has a simple and appealing form.

The importance of interval-proper scoring rules is their use in assessing the performance of forecasters issuing interval probabilistic forecasts. That is not to say that an interval-improper scoring rule necessarily results in a meaningful change in a forecaster's calculated skill; substituting an interval-improper scoring rule for an interval-proper scoring rule can have little quantifiable impact on a forecaster's skill. Rather, the egregious effect of impropriety is on the interpretation of a forecaster's computed skill. Under impropriety, a forecaster may hedge when issuing a forecast, giving a forecast that does not reflect their true opinion. In such cases the skill, being based on the published forecasts, no longer represents the forecaster's true views and gives only partial insight into their substantive ability.

We calculate the relative frequency of hedging using interval probabilistic forecasts simulated using precise probability of precipitation (PoP) forecasts provided by The Australian Bureau of Meteorology and the UK Meteorological Office. While hedging varies with site and forecast lead-time, the relative frequency of hedging in the cases we consider lies approximately in the range of $0 - 15\%$.

Interval-proper scoring rules depend explicitly on the set of intervals to which the interval forecasts refer. A change of the intervals used to express forecasts will influence the scoring rule and a natural question arises as to whether there is an optimal set of intervals. The question may be framed as a high-dimensional non-linear constrained optimisation problem and while we have not conducted a general investigation of this problem, in the particular case of the unadjusted interval-proper Brier scoring rule (equation (20)) it can be shown that the optimal partition is the equally-spaced partition, when 'optimal' is defined as the interval-proper Brier scoring rule being close in the squared-error sense to the precise-proper Brier scoring rule.

Acknowledgements

We would like to thank Prof. A. Abu-Hanna for introducing us to the problem of proper scoring rules for interval probabilistic forecasts. For their generous help in providing the data used in this paper, we would like to thank the Australian Burerau of Meteorology, in particular, Dr. D. Griffiths and Ms. I. Ioannou, and the UK Meteorological Office, specifically Dr. M. Mittermaier. Lastly, to the two anonymous referees, for their comments and suggestions, our sincerest thanks.

A. Appendix

For $0 < i \leq n$, write $\hat{p}_i = \frac{1}{2}(a_{i-1} + a_i)$, the mid-point of the interval I_i. The mid-point Brier scoring rule is defined by

$$s(I_i, X) = (\hat{p}_i - X)^2$$

and satisfies $s[I_i, q] = \mathbb{E}_q[s(I_i, X)] = \hat{p}_i^2 - 2\hat{p}_i q + q$. The mid-point Brier scoring rule is interval-proper if and only if

$$s[I_i, q] \leq s\big[I_j, q\big] \quad \forall i, j \text{ and } \forall q \in I_i.$$

Proof of Proposition 2.1: the mid-point Brier scoring rule is proper if and only if the partition is equally-spaced.

Proof. Suppose that the mid-point Brier scoring rule is interval-proper. Then

$$s[I_i, q] \leq s\big[I_j, q\big] \quad \forall i, j \text{ and } \forall q \in I_i$$

which holds if and only if, fixing i, $\forall q \in I_i$,

$$q \leq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j) \qquad i < j$$
$$q \geq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j) \qquad i > j. \qquad (24)$$

Condition (24) must hold for all $q \in I_i$ and so holds for $q = a_i$. In this case,

$$a_i \leq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j) \quad i < j$$

and, letting $j = i + 1$,

$$a_i \leq \frac{1}{4}(a_{i-1} + a_i + a_i + a_{i+1})$$

from which

$$a_i - a_{i-1} \leq a_{i+1} - a_i.$$

Also, condition (24) must hold for $q = \inf I_i = a_{i-1}$. Specifically,

$$a_{i-1} \geq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j) \quad i > j.$$

Letting $j = i - 1$,

$$a_{i-1} \geq \frac{1}{4}(a_{i-1} + a_i + a_{i-2} + a_{i-1})$$

giving

$$a_{i-1} - a_{i-2} \geq a_i - a_{i-1}.$$

As i was fixed arbitrarily, we have, for all $0 < i < n$, $a_i - a_{i-1} \leq a_{i+1} - a_i$ and for $1 < i \leq n$, $a_{i-1} - a_{i-2} \geq a_i - a_{i-1}$.

Now, for any $0 < k < n$, let $i = k$, to give $a_{k+1} - a_k \geq a_k - a_{k-1}$ and let $i = k + 1$ to give $a_k - a_{k-1} \geq a_{k+1} - a_k$, from which

$$a_{k+1} - a_k = a_k - a_{k-1}$$

and this holds for all $0 < k < n$, that is, the a_i are equally-spaced.

Conversely, suppose that the a_i are equally-spaced; $a_i = i/n$ for all $0 \leq i \leq n$. Then

$$\frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j) = \frac{1}{n}\left(\frac{i+j-1}{2}\right).$$

If $i < j$ then $i \leq j - 1$ and

$$a_i = \frac{i}{n} \leq \frac{1}{n}\left(\frac{i+j-1}{2}\right) = \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j)$$

so $q \leq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j)$ for all $q \in I_i$ with $i < j$. Equally, if $i > j$ then $i - 1 \geq j$ and

$$a_{i-1} = \frac{i-1}{n} \geq \frac{1}{n}\left(\frac{i+j-1}{2}\right) = \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j)$$

so that $q \geq \frac{1}{4}(a_{i-1} + a_i + a_{j-1} + a_j)$ for all $q \in I_i$ with $i > j$. Condition (24) is satisfied and therefore, the mid-point Brier scoring rule is interval-proper.

■

We remark in passing that the propriety of the more general λ-Brier scoring rule, defined by $s(I_i, X) = \{(1-\lambda)a_{i-1} + \lambda a_i - X\}^2$ also depends critically on the spacing of the partition, being proper if and only if, letting $\Delta_i = a_i - a_{i-1}$,

$$\begin{cases} \frac{\Delta_1}{\Delta_1+\Delta_2} \leq \lambda \leq 1 & \text{for } q \in I_1 \\ \frac{\Delta_k}{\Delta_k+\Delta_{k+1}} \leq \lambda \leq \frac{\Delta_{k-1}}{\Delta_k+\Delta_{k-1}} & \text{for } q \in I_k \\ & 1 < k < n \\ 0 \leq \lambda \leq \frac{\Delta_{n-1}}{\Delta_n+\Delta_{n-1}} & \text{for } q \in I_n \end{cases}$$

B. Appendix

We show under certain conditions on the partition $0 = a_0 < a_1 < \ldots < a_n = 1$, and on the functions s, S and h, that, letting $\xi_k \in I_{k+1}$, as n increases the equation

$$s(I_{k+1}, X) - s(I_k, X) = \\ -(h(\xi_k) - h(\xi_{k-1}))(X - a_k) \quad (25)$$

leads to the differential equation

$$\frac{\partial S(p,X)}{\partial p} = \frac{dh(p)}{dp}(p - X). \tag{26}$$

Definition B.1 *The partitions* $[a]_n = a_{n,0} < a_{n,1} < \ldots < a_{n,n}$, $0 = a_{n,0}$, $1 = a_{n,n}$, *are said to be increasingly refined as* $n \to \infty$ *if the mesh,* $\mu_n = \max\{a_{n,i} - a_{n,i-1} \mid i = 1, \ldots, n\}$ *tends to* 0.

Remark. When refering to subintervals of the partition $[a]_n = a_{n,0} < a_{n,1} < \ldots < a_{n,n}$, $0 = a_{n,0}$, $1 = a_{n,n}$, we shall use the notation $I_{n,1} = [a_{n,0}, a_{n,1}]$, $I_{n,k} = (a_{n,k-1}, a_{n,k}]$ for $k = 2, \ldots, n$.

Lemma B.1 *Let* $p \in [0,1]$. *If the partitions* $[a]_n$ *are increasingly refined then* $\forall \epsilon > 0$, $\exists N \geq 0$ *such that for each* $n > N$, *there is a* k *(depending on* n*) such that* $|a_{n,k} - p| < \epsilon$.

Proof. Fix $\epsilon > 0$. Since the partitions $[a]_n$ are increasingly refined, there is an $N \geq 0$ such that $\forall n > N$, $\mu_n < \epsilon$. Let $n > N$ so that $\mu_n < \epsilon$. If $p \in [0,1]$ then there is some k such that $p \in I_{n,k}$. Therefore, $|a_{n,k} - p| \leq |a_{n,k} - a_{n,k-1}| \leq \mu_n < \epsilon$. So for all $n > N$, there exists a k (depending on n) such that $|a_{n,k} - p| < \epsilon$. ■

Definition B.2 *We shall say that the interval scoring rule* s *converges in the Lipschitz sense to the precise scoring rule* S *at* p *if and only if* $\forall \epsilon > 0$, $\exists N \geq 0$ *such that for all* $n \geq N$, $|s(I_{n,k}, x) - S(p,x)| < \epsilon \min\{|a_{n,k-1} - p|, |a_{n,k} - p|\}$ *for all* x, $p \in I_{n,k}$. *If* s *converges to* S *in the Lipschitz sense at every* $p \in [0,1]$, *then we shall say simply that* s *converges to* S *in the Lipschitz sense.*

Proposition B.1 *Let* s *be an interval-proper scoring rule satisfying equation (25), with* h *continuously differentiable. Suppose that* s *converges to the precise scoring rule* S *in the Lipschitz sense, where* S *is continuously partially differentiable with respect to* p. *If the partitions* $[a]_n$ *are increasingly refined then*

$$\frac{\partial S(p,X)}{\partial p} = \frac{dh(p)}{dp}(p - X).$$

Proof. Let $\epsilon > 0$, $p \in [0,1]$. S is continuously partially differentiable with respect to p, so $\exists \delta_* > 0$ such that $\forall |r| < \delta_*$,

$$\left| \frac{S(p+r,X) - S(p,X)}{r} - \frac{\partial S(p,X)}{\partial p} \right| < \frac{\epsilon}{4}$$

and $\exists \delta' > 0$ such that if $|\xi - p| < \delta'$,

$$\left| \frac{\partial S(\xi,X)}{\partial \xi} - \frac{\partial S(p,X)}{\partial p} \right| < \frac{\epsilon}{4}.$$

Further, since h is continuously differentiable, $\exists \delta_{**}$ such that $\forall |r| < \delta_{**}$,

$$\left| \frac{h(p+r) - h(p)}{r} - \frac{dh(p)}{dp} \right| < \frac{\epsilon}{2}$$

and $\exists \delta'' > 0$ such that if $|\xi - p| < \delta''$, then

$$\left| \frac{dh(\xi)}{d\xi} - \frac{dh(p)}{dp} \right| < \frac{\epsilon}{2}.$$

Let $\delta = \min\{\delta_*, \delta', \delta_{**}, \delta'', \epsilon\}$.

As the partitions $[a]_n$ are increasingly refined, $\exists N_* \geq 0$ such that for $n \geq N_*$, $\mu_n < \frac{\delta}{2}$. The interval scoring rule s converges to S in the Lipschitz sense, so $\exists N' \geq 0$ such that $\forall n > N'$,

$$|s(I_{n,j}, X) - S(p,X)| < \frac{\epsilon}{4} \min\{|a_{n,j-1} - p|, |a_{n,j} - p|\}$$

for $p \in I_{n,j}$. Let $N = \max\{N_*, N'\}$, $n \geq N$ and let k (depending on n) satisfy $p \in I_{n,k}$. From equation (25),

$$\frac{s(I_{n,k+1}, X) - s(I_{n,k}, X)}{\xi_{n,k} - \xi_{n,k-1}} = - \left(\frac{h(\xi_{n,k}) - h(\xi_{n,k-1})}{\xi_{n,k} - \xi_{n,k-1}} \right) (X - a_{n,k}) \tag{27}$$

where, as above, $\xi_{n,k} \in I_{n,k+1}$.

Considering the left-hand side of equation (27),

$$
\begin{aligned}
&\left|\frac{s(I_{n,k+1},X)-s(I_{n,k},X)}{\xi_{n,k}-\xi_{n,k-1}}-\frac{\partial S(p,X)}{\partial p}\right| \\
=&\left|\frac{s(I_{n,k+1},X)-S(\xi_{n,k},X)}{\xi_{n,k}-\xi_{n,k-1}}\right. \\
&\quad-\frac{s(I_{n,k},X)+S(\xi_{n,k-1},X)}{\xi_{n,k}-\xi_{n,k-1}} \\
&\quad+\frac{S(\xi_{n,k},X)-S(\xi_{n,k-1},X)}{\xi_{n,k}-\xi_{n,k-1}} \\
&\quad\left.-\frac{\partial S(\xi_{n,k-1},X)}{\partial \xi_{n,k-1}}+\frac{\partial S(\xi_{n,k-1},X)}{\partial \xi_{n,k-1}}-\frac{\partial S(p,X)}{\partial p}\right| \\
\leq&\left|\frac{s(I_{n,k+1},X)-S(\xi_{n,k},X)}{\xi_{n,k}-\xi_{n,k-1}}\right| \\
&\quad+\left|\frac{s(I_{n,k},X)+S(\xi_{n,k-1},X)}{\xi_{n,k}-\xi_{n,k-1}}\right| \\
&\quad+\left|\frac{S(\xi_{n,k},X)-S(\xi_{n,k-1},X)}{\xi_{n,k}-\xi_{n,k-1}}-\frac{\partial S(\xi_{n,k-1},X)}{\partial \xi_{n,k-1}}\right| \\
&\quad+\left|\frac{\partial S(\xi_{n,k-1},X)}{\partial \xi_{n,k-1}}-\frac{\partial S(p,X)}{\partial p}\right|.
\end{aligned}
$$

But, S is partially continuously differentiable with respect to p, $\xi_{n,k}-\xi_{n,k-1}\leq a_{n,k+1}-a_{n,k-1}\leq 2\mu_n<\delta$, and $|\xi_{n,k-1}-p|<\mu_n<\delta$, from which it follows that

$$
\begin{aligned}
&\left|\frac{s(I_{n,k+1},X)-s(I_{n,k},X)}{\xi_{n,k}-\xi_{n,k-1}}-\frac{\partial S(p,X)}{\partial p}\right| \\
<&\frac{\epsilon}{4}\frac{\min\{|a_{n,k+1}-\xi_{n,k}|,|a_{n,k}-\xi_{n,k}|\}}{\xi_{n,k}-\xi_{n,k-1}} \\
&\quad+\frac{\epsilon}{4}\frac{\min\{|a_{n,k}-\xi_{n,k-1}|,|a_{n,k-1}-\xi_{n,k-1}|\}}{\xi_{n,k}-\xi_{n,k-1}}+\frac{\epsilon}{4}+\frac{\epsilon}{4} \\
<&\frac{\epsilon}{4}+\frac{\epsilon}{4}+\frac{\epsilon}{4}+\frac{\epsilon}{4} \\
=&\epsilon
\end{aligned}
$$

having noted too that

$$
\frac{\min\{|a_{n,k+1}-\xi_{n,k}|,|a_{n,k}-\xi_{n,k}|\}}{\xi_{n,k}-\xi_{n,k-1}}\leq\frac{|a_{n,k}-\xi_{n,k}|}{\xi_{n,k}-\xi_{n,k-1}}\leq 1
$$

and, similarly

$$
\frac{\min\{|a_{n,k}-\xi_{n,k-1}|,|a_{n,k-1}-\xi_{n,k-1}|\}}{\xi_{n,k}-\xi_{n,k-1}}\leq\frac{|a_{n,k}-\xi_{n,k-1}|}{\xi_{n,k}-\xi_{n,k-1}}\leq 1.
$$

Next,

$$
\begin{aligned}
&\left|\frac{h(\xi_{n,k})-h(\xi_{n,k-1})}{\xi_{n,k}-\xi_{n,k-1}}-\frac{dh(p)}{dp}\right| \\
=&\left|\frac{h(\xi_{n,k})-h(\xi_{n,k-1})}{\xi_{n,k}-\xi_{n,k-1}}-\frac{dh(\xi_{n,k-1})}{d\xi_{n,k-1}}+\frac{dh(\xi_{n,k-1})}{d\xi_{n,k-1}}-\frac{dh(p)}{dp}\right| \\
\leq&\left|\frac{h(\xi_{n,k})-h(\xi_{n,k-1})}{\xi_{n,k}-\xi_{n,k-1}}-\frac{dh(\xi_{n,k-1})}{d\xi_{n,k-1}}\right|+\left|\frac{dh(\xi_{n,k-1})}{d\xi_{n,k-1}}-\frac{dh(p)}{dp}\right| \\
<&\frac{\epsilon}{2}+\frac{\epsilon}{2} \\
=&\epsilon.
\end{aligned}
$$

Finally, $|(X-a_{n,k})-(X-p)|=|p-a_{n,k}|\leq\mu_n<\frac{\delta}{2}\leq\epsilon$. Combining these separate limit results, equation (27) gives

$$
\frac{\partial S(p,X)}{\partial p}=-\frac{dh(p)}{dp}(X-p).
$$

■

References

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

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* **135**: 1512–1519.

Dawid A. 1986. Probability forecasting. In: *Encyclopedia of Statistical Sciences*, vol. 7, Kotz S, Johnson N, Read C (eds), John Wiley & Sons, pp. 210–218.

Frongillo R, Kash I. 2014. General truthfulness characterizations via convex analysis. ArXiv:1211.3043v3.

Gneiting T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**: 746–762.

Gneiting T, Katzfuss M. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* **2014**(1): 125–151.

Gneiting T, Raftery A. 2007. Strictly proper scoring rules, prediction and estimation. *American Statistical Association* **102**: 359–378.

Good I. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**(1): 107–114.

Lambert N. 2013. Elicitation and evaluation of statistical forecasts. Preprint. Stanford University. (`web.stanford.edu/~nlambert/papers/elicitation.pdf`).

Lambert N, Pennock D, Shoham Y. 2008. Eliciting properties of probability distributions. *EC '08 Proceedings of the 9th ACM Conference on Electronic Commerce* : 129–138.

Lambert N, Shoham Y. 2009. Eliciting truthful answers to multiple-choice questions. *EC '09 Proceedings of the 10th ACM Conference on Electronic Commerce* : 109–118.

Murphy A. 1998. The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting* **13**: 5–15.

Murphy A, Epstein E. 1967. A note on probability forecasts and 'hedging'. *Journal of Applied Meteorology* **6**: 1002–1004.

Roby T. 1964. Belief states: A preliminary empirical study. Technical documentary report no. esd-tdr-64-238, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force.

Schervish M. 1989. A general method for comparing probability assessors. *The Annals of Statistics* **17**(4): 1856–1879.

Wilks D. 2006. *Statistical methods in the atmospheric sciences*. Elsevier, second edn.

Winkler R. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1): 1–60.

Winkler R, Murphy A. 1968. 'Good' probability assessors. *Journal of Applied Meteorology* **7**: 751–758.

