COMMUN. STATIST.—THEORY METH., 21(6), 1615-1632 (1992)

SOME ASPECTS OF PROBABILITY FORECASTING

Andrey Feuerverger and Sheikh Rahman

Department of Statistics
University of Toronto
Toronto, Canada M5S 1A1

Key Words and Phrases: probability forecasting; proper scoring rules; assessing, comparing, combining forecasts; Gaussian threshold model; elementary scores; graphical methods; Neyman-Pearson Lemma.

ABSTRACT

The problems of assessing, comparing and combining probability forecasts for a binary events sequence are considered. A Gaussian threshold model (analytically of closed form) is introduced which allows generation of different probability forecast sequences valid for the same events. Chi-squared type test statistics, and also a *marginal-conditional* method are proposed for the assessment problem, and an asymptotic normality result is given. A graphical method is developed for the comparison problem, based upon decomposing arbitrary proper scoring rules into certain *elementary scoring functions*. The special role of the logarithmic scoring rule is examined in the context of Neyman-Pearson theory.

1. INTRODUCTION

In the usual context of probability forecasting, we have a sequence of *binary events*, i.e. of Bernoulli random variables Z_i, $i = 1, 2\cdots$, which take values 0 or 1. We assume these variables to be independent, but not identically distributed. There also are one or more forecasters who -- presumably on the basis of suitable analyses -- each provide a corresponding sequence P_i, $i = 1, 2, \cdots$ of stated probabilities for the events. It is purported that $Pr[Z_i = 1] = P_i$, for all i.

Three distinct problems arise. Firstly, given one such sequence (Z_i, P_i), $i = 1, \cdots, n$, we seek to assess, nonparametrically, whether or not the Z_i can be considered

to have arisen from the stated P_i's. That is, we seek to test the null hypothesis

$$H_o: \ Pr[Z_i = 1] = P_i, \quad \text{for all } i. \tag{1.1}$$

This is the problem of *assessing* an individual forecaster. Secondly, given the forecasts P_i^A, P_i^B, $i = 1, \cdots, n$, from two (or more) forecasters, A and B, say, we wish to determine which of these forecasters provides the better forecast. This is the problem of *comparison* and it need not admit a unique resolution. Last, though not least, is the problem of *combining* the P_i^A's and P_i^B's to obtain a best possible combined forecast P_i^{AB}. The most common context in which these problems are considered is weather forecasting, but there are many other applications including probabilistic medical diagnosis, stock market prediction, measurement of knowledge, assessing fit in models having binary dependent variables, etc. Some points of entry to previous work are Clemen (1989), Dawid (1986), DeGroot and Fienberg (1986), Krzysztofowicz and Long (1991), Murphy and Winkler (1984), and Schervish (1989).

Our purpose is to explore some aspects of the problems mentioned above. In the section following we introduce a model of closed form which may be used for generating different sequences of probability forecasts valid for the same event sequence. This model is useful for application in simulation studies where suitable data sets must be generated for analysis. In sections 3, 4, and 5 respectively we present some new methods or results for each of the three main problems -- i.e. assessing, comparing and combining -- of probability forecasting. Section 6 explores connections between scoring rule methods and the Neyman-Pearson theory. Finally in section 7, some avenues for further work are indicated.

2. A GAUSSIAN THRESHOLD CROSSING MODEL.

To begin with, it is not immediately apparent that different sequences of probabilities can simultaneously all be *valid* (in the sense defined below) for the same sequence of binary event variables Z_i. A commonly cited type of example has the following form: half of the Z_i's are generated with probabilities 1/4 and the other half with probabilities 3/4, with these probabilities constituting one sequence; a second sequence of probabilities then consists of values all equal to 1/2. Strictly speaking,

however, the second sequence is not quite valid, for the *distribution* of $\sum_1^n Z_i$ is not binomial with parameters n and $1/2$ as tacitly implied, even though a typical test of fit will tend to accept H_0 too often. Interestingly, the second sequence is nevertheless *well calibrated* in the sense that the relative frequencies of the events do correspond correctly to the stated probabilities (see, for example, DeGroot and Fienberg, 1982, or Dawid, 1982). This leads us to define *validity* to mean, roughly, that the correct distributional properties (appropriate to the independence and Bernoulli P_i distributional assumptions) would be maintained if suitable replications of the system could be conducted. In any event, it is of interest to obtain other types of examples, especially ones allowing more substantive structure.

To this end, and for conceptual reasons, it is useful to consider an underlying data generation mechanism of the following kind: the outcome $Z_i = Z(X_i, Y_i, U_i, V_i)$ depends functionally upon the random vector (X_i, Y_i, U_i, V_i) which is sampled from some distribution, and the overall success probability is $P_i = E\, Z(X_i, Y_i, U_i, V_i)$. Further, we suppose that we have two forecasters, A and B say, and that A has prior access only to X_i while B has prior access only to Y_i. An entity called *nature* is considered to have access not only to X_i and Y_i, that is to all information available to any forecaster, but also to certain additional information U_i that is not available to them. For the moment, it is convenient to allow for the possibility that there is also a "component of information" V_i unknown to nature. Then it follows from the next Lemma that $P_i^A = P_i^A(X_i) = E[Z(X_i, Y_i, U_i, V_i) \mid X_i]$ is the best forecast available to A, $P_i^B = P_i^B(Y_i) = E[Z(X_i, Y_i, U_i, V_i) \mid Y_i]$ is the best forecast available to B, that $P_i^{AB} = P_i^{AB}(X_i, Y_i) = E[Z(X_i, Y_i, U_i, V_i) \mid X_i, Y_i]$ is the best forecast that could be made if the information available to both A and B were pooled, and likewise that $P_i^{nat} = P_i^{nat}(X_i, Y_i, U_i) = E[Z(X_i, Y_i, U_i, V_i) \mid X_i, Y_i, U_i]$ is nature's optimal forecast.

Lemma 2.1 P_i^A is well calibrated and is the best forecast for A in the sense of minimizing any proper scoring rule.

Proof: For the definition of proper scoring rules, see §4 below. The argument is conditional on the value of X_i. Then $P_i^A = P_i^A(X_i)$ is in fact the probability that $Z_i = 1$, so that if S is any proper scoring rule we have

$$E[S(Z_i, P_i^A) \mid X_i] \leq E[S(Z_i, \varphi(X_i)) \mid X_i]$$

for any function φ of X_i. Taking expectations then gives

$$E[S(Z_i, P_i^A)] \leq E[S(Z_i, \varphi(X_i))]$$

and the result follows. □

If the random variable V_i were omitted in the discussion above (or equivalently if nature were assumed to know V_i also) then nature's best forecast would always be categorical, taking on values 0 and 1 only (and equal to Z_i). From an applications standpoint, however, this viewpoint is often inconvenient. In any case, from these considerations, and by virtue of the properties of conditional expectation of indicator variables, it is clear now that different probability sequences can in fact be generated (for the same event sequence) that will be not only well-calibrated, but in fact fully *valid* in the sense defined above.

In the formulation above, the conditional expectation quantities P_i^A, P_i^B, P_i^{AB}, etc., are in principle determined from the joint distribution of the underlying variables X_i, Y_i, U_i, V_i, and from the form of $Z(\cdot)$. Unfortunately these probabilities will in general not have a closed functional form. For use in simulation studies and for other purposes, however, it is of some interest to have in hand models of closed form. To this end, we introduce the following Gaussian and threshold crossing based model. Thus suppose that (X, Y) is standardized bivariate normal with correlation ρ. We assume now that there are two forecasters A and B who know only X and Y respectively. Let

$$\Gamma = \mu + \alpha X + \beta Y + e, \tag{2.1}$$

where e is $N(0,1)$, and suppose that

$$Z = 1 \quad \text{if} \quad \Gamma > 0, \tag{2.2a}$$

while

$$Z = 0 \quad \text{if} \quad \Gamma \leq 0. \tag{2.2b}$$

It follows that

$$p^{AB} = P\{\Gamma > 0 \mid X = x, Y = y\} = \Phi(\mu + \alpha x + \beta y), \tag{2.3}$$

where Φ is the $N(0,1)$ cdf. Further, given $X = x$ we have $Y \sim N(\rho x, 1-\rho^2)$ or $Y = \rho x + \sqrt{1-\rho^2}\cdot\eta$ where $\eta \sim N(0,1)$ so that

$$\Gamma = \mu + (\alpha+\beta\rho)x + \beta\cdot\sqrt{1-\rho^2}\cdot\eta + e,$$

and consequently

$$p^A = P\{\Gamma > 0 \mid X = x\} = \Phi\left(\frac{\mu+(\alpha+\beta\rho)x}{\sqrt{1+\beta^2(1-\rho^2)}}\right). \qquad 2.4$$

Likewise

$$p^B = P\{\Gamma > 0 \mid Y = y\} = \Phi\left(\frac{\mu+(\beta+\alpha\rho)y}{\sqrt{1+\alpha^2(1-\rho^2)}}\right). \qquad 2.5$$

The expressions (2.4) and (2.5) provide the best possible forecasts, respectively, when only X or Y are known, while (2.3) provides the best possible forecast when both X and Y are known. By means of this model and its obvious multivariate extensions, and such closed form expressions as (2.3)-(2.5), we may generate arbitrary numbers of different probability sequences all valid for the same event sequence.

3. ASSESSING

Basic global statistics for testing the fit of a sequence of probability forecasts to a corresponding binary event sequence may be obtained for various choices of the weights W_i in the form

$$\xi = \sum_1^n W_i(Z_i - P_i), \qquad 3.1$$

which will be asymptotically normal provided that general conditions which limit the evolving proportion of P's close to 0 and 1 are met. The choice $W_i = 1$ for example was proposed in Murphy (1969) and leads to an overall calibration test statistic equivalent to the difference of means quantity $\overline{Z} - \overline{P}$. The choice $W_i = 1 - 2P_i$ was introduced in Seillier and Dawid (1987). A natural generalization is to consider a number of statistics $\xi_0, \xi_1, \cdots, \xi_s$ of the form (3.1) where ξ_j corresponds to the choice $W_i = L_j(P_i)$ for the weights, and the L_j are a sequence of polynomials respectively of degree j exactly, for

$j = 0, 1, \cdots, s$. A test statistic consistent against a wide class of alternatives can then be constructed using the vector $\xi = (\xi_0, \cdots, \xi_s)'$ and its covariance matrix Σ having entries

$$cov\,(\xi_j, \xi_k) = \sum_i L_j(P_i) L_k(P_i) P_i(1 - P_i). \tag{3.2}$$

This natural test will reject for large values of $\xi' \Sigma^{-1} \xi$ which under H_o will be χ^2_{s+1}.

Concerning the statistic $\xi' \Sigma^{-1} \xi$, it is worthwhile to note (from the identity $(A\xi)'(A\Sigma A')^{-1}(A\xi) = \xi' \Sigma^{-1} \xi$) that its value does not depend upon the choice of the polynomials L_j, as long as the degrees of the polynomials remain as stated. Hence the test statistic depends upon s only (and the selected value of s determines a tradeoff between power properties and the extent of consistency). We applied this procedure, using $s = 5$, to the cancer data set ($n = 306$) as reported in Haberman (1976), and studied by Landwehr, Pregibon and Shoemaker (1984). The latter authors examined two logit models for these data, and we used the probability sequences resulting from these two fits. The resulting χ^2 values (d.f.= 6) were 14.9 and 1.43 suggesting lack of fit in the first model, and overfitting in the second -- consistent with the other findings of these authors.

On examination of many data sets, we found that the "distribution" of the P_i's can frequently be adequately approximated by some beta density $f(p; \alpha, \beta)$. This allows an interesting decomposition of the χ^2 statistic into approximately orthogonal components: if the L_j are the Jacobi polynomials (Davis and Rabinowitz, 1984) defined on $[0,1]$, corresponding to estimates of α and β, we will then have approximate orthogonality in (3.2). Details may be found in a technical report by the authors.

Suppose now that in place of statistics of the form (3.1), we work instead with

$$\sum W_i \frac{Z_i - P_i}{\sqrt{P_i(1 - P_i)}}. \tag{3.3}$$

Then W_i based on the Legendre polynomials L_j (defined by orthogonality on $[0,1]$), for example using $W_j = L_j(i/(n+1))$, will lead to nearly orthogonal components. We also could consider Fourier weights

$$W_i = \sqrt{\frac{2}{n}} \cdot \cos \frac{2\pi ji}{n} \quad \text{and} \quad W_i' = \sqrt{\frac{2}{n}} \cdot \sin \frac{2\pi ji}{n} \tag{3.4}$$

with $j = 0, 1, \cdots, s$ which results in exact orthogonality.

The asymptotic normality of statistics such as (3.1) and (3.3) is of interest. By Lindeberg's condition (see for example Shiryayev, 1984, p.326) the asymptotic normality of (3.3), for instance, is equivalent to the approach to 0, as $n \to \infty$, for any $\epsilon > 0$, of the quantity

$$\sum_{i=1}^{n} W_i^2 \int_{|y| > \frac{\epsilon}{|W_i|}} y^2 dF_i(y) \qquad 3.5$$

where F_i is the cdf of $Y_i \equiv \dfrac{Z_i - P_i}{\sqrt{P_i(1-P_i)}}$. For weights (3.4) however, $|W_i| < \sqrt{\frac{2}{n}}$, so that (3.5) is

$$\leq \frac{2}{n} \sum_{i=1}^{n} \int_{|y| > \epsilon\sqrt{\frac{n}{2}}} y^2 dF_i(y) \qquad 3.6$$

$$\leq \frac{2}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} y^2 \left(\frac{|y|}{\epsilon\sqrt{\frac{n}{2}}} \right)^{\delta} dF_i(y) \qquad 3.7$$

$$= \frac{2^{1+\delta/2}}{\epsilon^{\delta} n^{1+\delta/2}} \sum_{i=1}^{n} |Y_i|^{2+\delta} \qquad 3.8$$

for any $\delta > 0$. We are thus led (using the Cramer-Wold devise, ibid. p.517) to the sufficient condition

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} [P_i(1-P_i)]^{-\delta} \to 0, \text{ as } n \to \infty, \qquad 3.9$$

for some $\delta > 0$, for the asymptotic normality of finite collections of terms (3.3) based on weights from (3.4). This condition limits the evolving proportion of P's close to 0 and 1. Similar calculations may be carried out for statistics of the form (3.1); for these we require, in addition to (3.9), also that the condition $\{W_j^2 P_j(1-P_j)\} \Big/ \{\sum_1^n W_i^2 P_i(1-P_i)\} = O(n^{-1})$ holds uniformly in j.

In the procedures discussed above, an excessive number of P_i too close to 0 or 1 can lead to unstable test statistics. Consequently, we may wish to carry out separate analyses for the extreme P segments of the data. The three independent p-values obtained for tests based on the data segments of $P \simeq 0$, $P \simeq 1$, and the remaining data, could then be combined (perhaps by Fisher's method). To this end we now indicate a significance testing procedure that may be applied to an extreme-P segment of a data set.

Thus suppose for illustration that (in some larger data set under investigation) we find that there are, say, 20 observation pairs (Z_i, P_i) corresponding to values of $P_i \leq .05$ given as follows:

(0, 0.0001), (0, 0.0001), (0, 0.0002), (1, 0.0003), (0, 0.0005), (0, 0.001), (0, 0.001), (0, 0.002), (0, 0.002), (0, 0.005), (1, 0.01), (0, 0.02), (0, 0.02), (0, 0.02), (0, 0.03), (0, 0.04), (0, 0.05), (0, 0.05), (0, 0.05), (0, 0.05).

Now, because all P_i here are small, the *marginal* (total) quantity $T = \sum_1^{20} Z_i$ in this segment will, under H_o, be approximately Poisson with mean $\lambda = \sum_1^{20} P_i = .3522$ so that the probability of obtaining a value at least as high as the $T = 2$ actually observed is approximately .0493. (The exact value is in fact .0471.) In many situations it will be appropriate to leave this p-value one-sided. In any case, *conditional* on $T = 2$, the configuration of the values shown appears unusual in that the $Z_i = 1$ occur in conjunction with relatively smaller P_i. In fact, for $T = 2$, there are $\binom{20}{2} = 190$ configurations for the vector $Z = (Z_1, \cdots, Z_{20})$, and the probability of any one, conditional upon $T = 2$, is given by

$$\frac{P_k P_l}{(1-P_k)(1-P_l)} \Bigg/ \sum_{i>j} \frac{P_i P_j}{(1-P_i)(1-P_j)} \qquad 3.10$$

where k, l reference the nonzero Z's. By means of a simple computer program (we used the S language) we may calculate all 190 probabilities, order them from smallest to largest, and so determine that the (conditional) probability of obtaining the configuration actually observed, or of any configuration of equal or lesser probability, is in fact 0.0257. Therefore altogether it appears that for the given data we have too many $Z = 1$ occurrences, and that these occur in a relatively unlikely configuration.

The question arises as to how the p-values -- obtained from the *marginal* and the *conditional* test procedures -- can be combined; in this respect, the viewpoint from the conditional inference model

$$f(x;\varphi,\lambda) = f(t;\varphi) \cdot f(s \mid t, \lambda), \qquad 3.11$$

where $x = (s,t)$, appears germane, at least approximately. The identifications to be made here are:

$$x \leftrightarrow Z, \quad t \leftrightarrow \sum_{i=1}^{n} Z_i, \quad \varphi \leftrightarrow \sum_{i=1}^{n} P_i,$$

$s \leftrightarrow$ observed configuration, $\lambda \leftrightarrow$ configuration probabilities.

In the model (3.11), the off-diagonal entries of the Fisher information matrix are null, so that asymptotically, inferences concerning φ and λ are independent. It is therefore tempting to treat the p-values from the *conditional* and *marginal* tests as approximately independent. However, in the present context, the given data are not being used to assess the plausibility of various values for φ and λ, but rather we are seeking to check the plausibility (goodness of fit) of the data using given values for φ and λ. We leave open the inevitable and interesting questions which now arise.

4. COMPARING

Comparative assessment of two or more forecasters is often carried out by means of *scoring rules* $S(Z,P)$. See, for example, Savage (1971). We adopt the convention that $\sum_1^n S(Z_i,P_i)$ is to be minimized. We remark here that we shall consider the scoring rules $S(Z,P)$ and $S(Z,P)+f(Z)$ to be equivalent. A scoring rule S is said to be *proper* if its expectation, $E_P S(Z,Q) \equiv P\cdot S(1,Q) + (1-P)\cdot S(0,Q)$, regarded as a function of Q, is minimized when $Q=P$, and is said to be *strictly proper* if $Q=P$ is the unique minimum. Typical rules are Brier's score $(Z-P)^2$, and the logarithmic score $-\log P^Z(1-P)^{1-Z}$. A characteristic difficulty is that the results of a comparison often depend upon which scoring rule is used.

In certain cases, one forecaster may in fact be better than (or at least as good as) another regardless of which scoring rule is used. For example, this will be the case if forecaster A is *sufficient* for forecaster B, so that B's forecasts behave stochastically like those of A plus an auxiliary randomization. See, for example, DeGroot and Fienberg (1982, 1986) We shall say that forecaster A is *uniformly at least as good* as forecaster B if A's expected score is at least as good as B's for any proper scoring rule. Related to the results of DeGroot and Fienberg we have:

Lemma 4.1 If forecaster A is sufficient for B, then A is uniformly at least as good as B.

Proof: Let S be any proper scoring rule and suppose that A stochastically dominates B. Then the forecasts of A and B are related as $P^B = f(P^A, R)$ where R is a random variable independent of P^A and Z. However since $P\{Z=1\} = P^A$ and S is proper we have $E[S(Z,P^A) \mid R] \leq E[S(Z,P^B) \mid R]$ and hence $ES(Z,P^A) \leq ES(Z,P^B)$. □

We now consider the problem of constructing a data analytic procedure for examining such questions as whether or not A is uniformly at least as good as B. It turns out that we may approach this by means of a graphical procedure based upon certain *elementary scoring rules* to be defined below. We first record the following representation result due to Shuford, Albert and Massengill (1966), Savage (1971) and Schervish (1989):

Lemma 4.2 $S(Z,P)$ is a proper scoring rule if and only if it may be represented in the form

$$S(1,P) = \int_P^1 (1-x)\cdot d\mu(x) \qquad 4.1a$$

$$S(0,P) = \int_0^P x\cdot d\mu(x) \qquad 4.1b$$

for some positive measure μ on $[0,1]$. The scoring rule will be strictly proper if and only if μ assigns positive measure to every nondegenerate subinterval of $[0,1]$.

We now define the *elementary scoring rules* $S_a(Z,P)$, for all $a \in [0,1]$, as those corresponding to (4.1) when μ is a unit point mass at a. Thus

$$\begin{aligned} S_a(1,P) &= 1-a \quad \text{if } P \leq a \\ &= 0 \quad \text{if } P > a, \end{aligned} \qquad 4.2a$$

and

$$\begin{aligned} S_a(0,P) &= a \quad \text{if } P > a \\ &= 0 \quad \text{if } P \leq a, \end{aligned} \qquad 4.2b$$

or equivalently,

$$S_a(Z,P) = (1-a)\cdot I(Z=1,\ P \leq a) + a\cdot I(Z=0,\ P>a), \qquad 4.2c$$

where I is the indicator function. Then clearly, if S is any proper scoring rule we will have

$$S(Z,P) = \int_0^1 S_a(Z,P)\, d\mu(a) \qquad 4.3$$

for some positive measure μ. In this sense the elementary scoring rules can be said to "generate" all the proper scoring rules and the following result is an immediate consequence:

Lemma 4.3 Forecaster A is uniformly at least as good as B if and only if A is at least as good as B for all elementary scoring rules.

Now suppose that we have two forecast sequences, P_i^A and P_i^B, for events Z_i. $i = 1, \cdots, n$. The comparison across all elementary rules can be carried out graphically as in Figure (4.1). For this Figure, the event sequence, and the forecasts for A and B were derived using the Gaussian threshold model with $n = 200$, and parameter values $\mu = .5$, $\alpha = .5$, $\beta = 1$, and $\rho = .25$. The forecasts used for A were the best forecasts based solely on the X's as given by (2.4), while the forecasts used for B were the best combined forecasts as given by (2.3). In this situation, we know that B's forecasts are uniformly better than those of A. In Figure (4.1), the two solid lines trace the empirical values of the totals of the elementary scores for A and for B as a ranges over $(0,1)$, except for an alteration in scaling as noted at the end of the following paragraph. The line for A is the one at the top. We remark here that by virtue of (4.2c) the computation of these score averages involves only simple counting procedures. For this data set we see that B's total score falls below A's for all α. We consequently know that for this data set, B's empirical score will be better than A's for any choice of a proper scoring rule.

To help assess the statistical significance of the differences between the two curves in Figure (4.1), we need to establish the variance of their difference

$$D(a) = \sum_{i=1}^{n} [S_a(Z_i, P_i^A) - S_a(Z_i, P_i^B)] \qquad 4.4$$

at each value of a. Now the n terms in this sum are assumed independent, but they are not identically distributed. We find

$$\text{var}(D(a)) = \sum_{i=1}^{n} [S_a(1, P_i^A) - S_a(1, P_i^B) - S_a(0, P_i^A) + S_a(0, P_i^B)]^2 \cdot P_i(1 - P_i). \qquad 4.5$$

To evaluate this last expression, we require values for the probabilities P_i^A, P_i^B. One possibility is to obtain these by fitting a logit model such as $\lambda_i = \alpha + \beta\lambda_i^A + \gamma\lambda_i^B$ to the data, where $\lambda_i = ln \frac{P_i}{1 - P_i}$ and likewise for λ_i^A and λ_i^B. There is however a more practical

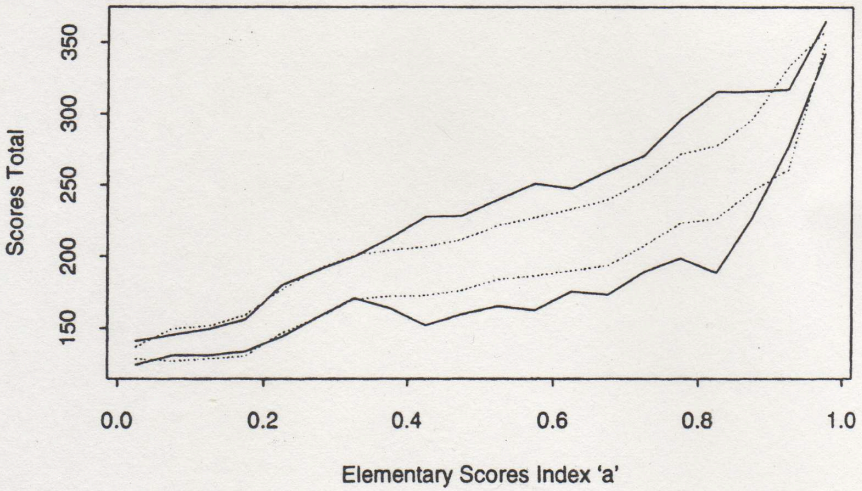

Fig. 1: Elementary Scores Plot

way to proceed. Under the null hypothesis that the forecasters are equivalent, the terms in the sum in (4.4) have zero means and can be regarded as having been sampled from a mixture distribution. But the variance of a mixture distribution is the corresponding linear combination of the variances of the component distributions, provided the means of these distributions are identical. Consequently we are led to the estimate

$$\hat{\text{var}}(D(a)) = \frac{n}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad 4.6$$

where $y_i = S_a(Z_i, P^A) - S_a(Z_i, P^B)$. Using this expression, we computed upper and lower pointwise confidence limits. These centered bands are shown plotted as dashed lines, each ± 1 standard error from the midpoint (for each a) of the score curves; thus when the total score curves (both) lie outside these bands, they are two standard errors apart and their difference is significant at the .05 level. These calculations were carried out using a combination of the S language and Fortran. Finally, we remark that we actually had used scores $S_a(Z,P)/a(1-a)$ instead of $S_a(Z,P)$ in producing Figure (4.1), with corresponding adjustments to the confidence bands; this provides a qualitatively simpler visual display for this data. Such rescaling, of course, does not affect the quantitative content.

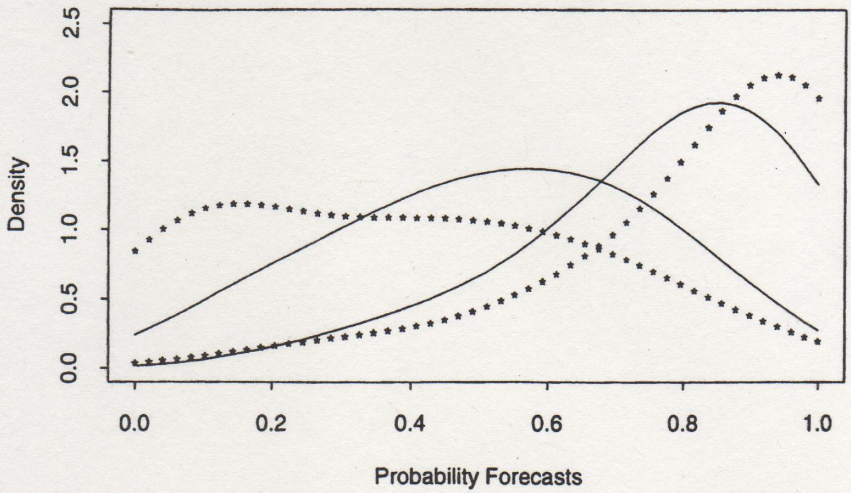

Fig. 2: Density Plots

A simpler, if less comprehensive, graphical procedure applied to the same data set is illustrated in Figure (4.2). Here we have plotted density estimates of the values of P corresponding to the events $Z = 0$ and $Z = 1$ for the two sets of forecasts. The density estimates were obtained using the "density" routine in the S computing language using the smallest kernel bandwidth that provided curves judged to be adequately smooth; each of the four density estimates is scaled so that the total area under each curve is proportional to the number of events in question. Here the starred lines correspond to forecaster B, and are seen to be more concentrated about the 0 and 1 extremes than the solid lines for forecaster A. We remark that when there is too little data to obtain reliable nonparametric density estimates, beta distributions may often be fitted to the data and plotted instead.

5. COMBINING

The problem of combining probability forecasts in an optimal way has an extensive literature. See for example Clemen (1989), Genest and Zidek (1985), Hogarth (1977), Lindley (1982).

Here we consider briefly the following problem. We suppose that the data (Z_i, P_i^A, P_i^B), $i = 1, \cdots, n$, can be considered to have been generated in accordance with the Gaussian threshold model as in equations (2.1)-(2.5), with the parameters μ, α, β, ρ being unknown. The sequences P_i^A, P_i^B of forecasts are assumed to be best possible based on knowledge, by A and B, of X_i and Y_i respectively. The underlying X_i, Y_i, e_i are unknown. We seek to estimate the best combined forecasts:

$$P_i^{AB} = \Phi(\mu + \alpha X_i + \beta Y_i). \tag{5.1}$$

(See equation (2.3).) Now, from (2.4) and (2.5) we have

$$\Phi^{-1}(P_i^A) = \mu_A + \sigma_A X_i \quad \text{and} \quad \Phi^{-1}(P_i^B) = \mu_B + \sigma_B Y_i \tag{5.2}$$

where

$$\mu_A = \frac{\mu}{\sqrt{1+\beta^2(1-\rho^2)}}, \quad \sigma_A = \frac{\alpha+\beta\rho}{\sqrt{1+\beta^2(1-\rho^2)}}, \tag{5.3a}$$

$$\mu_B = \frac{\mu}{\sqrt{1+\alpha^2(1-\rho^2)}}, \quad \sigma_B = \frac{\beta+\alpha\rho}{\sqrt{1+\alpha^2(1-\rho^2)}}. \tag{5.3b}$$

(We may assume σ_A, $\sigma_B > 0$ by reversing signs for the X's and Y's, if required. The degenerate cases σ_A, $\sigma_B = 0$ would mean that forecaster possesses no useful random information, and is not permitted.) It follows that μ_A, μ_B, σ_A, σ_B, ρ are the mean, standard deviation and correlation parameters of the bivariate Gaussian quantities $\Phi^{-1}(P_i^A)$, $\Phi^{-1}(P_i^B)$ and may be estimated in the obvious way. These estimates may then be substituted into (5.3a,b) to yield four equations (with redundancy) which may be solved (in closed form) to obtain estimates for μ, α and β. Once μ_A, μ_B, σ_A, σ_B are estimated, we may estimate X_i and Y_i by inverting (5.2) and finally estimate P_i^{AB} using (5.1). Note that one aspect of the solution is that knowledge of the Z_i's is not required. However because robustness properties have not been investigated, applicability of the Gaussian threshold model to the problem of combining forecasts may be limited to cases where the model assumptions, are met. Further details and investigations are given in the mentioned technical report available from the authors.

6. THE NEYMAN-PEARSON CONNECTION

Given an events vector $Z = (Z_1, \cdots, Z_n)$, and probability vectors denoted here as $P = (P_1, \cdots, P_n)$ and $Q = (Q_1, \cdots, Q_n)$ and corresponding to two forecasters, it is

seemingly natural to attempt to distinguish among them on the basis of the log likelihood ratio statistic

$$\log \prod_{i=1}^{n} \frac{Q_i^{Z_i}(1-Q_i)^{1-Z_i}}{P_i^{Z_i}(1-P_i)^{1-Z_i}}$$

$$= -\sum_{i=1}^{n} [S_{log}(Z_i, Q_i) - S_{log}(Z_i, P_i)],$$

where S_{log} is the logarithmic score. In the Neyman-Pearson context, this statistic is the most powerful for testing the hypothesis that the events Z were generated according to the model P versus the alternative that they were generated from Q. However, for the Neyman-Pearson Lemma to be in force, it is necessary that the vector Z have been generated according to either one or another of the hypotheses, and this condition generally need not be met in probability forecasting applications. Yet, although we know that different score functions are generally inconsistent amongst themselves, the logarithmic score appears somehow to occupy a special place here. A related issue is whether there is a particular score function that is optimal for distinguishing among any two given forecasters.

In this connection we record here three results:

Lemma 6.1 Negative log likelihood is a proper scoring rule.

Proof: This follows because the negative log likelihood for Z generated from P is

$$-\log \prod_{i=1}^{n} P_i^{Z_i}(1-P_i)^{1-Z_i}$$

which corresponds to the logarithmic scoring rule $S(Z,P)$ given by $S(1,P) = -\log P$ and $S(0,P) = -\log(1-P)$. □

Lemma 6.2 The likelihood ratio test corresponds to a comparison among forecasters based on the logarithmic scoring rule.

Lemma 6.3 If either of the two forecasters A or B is valid and stochastically dominates the other, then the likelihood ratio test based on Z, P^A, P^B is optimal for discriminating among them.

Proof: The validity implies that there is a conditioning under which Z can be regarded as having been generated from the valid probabilities sequence, while the domination implies that the other sequence is not valid under that conditioning. □

7. FINAL REMARKS

We indicate two main areas in which limited work has been done and which we believe will lead to further useful results. The first involves multivariate extension in the sense that each forecaster must provide a k-variate probabilities vector $(P_{i1},\cdots,P_{ik})$ for a k-variate binary occurences vector $(Z_{i1},\cdots,Z_{ik})$ at each time $i=1,2\cdots$. (But see DeGroot and Fienberg, 1986.) This situation arises, for example, when precipitation probabilities are required for several regions that are close enough to involve correlational effects. The other involves time series extension, as for example when daily precipitation occurences and their corresponding forecasts are in fact correlated over time. The case of stationarity is the natural one for initial study. The time series aspect has a second variant, namely the case when a sequence of probabilities is generated by providing, at regular time intervals, forecasts for the same future event. Finally, the multivariate and time series contexts can be considered in combination. It may readily be seen that the methods of the previous sections will not carry over without some modifications; in particular, the key concepts require appropriate reformulation and the test procedures must be suitably extended. We believe these to be fruitful avenues for further work.

ACKNOWLEDGEMENTS

The threshold model of §2 for generating probabilities was indicated to one of the authors by Professor D.R. Cox in conversation. We thank the referees for their valuable suggestions. It is a pleasure to acknowledge helpful conversations with M. Evans and N. Reid. We are grateful to F.W. Zwiers of the Atmospheric Environment Service for having provided us with a large volume of data that was very helpful to us in this research. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. *Int. J. Forecasting, 5,* 559-583.

Davis, P.J. and Rabinowitz, P. (1984). *Methods of Numerical Integration.* Academic Press, New York.

Dawid, A.P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc., 77,* 605-613.

Dawid, A. P. (1986). Probability forecasting. In: *Encyclopedia of Statistical Sciences.* S. Kotz, N.L. Johnson and C.B. Read, Editors. Wiley Interscience, New York.

DeGroot, M.H. and Fienberg, S.E. (1982). Assessing probability assessors: calibration and refinement. In *Statistical Decision Theory and Related Topics III, Vol 1,* 291-314, Academic Press, New York.

DeGroot, M.H. and Fienberg, S.E. (1986). Comparing probability forecasters: Basic binary concepts and multivariate extensions. In: *Bayesian Inference and Decision Techniques,* P. Goel and A. Zellner, editors, 247-264. Elsevier, Holland.

Genest, C. and Zidek, J.V. (1985). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science, 1,* 114-148.

Haberman, S.J. (1976). Generalized residuals for log-linear models. *Proc. 9th Internat. Biometrics Conference,* Boston. pp104-122.

Hogarth, R.M. (1977). Methods for aggregating opinions. In: *Decision Making and Change in Human Affairs,* H. Jungermann and G. de Zeeuw, Editors. pp.231- 235. Reidel, Dordrecht.

Krzysztofowicz, R. and Long, D. (1991). Forecast sufficiency characteristic: Construction and application. *Int. J. Forecasting, 7,* 39-45.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *J. Amer. Statist. Assoc., 79,* 61-71.

Lindley, D.V. (1982). The improvement of probability judgements. *J. Royal Statist. Soc., A, 145,* 117-126.

Murphy, A.H. (1969). Measures of the utility of probabilistic predictions in cost-loss ratio decision situation in which knowledge of the cost-loss ratio is incomplete. *J. Appl. Meteorol., 8,* 863-73.

Murphy, A.H. and Winkler, R.L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Assoc., 79,* 489-500.

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

Schervish, M.J. (1989). A general method for comparing probability assessors. *Ann. Statist., 17,* 1856-79.

Seillier, P. and Dawid, A.P. (1987). On testing the validity of probability forecasts. Research Report, Department of Statistical Science, University College, London.

Shiryayev, A.N. (1984). *Probability*. Springer-Verlag, New York.

Shuford, E.H., Albert, A. and Massengill, H.E. (1966). Admissible probability measurement procedures. *Psychometrika, 31*, 125-145.

Received September 1991; Revised January 1992