

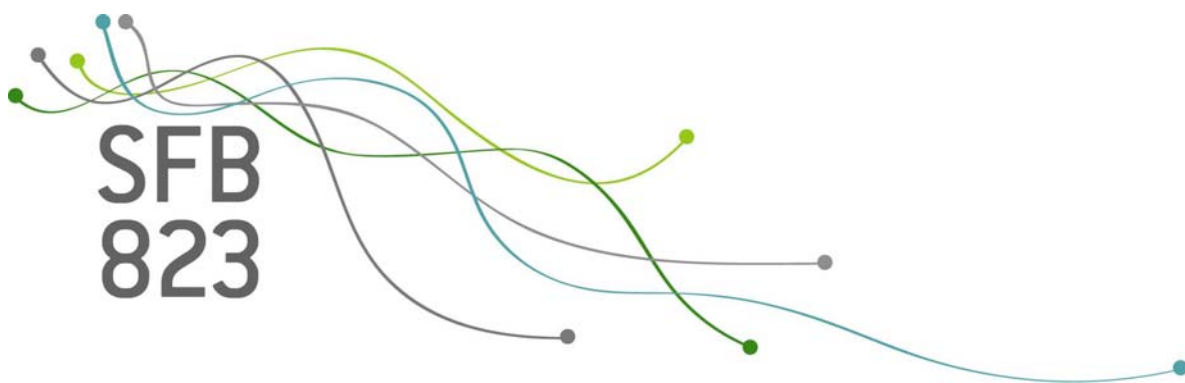
SFB
823

Partial orderings of default predictions

Walter Krämer, Peter N. Posch

Nr. 31/2015

Discussion Paper



Partial orderings of default predictions

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

Peter N. Posch

Fakultät WiSo, Technische Universität Dortmund, Germany

Phone: 0231/755-32735, Fax: 0231/755-8197

e-mail: peter.posch@tu-dortmund.de

*Corresponding author. Work supported by DFG-Sonderforschungsbereich 823. We are grateful to Simon Neumärker for expert research assistance and to Bernd Fitzenberger and other seminar participants at Berlin Humboldt University for helpful criticism and comments.

Abstract

We compare and generalize various partial orderings of probability forecasters according to the quality of their predictions. It appears that the calibration requirement is quite at odds with the possibility of some such ordering. However, if the requirements of calibration and identical sets of debtors are relaxed, comparability obtains more easily. Taking default predictions in the credit rating industry as an example, we show for a data base of 5333 (Moody's) and 6505 ten-year default predictions (S&P), that Moody's and S&P cannot be ordered neither according to their grade distributions given default or non-default or to their Gini-curves, but Moody's dominate S&P with respect to the ROC-criterion.

Keywords: probability forecasts, rating systems, partial ordering

1 Introduction

When talking about prediction, statisticians and econometricians usually consider point forecasts: the rate of inflation next year, the population of the earth in 2050, the coming budget deficit and so on. Of course, such point forecasts (usually: of the first moment of a distribution) are rather crude summaries of possible future events. Historically, this underlying insecurity has mostly been dealt with by attaching a probability distribution to these point forecasts, whereas forecasting the complete distribution of future events is of rather recent origin (See Gneiting (2008) or Gneiting and Katzfuss (2014) for a survey, in particular with respect to applications in meteorology, or Groen et al. (2013) for an application in economics).

This paper considers the special case of forecasting a Bernoulli variable with values 0 and 1. Other than in the case of continuous variables, forecasting the distribution has engendered an enormous literature here. Obvious applications are default predictions in consumer credit scoring and in the credit rating industry, or weather forecasts, where statements like: "The probability of rain in Chicago tomorrow is 20%" have been common for quite a while. Other applications are forecasting the probability of a recession (see Kauppi and Saikkonen (2008)). For concreteness, however, most of the discussion below will be phrased in terms of defaults and non-defaults. While the production of such forecasts has been heavily discussed both in the statistics and economics literature (see e.g. Creal et al. (2014) for a new method based on credit default swaps), much less is known about evaluating their relative performance. Section 2 summarizes previous results and suggests various extensions. It appears that the concept of calibration (DeGroot and Fienberg (1983)) is a rather tough requirement which prevents most probability forecasters from being unequivocally comparable. Section 3 extends previous comparability re-

sults to nonidentical sets of debtors and section 4 applies our results to default prediction made by the leading rating agencies Moody's and S&P.

2 Partial orderings of probability forecasts

There are two ways of comparing probability forecasters A and B . First, by computing scoring rules such as the well-known Brier-Score

$$S = \frac{1}{n} \sum_{i=1}^n (\theta_i - p_i)^2, \quad (1)$$

where n is the number of forecasts made, $\theta_i \in \{0, 1\}$ denotes whether the event in question has occurred ($\theta = 1$) or not ($\theta = 0$), and where p_i is the forecasted probability of the event in trial no. i (Winkler 1996). Second, by comparing complete distributions of forecasts, as will be done below. This allows statements such as " A is better than B irrespective of any scoring rule from some class of scoring rules" (see Krämer (2006)), but has the disadvantage that certain pairs of forecasters cannot be ranked (e.g. this method provides only a partial ordering). It can be shown that it is then possible to find sensible (in the sense of "proper", see Krämer (2006)) scoring rules such that A is better than B according to rule no. 1 and B is better than A according to rule no. 2. Krämer and Güttler (2008) provide an example where, for an identical set of private enterprises, Moody's provides better default forecasts than S&P for some scoring rules, while S&P is better for others.

Let $0 = a_1 < a_2 < \dots < a_k = 1$ be the probabilities that are available as forecasts for the future event in question. We take k to be small and finite here (for generalizations see Schervish (1989)). For instance, the US National Weather Service has only multiples of 10% as predicted probabilities of rain, so $k = 11$ (including 0% and 100%). In the credit rating industry, the major agencies have 7 different rating categories with up to three rating modifiers

each resulting in a maximum of 21 different grades including a default category. Typically, the worst four categories are combined into one, so there are usually 17 categories.

We are not concerned here with the methods by which default forecasts are produced. Rather, we take forecasts as given and take forecasters to be defined by the discrete bivariate probability function $r(\theta_i, a_j), i = 1, 2, j = 1, \dots, k$, resulting from some such method, whichever it may be, with $\theta = 1$ indicating default and $\theta = 0$ indicating non-default.

Of course, the true bivariate probability function $r(\theta_i, a_j)$ is known only after infinitely many trials. In practice we take its empirical counterpart as a suitable approximation. As this paper is focused on fundamental problems of comparability, this sampling issue is ignored below (see however Krämer and Güttler (2008) for some discussion of the statistical significance of observed differences in ranking quality).

Without loss of generality, the set $\mathcal{A} = \{a_1, \dots, a_k\}$ of available probabilities can be taken as identical for all forecasters involved. If not, define $\mathcal{A} := \mathcal{A}^A \cup \mathcal{A}^B$. Following Vardeman and Meeden (1983), the following additional notation will be used:

$$p(1) := \sum_j r(1, a_j) = \text{overall probability of default (PD)}$$

$$p(0) := \sum_j r(0, a_j) = \text{overall probability of no default (probability of survival PS)}.$$

$$q(a_j) := \text{probability with which default forecast } a_j \text{ is made.}$$

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

$$p(0|a_j) := \frac{r(0, a_j)}{q(a_j)} = \text{conditional probability of survival given forecast } a_j.$$

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

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

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 is equal to a_j . Formally:

$$a_j \stackrel{!}{=} p(1|a_j) = \frac{r(1, a_j)}{q(a_j)} \quad (2)$$

whenever $q(a_j) > 0$. Such forecasters are called "well calibrated" (DeGroot and Fienberg (1983)).

This calibration requirement has various consequences for the bivariate probability function $r(\theta_i, a_j)$. For instance, it is obvious from (2) that, given predicted a_j of the future event and the $r(1, a_j)$, the marginal frequencies $q(a_j)$ and therefore also the $q(a_j|0)$ and the $r(0, a_j)$ are fixed. These limited degrees of freedom for obtaining a bivariate probability function $r(\theta_i, a_j)$ that is compatible with calibration are further reduced by the requirement that

$$\sum q(a_j) = q(0) + \sum_{j=2}^k \frac{r(1, a_j)}{a_j} = r(0, 0) + \underbrace{r(1, 0)}_{=0} + \sum_{j=2}^k \frac{r(1, a_j)}{a_j} = 1. \quad (3)$$

These relationships are probably best clarified via a numerical example. Assume that among the a_j 's, there is 0.2 and 0.4 and that $r(1, 0.2) = 0.1$ and $r(1, 0.4) = 0.2$, as in the following table:

Table 1: A well calibrated probability forecaster

	a_j				
	...	0.2	...	0.4	...
$r(1, a_j)$	0	0.1	0	0.2	0
$q(a_j)$	0	0.5	0	0.5	0
$r(0, a_j)$	0	0.4	0	0.3	0
$q(a_j 1)$	0	1/3	0	2/3	0
$q(a_j 0)$	0	4/7	0	3/7	0

Then the first 4 entries in the table completely determine the rest. In particular, there can only be zeros in the columns where $a_j \notin \{0.2, 0.4\}$. Also, the marginal probabilities $p(1) = 0.3$ and $p(0) = 0.7$ follow immediately from

$$r(1, 0.2) = 0.1 \text{ and } r(1, 0.4) = 0.2.$$

These restrictions will be vital in establishing various relationships between partial orderings below.

The first such partial ordering relies on "refinement" (DeGroot and Fienberg (1983)). We say that A is *more refined* 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

$$q^B(a_i) = \sum_{j=1}^k M_{ij} q^A(a_j), \text{ and} \tag{4}$$

$$a_i q^B(a_i) = \sum_{j=1}^k M_{ij} a_j q^A(a_j), \quad i = 1, \dots, k. \tag{5}$$

Equation (4) means that, given A's forecast a_j , an additional independent randomization is applied according to the conditional distribution $M_{ij}(j =$

$1, \dots, k$) which produces forecasts with the same probability function as that of B . Condition 5 ensures that the resulting forecast is again well calibrated.

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 ,

$$q^B(a_i|\theta) = \sum_{j=1}^k M_{ij} q^A(a_j|\theta), \quad i = 1, \dots, k; \quad \theta = 0, 1. \quad (6)$$

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:

$$A \geq_{VM(d)} B :\Leftrightarrow \sum_{i=1}^j q^A(a_i|1) \leq \sum_{i=1}^j q^B(a_i|1), \quad j = 1, \dots, k. \quad (7)$$

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 survivors are more frequent in the "good" grades. Formally:

$$A \geq_{VM(nd)} B :\Leftrightarrow \sum_{i=1}^j q^A(a_i|0) \geq \sum_{i=1}^j q^B(a_i|0), \quad j = 1, \dots, k. \quad (8)$$

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

Table 2 provides an example. It extends table 1 to the case where an additional forecaster is involved. It is easily checked that both forecasters are well calibrated and that B dominates A in the Vardeman-Meeden-non-default-ordering. However, A and B cannot be ranked according to $VM(d)$ (it will emerge below that this is no coincidence).

Table 2: Two well calibrated forecasters

	a_j			
	0.2	0.25	0.4	1
	Forecaster A			
$r^A(1, a_j)$	0.1	0	0.2	0
$q^A(a_j)$	0.5	0	0.5	0
$r^A(0, a_j)$	0.4	0	0.3	0
$q^A(a_j 1)$	1/3	0	2/3	0
$q^A(a_j 0)$	4/7	0	3/7	0
	Forecaster B			
$r^B(1, a_j)$	0.1	0.05	0.1	0.05
$q^B(a_j)$	0.5	0.2	0.25	0.05
$r^B(0, a_j)$	0.4	0.15	0.15	0
$q^B(a_j 1)$	1/3	1/6	1/3	1/6
$q^B(a_j 0)$	8/14	3/14	3/14	0

It is also easily seen that B is more refined than A : Putting all of B 's forecasts from the 0.25 and 1 brackets into the 0.4 bracket yields forecasts which are identical in distribution to A .

A related criterion which seems to be favoured in the banking community is based on joining the points

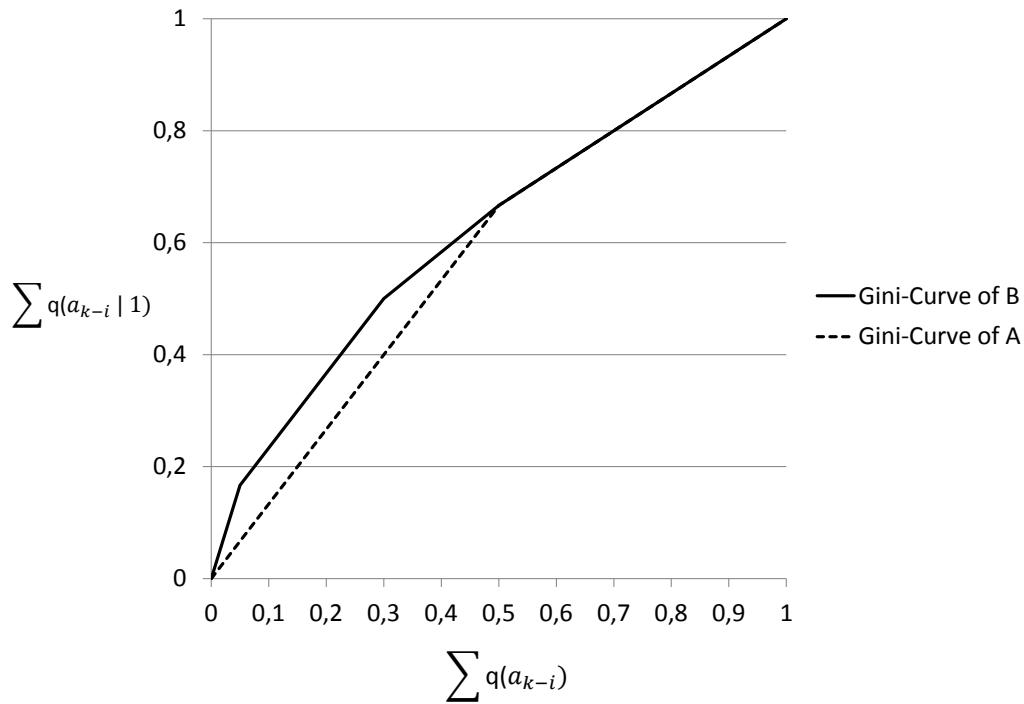
$$(0, 0), \left(\sum_{i=0}^{j-1} q(a_{k-i}), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, \dots, k \quad (9)$$

by straight lines. The resulting plot is variously called the power curve, the Lorenz curve, the Gini curve, or the cumulative accuracy profile (see e.g. Engelmann and Hayden and Tasche (2003)), and a forecaster A is considered better than a forecaster B in this - the Gini-default-sense (formally: $A \geq_G B$) - if A 's Gini curve is nowhere below that of B . It is easily checked that then

A 's Gini-curves for survivors is nowhere above that of B , so arguing in terms of survivors does not add anything new.

As figure 1 shows, B is in the above example also Gini-dominating A : Although both curves coincide from 0.5 upwards, A 's curve is below that of B to the left of 0.5.

Figure 1: Gini-ordering of forecasters A and B from table 2



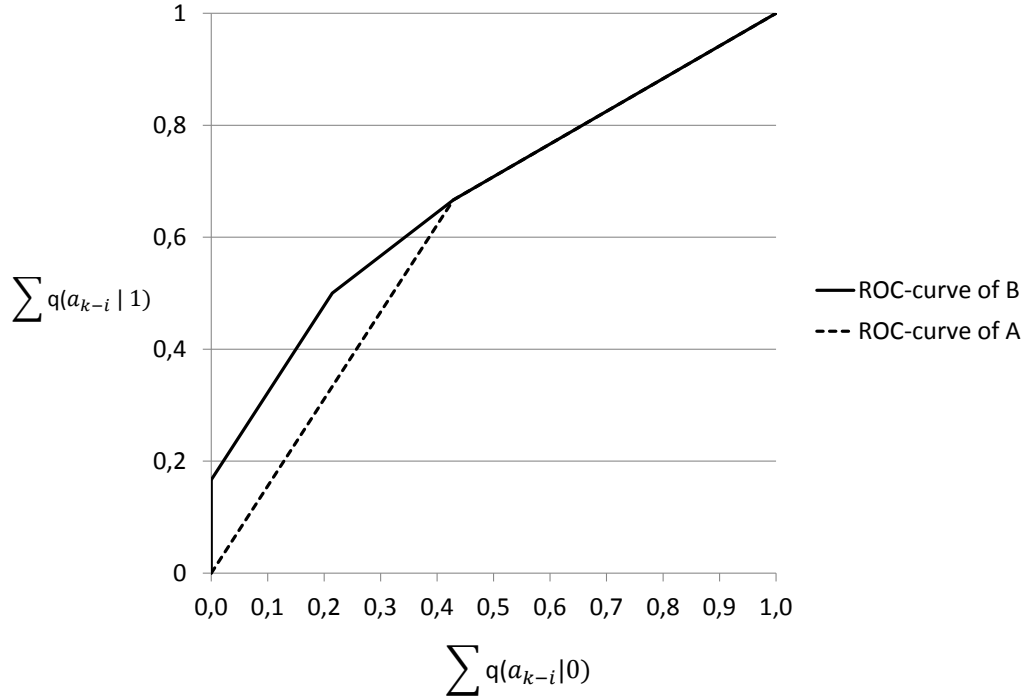
Finally, rather than looking at the Gini-curve, it is common in medical applications to consider the ROC-curve instead, defined by the points

$$(0, 0), \left(\sum_{i=0}^{j-1} q(a_{k-i}|0), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, \dots, k \quad . \quad (10)$$

With identical right marginals, this does not imply anything new. Figure 2 shows the ROC-curves corresponding to the Gini-curves from figure 1. It is seen that both Gini-curves shift leftwards, and that B keeps dominating A .

Our first result shows that this is no coincidence:

Figure 2: ROC-curves of forecasters A and B from table 2



THEOREM 1:

For probability forecasters A and B with $p^A(1) = p^B(1)$, the Gini-curves intersect if and only if the ROC-curves intersect.

PROOF:

For ease of notation, let p be the overall default probability and let $A_j = \sum_{i=0}^{j-1} q^A(a_{k-i})$, $A1_j = \sum_{i=0}^{j-1} q^A(a_{k-i} | 1)$, $A0_j = \sum_{i=0}^{j-1} q^A(a_{k-i} | 0)$.

Similarly for B . Then the Gini-curve and the ROC-curve are defined by the points $(A_i, A1_i)$ (Gini) and $(A0_i, A1_i)$ (ROC) respectively. Similarly for B .

And A dominates B in the Gini- and ROC-sense, respectively, if and only if

$$\frac{A1_i}{A_i} \geq \frac{B1_i}{B_i} \quad (\text{Gini}) \quad \text{or} \quad (11)$$

$$\frac{A1_i}{A0_i} \geq \frac{B1_i}{B0_i} \quad (\text{ROC}) \quad (12)$$

However, from

$$A_i = pA1_i + (1 - p)A0_i$$

we have

$$\begin{aligned} \frac{A1_i}{A_i} &= \frac{A1_i}{p(A1_i + (1 - p)A0_i)} \\ &= \frac{1}{p + (1 - p)\frac{A0_i}{A1_i}}, \end{aligned}$$

which is monotonously increasing in $\frac{A1_i}{A0_i}$. The same holds for $\frac{B1_i}{B0_i}$. Therefore, the relationships (11) and (12) are equivalent.

Both the Gini-curve and the ROC-curves do not require predicted default probabilities - sorting the debtors into classes with increasing default probability suffices. They are also both convex if forecasters are semi-calibrated, i.e. if $p(1 | a_i)$ is a non decreasing function of a_i . This is a minimum requirement we will stick to in what follows.

As both the Gini and the ROC-curve are invariant under monotone transformations of the predicted default probabilities a_j , the ordering implied by them is no longer a partial ordering: From $A \geq B$ and $B \geq A$ one can no longer infer that $A = B$ ("antisymmetry"). But transitivity persists. Orderings of this type are called pre-orderings, and it will be seen in the next section that the VM-ordering likewise violates the antisymmetry-condition if the restriction of identical right marginals is relaxed.

3 Generalizations and relationships among the orderings

Theorem 1 does not extend to $p^A(1) \neq p^B(1)$, as can be shown by simple counterexamples. Also, the calibration requirement severely restricts both the entries in the $r(\theta_i, a_j)$ -matrix and the chances that two probability forecasters can be compared in the first place. In particular, Krämer (2005) shows that for well calibrated forecasters A and B , if $q^A(0) = q^B(0) = 0$, then A and B cannot be strictly ordered according to $\geq_{VM(d)}$. And if $q^A(1) = q^B(1) = 0$, then A and B cannot be strictly ordered according to $\geq_{VM(nd)}$. The example in table 2 where $\geq_{VM(nd)}$ obtains is therefore an artefact of $q^B(1) > 0$.

Comparability is much easier if the calibration requirement is abandoned. Even for identical right marginals, Krämer (2005) shows that the unrestricted VM-ordering might then obtain, in which case it implies the Gini-ordering. Given semi-calibration (i.e. $p(1|a_j)$ is non decreasing in a), the Gini-ordering is also implied by sufficiency. The Gini-ordering is thus the least demanding of the bunch in the case of identical right marginals. Still, in practice, most forecasters do not seem to be comparable at all.

Therefore, we now also abandon the restriction $p^A(1) = p^B(1)$. In practice, this means that we can now consider non-calibrated forecasters for different populations with different overall default probabilities. In particular, it can now be shown via simple examples that the VM-ordering might obtain even in case of calibration.

For the Gini-ordering, it can now happen that B 's Gini-curve is better than A 's for defaults and worse for non-defaults, and we say that B dominates A in the Gini-sense if it does so both for defaults and non-defaults. In that case, it is easily seen that \geq_G keeps on defining a preordering.

As to the relationship between \geq_{VM} and \geq_{Gini} , neither implies the other under these more general circumstances, as can be shown by simple counterexamples. Also, unlike in the case of identical right marginals, the Gini and ROC orderings are no longer identical when right marginals are different. It is easy to find examples where $A \geq_{ROC} B$ but A and B cannot be ordered according to \geq_G . Rather, we now have the following result:

THEOREM 2:

For arbitrary bivariate probability functions $r^A(\theta_i, a_j), r^B(\theta_i, a_j)$ and semi-calibrated forecasters A and B , we have

$$A \geq_G B \Rightarrow A \geq_{ROC} B.$$

The converse does not hold.

PROOF:

In the notation from Theorem 1, $A \geq_G B$ is equivalent to

$$\frac{A1_i}{A_i} \geq \frac{B1_i}{B_i} \tag{19}$$

(i.e. A 's Gini-curve for defaults is above that of B 's) and

$$\frac{A0_i}{A_i} \leq \frac{B0_i}{B_i} \tag{20}$$

(i.e. A 's Gini-curve for non-defaults is below that of B 's). However, from (20) we have

$$\frac{A_i}{A0_i} \geq \frac{B_i}{B0_i} \tag{21}$$

and multiplying the left and right side of (19) with the left and right side of (21) yields

$$\frac{A1_i}{A0_i} \geq \frac{B1_i}{B0_i}, \tag{22}$$

which by the definition means $A \geq_{ROC} B$. That the converse is false can be shown by simple counterexamples.

From Theorem 2 it is clear that the ROC-ordering is the least demanding in practice, as will also be verified by our empirical example below.

4 Application

Our data are from the Moody's and S&P websites, respectively, from where we obtained the rating history of 5333 (Moody's) and 6505 private companies (S&P), covering the periods 1971-2014 (Moody's) and 1981 - 2014 (S&P) (see Standard & Poor's (2015) and Moody's (2015)). For each company, we recorded its first rating and its default state ten years after. Table 3 shows the results. PD is the percentage of defaults, and $q(a_i)$ denotes the relative frequency of rating class a_i , as defined in section 1. For instance, 24.26% of Moody's costumers and 22.94% of S&P's costumers were initially rated A (among which 2.09% defaulted within 10 years in case of Moody's and among which 1.71% defaulted in the case of S&P). Overall, we recorded 2301 defaults among debtors rated by S&P and 1938 defaults rated by Moodys, corresponding to $P^M(1) = 15.43\%$ and $P^{S\&P}(1) = 12.74\%$.

Table 3: Empirical ten year default rates (PD) and distribution of debtors among rating classes

Rating Class	Moody's		S&P	
	PD	$q(a_i)$	PD	$q(a_i)$
AAA/Aaa	0.49	3.41	0.71	1.07
AA/Aa	0.89	11.50	0.78	7.13
A	2.09	24.26	1.71	22.94
BBB/Baa	4.95	23.18	4.98	26.15
BB/Ba	19.79	14.23	16.38	17.37
B	40.25	17.86	29.97	22.77
CCC/Caa-C	65.97	5.54	51.35	2.56

Table 4 gives the resulting bivariate probability distribution if we view the empirical default rates in the various rating classes as predicted default probabilities (i. e. if we assume that both agencies are well calibrated), after incorporating the respective distribution of creditors across rating grades.

From table 4, one readily obtains the conditional probabilities $q^{Moody's}(a_i|1)$, $q^{S\&P}(a_i|1)$, $q^{Moody's}(a_i|0)$, $q^{S\&P}(a_i|0)$, and it emerges that none of the relationships (7) or (8) obtains (see table 5). Thus, Moody's and S&P cannot be ranked in either the VM -default nor non-default sense. However, as figures 3 and 4 show, the Gini-curve for defaults of S&P is below and the Gini-curve for survivors is above that of Moody's, so Gini-domination obtains.

Table 4: Bivariate probability distribution across rating classes and default states

a_i	Moody's			S&P		
	$q(a_i)$	$r(0, a_i)$	$r(1, a_i)$	$q(a_i)$	$r(0, a_i)$	$r(1, a_i)$
0.49	3.41	3.39	0.02	0.00	0.00	0.00
0.71	0.00	0.00	0.00	1.07	1.06	0.01
0.78	0.00	0.00	0.00	7.13	7.07	0.06
0.89	11.50	11.40	0.10	0.00	0.00	0.00
1.71	0.00	0.00	0.00	22.94	22.55	0.39
2.09	24.26	23.75	0.51	0.00	0.00	0.00
4.95	23.18	22.03	1.15	0.00	0.00	0.00
4.98	0.00	0.00	0.00	26.15	24.85	1.30
16.38	0.00	0.00	0.00	17.37	14.52	2.85
19.79	14.23	11.41	2.82	0.00	0.00	0.00
29.97	0.00	0.00	0.00	22.77	15.95	6.82
40.25	17.86	10.67	7.19	0.00	0.00	0.00
51.35	0.00	0.00	0.00	2.56	1.25	1.31
65.97	5.54	1.89	3.65	0.00	0.00	0.00

Table 5: Cumulated sums of defaults and non-defaults

a_i	Moody's		S&P	
	$\sum_{i=1}^j q(a_i 0)$	$\sum_{i=1}^j q(a_i 1)$	$\sum_{i=1}^j q(a_i 0)$	$\sum_{i=1}^j q(a_i 1)$
0.49	4.01	0.11	0.00	0.00
0.71	4.01	0.11	1.22	0.06
0.78	4.01	0.11	9.33	0.50
0.89	17.49	0.77	9.33	0.50
1.71	17.49	0.77	35.17	3.57
2.09	45.59	4.06	35.17	3.57
4.95	71.65	11.49	35.17	3.57
4.98	71.65	11.49	63.65	13.80
16.38	71.65	11.49	80.30	36.13
19.79	85.15	29.74	80.30	36.13
29.97	85.15	29.74	98.57	89.68
40.25	97.77	76.32	98.57	89.68
51.35	97.77	76.32	100.00	100.00
65.97	100.00	100.00	100.00	100.00

Figure 3: Gini-ordering (default)

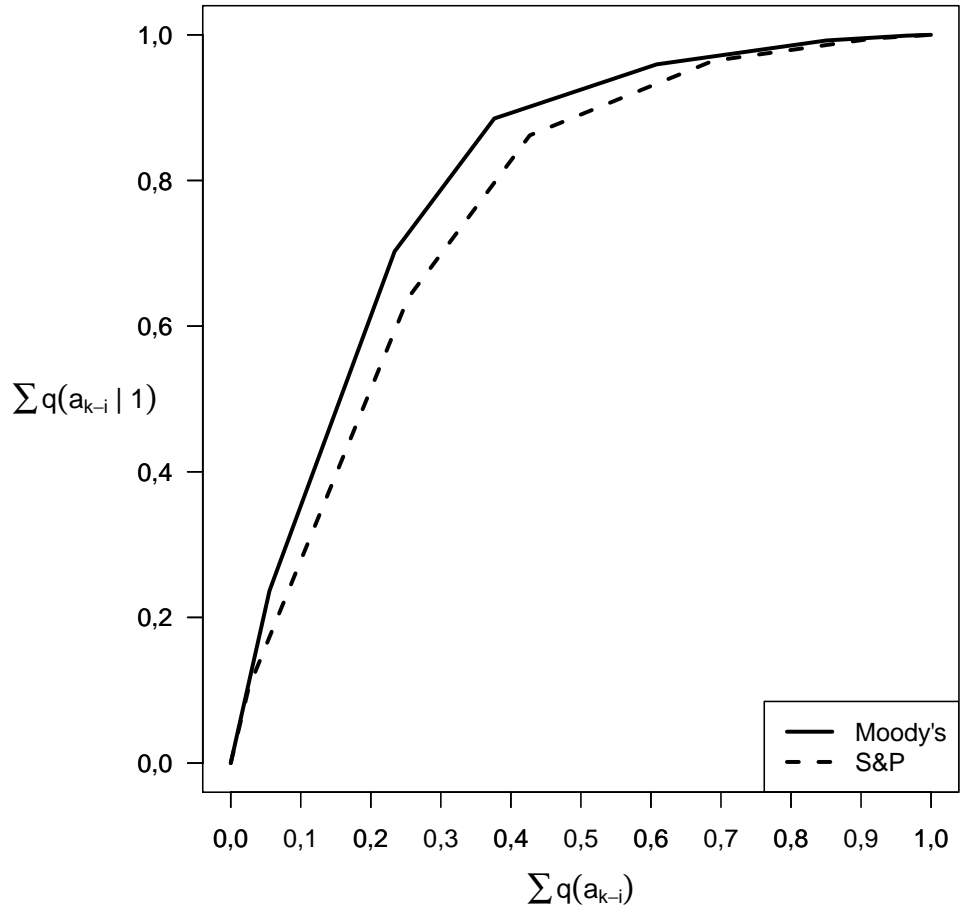
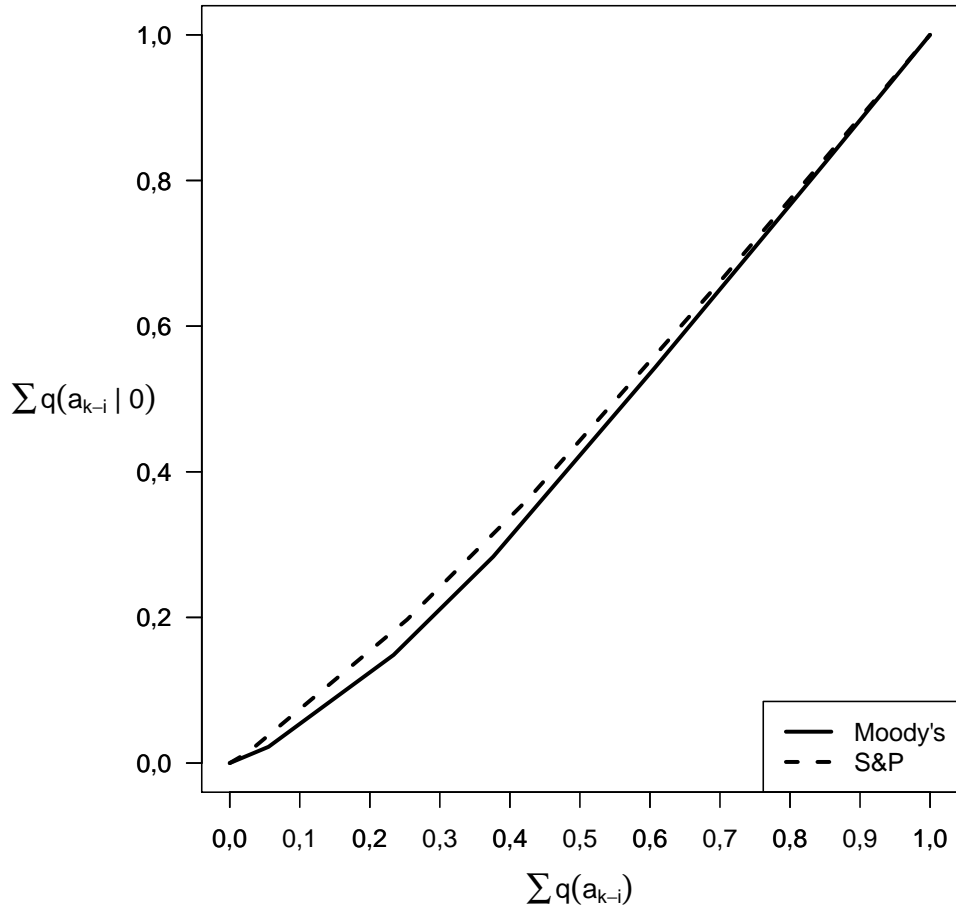
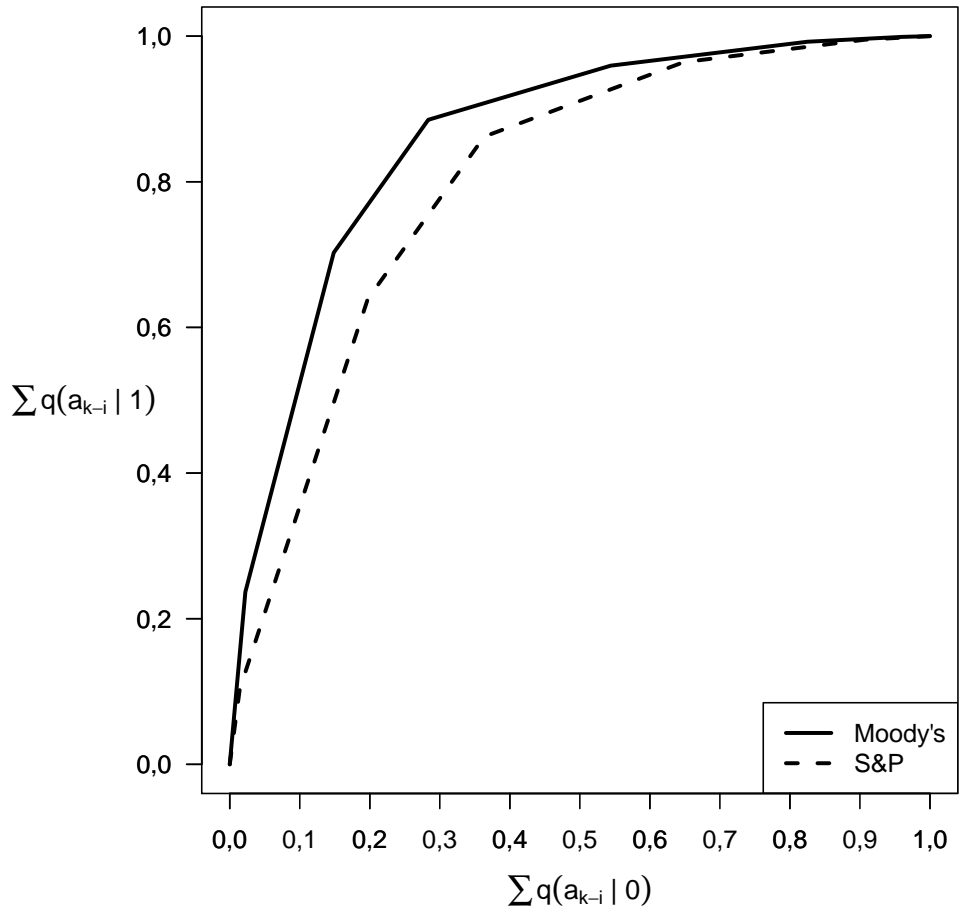


Figure 4: Gini-ordering (non-default)



From Theorem 2, it is therefore no surprise that Moody's dominates S&P as with respect to the ROC-criterion, as shown in figure 5.

Figure 5: ROC-curves



References

- CREAL, D. D. and GRAMACY, R.B. and TSAY, R. S. (2014). Market-Based Credit Ratings. *Journal of Business & Economic Statistics* **32**, 430-444.
- DEGROOT, M. and FIENBERG, S.E. (1983). The comparison and evaluation of forecasters. *The Statistician* **32**, 12-22.
- ENGELMANN, B. and HAYDEN, E. and TASCHE, D. (2003). Testing rating accuracy. *Risk* **1**, 82-86 .
- GNEITING, T. (2008). Editorial: Probabilistic forecasting. *J. R. Stat. Soc. A* **171**, 319-321.
- GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application* **1**.
- GROEN J.J.J, PAAP, R. and RAVAZZOLO, F. (2013). Real-time inflation forecasting in a changing world. *J. Bus. Econ. Stat.* **31**, 29-44.
- KAUPPI, H. and SAIKKONEN, P (2008). Predicting U.S. Recessions with Dynamic Binary Response Models. *The Review of Economics and Statistics* **90**, 777-791
- KRÄMER, W. (2005). On the Ordering of Probability Forecasts. *Sankhya: The Indian Journal of Statistics* **67**, 662-669.
- 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.
- MOODY'S (2015). Annual Default Study: Corporate Default and Recovery Rates, 1920-2014. *Moody's Investor Service*.

- SCHERVISH, M. (1989). A General Method for Comparing Probability Assessors. *The Annals of Statistics* **17**, 1856-1879.
- VAZZA, DIANE et. al. (2015). Default, Transition, and Recovery: 2014 Annual Global Corporate Default Study and Rating Transitions. *Standard & Poor's Ratings Direct*.
- VARDEMAN, S. and MEEDEN, G. (1983). Calibration, sufficiency and domination: Considerations for Bayesian probability assessor. *J. Amer. Statist. Assoc.* **78**, 808-816.
- WINKLER, R.L. (1996). Scoring rules and the evaluation of probabilities. *Test* **5**, 1-60.

