# Comparing the accuracy of default predictions in the rating industry for different sets of obligors*

by

**Walter Krämer**
Fakultät Statistik, Technische Universität Dortmund, Germany
Phone: 0231/755-3125, Fax: 0231/755-5284
e-mail: walterk@statistik.tu-dortmund.de
and
**Simon Neumärker**
Fakultät Statistik, Technische Universität Dortmund, Germany
Phone: 0231/755-3869, Fax: 0231/755-5284
e-mail: simon.neumaerker@tu-dortmund.de

**Abstract**

We generalize the refinement ordering for well calibrated probability forecasts to the case were the debtors under consideration are not necessarily identical. This ordering is consistent with many well known skill scores used in practice. We also add an illustration using default predictions made by the leading rating agencies Moody's and S&P.

Keywords: Moody's, S&P, probability forecasts, skill scores
JEL numbers: C4, G2

*Research supported by DFG-Sonderforschungsbereich 823. We are grateful by Peter Posch for providing the data for our example and to an unknown referee for most useful criticism and comments.

## 1. Introduction

Probability forecasting has a long tradition in many fields of application. In economics, the most popular ones are default predictions in the rating industry. According to the Basel-II and Basel-III accords for instance, banks have to attach predicted default probabilities to all outstanding loans. Although major rating agencies like Moody's or S&P are reluctant to identify their letter grades with predicted default probabilities, we will stick to this probability interpretation in what follows. Given two competing default predictors and the prevalence of split ratings in practice (see e.g. Hauck and Neyer (2014)), it is then natural to ask: Which one is better?

One option is to rely on some scalar measures of performance like the Brier Score. However, it is well known that different score functions might produce conflicting results (see e.g. Krämer and Güttler (2008) for an example). The present paper therefore is concerned with partial orderings which, if valid, will imply identical rankings with respect to all members from some suitable class of scoring functions. It extends Krämer (2006), which covers only identical sets of debtors, to cases where the two debtors under considerations are not necessarily identical. It is not concerned with the equally important issue of how ratings are produced in the first place (see Lahiri and Yang (2013) for an overview or Czarnitzki and Kraft (2004) or Boumparis et al. (2015) for relevant discussions in the present journal).

Section 2 below introduces a novel partial ordering based on Generalized Lorenz curves and section 3 provides an application to ten-year default predictions made by the leading rating agencies Moody's and S&P.

## 2. Modified Lorenz Dominance

Let $0 = a_1 < a_2 < ... < a_k = 1$ be a finite set of possible forecasts of default probabilities. Let $q^A(a_j)$ be the relative frequency with which the default probability $a_j$ is predicted by forecaster $A$ (similarly for $B$). This paper will only consider forecasts which are well calibrated, i.e. where

$$\mathbb{P}(\text{default}|a_j) = a_j \quad (j = 1, \ldots, k). \tag{1}$$

In addition, we will focus on theoretical distributions, i.e. we will not distinguish between relative default frequencies and default probabilities. Everything that follows will then depend only on the vectors $a := [a_1, ..., a_k]'$ and $q := [q(a_1), .., q(a_k)]'$.

For the special case where $A$ and $B$ are rating the same set of debtors, DeGroot and Fienberg (1983) suggest the concept of refinement to discriminate between the two. If, by applying a randomization to the probability forecasts of $A$, one obtains a new probability forecast with the same distribution as $B$, then $A$ is more refined than $B$. As shown by DeGroot and Eriksson (1985), this amounts to Lorenz-domination of the respective forecast distributions:

$$(2) \qquad A \geq_L B \Leftrightarrow \underbrace{\frac{1}{p}\int_0^x F^{A^{-1}}(t)dt}_{=L^A(x)} \leq \underbrace{\frac{1}{p}\int_0^x F^{B^{-1}}(t)dt}_{=L^B(x)}, \quad (0 \leq x \leq 1)$$

where $L^A(x)$ and $L^B(x)$ are the respective Lorenz curves,

$$(3) \qquad F^A(a) := \sum_{a_i \leq a} q^A(a_i)$$

is $A$'s default forecast distribution and where

$$(4) \qquad F^{A^{-1}}(t) := \inf\{a : F^A(a) \geq t\}$$

is the inverse of $A$'s default forecast distribution (similarly for $B$). The overall default probability can then be expressed as

$$(5) \qquad p = \int_0^1 F^{A^{-1}}(t)dt = \int_0^1 F^{B^{-1}}(t)dt$$

which equals the expectation of both $F^A$ and $F^B$. In view of calibration, $p = \sum a_i q^A(a_i) = \sum a_i q^B(a_i)$. This expectation could as well be dropped in equation (2), as it appears on both sides of the inequality, and mainly sees to it that both Lorenz curves end in $(1, 1)$.

Contrary to comparing income inequality, where Lorenz curves close to the diagonal are "good" (i.e. signal a more equal distribution of income), $A$ is in the present application considered better than $B$ if its Lorenz curve bends farther away from the diagonal, i.e. if its predicted default probabilities are more spread out. This is why we here, other than in the income distribution context, identify "domination" with a higher level of inequality. It can also easily be shown that the same ordering obtains if the ranking is based on predicted non-defaults:

$$(6) \qquad \int_0^x F^{A^{-1}}(t)dt \leq \int_0^x F^{B^{-1}}(t)dt \quad \Leftrightarrow \quad \int_0^x \tilde{F}^{A^{-1}}(t)dt \leq \int_0^x \tilde{F}^{B^{-1}}(t)dt$$

for $0 \leq x \leq 1$, where $\tilde{F}(a) := \sum_{\tilde{a}_i \leq a} \tilde{q}(\tilde{a}_i)$ is the distribution function of the predicted survival probabilities $\tilde{a}_i := 1 - a_i$ and $\tilde{q}(\tilde{a}_i) := q(a_i)$.

If $A$ and $B$ are rating different (possibly overlapping) sets of debtors, the overall probability of default will in general differ between the respective sets, and the refinement concept does no longer apply. However, the Lorenz-ordering is still possible, by replacing the overall default probability $p = p_A = p_B$ in (2) with $p_A$ and $p_B$, where appropriate. Other than in the case $p_A = p_B$, it now does matter whether we consider predicted default or predicted survival probabilities: It can be shown by simple counterexamples that $A$'s Lorenz curve for predicted default probabilities is better and $A$'s Lorenz curve for predicted survival probabilities is worse than that of $B$. Therefore the standard Lorenz order does not make much sense for nonidentical sets of debtors. Here is an extension:

DEFINITION: $A$ dominates $B$ in the modified Lorenz sense ($A \geq_{ML} B$) if $A \geq_L B$ (i.e. (2) obtains with $p_A$ and $p_B$ in place of $p$) and in addition,

$$0.5 \geq p_A \geq p_B \ (p_B < 0.5) \text{ or } 0.5 \leq p_A \leq p_B \ (p_B > 0.5).$$

For $p_A = p_B$, this reduces to the standard refinement ordering. Without loss of generality, we will confine ourselves to the empirically more relevant case $p_B < 0.5$ in what follows. The inequality $p_A > p_B$ then implies that the generalized Lorenz curve (defined as $p$ times standard Lorenz curve) of $A$ is larger than that of $B$ towards the right end of the $[0, 1]$-interval. Intuitively, this means that $A'$s predictions are both more spread out and on average closer to 0.5 at the same time.

It is well known from the theory of proper scoring rules (see e.g. Winkler (1996)) that it becomes harder to obtain good results as the overall default probability approaches 0.5. The well known Brier score for instance, given by

$$B(a, q) := \sum_{i=1}^{k} q(a_i) a_i (1 - a_i) \tag{7}$$

whenever a forecaster is well calibrated, approaches its optimal value of 0 even for the trivial forecast $a_i = p \ \forall i$ whenever $p \to 0$ or $p \to 1$. And the trivial forecast is worst in the Brier sense if $p = 0.5$ (always assuming that $p$ is among the available $a_i$'s). Two additional scoring rules often used in application are the logarithmic score

$$L(a, q) := \sum_{i=1}^{k} q(a_i) \left(a_i \ln(a_i) + (1 - a_i) \ln(1 - a_i)\right) \quad (\text{with } 0 \ln(0) := 0) \tag{8}$$

and the spherical score

$$S(a,q) := \sum_{i=1}^{k} q(a_i)\sqrt{a_i^2 + (1-a_i)^2}, \tag{9}$$

which are likewise producing good results for the trivial forecasts as $p \to 0$ or $p \to 1$.

In order to compensate for this intrinsic difference in difficulty, it is common to rely on skill scores rather than on ordinary scoring rules whenever $p_A \neq p_B$ (see Lahiri and Yang (2013) for additional motivation). Given any scoring rule $S(a,q)$, the corresponding skill score is given by

$$SS(a,q) := \frac{S(a,q) - S_t}{S_{opt} - S_t} \tag{10}$$

where $S_t$ is the trivial score obtained for $a_i = p \ \forall i$ and $S_{opt}$ is the optimal score where only $q(0)$ and/or $q(1)$ are different from zero (Winkler (1996)). A skill score then measures how close a forecaster is to the optimum. It takes its maximum value of 1 if defaults and non-defaults are both predicted with certainty; it takes the value zero for the trivial forecast, and it can even take on values less than zero if a forecaster is worse than the trivial forecast. For the Brier score, for instance, we have

$$BS(a,q) = \frac{B(a,q) - p(1-p)}{-p(1-p)}. \tag{11}$$

THEOREM: For two well calibrated probability forecasters $A$ and $B$, let $A \geq_{ML} B$. Then, for skill scores derived from the Brier score, the logarithmic score and the spherical score, $A$ is at least as good as $B$.

PROOF: The proof builds on Krämer (2006), who establishes the above result for the case $p_A = p_B$. Now, let without loss of generality, $0.5 > p_A > p_B > 0$, let $a_i^* := \frac{p_B}{p_A} a_i < a_i$ and consider a well calibrated forecaster $A^*$ with possible predictions $a_i^*$. Then $A^*$ has the same Lorenz curve as $A$, while, by construction, $p_{A^*} = p_B$. Therefore, $A^*$ cannot be worse than $B$ according to any strictly proper scoring rule.

Next we show that, for the Brier skill score, $A$ cannot be worse than $A^*$. Rewriting the Brier skill score as

$$BS(a,q) = 1 - \frac{B(a,q)}{p(1-p)}, \tag{12}$$

this amounts to

$$\frac{\sum q(a_i)ca_i(1-ca_i)}{cp(1-cp)} \geq \frac{\sum q(a_i)a_i(1-(a_i))}{p(1-p)} \tag{13}$$

where $c = p_B/p_A$. After several trivial reshufflings, this inequality is seen to be equivalent to

$$\sum q(a_i)a_i p \leq \sum q(a_i)a_i^2, \tag{14}$$

which in turn follows from $p = \sum q(a_i)a_i$, the general inequality $E(X^2) > [E(X)]^2$ and the fact that the $a_i$'s can be viewed as the values of a random variable with probability function $q(a_i)$.

In a similar fashion, it is seen that for the logarithmic skill score $LS(a^\star, q) \leq LS(a,q)$. For the purpose, rewrite $LS$ as a ratio of a convex and a concave function $P_N(c)$ and $P_D(c)$ of $c = p_B/p_A$ (ceteris paribus) and show that

$$LS(a^\star, q) = \frac{P_N(c)}{P_D(c)} \leq \frac{cP_N(1)}{cP_D(1)} = LS(a,q). \tag{15}$$

Given $a$ and $q$, one can likewise view the spherical score of $A^\star$ as a function of $c$ via

$$SS(a^\star, q) = \frac{\sum q(a_i)\sqrt{(ca_i)^2 + (1-ca_i)^2} - \sqrt{1-2cp(1-cp)}}{1-\sqrt{1-2cp(1-cp)}}, \tag{16}$$

where it can be shown by brute force calculation that $\frac{\partial SS}{\partial c} \geq 0$ for all $c \in (0,1)$, so

$$SS(a^\star, q) \leq SS(a,q). \quad \square$$

As an illustration, consider three well calibrated forecasters $A$, $A^*$ and $B$ with predicted default probabilities and distributions across predicted default probabilities as in table 1.

TABLE 1. Three well calibrated probability forecasters

| $a_i$ | $q^A(a_i)$ | $q^B(a_i)$ | $q^{A^*}(a_i)$ |
|---|---|---|---|
| 0 | 0.3 | 0.2 | 0.3 |
| $\frac{10}{11} \cdot 0.1$ | 0 | 0 | 0.5 |
| 0.1 | 0.5 | 0.6 | 0 |
| 0.2 | 0 | 0.2 | 0 |
| $\frac{10}{11} \cdot 0.3$ | 0 | 0 | 0.2 |
| 0.3 | 0.2 | 0 | 0 |

Then we have $p_B = p_{A^*} = 0.1 < p_A = 0.11$, with Lorenz curves of $A$ (equal to that of $A^*$) and $B$ as in figure 1.

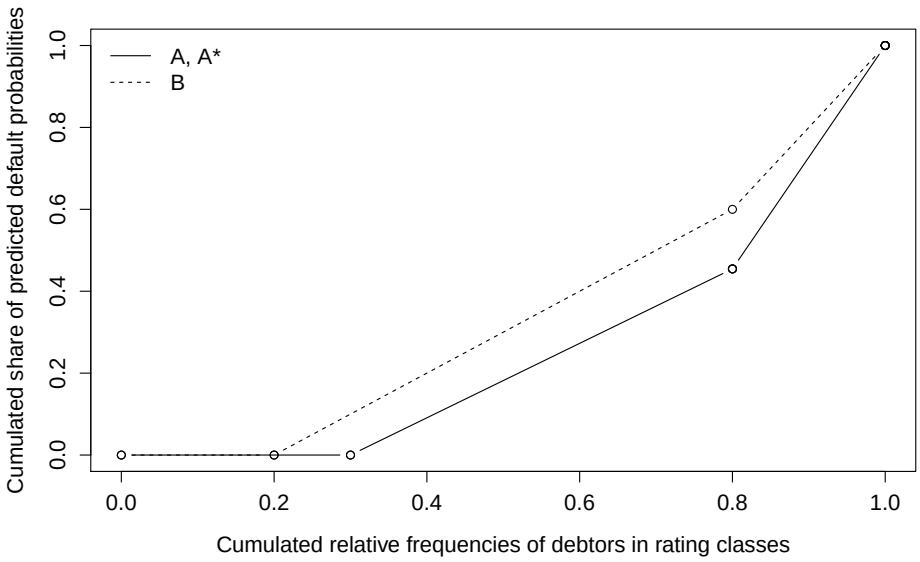


FIGURE 1. Lorenz curves of predicted default probabilities

It is seen that the Lorenz curve of $A$ is nowhere above that of $B$, so $A \geq_{ML} B$ in view of $p_A > p_B$. Table 2 reports the respective Brier scores, plus the logarithmic scores $L(a, q)$ and the spherical scores $S(a, q)$ where, contrary to the Brier score, large values of $L(a, q)$ and $S(a, q)$ are "good". It shows $A$ is inferior to $B$ in terms of the conventional Brier and spherical score, while it is uniformly superior in terms of all the skill scores considered here.

TABLE 2. Selected scores for predictions from table 1

| Rule | $A$ | $A^*$ | $B$ |
|---|---|---|---|
| Brier | 0.087 | 0.081 | 0.086 |
| Logarithmic | -0.285 | -0.270 | -0.295 |
| Spherical | 0.905 | 0.912 | 0.908 |
| Brier skill | 0.111 | 0.100 | 0.044 |
| Logarithmic skill | 0.178 | 0.171 | 0.092 |
| Spherical skill | 0.081 | 0.070 | 0.029 |

## 3. APPLICATION

As an illustration, table 3 shows ten-year default rates obtained from the web pages of Moody's and S&P (Moody's (2015) and Standard & Poor's (2015)).

TABLE 3. Empirical ten year default rates and distribution of debtors among rating classes

| | **Moody's** | | **S&P** | |
|---|---|---|---|---|
| Rating Class | $a_i^M$ | $q^M(a_i)$ | $a_i^S$ | $q^S(a_i)$ |
| AAA/Aaa | 0.0049 | 0.0341 | 0.0071 | 0.0107 |
| AA/Aa | 0.0089 | 0.1150 | 0.0078 | 0.0713 |
| A | 0.0209 | 0.2426 | 0.0171 | 0.2294 |
| BBB/Baa | 0.0495 | 0.2318 | 0.0498 | 0.2615 |
| BB/Ba | 0.1979 | 0.1423 | 0.1638 | 0.1737 |
| B | 0.4025 | 0.1786 | 0.2997 | 0.2277 |
| CCC/Caa-C | 0.6597 | 0.0554 | 0.5135 | 0.0256 |

As we equalize realized relative default frequencies and predicted default probabilities, both agencies are well calibrated by construction. Figure 2 presents the resulting Lorenz curves; it shows that Moody's predicted default probabilities are more spread out.

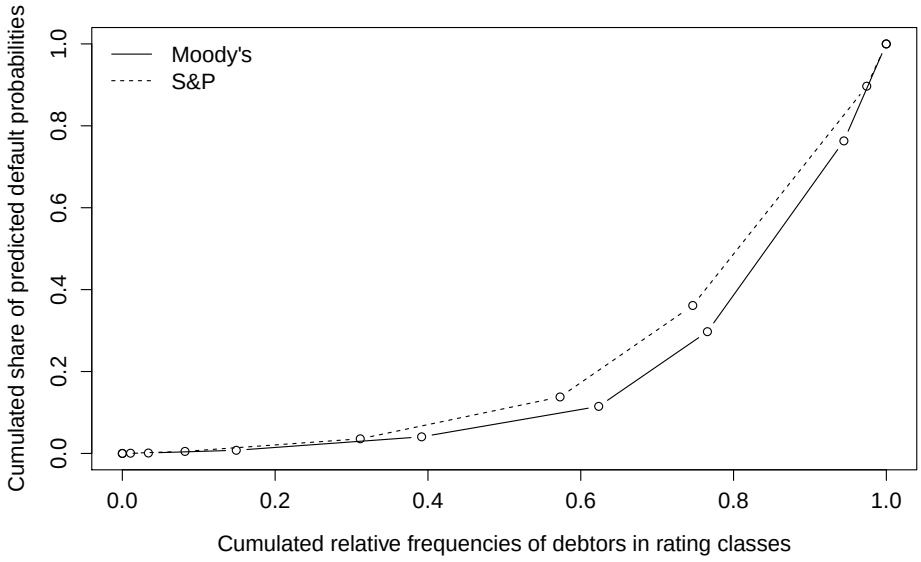


FIGURE 2. Lorenz curves of predicted default probabilities for Moody's and S&P

Since in addition

$$p_M = \sum_{i=1}^{7} a_i^M q^M(a_i) = 15.43\% \quad > \quad p_S = \sum_{i=1}^{7} a_i^S q^S(a_i) = 12.74\%, \tag{17}$$

Moody's dominate S&P in the modified Lorenz sense and are therefore also superior in terms of the skill scores discussed here (table 4). According to the unmodified spherical score and Brier score, however, S&P is better.

TABLE 4. Score values for Moody's and S&P predictions

| Rule | **Moody's** | **S&P** |
|---|---|---|
| Brier | 0.0950 | 0.0948 |
| Logarithmic | -0.3039 | -0.3095 |
| Spherical | 0.8935 | 0.8953 |
| Brier skill | 0.2719 | 0.1470 |
| Logarithmic skill | 0.2935 | 0.1885 |
| Spherical skill | 0.2411 | 0.1136 |

## References


Boumparis, P., Milas, C. and Panagiotidis, T. (2015). Has the crisis affected the behavior of the rating agencies? Panel evidence from the Eurozone. *Economics Letters* **136**, 118-124.

Czarnitzki, D. and Kraft, K. (2004). Innovation indicators and corporate credit ratings: evidence from German firms. *Economics Letters* **82**, 377-384.

DeGroot, M. and Eriksson, E.A. (1985). Probability forecasting, stochastic dominance, and the Lorenz curve. *Statistical decision theory and related topics III*, Vol **1**, S. S. Gupta und J. O. Berger (ed.), New York (Academic Press), 291-314.

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

Hauck, A. and Neyer, U. (2014). Disagreement between rating agencies and bond opacity: A theoretical perspective . *Economics Letters* **123**, 82-85.

Krämer, W. (2006). Evaluating probability forecasts in terms of refinement and strictly proper scoring rules. *Journal of Forecasting* **25**, 223-226.

Krämer, W. and Güttler, A. (2008). On comparing the accuracy of default predictions in the rating industry. *Empirical Economics* **34**, 343-356.

Lahiri, K. and Yang, L. (2013). Forecasting Binary Outcomes, *Handbook of Economic Forecasting*, Vol **2**, Part B, Elliott, G. and Timmermann, A. (ed.), Elsevier, 1025-1106.

Moody's (2015). Annual Default Study: Corporate Default and Recovery, 1920-2014. *Moody's Investor Service.*

Standard & Poor's (2015). Annual Global Corporate Default Study and Rating Transitions. *Standard & Poor's Ratings Services.*

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