Comments on *Proper Scoring Rules, Dominated Forecasts, and Coherence*

Martin Barrett
mbrrtt@wisc.edu

May 15, 2008

Introduction

References are to the bibliography in Schervish, Seidenfeld and Kadane.

The authors (henceforth: SSK) take as their point of departure results proved in [Lieb, et al 2007] concerning the coherence of unconditional forecasts with continuous scoring rules, and generalize them widely to cover conditional forecasts and a much larger class of scoring rules. As a consequence, we understand coherence$_1$ and coherence$_3$ are coextensive for a wide class of scoring rules, and not just for the classical rule called the Brier score, for which this was proved by De Finetti in 1974. This long paper marks a definite advance in the study of the mathematics of proper scoring rules. It is to the philosophical significance of these results, however, that I shall largely devote my remarks.

The Uses of Scoring Rules

Scoring rules were introduced originally to measure the empirical success of probabilistic forecasters. A paradigmatic example is provided by the weather forecaster who puts the chance of rain tomorrow at, say, 40%. A scoring rule provides a numerical measure of the forecaster's prediction by imposing a penalty which is a function both of the forecast for the event and whether it occurred or not.

Call this usage of scoring rules *evaluation*.

Let A be an uncertain event and x with $0 \leq x \leq 1$ be the agent's forecast probability for A. In the notation of SSK, the quadratic scoring rule, or Brier Score, is defined by $g_{0,A}(x) = x^2$, $g_{1,A}(x) = (1-x)^2$. When x is the forecast, $g_{1,A}(x)$ is the score (the penalty imposed) if A occurs, while $g_{0,A}(x)$ is the score if it does not. This rule has the intuitive property that when A occurs the penalty is relatively large if the prevision x is small but relatively small if x is large, and conversely when A does not occur. This is what makes the rule useful for evaluation.

It was understood early that to be effective for evaluation, scoring rules must have built in some incentive for the agent's forecasts to honestly reflect her opinions. Otherwise, we could have a scoring rule such as $g_{0,A}(x) = 0$, $g_{1,A}(x) = 1$ for $0 \leq x < 1$, $g_{1,A}(1) = 0$. This rule encourages the agent always to forecast the probability of A as 1, as long as her actual opinion of that probability is > 0.

Hence we have the stipulation that acceptable scoring rules must be *strictly proper* – that is, the agent's expected loss for a forecast x, using her own opinion p to compute the expectation, must be minimized exactly when she proffers $x = p$ as the forecast and only then. (This is to hold for all possible values of p, $0 \leq p \leq 1$.) It is readily shown that the quadratic scoring rule is strictly proper, while the rule discussed in the previous paragraph is not.

Because of its mathematical simplicity and tractability, quadratic scoring is still the most commonly encountered rule for evaluation, but there are myriad other strictly proper rules, both continuous and discontinuous. Some are used in actual evaluations.

The usage of scoring rules in [De Finetti 1974] and by subsequent investigators pursuing De Finetti's program is of a quite different sort. Here they serve as a basis for a method of *elicitation* of an agent's subjective probabilities, the specifics of which I will discuss shortly. Elicitation is distinct from evaluation because in the former, the concern is only with accurately extracting the probabilities themselves and not with whether they represent good or bad forecasts of the events in question.

De Finetti was quite explicit about the structure of the collection of subjective probabilities for a rational agent. He argued that they were given by a positive linear functional over the space of random variables (a.k.a. "uncertain quantities", "random quantities") presented to a given agent at a given time. Positivity means that $\mathrm{P}(X) \geq 0$ whenever $X(\omega) \geq 0$ for all $\omega \in \Omega$, the space of possibilities. He called $\mathrm{P}(X)$ a prevision for random quantities in general and a probability if $X = A$, the indicator function of an event.

He supposed that an agent may not have immediate introspective access to all of these previsions at any time (p. 84), and also (in my interpretation of his writing) that an agent may wrongly – inconsistently – judge the values of some $\mathrm{P}(X)$. She may judge, for example, that $\mathrm{P}(X) = 2$, $\mathrm{P}(Y) = 1$, and $\mathrm{P}(X + Y) = 0$, contradicting the linearity of P. These inconsistent judgments require correction of some sort; there are more subtle and less easily detectable ways to be inconsistent.

A Simplified Setup

In their paper, SSK consider possibly infinite collections of conditional forecasts. (Conditional forecasts are of quantities of the form $\Pr(A|B)$; unconditional forecasts are then just special cases $\Pr(A|\Omega)$). They allow each forecast in a given collection to be scored by a different scoring rule, which may be any continuous or discontinuous strictly proper scoring rule meeting some extra requirements (assumptions 1-3 on p. 5).

Now much of this generality is not required in order to consider the relation of SSK's results to De Finetti's program, nor is it needed to understand the very interesting examples which SSK offer to justify the various hypotheses of their theorems.

Therefore, order to keep the discussion manageable, I employ the following simplifications: consider situations which involve only a *finite* collection of mutually exclusive and exhaustive events (propositions if you prefer) $\{A_1, \ldots, A_n\}$, for $A_i \subset \Omega$. So we don't have to consider the distinction between finitely additive and countably additive probability. Following De Finetti, identify an event $A_i \subset \Omega$ with its indicator or characteristic function $A(\omega)$. All events thus become random variables. Constant functions $c(\omega)$ are denoted $\mathbf{c}$, so that we have the function identity $A_1 + \cdots + A_n = \mathbf{1} = \Omega$.

The collection $\mathcal{A}$ of all linear combinations of $A_1, \ldots, A_n$ forms a vector space of random quantities, and the only random quantities we consider are in this space. Note that by the function identity above, $\mathbf{1}$, the random quantity which has the value 1 with certainty, is in $\mathcal{A}$. These random quantities may be interpreted as representing uncertain gains or losses to the agent; we adopt SSK's convention that a positive quantity represents a loss. (De Finetti's convention is the opposite.) The agent's utilities are assumed to be linear in money so that the losses may be considered to be units either of dollars or utility.

The agent is faced with the task of making a sequence of *unconditional* probablistic forecasts $(x_1, \ldots, x_n)$: judgments of $\mathrm{P}(A_1), \ldots, \mathrm{P}(A_n)$. A minimum amount of coherence is built in to the forecasts at the outset: we assume that $0 \leq x_i \leq 1$ for all i.

When scoring rules are involved, the simplifying assumption is that only a single strictly proper scoring rule $g = (g_0, g_1)$ is employed, although it need not be smooth like the Brier Score or even continuous. The total score s_i when A_i is true is defined by $s_{g,i} = g_0(x_1) + g_0(x_2) + \cdots + g_1(x_i) + \cdots + +g_0(x_n)$. The total score can be written as a random quantity (function of ω):

$$S_g(x_1, \ldots, x_n) = s_1 A_1 + \cdots + s_n A_n = \sum_i (g_1(x_i) - g_0(x_i) A_i + \sum_i g_0(x_i)$$

Thus the total score function for the Brier score is the random quantity

$$S_B(x_1, \ldots, x_n) = \sum_i ((1 - x_i)^2 - x_i^2) A_i + \sum_i x_i^2 = \sum_i (1 - 2x_i) A_i + \sum_i x_i^2$$

In terms of the scoring rule g, a forecast $(x_1, \ldots, x_n)$ is strictly dominated by another forecast $(y_1, \ldots, y_n)$ if

$$S_g(y_1, \ldots, y_n) < S_g(x_1, \ldots, x_n).$$

Elicitation à la De Finetti

De Finetti [1974] describes two methods of operational elicitation of probabilities, with associated criteria of coherence. I view both of these methods as

"free choice with constraints". That the coherence requirements are supposed to function as constraints is obvious, but I think is also clear that De Finetti regarded agents in the forecasting (elicitation) situation as possibly possessing usable information about P in the form of judgments about the values of $\mathrm{P}(X)$ for some random quantities $X \in \mathcal{A}$, or possibly less specific information in the form of judgments of inequalities such as $\mathrm{P}(A_1) < 1/2$, etc. Such judgments also function as constraints on choice.

Brief exegetical argument for this point: his descriptions of both procedures involve the agent as being able to make some judgments about $\mathrm{P}(X)$ or at least express preferences between random quantities, which in his terminology means comparisons of expectations. For example, in his second (Brier score) procedure, the agent is able to judge (not stipulate), for some x_i, x, that the random quantity $(A_i - x_i)^2$ is preferable to $(A_i - x)^2$, i.e. that $\mathrm{P}((A_i - x_i)^2) < \mathrm{P}((A_i - x)^2)$. More telling, however, is that after describing an agent in the first procedure as making a choice (forecast) with the constraints of coherence$_1$ (the Dutch Book constraints) and the same agent in the second procedure as making a choice with the constraints of coherence$_2$ (=no dominating forecast), he then goes on to prove that these forecasts are numerically the same! Such a supposed proof would be absurd if the forecasts were in fact unrelated free selections of a coherent forecast. What he has in mind, I believe, is better understood when he likens these procedures to *measurement*; the same probability P is presupposed (but not completely known) in both cases. Then his proof amounts to the completely standard proof that the mean $a = \mathrm{E}(X)$ minimizes the quantity $\mathrm{E}(X - a)^2$.

Therefore, here is a somewhat formalized outline of what I take to be an agent in a De Finettian type-2 elicitation situation (for the simplified setup of mutually exclusive events A_i described above, with scoring rule g).

The agent's forecast is a choice of a vector in a class $\mathcal{X}$ of vectors. Each vector $x = (x_1, \ldots, x_n)$ is also a linear functional which operates on random quantities by $x : \sum_i a_i A_i \mapsto \sum_i x_i a_i$. (If $x_i \geq 0$ and $\sum_i x_i = 1$, x as a linear functional is a probability, and all probabilities on $\mathcal{A}$ have this form.)

$\mathcal{X}$ is constrained by the following criteria:

(coherence$_3$) If $S_g(x_1, \ldots, x_n) < S_g(y_1, \ldots, y_n)$, then $(y_1, \ldots, y_n) \notin \mathcal{X}$. The inequality is of random quantities and expresses dominance. It is a fully objective criterion, which depends on g, but in no way on any judgments.

(minimization) Any forecast which is judged to have larger expected loss than some other possible forecast is ruled out. In other words, if it is judged that $\mathrm{P}(S_g(x_1, \ldots, x_n)) < \mathrm{P}(S_g(y_1, \ldots, y_n))$, then $(y_1, \ldots, y_n) \notin \mathcal{X}$. (This is tantamount to the agent seeking the vector with the smallest expected loss. Nothing is assumed in the criterion, however, about which if any vectors the agent is able to make such judgments.)

(other) Any other constraints relating to initial judgments concerning P that the agent may make. These will vary from case to case.

Grand Consistency Requirement. After all the constraints are taken into account, $\mathcal{X}$ should contain at least one element, otherwise the coherence constraints plus judgments are inconsistent.

Elicitation. The agent chooses an element from $\mathcal{X}$, which defines P. For the ideal elicitation, $\mathcal{X}$ contains just one element. If there is more than one, the choice remains operationally unspecified.

For completeness, I record one other set of constraints, which I do not include explicitly among the constraints facing an agent in a type-2 De Finettian elicitation:

(coherence$_1$) $\mathcal{X}$ contains only probabilities. (equivalent to the no-Dutch Book conditions).

I am confident that it is possible to dispute this reconstruction of De Finetti's procedure. On the other hand, it has the virtues that (a) it is (I believe) consistent, and (b) it makes some sense out of what he writes. Doubters are invited to try to find another construal with these characteristics!

It is also possible to transform this procedure into a recipe to guide a machine's construction of "personal" probabilities.

Obviously, it is conceivable in an operational instance of elicitation that no possible forecast survives filtering out by the constraints, because an agent can conceivably judge anything. This is analogous to entertaining inconsistent beliefs. But it is worth noting that minimization and coherence$_3$ are necessarily consistent with each other in the following sense: if the judgments falling into the minimization-constraint category are all made according to some one probability P (i.e., they are consistent with this probability) then the forecast vector $(p_1, \ldots, p_n)$ (the vector for P) is not ruled out by minimization and is not ruled out by coherence$_3$ either. This is because under P, the vector $(p_1, \ldots, p_n)$ is the one which minimizes the expected total score (this follows from the strict properness of g) and so cannot be ruled out by minimization, and any vector which minimizes the expected total score for any P cannot be strictly dominated.

The relevant results from SSK:

Theorem 1 *Let g be a strictly proper scoring rule. ([De Finetti 1974]) If g is Brier score, then coherence$_3$ and coherence$_1$ are equivalent, i.e. coextensive. ([Lieb, et al 2007]) If g is continuous, then coherence$_3$ and coherence$_1$ are equivalent. (SSK) If g is continuous or discontinuous and satisfies assumptions 1-3 on page 5, then coherence$_3$ and coherence$_1$ are equivalent. (These three conclusions are in order of increasing generality.)*

Some examples:

(1) $n = 2$, Brier scoring rule B, no minimization constraints, no other constraints. The scoring rule is relevant only through the coherence$_3$ constraints. Because of the first clause of theorem 1, $\mathcal{X}$ is the class of all probabilities, and so the chosen forecast can be any probability. It is perhaps a misnomer to call a choice so undetermined by the procedure an *elicitation*.

(2) $n = 2$, any scoring rule g satisfying the assumptions 1-3 of SSK, no minimization constraints, one "other" constraint: $A_1 - 1/3 \sim 0$, i.e. $\mathrm{P}(A_1) = 1/3$. Because of this and coherence$_3$, which is equivalent to coherence$_1$ by SSK's clause of theorem 1, $\mathcal{X}$ contains just one forecast: $(1/3, 2/3)$, which is therefore chosen (elicited).

(3) $n = 2$, Brier score B, a collection of comparative judgments of the following form:

$$\mathrm{P}(\sum_i (1-2x_i)A_i + \sum_i x_i^2) \lesseqgtr \mathrm{P}(\sum_i (1-2y_i)A_i + \sum_i y_i^2)$$

where P is some probability which underwrites all these judgments, for every pair of vectors x, y which themselves are probabilities. Unrealistically comprehensive set of judgments to assume, perhaps, but a theoretically admissible assumption under my reconstruction of elicitation. Then coherence$_3$, which by Theorem 1 excludes everything except probabilities, and the postulated minimization constraints combine to exclude everything except $(p_1, \ldots, p_n)$, the probability vector corresponding to P. (This must be the minimizing vector.) $(p_1, \ldots, p_n)$ is therefore chosen. Something like this, in my view, is what De Finetti had in mind when he described the second procedure for elicitation. To the extent that it fails to be realistically operational, that may be a criticism of De Finetti.

Application to SSK

After this somewhat long setup, we can now raise, and posssibly answer, a question concerning the *import* (not the ingenuity, which is undeniable) of additional scoring rules and the theorems proved by SSK.

It is clear that in the case of evaluation (*not* the concern of SSK), scoring rules other than Brier score may be germane. For the form of the scoring rule may reflect the *evaluator's* utilities (losses), while the strict properness of the rule continues to provide the same incentive for the forecasters being evaluated to report their opinions honestly. For example, an evaluation of TV weather forecasters may reflect the evaluator's (station owner's) opinion that giving a really low probability of rain when it in fact rains is *really* bad, since viewers then tend to switch their loyalty to another station. There are other scoring rules which yield higher losses than Brier score when the forecast diverges greatly from the outcome.

But returning to the case of elicitation and foundations, does the extension of the proof of the coextensiveness of coherence$_1$ and coherence$_3$ to discontinuous g further illuminate the nature of personal probability? Although this extension is in the nature of an invariance result, I think not. De Finetti introduced the type-2 elicitation because of worries about the operational cogency of type-1 elicitation (fair betting quotients, no Dutch Books, etc.) He worried that the agent's expectations about the betting opponent's choices could skew the elicitation. Type-2 elicitation introduces objectively calculable scoring functions and judgments about their expected values, and neither of these depend on any betting opponent or the agent's expectations about such; so it arguably relieves these worries. Once De Finetti proved the first clause of Theorem 1, he had established that type-2 elicitation with the Brier score would yield the same class of coherent forecasts as type-1. This is the foundational result. Extension

to further scoring rules is akin to adding new measurement procedures for the same theoretical quantities and adds nothing to secure the foundation thus laid, in my view. This is already true even with respect to the earlier (and easier) results proved in [Lieb, et al 2007].

To view this from a non-positivist perspective (and I think De Finetti was excessively burdened with positivist baggage): it is though we had first devised a procedure (Einstein's) for synchronizing accurate clocks. Does this new procedure add to our understanding of the nature of time? Arguably, in connection with the special theory of relativity, it does. Does the provision of ever more accurate atomic clocks for this purpose add anything to this understanding? Arguably not, although of course the new clocks would be important to the practice of measurement. Einstein apparently believed that his ideal clocks and their method of synchronization *defined* time, but we can usefully employ SR and separate measurement from ontology without accepting this position.

Is Non-dominance Better Than Coherence For Search?

One of the virtues of subjective probability is that the elicitation procedure described above lends itself to the construction of machine algorithms, and thus perhaps an implementation of "personal probabilities" in the sphere of AI. Here, the equivalence of coherence$_1$ and coherence$_3$ might lead to an algorithm which searches for an acceptable forecast by checking non-dominance rather than coherence. In the simplified situation I described above, the simplest way to decide coherence is simply to add up n sumbers and see whether they sum to 1. But when we relax these simplifications, the machine might be faced with making a forecast for events $X_1, \ldots, X_n$ which are not mutually exclusive, and for which, therefore, checking coherence of a forecast is not so simple.

If, in fact, it were easier to calculate whether the forecast is undominated or not than it would be to calculate directly whether it is coherent$_1$ or not, then that would vindicate the importance of the theorems, albeit for a different purpose than the authors may have envisioned.

But would it be easier? I consider a simple example to see the issues involved. Imagine forecasting $n = 16$ events $X_1, \ldots, X_{16}$ which are not mutually exclusive. They might be complicated logical compounds of 16 basic variables $B_1, \ldots, B_{16}$ such as whether or not it rains, whether or not the temperature exceeds 70 degrees F, whether or not the dewpoint exceeds 68 degrees, whether or not the barometric pressure exceeds 30 inches Hg, at four different weather stations, all at some time t. The B_i events are (I assume) logically independent (any conjunction $B_1^{'} \wedge \ldots \wedge B_{16}^{'}$, where each $B_i^{'}$ is either B_i or B_i^c, is possible), but we allow for the logical relations between the compounds X_i to be arbitrarily complicated.

Let $C = \{C_j : 0 \leq j < 65536\}$ be the partition of logical space (with $2^{16} = 65536$ constituents) generated by the B_i. Represent each X_i as a function $C \rightarrow \{0, 1\}$. $X_i(C_j) = 1$ if and only if $C_j \subset X_i$. Picturesquely, each event X_i is a long string of 0's and 1's representing which consituents of the partition are included in X_i.

Suppose the forecast under consideration for $X_1, \ldots, X_{16}$ is $x_1, \ldots, x_{16}$. To calculate coherence (coherence$_1$) directly, the machine must solve a system of $n + 1 = 17$ equations in $2^n = 65536$ unknowns p_j:

$$1 = \sum_{j=0}^{65535} p_j, \; x_i = \sum_{j=0}^{65535} p_j X_i(C_j), 1 \leq i \leq 16.$$

The solution, if any, will give the probability p_j of each constituent of the partition. Using row reduction to solve this unwieldy system may require as many as $n^2 2^n$ operations (each row reduction operation involves 2^n operations).

Would a probabilistic algorithm to search for a dominating forecast do any better– be computationally less expensive? Assume for simplicity that the machine is using Brier score. To calculate the scoring function for the forecast, it must calculate $g_0(x_i) = x_i^2$ and $g_1(x_i) = (1 - x_i)^2$ for each of the 16 x_i, and then add the appropriate scores together constituent by constituent, which is $n = 16$ additions for each of the 2^n constituents. The number of operations in the general case is roughly $n2^n \sim 1$ million. To compare two forecasts for dominance, therefore, requires roughly $3n2^n$ operations, although since dominance is disproven at the first consituent for which the subtraction of the scores for two forecasts goes the wrong way, the typical number of operations is somewhat less.

The procedure would depend on the probability that for two forecasts chosen at random, one dominates the other, which I have found it difficult to assess, even for Brier score. Computer simulations might be enlightening. But the per-forecast comparability of the direct calculation of coherence to the cost of computing one dominance comparison suggests that it is unlikely that a dominance-based algorithm would be better.

Even assuming that in at least some cases it would, however, the goal of the algorithm would of course be to select a coherent forecast. I therefore fail to see the bearing of the part of theorems which deal with the classification of types of *in*coherent$_3$ forecasts.

For instance, consider SSK's example 4, which falls under my simplified setup. Here we find that for a certain discontinuous scoring rule, an incoherent$_3$ forecast, which by definition will be dominated by some other forecast, may be dominated *only* by another incoherent$_3$ forecast. (For example (SSK p. 14), the highly incoherent forecast (.6, .7) is dominated by (5.5, 6.5) but not by any coherent forecast) Interesting as this phenomenon is, it would seem to have no effect on the specifics of any actual dominance-based algorithm for generating (eliciting) probabilities with this scoring rule, for such an algorithm would not "know" which forecasts were coherent or not in the first place. For all the specificity of this example, the point is entirely general.

A Final Remark On a Related Matter

Foundational studies in personal probability may be imagined as a journey originally conceived to fly us safely from our native country to the Republic of

Probability. Unfortunately, the existing airlines, such as Air De Finetti, Savage Airways, and Jeffrey Airlines, all fly only to the poor farming village called Finitelyadditive Township, instead of to the more distant, but far more prosperous, Countablyadditive City (which should have been called Continuity City, except that the city fathers had a penchant for prosaic names).

Most of today's travelers who buy tickets on these airlines seem to have decided that if they can fly only to Finitelyadditive Township, this must have been their intended destination all along. A more sensible attitude, I believe, would lead them to return to the travel agent for a ticket on another airline. Indeed, there are other airplanes in the philosophical fleet, such as those belonging to Indispensability Airlines, which could fly to Countablyadditive City, although some of these have a rather poor safety record.

Perhaps some airline-building entrepreneurship is still called for.