On the ordering of probability forecasts[1]

by

Walter Krämer
Fachbereich Statistik, Universität Dortmund
D-44221 Dortmund, Germany
walterk@statistik.uni-dortmund.de

Version September 2002

Abstract

The paper explores the relationship between various orderings among probability forecasts that have been suggested in the literature. It is shown that well calibrated forecasters are in general not comparable according to the domination ordering suggested by Vardeman and Meeden (1983), that the orderings based on ROC-curves and Gini-curves are identical, and that the domination ordering in conjunction with semicalibration implies the rest.

Keywords: probability forecasts, calibration, refinement, domination.

[1]Research supported by Deutsche Forschungsgemeinschaft (DFG) under SFB 475. I am grateful to Andre Güttler, Martin Weber, Mark Wahrenburg, Jan Pieter Krahnen, Frank Bröker, Jörg Spannhoff and Roland Schultze for helpful criticism and comments.

1 The problem and notation

Let $0 = a_1 < a_2 < ... < a_k = 1$ be various probabilities of some event. In weather forecasting, the event could be: "It will rain tomorrow". In medicine, the event could be "Patient x will die". In the banking industry, the event could be "Borrower y will default". For concreteness, and to acknowledge the increasing importance of default predictions in the banking industry, the discussion will be couched in terms of defaults and non-defaults below. Otherwise, the notation follows Vardeman and Meeden (1983).

The paper takes the mechanism employed for the predictions as given. It is not concerned with the problem of how probability forecasts are produced (see e.g. Crouhy et al. 2001 for a survey of how risk rating systems operate in the banking industry). Rather, its point of departure is the discrete bivariate probability function $r(\theta_i, a_j)$, $i = 1, 2$, $j = 1, ..., k$, resulting from some such method, whichever it may be, with $\theta = 1$ indicating default and $\theta = 0$ indicating non-default. The following additional notation will be used:

$p(1) := \sum_j r(1, a_j)$ = overall relative frequency of default.

$p(0) := \sum_j r(0, a_j)$ = overall relative frequency of no default.

$q(a_j)$:= relative frequency with which default probability forecast a_j is made.

$p(1|a_j) := \frac{r(1,a_j)}{q(a_j)}$ = conditional relative frequency of default given probability forecast a_j.

$p(0|a_j) := \frac{r(0,a_j)}{q(a_j)}$ = conditional relative frequency of no default given probability forecast a_j.

$q(a_j|1) := \frac{r(1,a_j)}{p(1)}$ = conditional relative frequency of predicted default probability a_j given default.

$q(a_j|0) := \frac{r(0,a_j)}{p(0)}$ = conditional relative frequency of predicted default probability a_j given no default.

The problem is: given two forecasters A and B, characterized by their respective bivariate probability functions $r^A(\theta_i, a_j)$ and $r^B(\theta_i, a_j)$, which one is "better"?

One sensible requirement is that among borrowers with predicted default probability a_j, the relative percentage of defaults will be roughly equal to a_j. Formally:

$$a_j \stackrel{!}{=} p(1|a_j) = \frac{r(1, a_j)}{q(a_j)}$$

whenever $q(a_j) > 0$. Such forecasters are called "well calibrated" (Dawid 1982).

However, calibration, though desirable, is not sufficient for a useful forecast. For instance, a probability forecaster attaching default probability $p(1)$ to all borrowers is well calibrated but otherwise quite useless. Other criteria which have been suggested in the literature consider the concentration of default in the "bad" grades or the concentration of the non-defaults in the "good" grades, or whether A's forecasts can in some sense be derived from B's. Below we examine the relationships between theses orderings and show that most of them are equivalent for well calibrated forecasters, but can easily contradict each other otherwise.

2 Partial orderings among probability forecasters

Let $r^A(\theta_i, a_j)$ and $r^B(\theta_i, a_j)$ be the joint probability functions of forecasters A and B, respectively, with a common nondegenerate marginal distribution $p(\theta)$. First, we confirm ourselves to forecasters which are both well calibrated. Following DeGroot and Fienberg (1983), we say that A is <u>more refined</u> than B, in symbols: $A \geq_R B$, if there exists a $k \times k$ Markov matrix M (i.e. a matrix with nonnegative entries whose columns sum to unity) such that

$$q^B(a_i) \quad = \quad \sum_{j=1}^{k} M_{ij} q^A(a_j), \qquad \text{and} \tag{1}$$

$$a_i q^B(a_i) = \sum_{j=1}^{k} M_{ij} a_j q^A(a_j), \; i = 1, \ldots, k. \quad (2)$$

Equation (1) 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) ensures that the resulting forecast is again well calibrated.

We say that A is strictly more refined than B (in symbols: $A >_R B$) if $A \geq_R B$ and $r^A(\theta_i, a_j) \neq r^B(\theta_i, a_j)$ for some i and j. The same convention will also be used for the other partial orderings below.

DeGroot and Fienberg (1983, Theorem 1) show that, for well calibrated forecasters A and B,

$$A \geq_R B \iff \sum_{i=1}^{j-1} (a_j - a_i)[q^A(a_i) - q^B(a_i)] \geq 0, \quad \text{j=1,...,k-1.} \quad (3)$$

The concept of refinement easily extends to forecasters which are not necessarily well calibrated. Again following DeGroot and Fienberg (1983), we say that A is sufficient for B – in symbols: $A \geq_s B$ – if, for some Markov matrix M,

$$q^B(a_i|\theta) = \sum_{j=1}^{k} M_{ij} q^A(a_{ij}|\theta), \; i = 1, \ldots, k; \theta = 0, 1. \quad (4)$$

Vardeman and Meeden (1983) suggest to alternatively order probability forecasters according to the concentration of defaults in the "bad" grades. This will here be called the VM-default order. Formally:

$$A \geq_{VM(d)} B \quad :\iff \quad \sum_{i=1}^{j} q^A(a_i|1) \leq \sum_{i=1}^{j} q^B(a_i|1), \quad \text{j=1,...,k.} \quad (5)$$

Or to put this differently: A dominates B in the Vardeman-Meeden default ordering if its conditional distribution, given default, first-order stochastically dominates that of B.

The same can be done for the non-defaults. A is better than B in the VM-non-default sense if non-defaults are more frequent in the "good" grades. Formally:

$$A \geq_{VM(nd)} B \iff \sum_{i=1}^{j} q^A(a_i|0) \geq \sum_{i=1}^{j} q^B(a_i|0), \quad \text{j=1,...,k.} \tag{6}$$

Finally, A dominates B in the Vardeman-Meeden sense (in symbols $A \geq_{VM} B$) if both $A \geq_{VM(d)} B$ and $A \geq_{VM(nd)} B$.

A related criterion which seems to be favoured in the banking community (see e.g. Falkenstein et al. 2000) is based on joining the points

$$(0,0), \left(\sum_{i=0}^{j-1} q(a_{k-i}), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, ..., k \tag{7}$$

by straight lines. The resulting plot is variously called the power curve, the Lorenz curve, the Gini curve, or the cumulative accuracy profile, and a forecaster A is considered better than a forecaster B in this - the Gini-default-sense (formally: $A \geq_{G(d)} B$) - if A's Gini curve is nowhere below that of B.

Similar to the VM-criterion, this can likewise be done for non-defaults, by joining the points

$$(0,0), \left(\sum_{i=1}^{j} q(a_i), \sum_{i=1}^{j} q(a_i|0) \right), \quad j = 1, ..., k. \tag{8}$$

A is then considered better than B in the Gini-non-default sense (in symbols: $A \geq_{G(nd)} B$), if A's non-default Gini-curve is nowhere below that of B. And we say that A dominates B in the Gini sense (in symbols $A \geq_G B$) if $A \geq_{G(d)} B$ and $A \geq_{G(nd)} B$.

A final criterion is based on the receiver-operating-characteristic curve (ROC-curve) defined by the points

$$(0,0), \left(\sum_{i=0}^{j-1} q(a_{k-i}|0), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, ..., k. \tag{9}$$

This is often used in clinical medicine (see e.g. Zweig and Campbell 1993) to discriminate between competing diagnostic tests. A probability forecaster A is then better than probability forecaster B in the ROC-sense (in symbols: $A \geq_{ROC} B$) if its ROC-curve lies nowhere below that of B.

The area under the ROC-curve is an obvious indicator of the usefulness of a probability forecast: the larger the area, the better the forecast. It also has a nice interpretation: If all defaults and all non-defaults are paired, it is equal to the probability that in one such randomly chosen pair, the non-default is ranked higher than the default (with the provision that if default and non-default are ranked the same, a coin is tossed to resolve the tie).

3 Relationships among the partial orderings

We first confine ourselves to forecasters which are well calibrated. It is well known and easily seen that then $A \geq_R B \Leftrightarrow A \geq_s B$ (DeGroot and Fienberg 1983, Theorem 2). Also, both $A \geq_{VM(d)} B$ and $A \geq_{VM(nd)} B$ imply $A \geq_R B$ (Vardeman and Meeden 1983, Theorem 2.1), which in turn implies both $A \geq_G B$ and $A \geq_{ROC} B$, as will be seen from Theorem 5 below.

THEOREM 1: Let A and B be well calibrated forecasters. Then we have

a) If $q^A(0) = q^B(0) = 0$, then A and B cannot be strictly ordered according to $\geq_{VM(d)}$.

b) If $q^A(1) = q^B(1) = 0$, then A and B cannot be strictly ordered according to $\geq_{VM(nd)}$.

This theorem has implications for the usefulness of Theorem 2.1 in Vardeman and Meeden (1983, *p.* 809). Theorem 2.1 in Vardeman and Meeden states that with well calibrated forecasters either $A \geq_{VM(d)} B$ or $A \geq_{VM(nd)} B$ implies that $A \geq_R B$. While this is true, it is evident from Theorem 3 above that it is also trivial. If we disregard the cases where either of the frequencies $q^A(0)$, $q^B(0)$, $q^A(1)$ or $q^B(1)$ is positive (which, in conjunction with calibration, implies perfect foresight and is thus not very relevant in practice), then the only pairs of well calibrated forecasters where $A \geq_{VM(d)} B$ or $A \geq_{VM(nd)} B$ are those where A and B have the same probability functions. Then they are of course (weakly) ordered according to any of the criteria above.

PROOF OF THEOREM 1: Assume without loss of generality that $A \geq_{VM(d)} B$, $q^A(a_2) < q^B(a_2)$, and therefore $a_2 q^A(a_2) < a_2 q^B(a_2)$. Then $A \geq_{VM(d)} B$ implies that

$$\begin{aligned} a_2 q^A(a_2) + a_3 q^A(a_3) &\leq a_2 q^B(a_2) + a_3 q^B(a_3), \qquad \text{so} \qquad (10) \\ a_3 q^A(a_3) &\leq a_2[q^B(a_2) - q^A(a_2)] + a_3 q^B(a_3) \\ &< a_3[q^B(a_2) - q^A(a_2)] + a_3 q^B(a_3) \qquad (11) \end{aligned}$$

which yields

$$q^A(a_2) + q^A(a_3) \quad < \quad q^B(a_2) + q^B(a_3) \tag{12}$$

Again from $A \geq_{VM(d)} B$, we have

$$a_2 q^A(a_2) + a_3 q^A(a_3) + a_4 q^A(a_4) \quad \leq \quad a_2 q^B(a_2) + a_3 q^B(a_3) + a_4 q^B(a_4) \tag{13}$$

which can be rewritten as

$$\begin{aligned} a_4 q^A(a_4) &\leq a_2[q^B(a_2) - q^A(a_2)] + a_3[q^B(a_3) - q^A(a_3)] + a_4 q^B(a_4) \\ &< a_3[q^B(a_2) + q^B(a_3) - q^A(a_2) - q^A(a_3)] + a_4 q^B(a_4) \\ &< a_4[q^B(a_2) + q^B(a_3) + q^b(a_4) - q^A(a_2) - q^A(a_3)], \qquad (14) \end{aligned}$$

where the last inequality follows from (13) and $a_4 > a_3$. The upshot is that

$$q^A(a_2) + q^A(a_3) + q^A(a_4) \quad < \quad q^B(a_2) + q^B(a_3) + q^B(a_4).$$

Continuing along these lines, it is easily seen that

$$\sum_{i=1}^{k} q^A(a_i) < \sum_{i=1}^{k} q^B(a_i), \tag{15}$$

which is in contradiction to $\sum_i q^A(a_i) = \sum_i q^B(a_i) = 1$. This means that $a_2 q^A(a_2) < a_2 q^B(a_2)$ and $A \geq_{VM(d)} B$ cannot go together and thus proves part (a) of the theorem. The proof of part (b) is analogous. •

The requirement that $q^A(0) = q^B(0) = 0$ comes into play to rule out the possibility that $q^A(0) > q^B(0)$, and still $A >_{VM(d)} B$. It can be shown by simple examples that well calibrated forecasters with these properties exist. Similarly, $q^A(1) = q^B(1) = 0$ rules out the possibility that $q^A(1) < q^B(1)$ and still $A \geq_{VM(nd)} B$. Again, it can be shown by simple examples that well calibrated forecasters with these properties exist. But apart form these quite extraordinary cases, there is no hope of establishing a VM-ordering when both forecasters are well calibrated.

Next we consider the above partial orderings for forecasters which are not necessarily well calibrated. It is trivial that $A \geq_{VM(d)} B$ does not imply $A \geq_{VM(nd)}$ and vice versa. However, for the Gini-ordering, the default ordering and the non-default ordering are identical.

THEOREM 2: $A \geq_{G(a)} B \Leftrightarrow A \geq_{G(nd)} B$.

PROOF: Let $A \geq_{G(d)} B$ and (x, y^A) be on the Gini-curve of A's defaults. Let (x, y^B) be the correspondent point on the Gini-curve of B. Then, for the non-default-ordering, (x, y^A) and (x, y^B) translate into

$$\begin{aligned} \left(1 - x, \frac{1 - x(1 - y^A)p}{1 - p}\right) &= (x^*, y^{*A}) \quad \text{and} \\ \left(1 - x, \frac{1 - x(1 - y^B)p}{1 - p}\right) &= (x^*, y^{*B}), \end{aligned}$$

respectively, where p is the overall percentage of defaults, and where $y^B \leq y^A \Leftrightarrow y^{*B} \leq y^{*A}$. •

The ROC-ordering likewise does not produce anything new, as is shown in our next result:

THEOREM 3: $A \geq_G B \Leftrightarrow A \geq_{ROC} B$.

PROOF: Let (x^A, y) be on the Gini-curve of A, and let (x^B, y) be a point on the Gini-curve of B with identical y coordinate. These points translate into

$$\left(\frac{x^A - yp}{1-p}, y\right) = (x^{*A}, y) \text{ and} \tag{16}$$
$$\left(\frac{x^B - yp}{1-p}, y\right) = (x^{*B}, y) \tag{17}$$

on the ROC-curves of A and B, respectively. However,

$$x^A < x^B \Leftrightarrow x^{*A} \leq x^{*B},$$

so the Gini-curves intersect if and only if the ROC-curves intersect. •

By far the most stringent ordering among those considered here is VM. Also $VM(d)$ and $VM(nd)$, taken by themselves, do not imply anything as concerns the Gini-ordering (this can be shown by simple counterexamples), the unrestricted VM-ordering implies the Gini-ordering (and, by its equivalence will the Gini-ordering, the ROC-ordering as well).

THEOREM 4: $A \geq_{VM} B \Rightarrow A \geq_G B$. The converse does not hold.

PROOF: Let

$$(x^A, y^A) = \left(\sum_1^j q^A(a_i),\ \sum_1^i q^A(a_i|0)\right) \quad \text{and}$$
$$(x^B, y^B) = \left(\sum_1^j q^B(a_i),\ \sum_1^i q^B(a_i|0)\right)$$

be on the nd-Gini-curves of A and B, respectively. From $A \geq_{VM} B$, we have

$$\sum_1^j q^A(a_i|0) \;\geq\; \sum_1^j q^B(a_i|0) \tag{18}$$

so $y^A \;\geq\; y^B$. •

The VM-ordering does not imply sufficiency, as can again be shown by simple counterexamples, except when both forecasters are well calibrated. In fact, it can be shown (Vardeman and Meeden 1983, Theorem 2) that semi-calibration suffices (A is called semi-calibrated if $p(1|a_i)$ is nondecreasing in a_i). Semi-calibration, in conjunction with sufficiency, also implies the Gini-ordering:

THEOREM 5: If A is semi-calibrated, we have

$$A \geq_S B \Rightarrow A \geq_G B.$$

The converse does not hold.

PROOF: The implication of the theorem is best seen if one considers the variant of A's Gini-curve (7) where cumulation starts with the "good" grades i.e. by plotting and joining the points

$$(0,0), \left(\sum_{i=1}^j q(a_i), \sum_{i=1}^j q(a_i(1))\right) \quad j = 1, \ldots, k. \tag{19}$$

Obviously, $A \geq_G B$ if its Gini-curve, as defined in (19), is nowhere above that of B. In addition, it is easily checked that, if A is semicalibrated, its Gini-curve is equal to the standard Lorenz curve of a discrete random variable X with values

$$x_i = q^A(a_i|1)/q^A(a_i) \quad (i = 1, \ldots, k) \tag{20}$$

where

$$P(X = x_i) = q^A(a_i). \tag{21}$$

By assumption, we have

$$q^B(a_i|0) = \sum_{j=1}^{k} m_{ij} q^A(a_i|0) \quad \text{and } q^B(a_i|1) = \sum_{j=1}^{k} m_{ij} q^A(a_i|1)$$

which implies

$$q^B(a_i) = \sum_{j=1}^{k} m_{ij} q^A(a_i). \tag{22}$$

Let $k^* \leq k$ be the number of nonzero $q^B(a_i)$'s, and let Z be a discrete random variable with values $1, 2, \ldots, k^*$ with $P(Z = r) = q^B(a_r)$ such that the conditional distribution of X, given $Z = r$, is given by

$$P(X = a_i|Z = r) = \frac{m_{ri} q^A(a_i)}{q^B(a_r)}. \tag{23}$$

Then it is easily checked that the Gini-curve of B is equal to the Lorenz curve of $Y := E(X|Z)$ and the theorem follows from standard results on Lorenz-inferiority of conditional expectations (see e.g. Arnold, 1987, Theorem 3.4).

Again, one can show by simple couterexamples that $A \geq_G B$ does not imply $A \geq_S B$. •

4 Conclusion

Due to the stringency of the VM-ordering, it will rarely happen in practice that probability forecasters can be so compared. Therefore this ordering is of mainly academic interest, and one will refer to either the sufficiency or Gini-orderings when evaluating the relative performance of probability forecasters.

Alternatively, of course, one could use some scoring rule such as the Brier score. An interesting issue not touched upon here (see however Scherwish 1989) is whether domination in any of the above senses is equivalent to superiority according to some family of scoring rules.

References

Arnold, Barry C. (1978): *"Majorisation and the Lorenz Order." Berlin (Springer).*

Crouhy, Michael; Galai, Dan and Mark, Robert (2001): "Prototype risk rating systems." *Journal of Banking and Finance* 25, 47 – 95.

Dawid, A.P. (1982): "The well calibrated Bayesian." *Journal of the American Statistical Association* 77, 605 – 610.

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

DeGroot, M.H. 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.

Falkenstein, Eric; Boral, Andrew and Kocagic, Ahmed E. (2000): "Moody's rating methodology: Rise $Calc^{TM}$ for private componies II: more results and the Australian Model." *Moody's Report* No. 62265.

Scherwish, Mark, J. (1989): "A general method for comparing probability assessors." *The Annals of Statistics* 17, 1856 – 1879.

Vardeman and S., Meeden, G. (1983): "Calibration, sufficiency and domination: Considerations for Bayesian probability assessor." *Journal of the American Statistical Association* 78, 808 – 816.

Zweig, Mark H. and Campbell, Gregory (1993): "Receiver-Operating Characterisic (ROC) plots: A fundamental evaluation tool in clinical medicine." *Clinical Chemistry* 39, 561 – 577.