Evaluating probability forecasts in terms of refinement and strictly proper scoring rules[1]

by

Walter Krämer
Fachbereich Statistik, Universität Dortmund, Germany
Phone xx231/755-3125, Fax: xx231/755-5284
e-mail: walterk@statistik.uni-dortmund.de

Abstract

This note gives an easily verified necessary and sufficient condition for one probability forecaster to empirically outperform another one in terms of all strictly proper scoring rules.

Keywords: probability forecasts, scoring rules, refinement.

1 The problem and notation

Probability forecasting has a long and distinguished history in meteorology and medicine. Due to the increasing importance of default predictions in the credit industry, it has recently become important also in economics, so the subsequent discussion is couched in terms of default probabilities for corporate bonds.

Let $0 = a_1 < a_2 < \ldots < a_k = 1$ be k predicted probabilities of default. We circumvent the problem of converting conventional letter grades such as

[1] Research supported by Deutsche Forschungsgemeinschaft under SFB 475

AAA into predicted probabilities of default by equating the latter to historical default frequencies below. This note is not concerned with the intricacies of correctly mapping letter grades probabilities of default, but with assessing the empirical performance of competing rating agencies.

Let $q(a_i)$ be the relative frequency with which default probability forecast a_i is made and let $p(a_i)$ be the conditional relative frequency of default given probability forecast a_i. Given two rating agencies A and B who rate the same n borrowers, with frequency functions $q^A(a_i), q^B(a_i), p^A(a_i)$ and $p^B(a_i)$, it is then natural to ask whose forecasts have been better? Below it is shown that, in a sense, an unequivocal answer is possible if and only if A and B can be ranked according to the "empirical refinement ordering". Otherwise, there will always exist two strictly proper scoring rules such that one prefers A to B and the other prefers B to A.

2 The empirical refinement ordering

DeGroot and Fienberg (1983) introduce the refinement ordering among well calibrated probability forecasters. A probability forecaster is called well calibrated if, among borrowers with predicted default probability a_i, the long-run relative percentage of defaults is equal to a_i:

$$a_i = p(a_i). \tag{2.1}$$

A well calibrated forecaster A is called "more refined" than B, in symbols: $A \geq_R B$, if there exists a $k \times k$ Markov matrix M (i.c. a matrix with nonnegative entries whose columns seems to unity) such that

$$q^B(a_i) = \sum_{j=1}^{k} M_{ij} q^A(a_j) \tag{2.2}$$

and

$$a_i q^B(a_i) = \sum_{j=1}^{k} M_{ij} a_j q^A(a_j) \quad (i = 1, \ldots, k). \tag{2.3}$$

Equation (2.2) means that, given A's forecast a_j, an additional independent randomisation is applied according to the conditional distribution M_{ij} ($j = 1, ..., k$) which produces forecasts with the same probability function as that of B. Condition (2.3) ensures that the resulting forecast is again well calibrated.

Below, calibration is ensured by equating observed default rates to predicted ones. A forecaster who then dominates another one in the refinement sense is called "empirically more refined".

The crucial point for the subsequent discussion, first observed by DeGroot and Eriksson (1985), is that $A \geq_R B$ is equivalent to the fact that the distribution $q^A(a_i)$ second-order stochastically dominates the distribution $q^B(a_i)$. This allows to tap the vast literature on necessary and sufficient conditions for second order stochastic domination. In particular, we can use a theorem dating back to Hardy, Littlewood and Polya (1929) which states that $A \geq_R B$ if and only if

$$\sum_{i=1}^{k} g(a_i) q^A(a_i) \geq \sum_{i=1}^{k} g(a_i) q^B(a_i) \tag{2.4}$$

for all continuous, convex functions g on the unit interval. This key inequality is now related to scalar measures of forecasting performance known as scoring rules.

3 Strictly proper scoring rules

Let $\theta_i (i = 1, \ldots, n)$ be an indicator variable taking the value 1 if borrower i defaults and 0 otherwise, and let $P_i \in \{a_1, \ldots, a_k\}$ be the default probability attached to borrower i. A scoring rule is a function $F(\theta_1, \ldots, \theta_n; p_1, \ldots, p_n)$ which is designed to measure the performance of a forecast. Examples are the Brier-Score

$$B = -\frac{1}{n}\sum_{j=1}^{n}(p_i - \theta_i)^2, \tag{3.5}$$

the logarithmic score

$$L = -\frac{1}{n}\sum_{i=1}^{n}\ell n(|p_i + \theta_i - 1|) \tag{3.6}$$

or the spherical score

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i + \theta_i - a|}{\sqrt{p_i^2 + 1 - p_i)^2}} \tag{3.7}$$

(see e.g. Winkler 1996). A scoring rule can also be viewed as a random variable which takes a value $S_1(p)$ if the forecaster reports a predicted probability p for the event in question and the event actually occurs, and which takes a value $S_2(p)$ if the event in question does not occur. For the Brier-score, we have $S_1(p) = -(p-1)^2$ and $S_2(p) = -p^2$. A scoring rule is called "strictly proper" if its expectation, given the subjective probability distribution of the forecaster, is maximized if and only if the probability forecasts are equal to the subjective probabilities. All scoring rules above are strictly proper.

A key result about proper scoring rules, due to Savage (1971), states that a scoring rule is strictly proper if and only if the subjectively expected score for a forecaster who reports his true subjective probabilities, viewed as a function of p, is a strictly convex function. For the Brier-score, for instance, we have

$$E[B(p)] = -[p(p-1)^2 + p^2(1-p)] = -[p(1-p)]. \tag{3.8}$$

Also, any strictly convex function on the unit interval induces a strictly proper scoring rule via

$$S_1(p) = E[S(p)] + (1-p)dE[S(p)]/dp \tag{3.9}$$

and

$$S_2(p) = E[S(p)] - pdE[S(p)]/dp \tag{3.10}$$

(see Winkler 1996, section 3).

In the credit rating context, default probabilities are often equated to observed default frequencies. For this to make sense, the sample has to be quite large, of course. Then it is natural to evaluate scoring rules by attaching to borrower i the observed frequency of the grade borrower i has been sorted into.

THEOREM: If predicted default probabilities are equal to observed default rates, then forecaster A outperforms forecaster B according to all strictly proper scoring rules if and only if A is empirically more refined than B.

PROOF: The key to the proof of the theorem is to show that all empirically computed proper scoring rules, which are initially defined as functions of $\theta_1, \ldots, \theta_n$ and $p_1, \ldots, p_n$, depend on these inputs only via $a_1, \ldots, a_k$ and some strictly convex function g. To see this, note that p_i is by definition equal to the empirical default rate of grade $a_j \in \{a_1, \ldots, a_k\}$ which has been assigned to borrower i. Then the forecaster is by definition well calibrated, and a percentage $q(a_j)$ of the predicted p's are equal to a_j. For these p's and the corresponding θ's, the observed score is equal to the expected score, computed under the assumption that the realized default rate in class a_i corresponds to the predicted one:

$$S(\theta_1, \ldots, \theta_n; p_1, \ldots, p_n) = \sum_{j=1}^{k} q(a_j)[a_j S_1(a_j) + (1 - a_j) S_2(a_j)], \qquad (3.11)$$

where

$$g(a) := a S_1(a) + (1 - a) S_2(a) \qquad (3.12)$$

is a strictly convex function in view of Savage (1971). The assertion of the theorem then immediately follows from (2.4).

References

DeGroot, M. and Fienberg, S.E. (1983): "The comparison and evaluation of probability forecasters", *The Statistican* 32, 12 – 22.

DeGroot, M. and Eriksson, E.A. (1985): "Probability forecasting, stochastic dominance, and the Lorenz curve." in J.M. Bernardo et al. (eds.): *Bayesian Statistics* 2, Amsterdam, 99 – 118.

Hardy, G.H.; Littlewood, J.E. and Polya, G. (1929): "Some simple inequalities satisfied by convex functions." *Messenger Math.* 58, 145 – 152.

Savage, L.J. (1971): "Elicitation of personal probabilities and expectations." *Journal of the American Statistical Association* 66, 783 – 801.

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