A Family of Strictly Proper Scoring Rules Which Are Sensitive to Distance[1]

CARL-AXEL S. STAËL VON HOLSTEIN

University of Stockholm and The Economic Research Institute at the Stockholm School of Economics

(Manuscript received 16 January 1970)

ABSTRACT

An assessment which is "further away" from the "true" event than another assessment should receive a lower score. A definition of distance (from the true event) is considered as well as the sensitivity of scoring rules to distance. A family of scoring rules for $n \times n$ cost-loss ratio decision situations with incomplete knowledge is defined and is shown to be strictly proper. These scoring rules are modified to give a family of strictly proper scoring rules which are also sensitive to distance.

1. Introduction

The purpose of this paper is to describe a certain *family* of scoring rules which could be used to evaluate probability forecasts in meteorology. These scoring rules are particularly suitable when the variable of concern is ordered.

We shall assume that a forecast can be expressed by a probability vector, $\mathbf{r}=(r_1, \cdots, r_n)$, for n mutually exclusive and collectively exhaustive events. It is further assumed that the forecaster's true judgment can be represented by the vector, $\mathbf{p}=(p_1, \cdots, p_n)$, and $\mathbf{r}$ need not necessarily be equal to $\mathbf{p}$. (The forecaster might not be careful enough when formulating his forecast, or he might expect a better evaluation if he chooses a forecast which differs from his judgment, or his utility function might be nonlinear in the score.) It would be a desirable property of a scoring rule that it encourage the forecaster to be honest, i.e., to make $\mathbf{r}$ equal to $\mathbf{p}$.

A scoring rule is a function of the forecast $\mathbf{r}$ and the event which eventually turns out to be true. The forecaster receives a score $S_j(\mathbf{r})$ if the jth event occurs. His subjective expected score is then $S(\mathbf{r},\mathbf{p})$, where

$$S(\mathbf{r},\mathbf{p})=\sum_j p_j S_j(\mathbf{r}).$$

According to the terminology of Murphy (1969e),[2] a scoring rule is *proper* if

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

i.e., if no forecast receives a higher[3] score than $\mathbf{p}$. (There could, however, be forecasts $\mathbf{r} \neq \mathbf{p}$ which also receive the maximum score.) If the expected score is maximized only if $\mathbf{r}=\mathbf{p}$, then the scoring rule is *strictly proper*. We then have

$$S(\mathbf{p},\mathbf{p}) > S(\mathbf{r},\mathbf{p}), \text{ if } \mathbf{r} \neq \mathbf{p}.$$

For a general discussion of scoring rules see, for instance, Winkler and Murphy (1968) and Staël von Holstein (1970), who also give some examples of strictly proper scoring rules.

In many practical applications the events will represent an ordered measure of some quantity. For instance, the event k may mean more precipitation than event $k-1$ for each k. It would then be a desirable property of a scoring rule that it be a function not only of the probabilities $\mathbf{r}$ but also of the event numbers. Let us consider the following two forecasts (0.1, 0.1, 0.3, 0.3, 0.2) and (0.3, 0.1, 0.1, 0.3, 0.2). Most strictly proper scoring rules in use today will give the two forecasts the same score if the fourth event turns out to be true, although in most applications the first forecast would probably be regarded as better than the second one.

It would therefore be valuable to find strictly proper scoring rules that discriminate between such forecasts. We can formalize this kind of discriminating ability in the following way.

A forecast $\mathbf{r}'$ is said to be *more distant from the true event* than another forecast $\mathbf{r}$ if $\mathbf{r}' \neq \mathbf{r}$ and

$$R_i' \geq R_i, \qquad i=1, \cdots, j-1,$$
$$R_i' \leq R_i, \qquad i=j, \cdots, n-1,$$

[1] The preparation of this paper was supported by a grant from The Bank of Sweden Tercentenary Fund.

[2] The literature on scoring rules has used the term "proper" instead of "strictly proper." Staël von Holstein (1970) therefore introduced the term "quasi-proper" (equivalent to "proper" as defined in this paper) to describe certain scoring rules based on utility matrices. This terminology is now abandoned in order to avoid confusion with Murphy's definitions.

[3] Some scoring rules may be defined in such a way that a low score is preferred to a high score. They are then said to have a negative rather than a positive orientation (Winkler and Murphy, 1968). Such rules will be proper if the forecaster minimizes his subjective expected score by setting $\mathbf{r}=\mathbf{p}$. The following discussion could easily be modified to include such rules.

where event j is the true event and

$$R_i=\sum_{k=1}^{i} r_k.$$

In other words, if we consider any tail (on either side of event j) then $\mathbf{r}'$ has at least the same probability mass in that tail as $\mathbf{r}$. It may be easier to recognize this if the second inequality is rewritten as

$$1-R_i'\geq 1-R_i,$$

or

$$\sum_{k=i+1}^{n} r_k'\geq \sum_{k=i+1}^{n} r_k, \quad i=j,\cdots,n-1.$$

A scoring rule is said to be *sensitive to distance* if

$$S_j(\mathbf{r})>S_j(\mathbf{r}'),$$

whenever $\mathbf{r}'$ is more distant from the true event than $\mathbf{r}$.

All reasonable scoring rules are sensitive to distance when $n=2$. For $n\geq 2$, the first rule of this kind was the *ranked probability score*, defined by Epstein (1969) as

$$S_j(\mathbf{r})=\tfrac{3}{2}-\frac{1}{2(n-1)}\sum_{i=1}^{n-1}\left[\left(\sum_{k=1}^{i} r_k\right)^2+\left(\sum_{k=i+1}^{n} r_k\right)^2\right]-\frac{1}{n-1}\sum_{i=1}^{n}|i-j|r_i.$$

Murphy (1969d) has shown that this scoring rule is strictly proper. It will be proved in Section 5 that it is also sensitive to distance.

In this paper we shall relax one of the assumptions underlying the derivation of the ranked probability score (namely, that the distribution of the cost-loss ratio is uniform) and show that this leads to a family of strictly proper scoring rules which are sensitive to distance. These rules are intended for evaluating forecasts of ordered variables.

2. Utility measures

A forecast will generally be used as the basis for some decision. Therefore, the evaluation of a forecast ought to be related to its economic consequences. Most scoring rules, however, do not explicitly do this.

We shall assume that the decision situation can be formulated in the following way. A decision maker has to choose one of m alternatives, $A_1, \cdots, A_m$. This choice depends on which of n events, $s_1, \cdots, s_n$, is the true event. The utility of the choice of A_i when s_j is the true event is u_{ij}. The probability forecast[4] for the n events is $\mathbf{r}=(r_1, \cdots, r_n)$. We shall assume that the decision maker wants to maximize the expected utility.

[4] The forecast is made either by the decision maker or by an independent forecaster, e.g., a meteorologist. In the latter case it is assumed that the forecaster's utility function coincides with the decision maker's utility function. This is a strong assumption but it can be justified, at least approximately, in many practical situations.

Let W_i be the set of forecasts for which A_i is at least as good as all other alternatives, i.e.,

$$W_i=\{\mathbf{r}\,|\sum_j r_j u_{ij}\geq\sum_j r_j u_{hj}, \quad \text{for each } h\neq i\}.$$

If we use the utility as an evaluation of the forecast, we have the scoring rule

$$S_j(\mathbf{r})=u_{ij}, \quad \text{when } \mathbf{r}\notin W_i.$$

The forecaster's expected score is then

$$S(\mathbf{r},\mathbf{p})=\sum_j p_j u_{ij}, \quad \text{when } \mathbf{r}\notin W_i.$$

We now let $p\notin W_a$. In this case we have

$$S(\mathbf{p},\mathbf{p})-S(\mathbf{r},\mathbf{p})=\sum p_j u_{aj}-\sum p_j u_{ij}\geq 0,$$

with equality only when $\mathbf{r}\notin W_a$. The utility measure is thus proper but not strictly proper (any $\mathbf{r}\notin W_a$ will maximize the expected score). This result[5] holds irrespective of the utility matrix $\{u_{ij}\}$.

3. A family of scoring rules for $n\times n$ cost-loss ratio decision situations with incomplete knowledge

We shall now restrict our attention to a particular family of utility matrices, which are related to the following: Let $s_1, \cdots, s_n$ be weather states which require successively less protection and let $A_1, \cdots, A_n$ be choices of the amount of protection. The cost of protection with alternative A_i is $(n-i)C/(n-1)$, where C is the cost of the maximum protection available. Let s_j be the true weather state. If $j\geq i$ the protection is sufficient, otherwise there is a cost of $(i-j)L/(n-1)$, where L is the cost of the greatest possible loss. The total cost of choosing A_i when s_j occurs is therefore

$$c_{ij}=\begin{cases}\dfrac{n-i}{n-1}C, & j\geq i\\[2ex] \dfrac{n-i}{n-1}C+\dfrac{i-j}{n-1}L, & j<i\end{cases}$$

where $i, j=1, \cdots, n$. Introducing the utility u_{ij} by

$$u_{ij}=1-c_{ij}/L,$$

and writing X for the cost-loss ratio C/L, we have

$$u_{ij}=\begin{cases}1-\dfrac{n-i}{n-1}X, & j\geq i\\[2ex] 1-\dfrac{i-j}{n-1}-\dfrac{n-i}{n-1}X, & j<i\end{cases}$$

[5] Murphy (1969a) has given an earlier proof which appeared in a dissertation and was therefore not known to this author.

where $i, j=1, \cdots, n$. The expected value of choosing A_i, given the forecast $\mathbf{r}$, is

$$1-\frac{n-i}{n-1}X-\sum_{j=1}^{i-1} r_j\frac{i-j}{n-1}.$$

It can easily be shown that the decision maker should choose A_i whenever $\mathbf{r} \notin W_i$, where

$$W_i=\{\mathbf{r}|\sum_1^{i-1} r_j<X\leq\sum_1^{i} r_j\}.$$

Introduce

$$R_i=\sum_1^{i} r_j, \quad (R_0=0),$$

and

$$\delta(W_i)=\begin{cases}1, & \text{if} \quad X\notin W_i\\ 0, & \text{if} \quad X\notin W_i\end{cases}$$

We shall write $U_j(\mathbf{r},X)$, or simply $U_j(X)$, for $S_j(\mathbf{r})$ to indicate the dependence of the score on X. Then,

$$U_j(X)=\sum_{i=1}^{n} u_{ij}\delta(W_i).$$

The cost-loss ratio X will not be known with certainty in many applications and it is then appropriate to treat X as a random variable with some distribution function F, defined on the interval $[0,1]$ (see, e.g., Murphy, 1969b,c). This distribution represents the decision maker's knowledge of X. Different values of X will lead to different choices of alternatives A_i. The score selected is the expected utility $E[U_j(X)]$, or simply EU_j, where the expectation is taken with respect to F. Then,

$$EU_j=\int U_j(x)dF(x)=\int\left[\sum_{i=1}^{n} u_{ij}\delta(W_i)\right]dF(x),$$

$$=\sum_{i=1}^{n}\int_{R_{i-1}}^{R_i} u_{ij}dF(x),$$

$$=\sum_{i=1}^{n}\int_{R_{i-1}}^{R_i}\left(1-\frac{n-i}{n-1}x\right)dF(x)-\sum_{i=j+1}^{n}\int_{R_{i-1}}^{R_i}\frac{i-j}{n-1}dF(x).$$

Murphy (1969c) studied this scoring rule for $n=2$ and proved that the resulting rule (which he calls the *expected-utility measure*) is strictly proper. Epstein (1969) has studied the rule for a general n but with $F(x)=x$, $0\leq x\leq 1$ (see also Section 5).

4. Proof that the scoring rule EU_j is strictly proper

We shall now prove that this rule is strictly proper for a large family of distributions, i.e., those distributions which have a positive density function f. Then $f(x)=F'(x)$. It will be convenient to use the following notations:

$$G(x)=\int_0^x tf(t)dt \quad [\text{this implies } G'(x)=xf(x)],$$

$$\mu=G(1)[=E(X)].$$

We can write EU_j as

$$EU_j=1-\frac{1}{n-1}\{\sum_{i=1}^{n}(n-j)[G(R_i)-G(R_{i-1})]$$

$$+\sum_{i=j+1}^{n}(i-j)[F(R_i)-F(R_{i-1})]\},$$

$$=\frac{1}{n-1}[j-1-\sum_{i=1}^{n-1}G(R_i)+\sum_{i=j}^{n-1}F(R_i)].$$

The forecaster's expected score is

$$S(\mathbf{r},\mathbf{p})=\sum p_jEU_j.$$

We shall study $S^*(\mathbf{r},\mathbf{p})=(n-1)S(\mathbf{r},\mathbf{p})$ for notational convenience. Then,

$$S^*(\mathbf{r},\mathbf{p})=\sum_{j=1}^{n} jp_j-1-\sum_{i=1}^{n-1}G(R_i)+\sum_{j=1}^{n} p_j\sum_{i=j}^{n-1}F(R_i).$$

Introduce

$$\Phi(\mathbf{r})=S^*(\mathbf{r},\mathbf{p})-\lambda(\sum r_i-1),$$

where λ is a Lagrangian multiplier. Maximizing $\Phi(\mathbf{r})$ is then equivalent to maximizing $S^*(\mathbf{r},\mathbf{p})$ under the condition that $\sum r_i=1$. Taking the partial derivative with respect to r_k, say, and setting the derivative equal to zero gives

$$\Phi_k'=\frac{\partial\Phi(\mathbf{r})}{\partial r_k}=-\sum_{i=k}^{n-1}G'(R_i)+\sum_{j=1}^{k}p_j\sum_{i=k}^{n-1}f(R_i)$$

$$+\sum_{j=k+1}^{n}p_j\sum_{i=j}^{n-1}f(R_i)-\lambda=0, \quad \text{for } k=1,\cdots,n.$$

Eliminating λ between two adjacent equations results in

$$\Phi_{k+1}'-\Phi_k'=0, \quad \text{for } k=1,\cdots,n-1.$$

We thus have

$$G'(R_k)+\sum_{j=1}^{k+1}p_j\sum_{i=k+1}^{n-1}f(R_i)$$

$$-\sum_{j=1}^{k}p_j\sum_{i=k}^{n-1}f(R_i)-p_{k+1}\sum_{i=k+1}^{n-1}f(R_i)=0,$$

or

$$G'(R_k)-\sum_{j=1}^{k}p_jf(R_k)=0,$$

or

$$R_k f(R_k) - \sum_{j=1}^{k} p_j f(R_k) = 0,$$

or

$$f(R_k)(\sum_{j=1}^{k} r_j - \sum_{j=1}^{k} p_j) = 0, \quad \text{for } k = 1, \cdots, n-1.$$

Since $f(x) > 0$ for $0 \leq x \leq 1$, it follows that

$$\sum_{j=1}^{k} r_j = \sum_{j=1}^{k} p_j, \quad \text{for } k = 1, \cdots, n-1,$$

or

$$r_k = p_k, \qquad \text{for } k = 1, \cdots, n.$$

It can easily be shown that this extremum point is a maximum point. In order to maximize his expected score $S(\mathbf{r},\mathbf{p})$, the forecaster has to make his forecast $\mathbf{r}$ agree with his judgment $\mathbf{p}$. The scoring rule EU_j is therefore a strictly proper scoring rule.

5. A family of strictly proper scoring rules which is sensitive to distance

The score EU_j has a maximum (when $r_j = 1$) equal to $1-(n-j)\mu/(n-1)$ [where $\mu = E(X)$]. The minimum is obtained either when $r_1 = 1$ or $r_n = 1$ and it is equal to $1 - \max\{\mu, (n-j)/(n-1)\}$. The goodness of a forecast, as reflected by its score, is therefore dependent on what weather occurs. Even a perfect forecast, i.e., $r_j = 1$, may receive a rather poor score because the weather was poor.

In order to have a score that is less dependent on which event actually occurs we can proceed as follows. Define a new score EU_j^+ which is similar to EU_j but based on the utility matrix $\{u_{ij}^+\}$ where

$$u_{ij}^+ = u_{i,n+1-j}.$$

This is equivalent to making state 1 (instead of state n) the state requiring the least protection of all states. EU_j^+ is evidently strictly proper. The final score is defined as

$$S_j(\mathbf{r}) = EU_j + EU_j^+ - (1-\mu),$$

or

$$S_j(\mathbf{r}) = 1 + \mu - \frac{1}{n-1}\Big[\sum_{i=1}^{n} (n-i) \int_{R_{i-1}}^{R_i} x f(x)\,dx + (i-1) \times \int_{1-R_i}^{1-R_{i-1}} x f(x)\,dx\Big] - \frac{1}{n-1}\Big[\sum_{i=1}^{j} (j-i) \times \int_{1-R_i}^{1-R_{i-1}} f(x)\,dx + \sum_{i=j+1}^{n} \int_{R_{i-1}}^{R_i} (i-j) f(x)\,dx\Big],$$

or

$$S_j(\mathbf{r}) = \mu - \frac{1}{n-1} \sum_{i=1}^{n-1} [G(R_i) + G(1-R_i)] + \frac{1}{n-1}\Big[\sum_{i=1}^{j-1} F(1-R_i) + \sum_{i=j}^{n-1} F(R_i)\Big].$$

It is easily seen that the maximum is 1 irrespective of the true state.[6] This scoring rule is strictly proper since it is the sum of two strictly proper scoring rules.

We shall now prove that the scoring rule is sensitive to distance. Assume first that $\mathbf{r}$ and $\mathbf{r}'$ are two forecasts which only differ for two states a and b, i.e.,

$$r_a' = r_a + \epsilon \text{ and } r_b' = r_b - \epsilon, \qquad a < b \leqslant j, \qquad \epsilon \geq 0.$$

Let $H(\epsilon) = (n-1)[S_j(\mathbf{r}) - S_j(\mathbf{r}')]$, or

$$H(\epsilon) = \sum_{i=a}^{b-1} [-G(R_i) - G(1-R_i) + G(R_i+\epsilon) + G(1-R_i-\epsilon) + F(1-R_i) - F(1-R_i-\epsilon)].$$

Then

$$\frac{\partial H(\epsilon)}{\partial \epsilon} = \sum_{i=a}^{b-1} [G'(R_i+\epsilon) - G'(1-R_i-\epsilon) + f(1-R_i-\epsilon)],$$

$$= \sum_{i=a}^{b-1} [(R_i+\epsilon) f(R_i+\epsilon) - (1-R_i-\epsilon) \times f(1-R_i-\epsilon) + f(1-R_i-\epsilon)],$$

$$= \sum_{i=a}^{b-1} (R_i+\epsilon)[f(R_i+\epsilon) + f(1-R_i-\epsilon)].$$

Since f is positive, we have

$$\frac{\partial H(\epsilon)}{\partial \epsilon} \geq 0.$$

$H(\epsilon)$ is therefore minimized when $\epsilon = 0$, i.e., when $\mathbf{r}' = \mathbf{r}$. The result would be similar if $j \geq b > a$.

We could have written $\mathbf{r}'$ as

$$\mathbf{r}' = \mathbf{r} + \epsilon \mathbf{y}(a,b),$$

where

$$\mathbf{y}(a,b) = (y_1, \cdots, y_n), \text{ and } y_i = \begin{cases} 1, \text{ when } i = a \\ -1, \text{ when } i = b \\ 0, \text{ otherwise} \end{cases}$$

Generally, if $\mathbf{r}'$ is more distant from the true event than $\mathbf{r}$, then $\mathbf{r}'$ can be written as

$$\mathbf{r}' = \mathbf{r} + \Sigma \epsilon_i \mathbf{y}(a_i,b_i), \ \epsilon_i > 0; \ a_i < b_i \leq j \text{ or } j \leq b_i < a_i.$$

That is, we can "get from $\mathbf{r}$ to $\mathbf{r}'$" in a finite number of steps. By the preceding argument, each step will reduce the score $S_j(\mathbf{r})$ since $\epsilon_i > 0$. The scoring rule is therefore sensitive to distance.

6. An alternative definition of distance

It is easy to think of definitions of other distance concepts, which in turn might lead to other definitions of sensitivity to distance. One suggestion is as follows

[6] Epstein (1969) derived the ranked probability score in this way by using $f(x) = 1$ for $0 \leq x \leq 1$.

(Murphy, 1969, personal communication): $\mathbf{r}'$ is symmetrically more distant than $\mathbf{r}$ (from the true event j) if $C_i' \leq C_i$ for all i $[i=0, 1, \cdots, \max(j-1, n-j)]$, where

$$C_i = \sum_{k=j-i}^{j+i} r_k$$

(i.e., the C_i represent symmetric sums of the r_k centered on r_j). One implication of this definition is that if you are wrong, the direction in which you are wrong does not matter. This is a strong assumption but there may be applications where it is reasonable.

Consider the following two forecasts, $\mathbf{r}=(0.05, 0.10, 0.20, 0.35, 0.30)$ and $\mathbf{r}'=(0.10, 0.10, 0.20, 0.30, 0.30)$. Assume that the third event is true. The ranked probability score assigns the scores 0.866 and 0.875 to $\mathbf{r}$ and $\mathbf{r}'$, respectively; thus $\mathbf{r}'$ receives the higher score although it is more distant than $\mathbf{r}$ according to the symmetric definition. Therefore, we cannot use this definition of distance if we want the ranked probability score to be sensitive to distance. The definition used in this paper is more restrictive than the symmetric definition and it cannot say that one of the two forecasts is more distant than the other.

7. Conclusion

The choice of the density function f for the scoring rule defined in Section 5 is arbitrary as long as $f(x)>0$, for $0 \leq x \leq 1$. This means that we can generate a family of scoring rules by varying f. It is evident that the calculations will be rather cumbersome if f is other than the uniform distribution which leads to the ranked probability score. The density function f is a quantification of the decision-maker's judgment about the cost-loss ratio. It is conceivable that this could perhaps be approximated by a density function of the beta type. The computations would then be reduced to a set of incomplete beta integrals, which exist in tabulated form. The resulting scoring rules would be an extension of the *beta measures* as defined by Murphy (1969b).

The family of scoring rules defined in this paper was derived from a particular utility matrix. A subject for further research is to see what general conditions must be imposed on a utility matrix and on the formulation of incomplete knowledge about its components in order to arrive at scoring rules that are strictly proper and/or sensitive to distance.

Acknowledgments. The author would like to thank Dr. Allan H. Murphy and Mr. Björn Leonardz for valuable comments on an earlier version of this paper.

REFERENCES

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Murphy, A. H., 1969a: The evaluation of probabilistic predictions in meteorology. Ph.D. thesis, Dept. of Meteorology and Oceanography, University of Michigan.

——, 1969b: Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteor.*, **8**, 863–873.

——, 1969c: On expected-utility measures in cost-loss ratio decision situations. *J. Appl. Meteor.*, **8**, 989–991.

——, 1969d: On the "ranked probability score." *J. Appl. Meteor.*, **8**, 988–989.

——, 1969e: A note on proper and strictly proper scoring rules. Unpubl. manuscript, Dept. of Meteorology and Oceanography, University of Michigan.

Staël von Holstein, C.-A. S., 1970: Some problems in the practical application of Bayesian decision theory. *Behavioral Approaches to Modern Management*, Gothenburg, The Graduate School of Economics and Business Administration (in press).

Winkler, R. L., and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.*, **7**, 751–758.