MANAGEMENT SCIENCE
Vol. 55, No. 4, April 2009, pp. 582–590
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5504 | 0582

informs

DOI 10.1287/mnsc.1080.0955

Sensitivity to Distance and Baseline Distributions in Forecast Evaluation

Victor Richmond R. Jose, Robert F. Nau, Robert L. Winkler
Fuqua School of Business, Duke University, Durham, North Carolina 27708
{vrj@duke.edu, rnau@duke.edu, rwinkler@duke.edu}

Scoring rules can provide incentives for truthful reporting of probabilities and evaluation measures for the probabilities after the events of interest are observed. Often the space of events is ordered and an evaluation relative to some baseline distribution is desired. Scoring rules typically studied in the literature and used in practice do not take account of any ordering of events, and they evaluate probabilities relative to a default baseline distribution. In this paper, we construct rich families of scoring rules that are strictly proper (thereby encouraging truthful reporting), are sensitive to distance (thereby taking into account ordering of events), and incorporate a baseline distribution relative to which the value of a forecast is measured. In particular, we extend the power and pseudospherical families of scoring rules to allow for sensitivity to distance, with or without a specified baseline distribution.

Key words: forecast verification; ranked categories; scoring rules; sensitivity to distance; baseline distributions
History: Received March 23, 2008; accepted September 30, 2008, by David E. Bell, decision analysis. Published online in *Articles in Advance* January 5, 2009.

1. Introduction

In many applications, we are interested in probability assessments for ordered events. For example, we may be interested in a probabilistic forecast for a team's three possible outcomes in a soccer match (win, tie, lose). These outcomes are ordered in the sense that a tie is a result that is better than a loss but worse than a win. Or we might want to quantify the chances that the percentage change in the NASDAQ index over the next month will fall into certain binned categories (e.g., below −5%, −5% to 0%, 0% to 5%, above 5%).

Scoring rules are used in probability assessment to provide an ex ante incentive for truthful reporting and in probability evaluation to measure how informative the probabilities look ex post after the outcome is known. However, virtually all scoring rules that are commonly used in practice do not take account of any ordering in the events of interest. A notable exception is the ranked probability score (RPS) introduced by Epstein (1969). The RPS, which has been used in meteorology, is a quadratic rule that is strictly proper (thereby encouraging honest reporting) and considers the ordering of events in the sense that it gives higher scores for assessments that are "closer" to the event that occurs based on a particular concept of distance. For example, if two forecasts assign the same probability to the event that actually occurs but the first forecast gives higher probabilities for events "close to" the actual event and lower probabilities for events "more distant" from the actual event, the first forecast will receive a higher score. In the NASDAQ example, if the percentage change turns out to be +2% (in the third category), the probability assessment (0.1, 0.4, 0.3, 0.2) would get a higher score than (0.2, 0.3, 0.3, 0.2) because it differs only in having a larger probability for the second category and a smaller probability for the first category. This property of the RPS is called sensitivity to distance; in the example, the first category is "more distant" from the actual result than is the second category. Because sensitivity to distance may be viewed as a desirable property by someone obtaining and evaluating probability assessments, it would be useful to have a wider choice of rules satisfying that property.

Another issue in the design of scoring rules concerns the appropriate baseline distribution against which the assessed probabilities are evaluated. For the standard scoring rules commonly used in practice, there is no provision for the consideration of a baseline distribution, which is a uniform distribution by default. For example, the informativeness of a probability of precipitation is evaluated relative to a baseline probability of one-half because there are two possible outcomes, precipitation and no precipitation. In a location where precipitation occurs on average only 5% of the time, a baseline probability of one-half seems inappropriate. In the case of scoring rules for unordered categories, this issue has been addressed by looking at "asymmetric" scoring rules that allow for the consideration of a nonuniform baseline distribution

(Winkler 1994, Jose et al. 2008). But for spaces where outcomes are ordered, no algorithm has been developed that provides for the creation of a rule with any desired baseline distribution.

In this paper, we develop a rich family of strictly proper scoring rules that are sensitive to distance and allow probability assessments to be evaluated relative to any chosen baseline distribution. In §2, we review the notion of strictly proper scoring rules and some properties that such rules could possess. The property of sensitivity to distance is considered in §3, and the RPS is extended to forms other than quadratic, creating a discrete version of a family of continuous rules developed in Matheson and Winkler (1976) and allowing considerable flexibility in the choice of a scoring rule that is sensitive to distance in the ordered n-state case. Section 4 then shows how the rules from §3 can be extended further to incorporate a baseline distribution. Some concluding comments and a brief discussion are given in §5.

2. Scoring Rules

Suppose there are n states of the world. Let the forecaster's belief be denoted by $\mathbf{p} = (p_1, \ldots, p_n)$, and denote his reported distribution (his forecast) by $\mathbf{r} = (r_1, \ldots, r_n)$. Both $\mathbf{p}$ and $\mathbf{r}$ are assumed to be proper probability distributions (i.e., the probabilities are nonnegative and sum to one). Moreover, in the case where the states of the world are ordered, we denote the corresponding cumulative probabilities by capital letters: $P_i = \sum_{j \le i} p_j$ and $R_i = \sum_{j \le i} r_j$.

A *scoring rule* is a function S that assigns a real number (a score) based on a reported distribution $\mathbf{r}$ and an outcome (the state that actually occurs). Ex post, the forecaster receives the score $S(\mathbf{r}, \mathbf{e}_j)$ when event j is observed, as indicated by the jth standard basis vector $\mathbf{e}_j$. From an ex ante perspective, the forecaster's expected score is

$$S(\mathbf{r}, \mathbf{p}) = \sum_{j=1}^{n} p_j S(\mathbf{r}, \mathbf{e}_j).$$

Scoring rules of greatest interest are those for which the expected score is maximized only by truthful reporting. That is, for any $\mathbf{p}$ and $\mathbf{r}$,

$$S(\mathbf{p}, \mathbf{p}) \ge S(\mathbf{r}, \mathbf{p}),$$

where the equality holds if and only if $\mathbf{r} = \mathbf{p}$. Such rules are called *strictly proper*. Any positive linear transformation $aS + b$, $a > 0$, of a strictly proper S is also strictly proper, so scoring rules can be scaled as desired. For general discussions of some scoring rules and properties of scoring rules, see Winkler (1996) and Gneiting and Raftery (2007).

Example 1. Two rich families of strictly proper rules are the power and pseudospherical families of scoring rules with parameter β (Jose et al. 2008):

$$S_\beta^P(\mathbf{r}, \mathbf{e}_j) = \frac{r_j^{\beta-1} - 1}{\beta - 1} - \frac{\mathbb{E}_{\mathbf{r}}[\mathbf{r}^{\beta-1}] - 1}{\beta} \quad \text{and} \tag{1}$$

$$S_\beta^S(\mathbf{r}, \mathbf{e}_j) = \frac{1}{\beta - 1}\left[\left(\frac{r_j}{\mathbb{E}_{\mathbf{r}}[\mathbf{r}^{\beta-1}]^{1/\beta}}\right)^{\beta-1} - 1\right], \tag{2}$$

where $\mathbb{E}_{\mathbf{r}}[\mathbf{r}^{\beta-1}] = \sum_{i=1}^{n} r_i(r_i^{\beta-1})$. These families are continuous in β and strictly proper for all real β and include the most-commonly used scoring rules. Contemplating a spectrum of possible choices for β offers great flexibility in the choice of a rule. For example, when $\beta = 2$, these families yield the well-known quadratic and spherical rules, respectively. Moreover, when $\beta \to 1$, both rules converge to the logarithmic rule. This rule, $S(\mathbf{r}, \mathbf{e}_j) = \log r_j$, is the only strictly proper rule that satisfies *locality*, the property that the score $S(\mathbf{r}, \mathbf{e}_j)$ depends only on r_j, the reported probability for the state that is observed, not on r_i for $i \ne j$, the other elements of $\mathbf{r}$.

Also, each of these families of scoring rules can be generated by using the forecast to solve a canonical utility maximization problem with linear-risk-tolerance (also called HARA) utility functions $u_\beta(x) = [(1 + \beta x)^{(\beta-1)/\beta} - 1]/(\beta - 1)$ and to give the forecaster a score equal to the utility of the payoff that is received when the optimal solution is implemented. This canonical form of the linear-risk-tolerance utility function, which is introduced by Jose et al. (2008), has the following key properties: (i) it is a continuous function of β on the entire real line for fixed x, (ii) the forecaster's marginal utility of wealth is normalized to be 1 at $x = 0$ for all β, and (iii) the forecaster's local risk tolerance at wealth x is normalized to be $1 + \beta x$. In this context, different values of β represent different risk attitudes on the part of a decision maker who must act on the basis of the forecast and who wishes to reward the forecaster in proportion to the decision maker's own utility gain or loss. The use of $\beta = 1$ corresponds to logarithmic utility, whereas $\beta = 2$ corresponds to square-root utility. The cases $\beta = 0$ and $\beta = 0.5$ are also of interest, corresponding to exponential and reciprocal utility, respectively.

Most strictly proper scoring rules typically used in practice are *symmetric* in the sense that they are invariant with respect to the labeling of the states of the world. If $n = 2$, for example, this means that $S(\mathbf{r}, \mathbf{e}_1) = S(\mathbf{1} - \mathbf{r}, \mathbf{e}_2)$. The members of the power and pseudospherical families defined in (1) and (2) are symmetric.

3. Sensitivity to Distance

When the state space is ordered, a property of interest for scoring rules is *sensitivity to distance*. We say that a

forecast $\mathbf{r}'$ is *more distant from event* j than a forecast $\mathbf{r}$ if $\mathbf{r}' \neq \mathbf{r}$ and

$$R'_i \geq R_i \quad i = 1, \ldots, j-1,$$
$$R'_i \leq R_i \quad i = j, \ldots, n-1$$

(Staël von Holstein 1970a, b). This means that $\mathbf{r}$ can be obtained from $\mathbf{r}'$ by successive movements of probability mass toward event j from other events "more distant" from event j.

The scoring rule S is said to be *sensitive to distance* if $S(\mathbf{r}, \mathbf{e}_j) > S(\mathbf{r}', \mathbf{e}_j)$ whenever $\mathbf{r}'$ is more distant from $\mathbf{r}$ for all j. Nau (2007) provides a characterization for strictly proper rules that are sensitive to distance. Note that scoring rules that are sensitive to distance are not symmetric; any relabeling of the states of the world will destroy the natural ordering required for sensitivity to distance. Also, all strictly proper scoring rules are technically but trivially sensitive to distance when $n = 2$ because any ordering of the states does not affect the score. The spirit of the notion of sensitivity to distance kicks in only for $n \geq 3$.

Example 2. Epstein (1969) introduced the RPS, a strictly proper rule that is sensitive to distance:

$$\mathrm{RPS}(\mathbf{r}, \mathbf{e}_j) = \frac{3}{2} - \frac{1}{2(n-1)} \sum_{i=1}^{n-1} [R_i^2 + (1 - R_i)^2] - \frac{1}{n-1} \sum_{i=1}^{n} |i - j| r_i. \quad (3)$$

Epstein's development of the RPS is based on a decision-making situation known in meteorology as the cost-loss ratio problem, and it is in the "business sharing" spirit of McCarthy (1956) and Savage (1971). Staël von Holstein (1970a, b) generalizes the RPS by considering uncertainty about the cost-loss ratio and generating a family of scoring rules for which each member corresponds to a probability distribution for the cost-loss ratio.

Murphy (1971) points out that (3) is equivalent to

$$\mathrm{RPS}(\mathbf{r}, \mathbf{e}_j) = -\sum_{i=1}^{j-1} R_i^2 - \sum_{i=j}^{n-1} (1 - R_i)^2. \quad (4)$$

The idea underlying (4) is that we are considering the $n-1$ dividing lines between adjoining states, using a quadratic score for the dichotomy between the region below the dividing line and the region above the dividing line in each case, and adding the resulting $n-1$ quadratic scores. These quadratic scores are $S((R_i, 1 - R_i), \mathbf{e}_2) = -R_i^2$ for $i < j$ and $S((R_i, 1 - R_i), \mathbf{e}_1) = -(1 - R_i)^2$ for $i \geq j$.

A similar approach is used in Matheson and Winkler (1976) to generate strictly proper sensitive-to-distance scoring rules in the continuous case. If S is strictly proper for a dichotomy ($n = 2$ states), R is a continuous cumulative distribution function (cdf) with corresponding density function r for a random variable of interest, and x is the value of that variable that is actually observed, define a new scoring rule

$$S^*(r, x) = \int_{-\infty}^{x} S((R(u), 1 - R(u)), \mathbf{e}_2)\, du + \int_{x}^{\infty} S((R(u), 1 - R(u)), \mathbf{e}_1)\, du. \quad (5)$$

Extending the notion of "more distant" to the continuous case, we say that a continuous cdf R' is *more distant from the value* x than another continuous cdf R if $R' \neq R$, $R'(u) \geq R(u)$ for $u \leq x$, and $R'(u) \leq R(u)$ for $u \geq x$. Then (5) defines a family of continuous, sensitive-to-distance scoring rules corresponding to the infinite number of possible rules S that are strictly proper for a dichotomy.

The strategy used to generate the RPS for the n-state case from a standard quadratic scoring rule, as illustrated in (4), and to generate a family of sensitive-to-distance rules for the continuous case, as illustrated in (5), can be used to generate a family of sensitive-to-distance scoring rules for the n-state case. If S is strictly proper for a dichotomy ($n = 2$ states), then we define a new rule $\tilde{S}$ as follows:

$$\tilde{S}(\mathbf{r}, \mathbf{e}_j) = \sum_{i=1}^{j-1} S((R_i, 1 - R_i), \mathbf{e}_2) + \sum_{i=j}^{n-1} S((R_i, 1 - R_i), \mathbf{e}_1). \quad (6)$$

This is a discrete version of Matheson and Winkler's (1976) family of continuous rules. Note that when $n = 2$, $\tilde{S} = S$. As mentioned above, the spirit of the notion of sensitivity to distance kicks in only for $n \geq 3$.

Proposition 1. *If S is strictly proper, the scoring rule $\tilde{S}$ given by* (6) *is a strictly proper scoring rule.*

Proof. The expected score is

$$\tilde{S}(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^{n-1} P_i S((R_i, 1 - R_i), \mathbf{e}_1) + (1 - P_i) S((R_i, 1 - R_i), \mathbf{e}_2). \quad (7)$$

For each i in (7), the maximizing solution can be obtained from the following system of equations:

$$\sum_{k \leq i} r_k = \sum_{k \leq i} p_k \quad \text{and} \quad \sum_{k > i} r_k = \sum_{k > i} p_k.$$

Iteratively, we get $r_i = p_i$ for all i. Hence, $\mathbf{r} = \mathbf{p}$ is a maximizing solution for the aggregate problem. To show that it is unique, suppose that $\mathbf{r}'$ is a maximizing solution distinct from $\mathbf{p}$. Let m be the first

state for which $\mathbf{r}$ differs from $\mathbf{p}$. Then $P_m S((R'_m, 1-R'_m), \mathbf{e}_1) + (1-P_m)S((R'_m, 1-R'_m), \mathbf{e}_2)$ is strictly lower than $P_m S((P_m, 1-P_m), \mathbf{e}_1) + (1-P_m)S((P_m, 1-P_m), \mathbf{e}_2)$ because S is strictly proper, and the other terms in the summation can do no better. Hence, we have a contradiction: $\mathbf{r}'$ cannot be a maximizing solution. □

Example 3. We illustrate the application of (6) using the logarithmic score. When event j occurs,

$$\tilde{S}(\mathbf{r}, \mathbf{e}_j) = \sum_{i=1}^{j-1} \log(1-R_i) + \sum_{i=j}^{n-1} \log R_i.$$

If $n=4$, for example, the expected score is

$$\tilde{S}(\mathbf{r}, \mathbf{p}) = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}^T \begin{bmatrix} \log R_1 & \log R_2 & \log R_3 \\ \log(1-R_1) & \log R_2 & \log R_3 \\ \log(1-R_1) & \log(1-R_2) & \log R_3 \\ \log(1-R_1) & \log(1-R_2) & \log(1-R_3) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Similar to the RPS, this scoring rule is sensitive to distance. Consider two distributions $\mathbf{r}$ and $\mathbf{r}'$ that differ in only two states. For example, let $a < b \le j$, with $r'_a = r_a + \epsilon$ and $r'_b = r_b - \epsilon$, where $\epsilon > 0$. This implies that r' is more distant than r if state j occurs. Now, if we let $H_j(\mathbf{r}, \mathbf{r}') = \tilde{S}(\mathbf{r}, \mathbf{e}_j) - \tilde{S}(\mathbf{r}', \mathbf{e}_j)$, the improvement in score in moving from $\mathbf{r}'$ to $\mathbf{r}$, we then have

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=a}^{b-1} \log(1-R_i) - \sum_{i=a}^{b-1} \log(1-R_i-\epsilon) = \sum_{i=a}^{b-1} \log \frac{(1-R_i)}{(1-R_i-\epsilon)}.$$

Note that $(1-R_i)/(1-R_i-\epsilon) > 1$, so that $\log[(1-R_i)/(1-R_i-\epsilon)] > 0$ for each of the terms in the summation. Therefore, $H_j(\mathbf{r}, \mathbf{r}') > 0$. A similar argument can be made when $a > b \ge j$. Thus, as shown more generally in Proposition 2, the rule is sensitive to distance.

Proposition 2. *Let $\tilde{S}$ be a ranked scoring rule generated from* (6) *by a strictly proper scoring rule S. Then $\tilde{S}$ is sensitive to distance.*

Proof. Using the notation from Example 3, we have the following form for $H_j(\mathbf{r}, \mathbf{r}')$, with $r'_a = r_a + \epsilon$ and $r'_b = r_b - \epsilon$:

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=a}^{b-1} S((R_i, 1-R_i), \mathbf{e}_2) - S((R_i+\epsilon, 1-R_i-\epsilon), \mathbf{e}_2) \quad \text{when } a < b \le j,$$

$$H_j(\mathbf{r}, \mathbf{r}') = \sum_{i=b}^{a-1} S((R_i, 1-R_i), \mathbf{e}_1) - S((R_i-\epsilon, 1-R_i+\epsilon), \mathbf{e}_1) \quad \text{when } a > b \ge j.$$

Because S is strictly proper, $S((r, 1-r), \mathbf{e}_1)$ must be increasing in r whereas $S((r, 1-r), \mathbf{e}_2)$ should be decreasing in r. Hence, H_j must be positive in both cases. Finally, any distribution that is more distant with respect to an event j can be created through a step-by-step transfer of probability mass away from the event of concern. Hence, $\tilde{S}$ is sensitive to distance. □

Example 4. Using (6), the power and pseudospherical families in (1) and (2) can be extended to take into account sensitivity to distance:

$$\tilde{S}^P_\beta(\mathbf{r}, \mathbf{e}_j) = \sum_{i=1}^{j-1} S^P_\beta((R_i, 1-R_i), \mathbf{e}_2) + \sum_{i=j}^{n-1} S^P_\beta((R_i, 1-R_i), \mathbf{e}_1) \quad \text{and} \tag{8}$$

$$\tilde{S}^S_\beta(\mathbf{r}, \mathbf{e}_j) = \sum_{i=1}^{j-1} S^S_\beta((R_i, 1-R_i), \mathbf{e}_2) + \sum_{i=j}^{n-1} S^S_\beta((R_i, 1-R_i), \mathbf{e}_1). \tag{9}$$

This construction allows us to have greater flexibility in the type of scoring rule that we can use in problems with ranked categories. It provides an extensive set of strictly proper scoring rules $\tilde{S}$ that are sensitive to distance. Propositions 1 and 2 can also be extended to allow for the use of a different S for each j in (6), increasing the generality of the results, but using the same S for each j should provide sufficient flexibility in most cases. Also, in choosing a scoring rule S to use in generating $\tilde{S}$, it is important to realize that properties of S need not carry over to $\tilde{S}$. As noted in §2, for example, among rules that are not sensitive to distance, the logarithmic rules $S(\mathbf{r}, \mathbf{e}_j) = \log r_j$ in the discrete case and $S(r, x) = \log r(x)$ in the continuous case, along with any positive linear transformations of these rules, are the only strictly proper rules satisfying locality. When a logarithmic S is used in (6), however, the resulting $\tilde{S}(\mathbf{r}, \mathbf{e}_j)$ clearly depends on elements of $\mathbf{r}$ other than r_j and therefore does not satisfy locality. Similarly, when a logarithmic S is used in (5), the resulting $S^*(r, x)$ clearly depends on $r(u)$ for $u \ne x$. Except in the trivial case when $n=2$, no sensitive-to-distance rules satisfy locality because the locality property is inherently inconsistent with the sensitive-to-distance property.

4. Scoring Rules with Baselines

Often we are interested in evaluating probabilities relative to a baseline distribution. For example, in a location where precipitation occurs on average only 5% of

the time we might want to evaluate a probability of precipitation relative to a baseline probability of 0.05. A score used often in weather forecasting with this notion in mind is the skill score (Skill):

$$\text{Skill}(\mathbf{r}, \mathbf{e}_j) = \frac{S(\mathbf{r}, \mathbf{e}_j) - S(\mathbf{q}, \mathbf{e}_j)}{S(\mathbf{e}_j, \mathbf{e}_j) - S(\mathbf{q}, \mathbf{e}_j)}, \tag{10}$$

where S is a strictly proper scoring rule and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ is a baseline distribution that is strictly positive (i.e., $q_i > 0$ for $i = 1, \ldots, n$). Skill measures the improvement in score from the baseline $\mathbf{q}$ to $\mathbf{r}$, divided by the improvement from $\mathbf{q}$ to a perfect forecast. The denominator of Skill is positive, and the numerator can be positive or negative, corresponding to positive or negative skill.

Unfortunately, the skill score does not encourage truth telling. Murphy (1973) shows that because the skill score is not strictly proper, hedging (untruthful reporting) could be a problem. On the other hand, strictly proper scoring rules typically used in practice do not allow evaluation relative to a chosen baseline distribution. For standard (nonranked), symmetric, strictly proper scoring rules such as the members of the families defined in (1) and (2), a uniform distribution $(1/n, \ldots, 1/n)$ provides the lowest possible expected score under honest reporting and is the default baseline distribution in the sense that the score obtained with a distribution $\mathbf{r}$ can be viewed as being relative to this minimum score with the uniform distribution. In contrast, probabilities of one-half for the two extreme events of an ordered space and zero for all other events yield the lowest expected score under honest reporting for the RPS, so this is the default baseline distribution for the RPS. These distributions might be thought of as the "least informative" beliefs that one could have in the two situations, where the difference between the two is that with the RPS we care about sensitivity to distance. Proposition 3 shows that the baseline distribution for the RPS holds for a wide class of strictly proper sensitive-to-distance scoring rules.

Proposition 3. *For a scoring rule $\tilde{S}$ generated from* (6) *using a symmetric, strictly proper S, the expected score under honest reporting is minimized when* $\mathbf{r} = \mathbf{p} = (0.5, 0, \ldots, 0, 0.5)$.

Proof. From (7), the expected score $\tilde{S}(\mathbf{r}, \mathbf{p})$ is the sum of expected scores using S for $n-1$ dichotomies. Because S is symmetric, the expected score when $R_i = P_i$ is minimized at $P_i = 0.50$. Setting $\mathbf{p} = (0.5, 0, \ldots, 0, 0.5)$ yields $P_i = 0.5$ for $i = 1, \ldots, n-1$, so each of the $n-1$ expected scores is minimized individually. Therefore, the overall expected score is minimized. □

Proposition 3 helps to explain a seemingly strange result, which is that for the RPS as well as for other scores generated from (6), the score obtained with a uniform $\mathbf{r}$ depends on the event that actually occurs. In contrast, a symmetric, strictly proper scoring rule that is not sensitive to distance gives the same score to a uniform $\mathbf{r}$ regardless of which event occurs. For the rules generated from (6), the score obtained is the same for all events when $\mathbf{r} = (0.5, 0, \ldots, 0, 0.5)$. In each case, the "least informative" baseline distribution yields the same score for all events. Like locality, the desire to have a uniform $\mathbf{r}$ receive the same score for all events is inherently inconsistent with sensitivity to distance. When sensitivity to distance is taken into account, events farther from the middle of the set of ordered events receive lower scores under a uniform $\mathbf{r}$ because there is more probability that is more distant from such events. However, note that when $\mathbf{r} = (0.5, 0, \ldots, 0, 0.5)$, the cumulative probabilities that are used in (6) are uniform: $(R_1, \ldots, R_{n-1}) = (0.5, 0.5, \ldots, 0.5, 0.5)$. We might think of the choice between these two "least informative" baseline distributions as analogous to the choice of a diffuse prior in Bayesian analysis. With a Bernoulli process, for which the natural-conjugate family is beta on $[0, 1]$, with density proportional to $p^a(1-p)^b$, two commonly used diffuse priors are $a = b = 0$ and $a = b = -1$. The former is uniform, and the latter is U-shaped with a density that becomes infinite as $p \to 0$ or 1.

When there is a baseline distribution that represents the notion of "least informative" in a given situation or is a relevant reference distribution for some other reason (e.g., when it represents the prior beliefs of a decision maker who must act on the basis of the forecast), it is useful to be able to scale a scoring rule such that it attains a minimum expected score for honest reporting at this reference distribution. One way to address this issue is to create asymmetric strictly proper rules that are related to a baseline distribution while retaining the property of strict propriety, as done in Winkler (1994) and Jose et al. (2008) for unordered spaces. Using the construction in the previous section, we can create similar scores for ordered state spaces. In particular, by using the families of ranked power and pseudospherical scores, we can create new scoring rules that are strictly proper, are sensitive to distance, and permit the choice of a baseline distribution that is not necessarily uniform, thereby further generalizing (8) and (9).

The power and pseudospherical scores with baseline $\mathbf{q}$ are as follows:

$$S_\beta^P(\mathbf{r}, \mathbf{e}_j \| \mathbf{q}) = \frac{(r_j/q_j)^{\beta-1} - 1}{\beta - 1} - \frac{\mathbb{E}_\mathbf{r}[(\mathbf{r}/\mathbf{q})^{\beta-1}] - 1}{\beta} \quad \text{and} \tag{11}$$

$$S_\beta^S(\mathbf{r}, \mathbf{e}_j \| \mathbf{q}) = \frac{1}{\beta - 1}\left[\left(\frac{r_j/q_j}{\mathbb{E}_\mathbf{r}[(\mathbf{r}/\mathbf{q})^{\beta-1}]^{1/\beta}}\right)^{\beta-1} - 1\right], \tag{12}$$

where $\beta \in \mathbb{R}$ and $\mathbb{E}_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}] = \sum_{i=1}^{n} r_i(r_i/q_i)^{\beta-1}$. We can construct corresponding families of power and pseudospherical rules that are sensitive to distance by applying (6) to dichotomous versions of the ranked power and pseudospherical scores. As with $\mathbf{p}$ and $\mathbf{r}$, we denote the cumulative probabilities corresponding to $\mathbf{q}$ by an uppercase $\mathbf{Q}$. Then we can write the ranked power and pseudospherical scores with baseline $\mathbf{q}$ as follows, where S_β^P and S_β^S are the power and pseudospherical scores in (11) and (12) for a dichotomy, $\mathbf{R}_j = (R_j, 1-R_j)$, and $\mathbf{Q}_j = (Q_j, 1-Q_j)$:

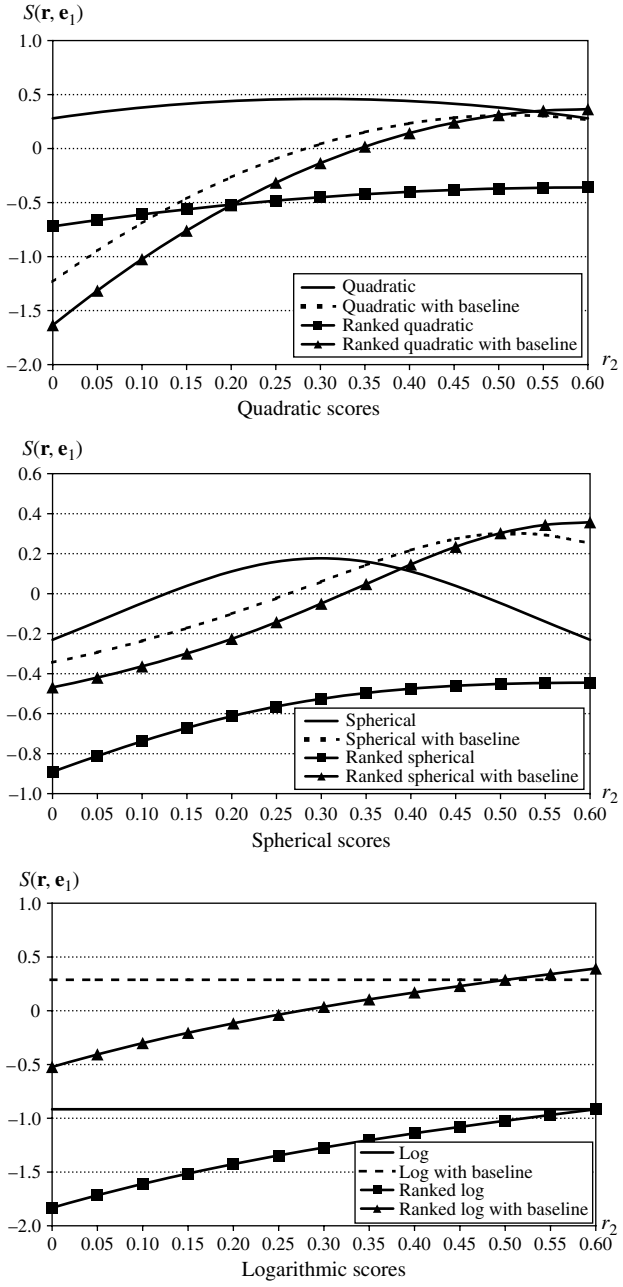
$$\tilde{S}_\beta^P(\mathbf{r}, \mathbf{e}_j \| \mathbf{q}) = \sum_{i=1}^{j-1} S_\beta^P(\mathbf{R}_i, \mathbf{e}_2 \| \mathbf{Q}_i) + \sum_{i=j}^{n-1} S_\beta^P(\mathbf{R}_i, \mathbf{e}_1 \| \mathbf{Q}_i) \quad \text{and} \tag{13}$$

$$\tilde{S}_\beta^S(\mathbf{r}, \mathbf{e}_j \| \mathbf{q}) = \sum_{i=1}^{j-1} S_\beta^S(\mathbf{R}_i, \mathbf{e}_2 \| \mathbf{Q}_i) + \sum_{i=j}^{n-1} S_\beta^S(\mathbf{R}_i, \mathbf{e}_1 \| \mathbf{Q}_i). \tag{14}$$

Example 5. Figure 1 shows plots of quadratic, spherical, and logarithmic scores as a function of r_2 for a three-event example with $\mathbf{r} = (0.4, r_2, 0.6 - r_2)$ and $j = 1$. In each case, four scores are considered: the standard score from (1) or (2), the ranked score from (8) or (9), the standard score from (11) or (12) with baseline $\mathbf{q} = (0.3, 0.6, 0.1)$, and the ranked score from (13) or (14) with the same baseline. These plots illustrate how scores can differ when a ranked score is used as opposed to a standard (nonranked) score and when a baseline different from the default baseline is considered, and they also demonstrate that the scores generated by different values of β have different properties. For example, the standard quadratic and spherical scores are maximized when $r_2 = 0.3$, the value for which $r_2 = r_3$. The corresponding scores with the baseline are maximized near $r_2 = 0.5$. The logarithmic scores with and without the baseline are constant in r_2 because they depend on $\mathbf{r}$ only through r_1, although they differ from each other because the score with the baseline also depends on q_1. All of the ranked scores, with or without the baseline, are increasing in r_2; with $j = 1$ and r_1 fixed, the ranked scores are higher as more probability is shifted to event 2, the event closest to event 1.

Note that the shapes of the curves in Figure 1 and the relationships among these curves differ not only depending on whether the score is ranked or whether a nondefault baseline is used, but also for the different scoring rules. Of course, this is a single example, and the curves will differ as the details change. It is clear, however, that considering sensitivity to distance and making comparisons with a nondefault baseline can significantly change the nature of the scores.

Figure 1 **Scores as a Function of r_2 for Different Scoring Rules When $\mathbf{r} = (0.4, r_2, 0.6 - r_2)$ and $j = 1$**



We can develop a general approach for creating strictly proper rules that are sensitive to distance and attain a minimum expected score when the forecaster's distribution is equal to a baseline distribution $\mathbf{q}$. Let f be any function on the unit square such that (1) $|f(x, y)| < \infty$ for all $(x, y) \in [0, 1]^2$; (2) $f(x, y^*)$ is strictly convex in x for any fixed $y^* \in (0, 1)$ and attains a minimum at $x = y^*$; and (3) for any fixed $y^* \in (0, 1)$ and for every $x \in [0, 1]$, f has a subgradient f' at x, i.e., there exists f' such that $f(x^*, y^*) \geq f(x, y^*) + f'(x, y^*)(x^* - x)$ for all $x^* \in [0, 1]$. Using the Schervish (1989) representation of strictly proper scoring rules,

we can generate a strictly proper binary scoring rule for a fixed baseline $(q, 1-q)$ as follows:

$$S^f((p,1-p),\mathbf{e}_1\|(q,1-q)) = f(p,q)+(1-p)f'(p,q) \quad \text{and} \tag{15}$$

$$S^f((p,1-p),\mathbf{e}_2\|(q,1-q)) = f(p,q)-pf'(p,q). \tag{16}$$

Proposition 4. *The scoring rule*

$$\tilde{S}^f(\mathbf{r},\mathbf{e}_j\|\mathbf{q}) = \sum_{i=1}^{j-1} S^f(\mathbf{R}_i,\mathbf{e}_2\|\mathbf{Q}_i)+\sum_{i=j}^{n-1} S^f(\mathbf{R}_i,\mathbf{e}_1\|\mathbf{Q}_i) \tag{17}$$

generated using the binary scoring rule S^f from (15) *and* (16) *attains a minimum possible expected score for honest reporting if and only if* $\mathbf{p}=\mathbf{q}$.

Proof. We can write the expected scores for honest reporting given a fixed $\mathbf{q}$ as follows, where $\mathbf{P}_j = (P_j, 1-P_j)$ and $\mathbf{Q}_j = (Q_j, 1-Q_j)$:

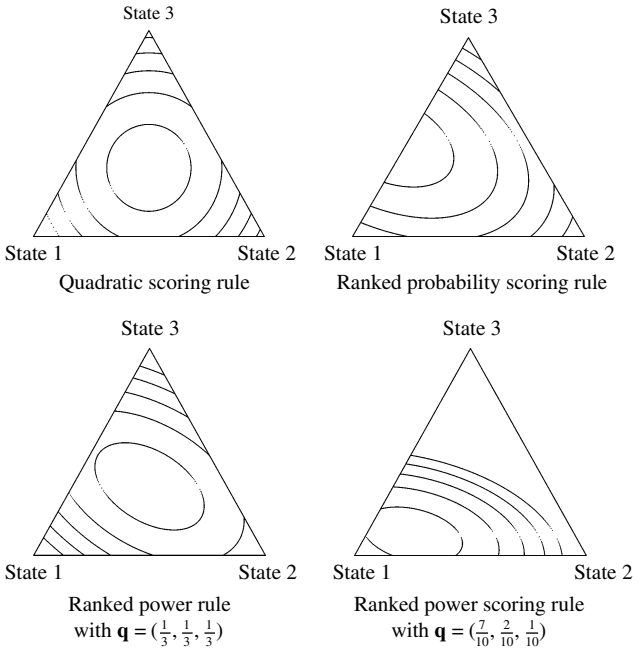
$$\begin{aligned}\tilde{S}^f(\mathbf{p},\mathbf{p}\|\mathbf{q}) &= \sum_{j=1}^{n} p_j\left[\sum_{i=1}^{j-1} S^f(\mathbf{P}_i,\mathbf{e}_2\|\mathbf{Q}_i)+\sum_{i=j}^{n-1} S^f(\mathbf{P}_i,\mathbf{e}_1\|\mathbf{Q}_i)\right]\\ &= \sum_{j=1}^{n-1}\left[P_j S^f(\mathbf{P}_j,\mathbf{e}_1\|\mathbf{Q}_j)+(1-P_j)S^f(\mathbf{P}_j,\mathbf{e}_2\|\mathbf{Q}_j)\right]\\ &= \sum_{j=1}^{n-1} f(P_j,Q_j).\end{aligned}$$

Because f is minimized when $P_j = Q_j$, $\tilde{S}^f$ attains a minimum when $\mathbf{p}=\mathbf{q}$. Uniqueness follows from strict convexity. □

The power and pseudospherical rules in (13) and (14) are examples of the general rule in (17). For example, if f is the power divergence (Havrda and Charvrát 1967), then $\tilde{S}^f$ from (17) is the ranked power score with baseline $\mathbf{q}$. Proposition 4 provides a way to generate a wide variety of strictly proper rules that are sensitive to distance and attain a minimum expected score for a distribution equal to a baseline distribution $\mathbf{q}$. The power and pseudospherical forms are included in this variety of rules and provide flexible families of rules that have intuitive appeal, with special cases that are related to well-known scoring rules that are commonly used in decision and risk analysis applications and as convenient analytical forms for theoretical work. Specifically, (13) and (14) extend those well-known rules to include sensitivity to distance and evaluation relative to a baseline distribution. Therefore, we feel that the power and pseudospherical rules offer sufficient richness for all practical purposes.

Example 6. Figure 2 shows contours of the expected score when $n=3$ for the quadratic scoring rule, the RPS, and the power scoring rule with $\beta=2$ and baselines $(1/3, 1/3, 1/3)$ and $(7/10, 2/10, 1/10)$. The quadratic score attains a minimum expected score at the point $(1/3, 1/3, 1/3)$, similar to the ranked power scoring rule centered at the same point. Unlike the quadratic score, however, the latter score is only symmetric along the middle state because this score is sensitive to distance. For the RPS, Figure 2 shows that a minimum is attained at $(1/2, 0, 1/2)$. The two plots at the bottom of Figure 2 indicate how the expected score changes when a baseline other than $(1/2, 0, 1/2)$ is selected. The minimum expected score for the ranked power scoring rule is attained at the baseline distribution, as shown in Proposition 3. Though the baseline is restricted to be strictly positive, we can approximate the RPS by choosing a baseline $(1/2-\epsilon/2, \epsilon, 1/2-\epsilon/2)$ and making $\epsilon > 0$ arbitrarily small.

Figure 2 **Expected Score Contours for Four Different Quadratic Scores When $n=3$**



The families of scoring rules generated in (11) and (12) measure the skillfulness of a forecaster in improving upon a relevant baseline distribution, in the spirit of the skill score in (10), and do it in a way that maintains strict propriety and therefore encourages honest reporting. In turn, the scoring rules (13) and (14) generated by the power and pseudospherical families yield strictly proper scoring rules that are sensitive to distance and attain a minimum expected score when the forecaster's belief is equal to a baseline distribution $\mathbf{q}$. This provides a rich set of scoring rules $\tilde{S}$ that are strictly proper, are sensitive to distance, and incorporate a baseline distribution.

5. Summary and Discussion

Most theoretical studies and real applications of scoring rules focus on rules that ignore any inherent

ordering of the states and implicitly evaluate probability distributions relative to a uniform baseline distribution. There may be good reasons to take into account the ordering of states, so that assigning more probability "close to the state that actually occurs" is a sign of a better forecast. Similarly, because scores provide a comparative evaluation relative to a reference distribution, there may be good reason to use a baseline other than the default uniform distribution in order to reward forecasters who provide more informative forecasts relative to that baseline distribution. In practice, interest in these issues has been confined primarily to weather forecasting, where the RPS was developed to take into account ordering by being sensitive to distance, and the skill score was developed to allow comparisons to meteorological baselines such as climatology or persistence. The RPS is strictly proper and a good candidate when the ordering of the states is of interest. Its quadratic nature is arbitrary, however, and the default baseline distribution $(0.5, 0, \ldots, 0, 0.5)$ is restrictive. As for the skill score, it has the disadvantage of not being strictly proper. Jose et al. (2008) provide a rich set of strictly proper scoring rules, the power and pseudospherical families, that can incorporate the notion of a baseline distribution but do not take into account any ordering of the states.

In this paper, we develop a convenient way of constructing scoring rules that are strictly proper, sensitive to distance, and able to incorporate a baseline distribution relative to which the value of a forecast is measured. This enables us to further enrich the power and pseudospherical families of scoring rules to allow for sensitivity to distance, with or without a specified baseline distribution.

Of course, with greater flexibility in the choice of scoring rules comes a greater challenge in choosing an appropriate rule, including the basic form of the rule (e.g., power versus pseudospherical with a choice of β in either case) and the baseline distribution if one is desired. Some thoughts about the choice between power and pseudospherical rules and the choice of β for rules not sensitive to distance are given in Jose et al. (2008). A baseline distribution could be based on the decision maker's prior distribution, on what is viewed as a "least informative" distribution in a given situation, or on some other relevant benchmark. We feel that the choice of a baseline distribution can be important because not giving a specific baseline means reverting to the default baseline, which often seems unsuitable. The choice of whether to consider a rule that is sensitive to distance depends on whether having more probability close to the event that occurs is thought to be valuable in the sense of making it more likely that a decision made on the basis of the probabilities will yield a better outcome. The most suitable rule to use in a given case is highly dependent on the context of the application and the properties that the decision maker thinks are important, which makes it difficult to give specific rules of thumb. Keep in mind that not all properties of standard scoring rules (rules that are not sensitive to distance and do not incorporate a baseline distribution) necessarily carry over to the corresponding ranked scoring rules and/or scoring rules with baselines. Fortunately, the most important property, strict propriety, does carry over with our construction.

The sensitive-to-distance scoring rules developed to date, including those in this paper, are based on the definition of "more distant than" given in §3. That definition treats the probabilities separately for events on the two sides of the event that occurs. Murphy (1970) formulates a different definition of "more distant than" based on symmetric sums of probabilities for sets of events centered at the event that occurs and demonstrates that the RPS is not sensitive to distance according to this symmetric definition. The continuous rule in (5) and the n-state rule in (6) are generated in the same manner as the RPS is in (4), and these rules are also not sensitive to distance in the symmetric sense. Whether there are strictly proper rules that are sensitive to distance according to the symmetric definition or, even stronger, to both definitions, is an open question. We have tried the most obvious way to modify (6) to deal with symmetric sums, and the result is not sensitive to distance in the symmetric sense. As Staël von Holstein (1970b) points out, the symmetric definition implies that if you are wrong, the direction in which you are wrong does not matter.

If desired, the results here can be extended to yield even broader families of scoring rules. As noted in §3, (6) can be generalized to allow for a different binary rule S for each j. Similarly, we could incorporate weights for each element of the sums in (6), creating a weighted sum of scores that could put greater weight on certain outcomes. This can be done for the new rules in (11)–(14) and (17) also, and it is analogous to the use of the weighting function $dG(u)$ by Matheson and Winkler (1976) in the continuous case. For example, if the extreme categories are viewed as more important in some sense than central categories, as might be the case in investing when the extreme categories represent unusually large positive or negative returns, those categories could be weighted more heavily.

Acknowledgments

The authors are grateful to the reviewers for helpful comments.

References

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**(6) 985–987.

Gneiting, T., A. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102**(477) 359–378.

Havrda, J., F. Chavrát. 1967. Quantification method of classification processes: The concept of structural α-entropy. *Kybernetika* **3** 30–35.

Jose, V. R. R., R. F. Nau, R. L. Winkler. 2008. Scoring rules, generalized entropy, and utility maximization. *Oper. Res.* **56**(5) 1146–1157.

Matheson, J., R. L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Sci.* **22**(10) 1087–1096.

McCarthy, J. 1956. Measures of the value of information. *Proc. Natl. Acad. Sci. USA* **42** 654–655.

Murphy, A. H. 1970. The ranked probability score and the probability score: A comparison. *Monthly Weather Rev.* **98**(12) 917–924.

Murphy, A. H. 1971. A note on the ranked probability score. *J. Appl. Meteorol.* **10**(1) 155–156.

Murphy, A. H. 1973. Hedging and skill scores for probability forecasts. *J. Appl. Meteorol.* **12**(1) 215–223.

Nau, R. F. 2007. Scoring rules that are sensitive to distance and dominance. Working paper, Duke University, Durham, NC.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**(336) 783–801.

Schervish, M. J. 1989. A general method for comparing probability assessors. *Ann. Statist.* **17**(4) 1856–1879.

Staël von Holstein, C.-A. S. 1970a. A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteorol.* **9**(3) 360–364.

Staël von Holstein, C.-A. S. 1970b. *Assessment and Evaluation of Subjective Probability Distributions*. The Economic Research Institute at the Stockholm School of Economics, Stockholm.

Winkler, R. L. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Sci.* **40**(11) 1395–1405.

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