Discriminatory Power – an Obsolete Validation Criterion?

Manuel Lingo[a] **Gerhard Winkler**[a,b]

February 2008

Abstract

In this paper we analyse two common measures of discriminatory power - the Accuracy Ratio and the Area Under the Receiver Operator Characteristic - in a probabilistic framework. Under the assumption of a random default event, we verify that the measures will be portfolio dependent as discovered by Hamerle et al. (2003) and furthermore stochastic as indicated by Blochwitz et al. (2005). As an extension of these two papers we first study how the structure of a portfolio influences the measures. Furthermore we demonstrate that the measures contain information about a rating system's calibration quality. To this end a testing procedure is developed that also allows for a comparison of the measures across different portfolios. Our analysis leads to the final conclusion that high granularity and good calibration quality are sufficient to maximize attainable discriminatory power and are thus the sole criteria to be considered in any validation exercise.

JEL classification: G20, C49.

Keywords: Credit Risk, Validation, Discriminatory Power, Calibration Quality, ROC, CAP.

[a] *Department of Finance and Accounting, Vienna University of Economics and Business Administration, Heiligenstaedterstrasse 46, A-1190 Vienna, Austria;*

[b] *Oesterreichische Nationalbank (OeNB), Otto Wagner Platz 3, A-1090 Vienna, Austria.*

Any views expressed represent those of the authors only and not necessarily those of Oesterreichische Nationalbank.

1. Introduction

Sound rating systems are important for all financial institutions as they form the basis for calculating risk premia, pricing credits, and allocating economic capital. Furthermore, under both IRB approaches, the New Basel Capital Accord will allow banks to determine their capital requirements based on their own internal ratings (Basel Committee on Banking Supervision, 2005a). Hence banks and their supervisors are required to develop appropriate validation methodologies to be able to meaningfully assess a rating system's performance.

Historically rating systems have been mainly designed to establish a ranking among obligors based upon their relative credit risk (see e.g. Altman, 1968 and 1984, or Dimitras et al., 1996 for an overview of different modelling techniques suggested for this purpose). Hence the power to discriminate between different levels of credit risk has always been regarded as a key attribute of a good credit rating system (Sobehart and Keenan, 2001 and Basel Committee on Banking Supervision, 2005b). It follows, that academic research on the correct measurement of discriminatory power has been quite extensive. The Cumulative Accuracy Profile or the Receiver Operator Characteristic and their corresponding summary statistics, the Accuracy Ratio and the Area Under the Receiver Operator Characteristic, are among the most prominent methods which have been suggested to measure discriminatory power. They are thus heavily used in practice in the evaluation of a rating system’s performance (see e.g. Sobehart et al., 2000 and Stein, 2007). Furthermore, their popularity may be also partly due to the fact that these techniques can be easily employed to enforce a bank’s desired lending policy in the attempt to maximize the economic benefit of powerful credit scoring (see e.g. Stein and Jordao, 2003 and Blöchinger and Leippold, 2006).

More recently, however, it has become common to quantify the ordinal rating scales by attaching cardinal measures for the absolute level of credit risk to each rating. This is in fact the process also envisioned within the new Basel regulatory framework, where for each rating an estimate of the probability of default must be provided. Subsequently this estimate must be used in the computation of minimum capital requirements for all exposures (Basel Committee on Banking Supervision, 2005a). Hence, it appears as a “generally accepted rating principle” that the prime aim of a rating system is to generate estimates of the probability of default for all obligors (Krahnen and Weber, 2001). In this context the quality of a rating system can additionally be characterized by its ability to correctly assess the probability of default for each obligor. It follows that calibration quality - referring to the extent to which the forecasted probabilities of default match the subsequently realized default rates - is the second important criterion, apart from discriminatory power, which a sophisticated rating system should exhibit. As a result supervisory bodies continue emphasizing the importance to account for both aspects in validation (Basel Committee on Banking Supervision, 2005b).

Against this background we shall reconsider the most prominent measures of discriminatory power, the Accuracy Ratio and the Area under the Receiver Operator Characteristic. We begin by defining these measures in a probabilistic framework. Based on that, we study their properties in light of the fact that nowadays credit risk is mostly expressed in absolute terms by means of probabilities of default. We confirm the results of Blochwitz et al. (2005) and Hamerle et al. (2003) that the measures are portfolio dependent and stochastic and can thus not be directly compared across different portfolios. Extending these works, the main contributions of this paper are the following:

- We study how the structure of the portfolio influences the measures.
- Thereafter we analyze some important properties of the distributions of the measures by applying common models for the default process.
- Based on the results of these analyses we show how a comparative analysis of discriminatory power across different portfolios may still be achieved.
- Finally we demonstrate that the proposed measures for discriminatory power contain information on a rating system’s calibration quality if the rating system under evaluation provides estimates for probabilities of default. To this end we

develop a simple testing procedure, discuss its merits and pitfalls and apply it in an empirical example using data from two Austrian banks.

The outcomes of our analysis lead to the conclusion that it is counterintuitive to validate discriminatory power if the goal of a rating is to estimate probabilities of default. In this case, high granularity and good calibration quality are sufficient to maximize the attainable discriminatory power and are thus the sole criteria to be considered.

The paper is organized as follows. In a reconsideration of some of the most prominent measures of discriminatory power in Section 2 we study some of their properties under the assumption of a random default event. Section 3 discusses the implications of these properties. In Section 4 we provide an example using data of two Austrian banks to empirically demonstrate our main findings. Section 5 concludes the paper.

2. Methodological Reconsideration of Common Measures of Discriminatory Power

2.1 Assessment of a rating system's discriminatory power under the assumption of a random default event

Generally speaking, every lender faces a dichotomous classification problem in the attempt to ex-ante distinguish whether an obligor will default or not. In this context, the ability to discriminate in advance between subsequently defaulting and non-defaulting entities is referred to as the rating system's discriminatory power (see Basel Committee on Banking Supervision, 2005b). If a rating system could perfectly discriminate, it would be able to classify obligors into two distinct groups: obligors who will default with probability of 1 and obligors who will certainly not default, i.e. those with default probability of 0.

In practice, however, one is usually not able to unequivocally and deterministically classify all entities, ex-ante, into two – the 'will default' and the 'will not default' – categories only, due to incomplete information (see Basel Committee on Banking Supervision, 2000, and Krahnen and Weber, 2001). With the help of a rating system, a rater will thus try to establish a ranking among obligors depending on the estimates of some measure of creditworthiness in order to obtain indications about each obligor's future state. Based on this argument we introduce

POSTULATE 1:

The default event shall be understood as being of stochastic nature. Hence it can be represented by a random indicator variable $Y_{n,\Delta t}$ which may take the following values:

$Y_{n,\Delta t} = 1$ if the obligor n defaults within a certain period of time Δt

$Y_{n,\Delta t} = 0$ if the obligor survives the period Δt

Postulate 1 has two immediate consequences which are of fundamental importance for the remainder of this paper:

Corollary 1:

The realization of the random variable $Y_{n,\Delta t}$ will be governed by a (ex-ante) probability of default

$$PD_{n,t,\Delta t} = P\left(Y_{n,\Delta t} = 1\right) \tag{1}$$

which can take values in the open interval (0%;100%).

Corollary 2:

At each point in time t and for any time horizon Δt, a unique, true probability of default exists for each and every obligor.

Explicitly acknowledging postulate 1 and its two corollaries, the probability of default is one of the key risk parameters of an obligor's creditworthiness under the new regulatory

Basel II framework. All financial institutions adopting an IRB-approach are obliged to estimate this probability for each and every obligor (see Basel Committee on Banking Supervision, 2005a). Thus it follows that the prime aim of any rating system is to produce an estimate of the probability of default $\hat{PD}_{n,t,\Delta t}$ (see Krahnen and Weber, 2001). This is commonly achieved by making use of some kind of link function f that transforms a (ordinal or metric) rating score $R_{n,t}$ into the probability of default estimate[1]:

$$\hat{PD}_{n,t,\Delta t} = f\left(R_{n,t}\right) \tag{2}$$

This transformation, the mapping of a score to a probability of default, is called calibration (see Stein, 2007). Corollary 2, similarly to Krahnen and Weber (2001), suggests that a unique true probability of default exists for every obligor. According to Krahnen and Weber (2001) this is a "generally accepted rating principle". Note, however, that any estimate, $\hat{PD}_{n,t,\Delta t}$, produced by a rating system will naturally be noisy, as the true $PD_{n,t,\Delta t}$ given in (1) practically remains unknown. The deviation of the estimate from the true probability of default is thus the rating system's calibration error.

Against this background, apparently both facts

i. the stochastic nature of the default event (i.e., postulate 1),

and its consequence that

ii. rating systems aim at estimating the true probabilities of default (i.e., corollaries 1 and 2),

have to be adequately accounted for in the assessment of a rating system's discriminatory power. However, according to our experience the current practice in the application of measures of discriminatory power shows inconsistencies with the suggested propositions. This is due to the fact that the most prominent methods proposed for the evaluation of

[1] Note the general meaning of „link function" which can be of both types, theoretical, such as e.g. logit or probit, or empirical based on historical default experiencies.

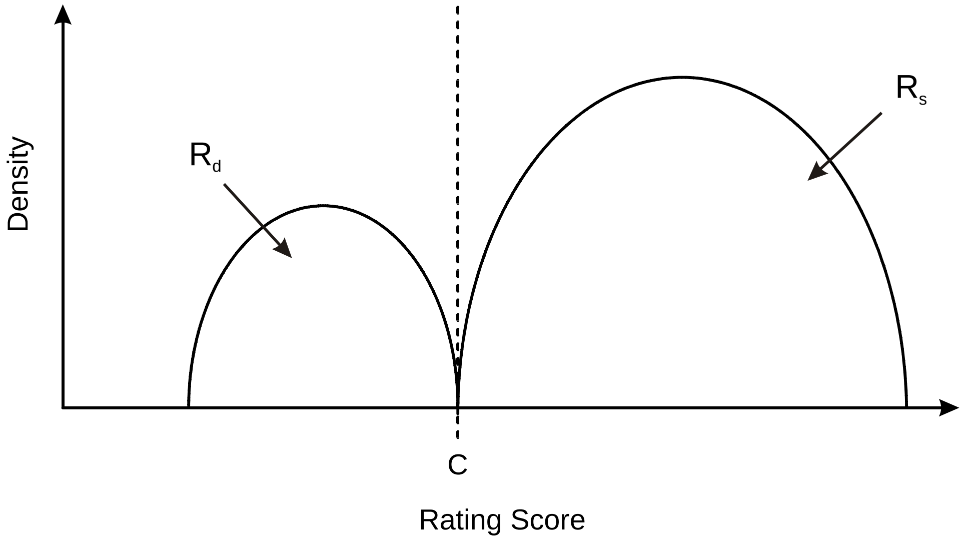
discriminatory power rely on backtesting (see Basel Committee on Banking Supervision, 2005b). At its core, any backtesting analysis involves the comparison of the forecast values for the future states, of all obligors, to the subsequently realized states. Observe, however, that if measured in a backtesting experiment, the extent to which a rating system was able to differentiate between subsequently defaulting and non-defaulting entities does not depend on the quality of the estimates of the obligors' credit quality only (i.e. the goodness of the ranking which is established upon probabilities of default derived from the rating scores). Even if one knew about the true probabilities of default governing the default processes of the rated entities, the realizations of the corresponding indicator variables $Y_{n,\Delta t}$ used in the calculation of measures of discriminatory power would remain random as the true probabilities are always greater than 0% and lower than 100%. As a consequence the measures of a rating system's discriminatory power derived from backtesting experiments will be of stochastic nature themselves. In light of this essential conclusion we will now review the most common measures of discriminatory power and then reconsider some of their theoretical properties in the next sections.

2.2 Review of common measures of discriminatory power

The most popular concepts to assess the discriminatory power of a rating system are the cumulative accuracy profile (CAP) and the receiver operating characteristic (ROC) with their corresponding summary statistics, the accuracy ratio (AR), and the area under the ROC curve (AUROC), respectively (see e.g. Basel Committee on Banking Supervision, 2005b, which contains a recommendation of these tools for the assessment of a rating system's discriminatory power). Detailed explanations how one should use these methods in a validation context can be found in Sobehart et al. (2000). We will first give a definition for these concepts in a probabilistic framework and discuss some of their well-known properties before moving on to studying the implications of the insights gained in the previous section.

To begin with, suppose that at a certain point in time t a sample consisting of N entities undergoes a credit assessment and that the result of each assessment may be summarized in a continuously-valued rating score $R_n \in \Re$ for each obligor n (Note that the time index will be skipped from now on for convenience). We will follow the convention that lower values of the score variable R are assumed to be associated with (a greater likelihood of) the default event. Also suppose that after a period Δt has elapsed, it can be determined at point $t + \Delta t$ independent of the test, that D entities have defaulted whereas S = N - D entities have survived. Without any loss of generality we can assume that the defaulting companies are numbered 1 to D, while the non-defaulting ones are numbered D+1 to N. Given the information on defaults available at point $t + \Delta t$, a rating system is said to discriminate perfectly if the rating scores of all defaulted entities, R_d , $d \le D$, would fall below a certain cut-off value of the rating score C whereas the scores of all non-defaulting entities, R_s , $s > D$, would be above this threshold. Exemplary score distributions of a perfect rating system are shown in Figure 1.

Figure 1: Score distributions of a perfectly discriminating rating system



In an imperfect rating system, the distributions of the scores of the defaulters and the non-defaulters would not be completely separated as in Figure 1, but overlapping. One can then evaluate the model's performance by the hit rate HR(C).

Denote the cumulative probability that the rating score of an entity from the entire portfolio is lower than or equal to a cut-off value C as:

$$F(C) = P(R \leq C) \tag{3}$$

Given the information available at point $t + \Delta t$, the hit rate is defined as the cumulative probability that the score of a defaulting entity R_d will be lower than a threshold C and is thus given by:

$$HR(C) = F_d(C) = P(R \leq C \mid Y = 1) \tag{4}$$

The hit rate measures a model's ability to correctly predict defaulters depending on the cut-off value C (Basel Committee on Banking Supervision, 2005b). As a consequence, 1-HR(C) will be the rating model's Type I error.

Analogously we can define a false alarm rate FAR(C) as the cumulative probability that a non-defaulter has been mistakenly classified as a defaulter depending on the cut-off value C:

$$FAR(C) = F_s(C) = P(R \leq C \mid Y = 0) \tag{5}$$

The false alarm rate is also referred to as the model's Type II error (Sobehart et al, 2000). It follows that 1-FAR(C) measures the model's ability to correctly identify survivors. Table 1 sums up these definitions in a contingency table also referred to as confusion matrix (see Stein, 2007).

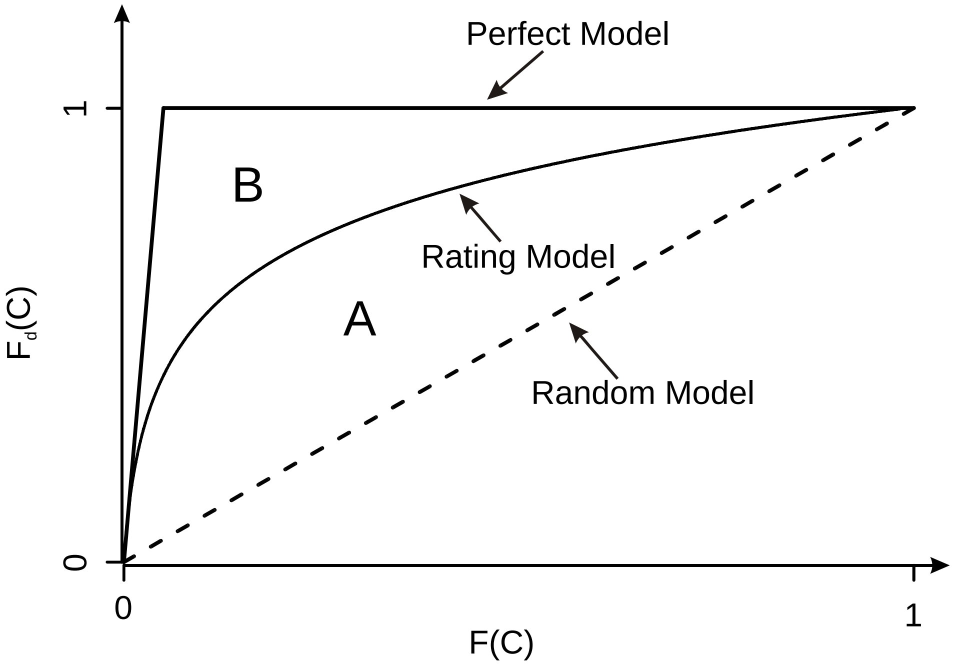
Table 1: Confusion Matrix

Prediction (dependent on cut-off value C)	Realisation	
	Default	No Default
Default	HR(C)	Type-II-Error = FAR(C)
No Default	Type-I-Error = 1-HR(C)	1-FAR(C)

Cumulative accuracy profiles (CAP) and receiver operating characteristics (ROC) - with their corresponding summary statistics, the accuracy ratio (AR), and the area under the ROC curve (AUROC) - incorporate the information of the hit rate and of the false alarm rate (see Sobehart et al, 2000) and are thus both calculated using the information available at point $t + \Delta t$:

- The cumulative accuracy profile (CAP) is defined as a plot of $HR(C) = F_d(C)$ (the cumulative probability that the rating score of a defaulting entity is lower than a threshold C) against $F(C)$ (the cumulative probability that the rating score of all the entities will be lower then C) for all values of C. Figure 2 demonstrates graphically that the CAP curve of any informative rating model will lie within two extremes. A rating model that can, ex-ante, perfectly discriminate between defaulting and non-defaulting entities would assign lower scores to all the defaulting entities than to all the non-defaulting ones (bold line in Figure 1). In contrast, for a random model without any discriminatory power, the cumulative probability that the rating score of the defaulted entities is lower than or equal to C will be equal to the cumulative probability that the rating score of all the entities is lower than or equal to C, i.e. $\forall C \in \Re : F_d(C) = F(C)$ (dashed line in Figure 2). Thus the closer the CAP curve of a real rating model is to that of the perfect model, the better its predictive power.

Figure 2: The Cumulative Accuracy Profile (CAP)



A one-dimensional summary measure of a model’s CAP curve is the so-called Accuracy Ratio (AR). AR is defined as the area between the CAP curve of the model under consideration and the CAP curve of the random model divided by the area between the perfect model and the CAP curve of the random model (i.e., A/(A+B) in Figure 2). As the probability to observe a default is P(Y=1), it follows based on geometrical considerations, that the area between the CAP curve of a perfect model and the CAP curve of the random model is:

$$\begin{aligned} A + B &= P(Y = 1) \cdot 0.5 + P(Y = 0) - 0.5 \\ &= 0.5 \cdot P(Y = 0) \end{aligned} \tag{6}$$

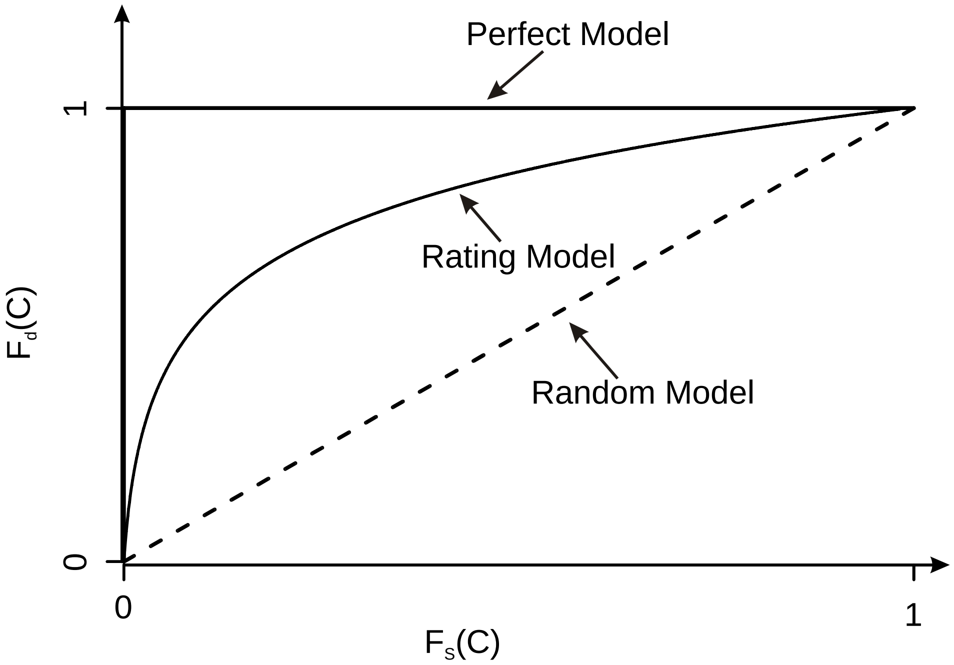
The accuracy ratio of any real rating system will then be:

$$AR_{RS} = \frac{A}{A+B} = \frac{\int_0^1 F_d(C)dF(C) - 0.5}{0.5 \cdot P(Y=0)} \tag{7}$$

In principle, the AR can take any value in the interval [-1 , +1]. For any informative model, however, the value will lie between 0 representing the random model and 1 for the perfect model.

- A similar concept is the receiver operator characteristic (ROC) which is defined as a plot of $HR(C) = F_d(C)$ (the cumulative probability that the rating score of a defaulting entity is lower than a threshold C) against $FAR(C) = F_s(C)$ (the cumulative probability that the rating score of a non-defaulting entity will be lower then C) for all values of C. Similar to CAP curves, Figure 3 shows that the ROC curve of any informative rating model also lies between the curve of a perfect model (which contains the point (0,1)), and the curve of a random model (represented again by a 45° line).

Figure 3: The Receiver Operator Characteristic (ROC)



A corresponding summary statistic of the ROC curve is given by the Area Under the ROC curve (AUROC) which is defined as the area below the ROC curve and is thus computed as follows:

$$AUROC_{RS} = \int_{0}^{1} F_d(C) dF_s(C) \tag{8}$$

The AUROC can take any value in the interval [0 , 1]. For any informative model the value will lie between 0.5 representing the random model and 1 for the perfect model.

Note that AR and AUROC do indeed contain the same information and are unambiguously related to one another (see Engelmann et al., 2003). As shown in Appendix 1, the relation between AR and AUROC is given by the following equation:

$$AR_{RS} = 2 \cdot AUROC_{RS} - 1 \tag{9}$$

2.3 The stochastic nature of measures of discriminatory power

After a review of common measures of discriminatory power as well as some of their properties we shall now discuss how postulate 1 (i.e., the stochastic nature of the default event), and its corollaries (the fact that rating systems aim at estimating the true obligor-specific probabilities of default), impact the interpretation of these measures. As shown in Appendix 2, the AR may also be expressed as follows:

$$AR_{RS} = 2 \cdot \left\langle \frac{1}{P(Y=1) \cdot [1 - P(Y=1)]} \cdot \left[\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{C_1} f(Y=1|C_2) \cdot f(C_2) dC_2 \right) \cdot f(Y=0|C_1) \cdot f(C_1) dC_1 \right] \right\rangle - 1 \tag{10}$$

Obviously, the AR (and thus also the AUROC which is the term in angle brackets in (10)) is a function of the following variables:

i. the probability (density) to observe a certain rating score given by $f(C)$,

ii. the probability of default associated with a rating score as given by $f(Y=1|C)$,

iii. the probability to survive associated with a rating score as given by $f(Y=0|C)$, and

iv. the overall probability $P(Y=1)$ to observe a default event in the entire portfolio.

The probability to survive is related to the probability of default by means of the relation $f(Y=0|C)=1-f(Y=1|C)$. The overall probability of default is defined as $P(Y=1)=\int_{-\infty}^{+\infty} f(Y=1|C)\cdot f(C)dC$. Hence both summary measures of discriminatory power ultimately depend on two aspects,

i. the distribution of the rating scores as given by the density $f(C)$, and

ii. the probabilities of default assigned to the rating scores $f(Y=1|C)$.

These findings clearly have important consequences for the interpretation of measures of discriminatory power in the validation of rating systems:

As first described by Hamerle et al. (2003), it is evident from (10) that the measures of discriminatory power will be portfolio dependent. This is due to the fact that the composition of the portfolio predetermines a rating system's attainable discriminatory power (see Hamerle et al., 2003 and Stein, 2007). To recognize this, remember that the AUROC and the AR are both determined by the distribution of the rating scores and their associated probabilities of default. If one accounts for the stochastic nature of the default event and accepts corollary 2, that for each and every obligor a unique true probability of default exists, it is evident that the AUROC and the AR will be dependent on the credit quality of the portfolio under consideration: The true probabilities of default of the obligors contained in a portfolio will determine the achievable discriminatory power. In addition it follows that a comparison of the discriminatory power of two rating systems for two different portfolios is not straightforward as the values for AR and AUROC are not comparable across different universes (see Stein, 2007).

Second, both the measures, the AR and the AUROC, can be regarded as being of stochastic nature as indicated by Blochwitz et al. (2005). To see this more clearly, note that the probabilities of default associated with the rating scores $f(Y=1|C)$ used in the calculation of the AR and the AUROC are empirical (a posteriori) probabilities, calculated based on the information available at point $t+\Delta t$. Their values depend on the realizations of $Y_{n,\Delta t}$, and in addition also on the size of the portfolio N and the distribution of the rating scores $f(C)$. From (1), however, we know that these realizations will be governed by the (true) ex-ante probabilities of default $PD_{n,t,\Delta t}=P(Y_{n,\Delta t}=1)$ as given at point t. These facts have two important consequences that have to the best of our knowledge not yet been fully discussed in literature.

i. Even if a rating system has been perfectly calibrated, the predicted probabilities of default (i.e., the ex-ante probabilities given at point t) do not necessarily have to match the actual default rates (i.e., the a posteriori empirical probabilities as given at point $t+\Delta t$) exactly. In a practical example this would only be the case if the number of obligors experiencing a certain rating given by $f(C)\cdot N$ is infinitely large. Only then will the ex-ante probability be identical with the a posteriori empirical probability for this rating. As a consequence the observed values for measures of discriminatory power may vary from one experiment to another even if a rating system is perfectly calibrated and the structure and quality of the portfolio remain constant. Thus the proposed measures of discriminatory power are stochastic themselves.

ii. As already mentioned in Stein (2007) discriminatory power and calibration are related. That said; note that in light of a stochastic default event, the proposed measures of discriminatory power can also be applied to analyze a rating system’s calibration quality: A concrete value observed in an experiment, for any of the proposed measures of discriminatory power, which is based on empirical probabilities, can be compared to the distribution of values obtained using the ex-ante probabilities of default in the calculation of the measures. That, however, means that tests for discriminatory power and tests for calibration quality of a rating system – although often treated as two different concepts in practice – are in fact closely related. We will elaborate further on this idea in the following sections.

To sum up, if one acknowledges the stochastic nature of the default event, important properties of the proposed measures of discriminatory power become apparent. Moreover, these measures may have to be interpreted differently as it is current practice.

3. Implications of the Previously Identified Properties of the Proposed Measures of Discriminatory Power

3.1 First set of preliminary examples

To elaborate on the key properties of the proposed measures of discriminatory power derived in the previous section consider the following example:

Assume that a bank A uses a rating system which assigns obligors to only two rating classes R_1 and R_2. Information about the distribution of obligors across the two rating classes and the ex-ante probabilities of default (PD), derived from calibration of the model(s), as given at point t are given in Table 2.

Table 2: Portfolio of sample bank A

Rating Class R	Obligors	ex-ante PD
1	1500	2.5%
2	1500	5.5%

Suppose that the rating system of bank A is perfectly calibrated (i.e., the ex-ante PDs equal the true probabilities of default). Furthermore suppose for the moment that the default rates observed at point $t + \Delta t$ (i.e. the empirical probabilities of default) are identical to the ex-ante PDs for the moment.

For the calculation of the AR and AUROC measures, note that the continuous rating score has now been replaced by a discrete number of distinct rating classes (i.e., 2 in our example). As already mentioned in Fawcett, 2004, any ROC- or CAP-curve generated from a finite set of scores is a step function, which approaches the true curve as the number of (rating) scores (and thus the number of classes) approaches infinity. Therefore we have to discretize the formula for our summary measures given in (10) which then becomes:

$$AR_{RS} = 2 \cdot \left\langle \frac{1}{P(Y=1) \cdot [1 - P(Y=1)]} \cdot \left[\sum_{i=1}^{M} \sum_{j=1}^{M} P(Y=1|R=i) \cdot P(R=i) \cdot P(Y=0|R=j) \cdot P(R=j) \right] \right\rangle - 1 \quad (11)$$

where i,j = 1,…,M denote the rating (class membership) and the term in angle brackets is again the AUROC.

A second problem may very likely arise in practice when there are equally scored obligors (e.g., in our example we assumed for the moment that all the 1500 obligors in rating class R_1 have the same score and are thus assigned the same probability of default), since one has to establish an ordering among them for the calculation of AUROC and AR. The formula given in (11) considers the most “optimistic” scenario supposing that all the defaulters end up at the beginning of the sequence of “equally” ranked obligors. The other, most “pessimistic”, extreme would be to assume that all defaulters end up at the end of the sequence. Lacking information on an ordering of obligors within a rating class, it seems reasonable to compute the expected performance for each and every class as an average of the “optimistic” and the “pessimistic” scenario (Fawcett, 2004). Formula (11) then has to be adjusted as follows:

$$AR_{RS} = 2 \cdot \frac{1}{P(Y=1) \cdot [1 - P(Y=1)]} \cdot \left\langle \begin{array}{l} \left[\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} P(Y=1|R=j) \cdot P(R=j) \cdot P(Y=0|R=i) \cdot P(R=i) \right] + \\ + \frac{1}{2} \cdot \left[\sum_{i=1}^{M} P(Y=1|R=i) \cdot P(R=i) \cdot P(Y=0|R=i) \cdot P(R=i) \right] \end{array} \right\rangle - 1$$

(12)

Formula (12) provides the same results as when calculating the summary measures of discriminatory power based on geometrical considerations using the trapezoidal rule and is thus conform with the definition given by supervisory agencies (see Basel Committee on Banking Supervision, 2005b).

For sample bank A the variables required for the calculation of AUROC and AR take the following values:

$$P(Y=1|R=1)=0.025$$
$$P(R=1)=\frac{1500}{3000}=0.5$$

$$P(Y=1|R=2)=0.055$$
$$P(R=2)=\frac{1500}{3000}=0.5$$

$$P(Y=1)=0.5\cdot 0.025+0.5\cdot 0.055=0.04$$

Using formula (12) the AR for sample bank A is thus:

$$AR_{bank\ A}=2\cdot\left\langle\frac{1}{0.04\cdot 0.96}\cdot\left[0.055\cdot 0.5\cdot 0.975\cdot 0.5+\frac{1}{2}\left(0.025\cdot 0.5\cdot 0.975\cdot 0.5+0.055\cdot 0.5\cdot 0.945\cdot 0.5\right)\right]\right\rangle-1= \\ =0.1953...$$

The corresponding AUROC will hence be:

$$AUROC_{bank\ A}=\frac{AR_{RS}+1}{2}= \\ =0.5976...$$

In addition consider now a second bank B which also employs a rating scale consisting of two rating classes. Information about the distribution of obligors across the two rating classes and the ex-ante probabilities of default are given in Table 3.

Table 3: Portfolio of sample bank B

Rating Class R	Obligors	ex-ante PD
1	1500	2.5%
2	1500	10.0%

Like bank A, bank B's portfolio also comprises 3000 obligors, 1500 of which are again assigned to rating class 1, the remaining 1500 to rating class 2. However, the credit quality of bank B's portfolio is obviously worse than the one of bank A. The ex-ante PD of rating class 1 is again 2.5% whereas the ex-ante PD of rating class 2 is 10% this time. Again we assume that the rating system of bank B is perfectly calibrated and that the observed default rates are identical to the ex-ante PDs.

By application of formula (12) we obtain:

$AR_{bank\ B} = 0.32$ and $AUROC_{bank\ B} = 0.66$

As we infer from the two examples the value of AR and AUROC will be higher, the steeper the PD-function of the rating classes. With a steep PD-function, the relative share of defaulters contained in the worse rating classes will ceteris paribus be much larger than the relative share of defaulters contained in the better rating classes (see Stein, 2007). As a consequence, in our example the rating system of bank B apparently does a better job in segregating the defaulting entities from the non-defaulting ones. It is thus important to note, that any (supervisory) recommendation on some minimum discriminatory power that a rating system has to achieve, implicitly comprises a specific requirement on the structure of the rater's rating scale. More specifically, it constrains the specification of the PD-function.

In addition to the steepness of the PD-function, the structure of a system's rating scale impacts the discriminatory power also via the scale's granularity. As the CAP- and ROC-curves will be step functions, when the number of different rating scores / classes is finite, the AUROC and the AR will ceteris paribus increase with the rating system's granularity. To see this, we consider again bank A and modify the first example. Assume that the bank's rating system provides an individual probability of default for each and every obligor. A common practice in reality, however, is pooling obligors with similar credit quality in one rating class, for simplification purposes. Accordingly, we now suppose that the ex-ante PDs reported in Table 2 are not the true probabilities of default of the obligors in the respective rating classes, but the averages of the obligor specific true probabilities of default. Furthermore, we assume that the averages have been computed on the individual probabilities of default of the obligors which are evenly distributed in both rating classes, for rating class 1 in the interval (1% ; 4%) with an average of 2.5% as reported in Table 2 and for rating class 2 in the interval (4% ; 7%) with an average of 5.5%. Bank A could then revise its rating scale and create, e.g., three rating classes, where the first one comprises the 1000 obligors with a probability of default in the 1%-3% range, the second class the 1000 obligors with a probability of default in the 3%-5% range, and the third one the remaining 1000 obligors whose probability of default is in the 5%-7% range. The new average probabilities of default of the three distinct rating classes will then be 2% for 1, 4% for 2 and 6% for 3. The regrouping of the obligors into a larger number of rating classes will then ceteris paribus lead to an increase in the values of the summary measures of discriminatory power for which we now obtain (by using (12)):

$AR_{bank\ A} = 0.2314...$ and $AUROC_{bank\ A} = 0.6157...$

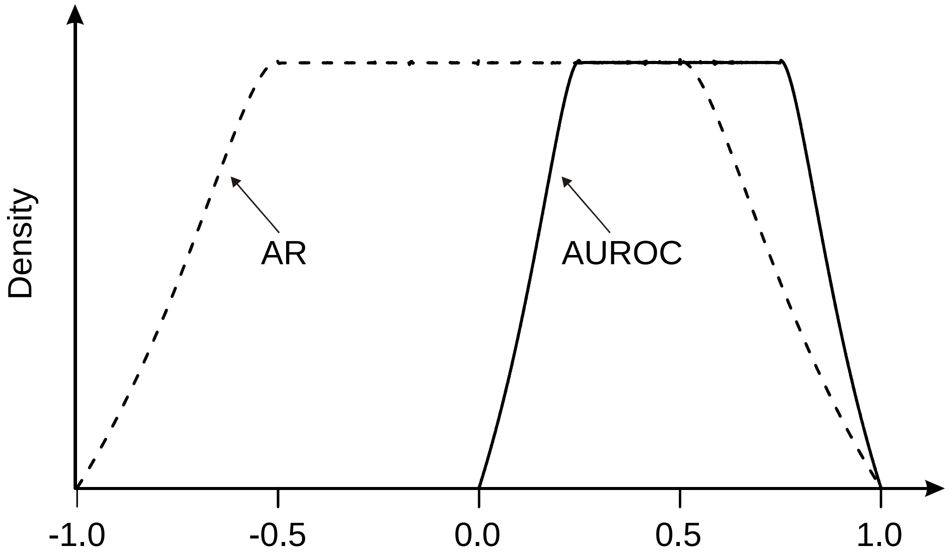
We hence conclude that a rating system's discriminatory power strongly depends on the structure of the rating scale. That said, however, if one accepts corollaries 1 and 2, i.e. the view that the random default events of obligors are governed by true and portfolio independent probabilities of default (which are themselves dependent on obligor characteristics only), the rater has no influence on the shape of the PD-function. In this case the true credit quality of the portfolio as given by the true probabilities of default determines the discriminatory power which a rating system may achieve. In our examples, the values of the proposed measures of discriminatory power are higher for bank B only because bank B manages a different portfolio whose credit quality is lower than the one of bank A. Obviously, it would thus be premature to conclude that bank A's rating system is inferior to the one of bank B. As a consequence, measures of discriminatory power are not directly comparable across different portfolios when the credit quality of these portfolios is different. Thus, any performance evaluation of two competing models / rating systems has to be undertaken in the same portfolio based on the same time window.

3.2 Distribution of AR and AUROC assuming independent default events

We now extend the previous examples in order to elaborate on the stochastic nature of the summary measures of discriminatory power. We continue assuming that the ex-ante PDs for both rating classes exactly match the true probabilities of default. The default event itself, however, is random as previously noted. As a consequence the number of realized defaults may vary.

The portfolio structure of both banks mentioned in our introductory example, suggests that the total number of realized defaults must lie within the 0 – 1500 bounds for both rating classes 1 and 2. Thus, depending on the observed defaults in both rating classes, there are a total number of $1501^2 = 2253001$ different default patterns. Figure 4 shows the distribution of the AR and AUROC values, obtained while computing all the possible default patterns. Evidently, AR can take any value between -1 and +1 (and 0 to +1 for AUROC respectively) depending on the realized default pattern. The expected value for the AR is 0 and 0.5 for AUROC.

Figure 4: Distribution of AR and AUROC assuming that all default patterns are equally likely



The distributions shown in Figure 4 are based on the assumption that all default patterns are equally likely. However, as the default patterns are determined by the ex-ante probabilities of default, each of them will have an individual likelihood of occurrence. To compute this likelihood, we begin by applying a binomial model to the defaults in every rating class *i*, i.e.

$$P(D_i = d_i) = b(d_i; N_i; PD_i) = \binom{N_i}{d_i} PD_i^{\,d_i} (1 - PD_i)^{N_i - d_i} \tag{13}$$

where

- *D*D_i, d_i denote the number of defaults observed within a period of length Δt as given by $\sum_{n=1}^{N} Y_{n,\Delta t}$,
- the 'success' probability PD_i, is the ex-ante (true) probability of default for the respective rating class *i* as given at point t,
- and the number of experiments N_i is the number of obligors assigned to the respective rating class i.

Mind that the adequacy of a binomial model is based on two critical assumptions:

i. the (ex-ante) probability of default being exactly the same for all obligors within a rating class,

   (This appears to be a reasonable assumption since a rating class should, by definition, contain obligors which are homogenous with respect to their credit quality, see Basel Committee on Banking Supervision, 2005a and 2006.)

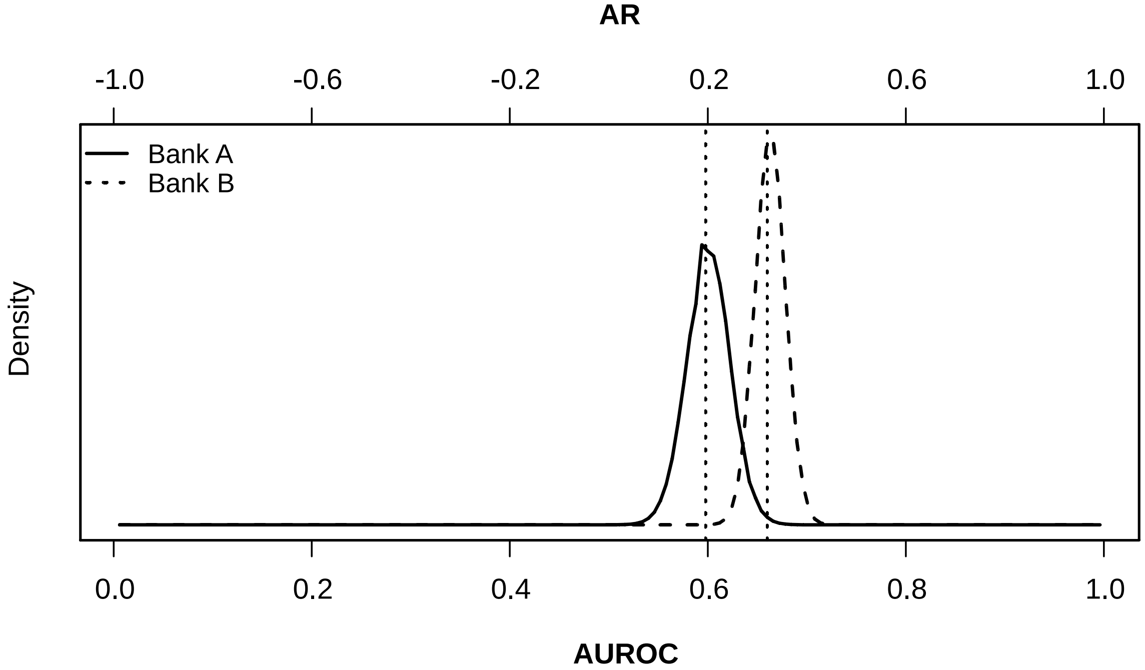
ii. the default events being independent.

   (In practice this assumption is more disputable as it implies that there is no default correlation. It shall therefore be relaxed at a later stage.)

The binomial distribution function yields the probability to observe a number of D_i defaults at point on $t + \Delta t$ depending on a given ex-ante PD_i at point t and a given size of the portfolio N. Using D_i and N the empirical probabilities of default (i.e. the realized default frequencies at point $t + \Delta t$) can be derived for every rating class. These enable one to compute the AR and AUROC for each and every default pattern, as well as the likelihood of its occurrence. Depending on the number of rating classes, this likelihood will be the joint probability that the number of defaults D_1 in rating class 1 is equal to d_1; the number of defaults D_2 in rating class 2 is equal to d_2; etc.

The computation of the AR and AUROC measures for all of the different default patterns, as well as their corresponding likelihoods of occurrence, enables us to construct the distributions of these summary measures. Figure 5 shows the distribution of the AR and AUROC measures for our sample banks A and B.

Figure 5: Distribution of AR and AUROC for both sample banks



Note, that the expected values obtained for AR and AUROC, as illustrated in Figure 5, are identical to the values computed in the examples in the preceding section, as we previously assumed that the empirical probabilities of default equal the ex-ante probabilities of default. Figure 5 clearly shows that similarly to the realized default frequencies, which may differ from the ex-ante probabilities of default, the AR and AUROC measures may also differ from their expected values. Thus, even though we assumed that the ex-ante probabilities of default are the true probabilities of default (meaning that the rating systems are free of any calibration errors) we can still observe considerable variation in the realized AR and AUROC values around their expected values. E.g. for bank A whose expected value for the AR is 0.1953, the AR will lie in the interval [0.1230 , 0.2665] with 90% probability. Furthermore note that there is even a chance – albeit very small – that the AR of bank A will be higher than the AR of bank B. For example the probability to observe an AR > 0.2665 is 5.00% for bank A, whereas the probability to obtain an AR < 0.2665 is 4.32% for bank B.

3.3 Distribution of AR and AUROC in the presence of default correlation

Up to now we have assumed that defaults occur independently. To study the effects of dependency among defaults on the distribution of the proposed measures for discriminatory power, we will now relax this assumption. In our analysis we will employ a one-factor model of firm's value to model dependency in defaults. Let V_n denote the value of the assets of the n'th obligor and assume that it is standard normally distributed $V_n \sim \Phi(0,1)$. Consistent with standard structural credit risk models (see Merton, 1974) we suppose that the firm will default if its value falls below a pre-specified threshold $V_n < K_n$. Furthermore, we assume that the values of the assets of the obligors within one rating class are driven by a common, standard normally distributed factor X and an idiosyncratic standard normal noise term ε_n:

$$V_n = \sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_n \quad \forall n \le N \tag{14}$$

Consequently, the values of the assets of any two distinct obligors are correlated with linear correlation coefficient ρ. However, the defaults will be independent conditional on the realisation x of the common factor X. Using (14) the conditional default probability $PD_n(x) = P(Y_n = 1 | X = x)$ for obligor n is given by:

$$\begin{aligned} PD_n(x) &= P(Y_n = 1 | X = x) \\ &= P\left(\sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_n < K_n | X = x\right) \\ &= P\left(\varepsilon_n < \frac{K_n - \sqrt{\rho}X}{\sqrt{1-\rho}} | X = x\right) \\ &= \Phi\left(\frac{K_n - \sqrt{\rho}x}{\sqrt{1-\rho}}\right) \end{aligned} \tag{15}$$

Hence, in contrast to (13), where we assumed independence of the default events, the probability mass function for the number of defaults in rating class *i* (allowing for default correlation) is defined by a generalized Bernoulli mixture model (see e.g. Schoenbucher, 2000 or Bluhm et al., 2003) and thus given by:

$$P(D_i = d_i) = \int_{-\infty}^{+\infty} \binom{N_i}{d_i} \Phi\left(\frac{K_i - \sqrt{\rho}x}{\sqrt{1-\rho}}\right)^{d_i} \left[1 - \Phi\left(\frac{K_i - \sqrt{\rho}x}{\sqrt{1-\rho}}\right)\right]^{N_i - d_i} \phi(x)dx \tag{16}$$

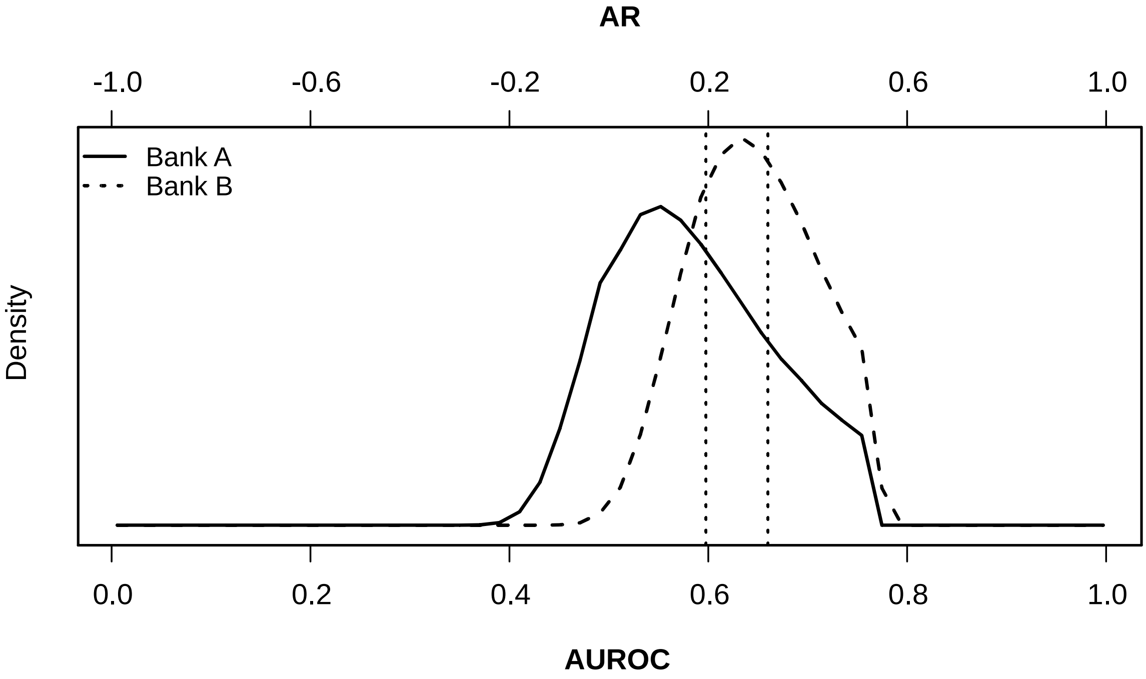
Note that our approach to model dependency in defaults is similar in nature to the CreditMetrics model (see J. P. Morgan, 1997). Furthermore, it is also closely related to the Basel Committee on Banking Supervision's approach to incorporate correlation into the calculation of risk weighted assets (see Basel Committee on Banking Supervision, 2005a).

Based on (16) we can now modify our previous examples to account for correlation. To assess the correlation effects on the distribution of the AR and AUROC measures, we estimate correlation according to its specification in the calculation of risk-weighted assets as defined by the new Basel II framework. According to paragraph 272 in Basel Committee on Banking Supervision (2005a) the relationship between the probability of default and correlation is given by:

$$\rho = 0.12\frac{1-e^{-50\cdot p_i}}{1-e^{-50}} + 0.24\left(1 - \frac{1-e^{-50\cdot p_i}}{1-e^{-50}}\right) \tag{17}$$

In (17) correlation only depends on the probability of default PD_i assigned to a certain rating class *i*. It can take on any value between 0.12 and 0.24 depending on the probability of default. For our sample banks correlation is thus 0.1544 in rating class 1, 0.1277 in rating class 2 of bank A and 0.1208 in rating class 2 of bank B. Using these values, we can compute the probabilities of occurrence for every attainable AR and AUROC value based on (16). The resulting distributions of AR and AUROC are shown in Figure 6.

Figure 6: Distribution of AR and AUROC for both sample banks in the presence of default correlation



As one would expect the variance is higher than in the case with independent default. This suggests that extreme values for AR and AUROC more likely if defaults are correlated. Consequently, the chance that the observed values for AR and AUROC will be higher for bank A (whose expected values are lower) than for bank B, also increases considerably. E.g., the probability to observe an AR > 0.2665 is now 32.75% for bank A, whereas the probability to obtain an AR < 0.2665 is now 44.13% for bank B

Evaluating the findings from our examples, we may conclude that due to the portfolio dependence and the stochastic nature of the proposed measures of discriminatory power, it is not admissible to base a comparative analysis of the discriminatory power of two rating systems, for two different portfolios, upon a sole comparison of realized values of AR or AUROC. Rather, each realization should be interpreted in the context of the rating system's portfolio dependent AR and AUROC distributions. As we will show in the following section, the deviation of the realization from the expected value may then serve as a comparable indicator for the quality of rating systems across different portfolios.

3.4 The information about calibration inherent in measures of discriminatory power

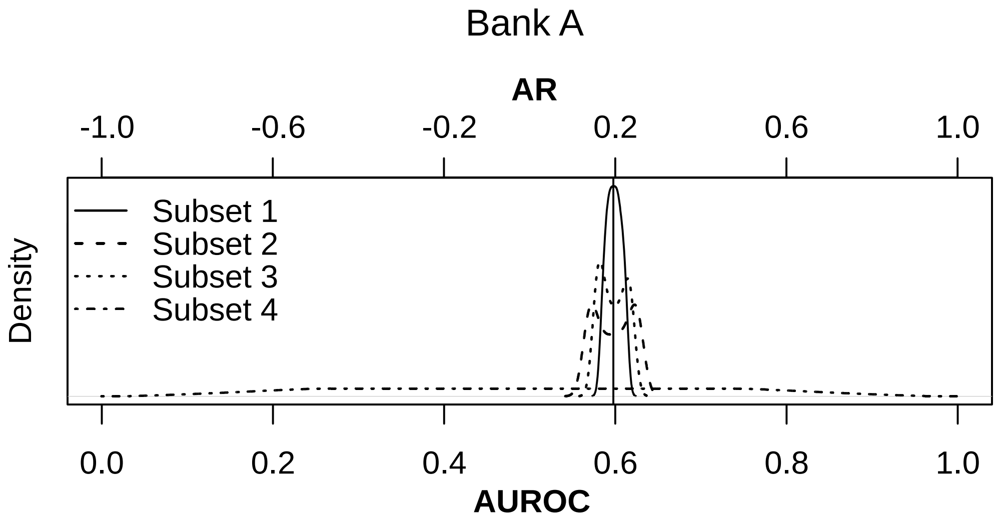
Looking at the distributions of the AR and AUROC measures as shown in Figures 5 and 6, it appears important to investigate the relation between the likelihood of a certain default pattern and the corresponding likelihood to observe a certain AR or AUROC value in a next step. To get first indications on the density properties of the AR and AUROC we proceed as follows:

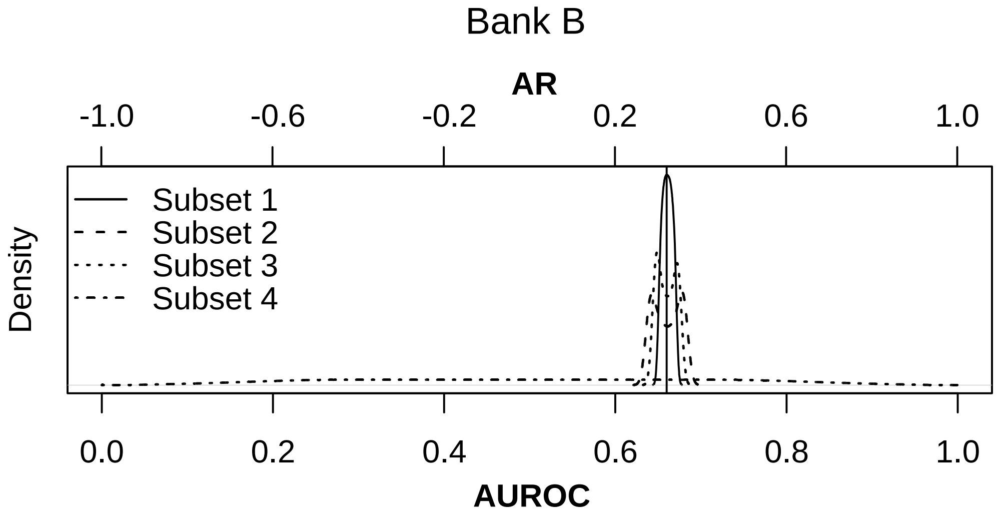
- order the values obtained for both measures based on their probability of occurrence in a descending manner (i.e. from the default pattern with the greatest likelihood to the one with the lowest likelihood)
- divide these ordered values into four subsets

Each of the subsets will now represent the respective 25% of cumulative probability of occurrence. The results are depicted in Figure 7 for both scenarios (i.e., with and without accounting for correlation).

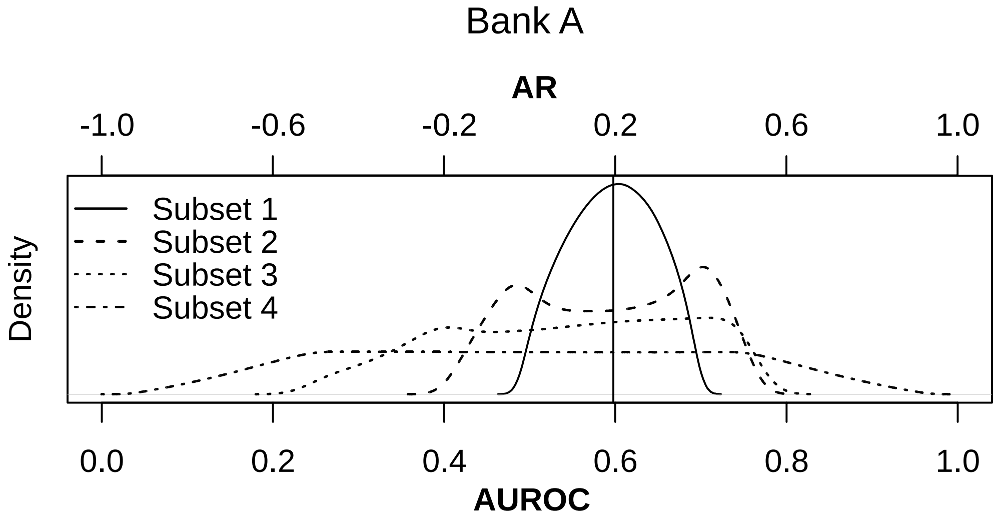
Figure 7: Analysis of AR and AUROC distributions

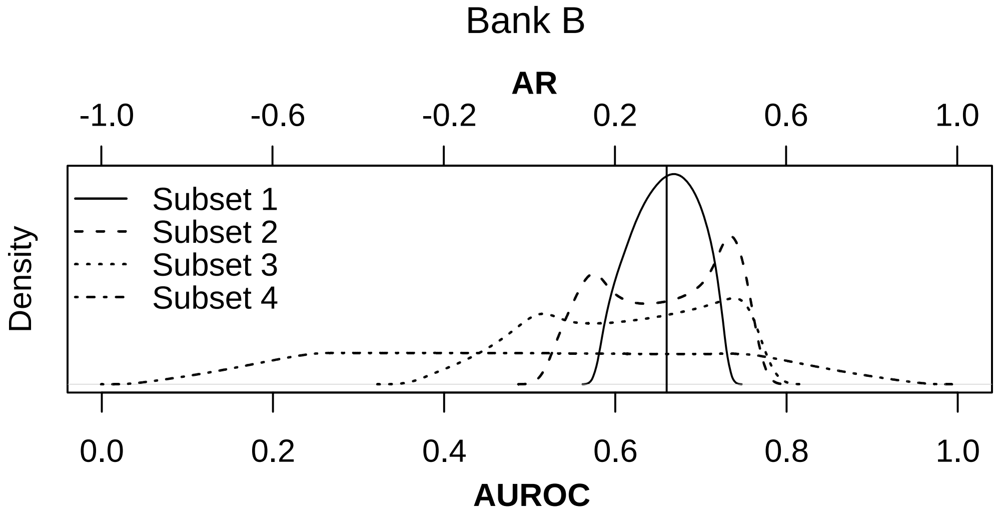
Without Correlation





With Correlation





Analysis of Figure 7 indicates that the default patterns with the greatest likelihood will yield AR and AUROC measures close to the expected value whereas unlikely default patterns can result in extreme values for the two measures. However, this does not necessarily imply that an unlikely default pattern might not also result in an AR or AUROC value close to its expected value as we infer from looking at the chain dotted line in Figure 7. Conversely, the observation of a value differing strongly from the expected value appears sufficient to conclude that the default pattern yielding this value is very unlikely.

Figure 7 thus reveals that discriminatory power and calibration quality are indeed closely related. In fact, the proposed measures of discriminatory power obviously incorporate information about a rating system's calibration quality. The measures may thus also be used to jointly assess the accuracy of all probability of default estimates. Given a specific portfolio and a certain structure of the rating scale, if the realized values for AR or AUROC on that portfolio differ strongly from their expected values, the rating system is conclusively miscalibrated. This arises from the fact that the default patterns yielding extreme values for AR and AUROC would be very unlikely if the probabilities of default assigned to the obligors contained in the portfolio were correct. A basic test procedure of the Null Hypothesis that the probabilities of default assigned to obligors are the true probabilities of default against the alternative that the rating system is miscalibrated would consist of:

- the construction of confidence intervals for the portfolio dependent values of AR and AUROC,
- the computation of the realized value of AR and AUROC,
- and the comparison of this realized value against the confidence intervals at the desired level of significance.

An evaluation procedure of that kind belongs to the family of tests that allow for the joint assessment of the calibration quality of all rating grades such as the Hosmer-Lemeshow Test (Hosmer and Lemeshow, 1980 and Hosmer et al., 1988), the Spiegelhalter Test (1986) or so-called "Traffic Light"-approaches (e.g. Tasche, 2003 and Coppens et al., 2007 for a general overview). Mind, however, that the suggested test based on measures of discriminatory power is not as robust and powerful as the above mentioned alternatives. This follows from the finding that a value for AR and AUROC close to its respective expected value is not sufficient to conclude that the rating system is calibrated well. (Unlikely default patterns may also result in values close to the expected values as demonstrated in Figure 7.) Still it appears important for a comprehensive understanding of the proposed measures of discriminatory power to note that the measures can provide concrete indications about a rating system's calibration quality. The observation of an extreme value for AR or AUROC differing strongly from the respective expected value constitutes a definite sign for erroneous calibration.

4. Empirical Example

4.1 Data Description

To illustrate our findings, as derived in the previous sections, we empirically investigate the discriminatory power using data of two Austrian banks. Our dataset stems from the monthly reporting of the two banks to the Credit Register hosted by the Austrian National Bank. This database contains information on all major loans (in excess of EUR 350,000) granted by credit and financial institutions as well as contract insurance companies in Austria. According to the Austrian Banking Act, all financial institutions have to supply monthly information to the credit register on any changes in the status of their outstanding credits as well as information on the new loans granted during the period. Of importance for our empirical study is the fact that the reports also contain data about the banks' own assessments of obligor's creditworthiness. This assessment is expressed in a form of a rating and includes information on the default state. In addition, banks have to provide the central bank with detailed information on the set-up of their rating scales concerning granularity (i.e., the number of rating classes) and the level of risk associated with each class in terms of a 1-year probability of default. Recording of this kind of information, in the Austrian Credit Register, has started at the beginning of 2003.

Our sample comprises time series data on the major loan portfolio of two banks over a period lasting from August 2004 to December 2006. It thus consists of a total number of 29 monthly reports for both banks. The information available for bank 1 consists of a total number of 36,040 ratings on a total of 3,133 obligors. For bank 2 the total number of ratings is 52,041, while the corresponding number of obligors is 7,900. Regarding granularity, bank 1 uses a rating scale consisting of 9 rating classes for non-defaulted obligors whereas bank 2 employs a scale comprising 19 different buckets for its non-defaulted obligors.

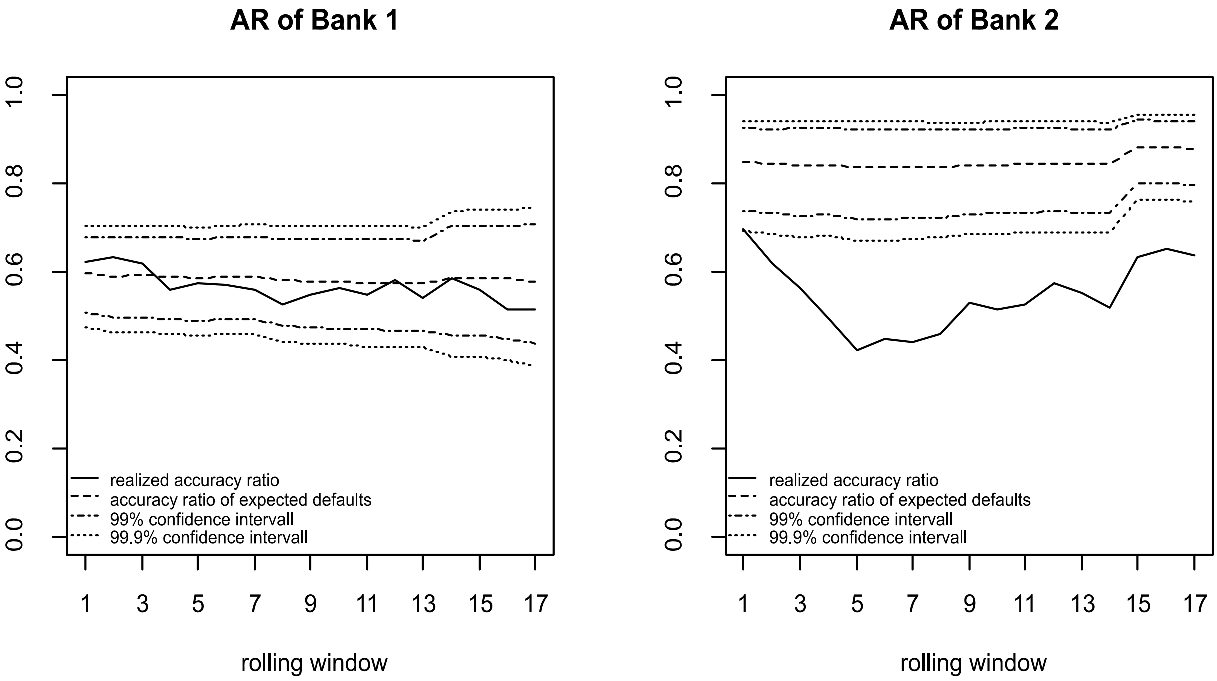
4.2 Results

In our empirical analysis we assess the discriminatory power of the rating systems of the two sample banks over a one-year horizon. This horizon is consistent with the aforementioned reporting criteria. To this end, we split up the data into a sequence of 17 roll-over time windows. The length of each window is one year and the increment between the consecutive annual time windows is one month. Thus the first window starts in August 2004 and ends in August 2005, the second window starts in September 2004 and ends in September 2005. The last window comprehends the period between December 2005 and December 2006.

For each time-window we define a static pool consisting of those obligors who possessed a rating at the beginning of the time period and were still contained in the portfolio at its end. Each obligor within in the static pool is a member of a rating class that denotes his 1-year probability of default for the upcoming year. Based on this information, we can construct the ex-ante distributions of the AR (For the sake of brevity, we skip an analysis of the AUROC measure in our empirical examples. This will not have any impact on our conclusions, since the two measures are unambiguously related to each other as shown in section 2.1.) In the next step, we compute the realized default frequency for each rating class. For this purpose we count the number of defaults observed within the respective year, using data from the 12 consecutive monthly reports to the Credit Register, and divide it by the size of the static pool. This enables us to compute the realized values of AR.

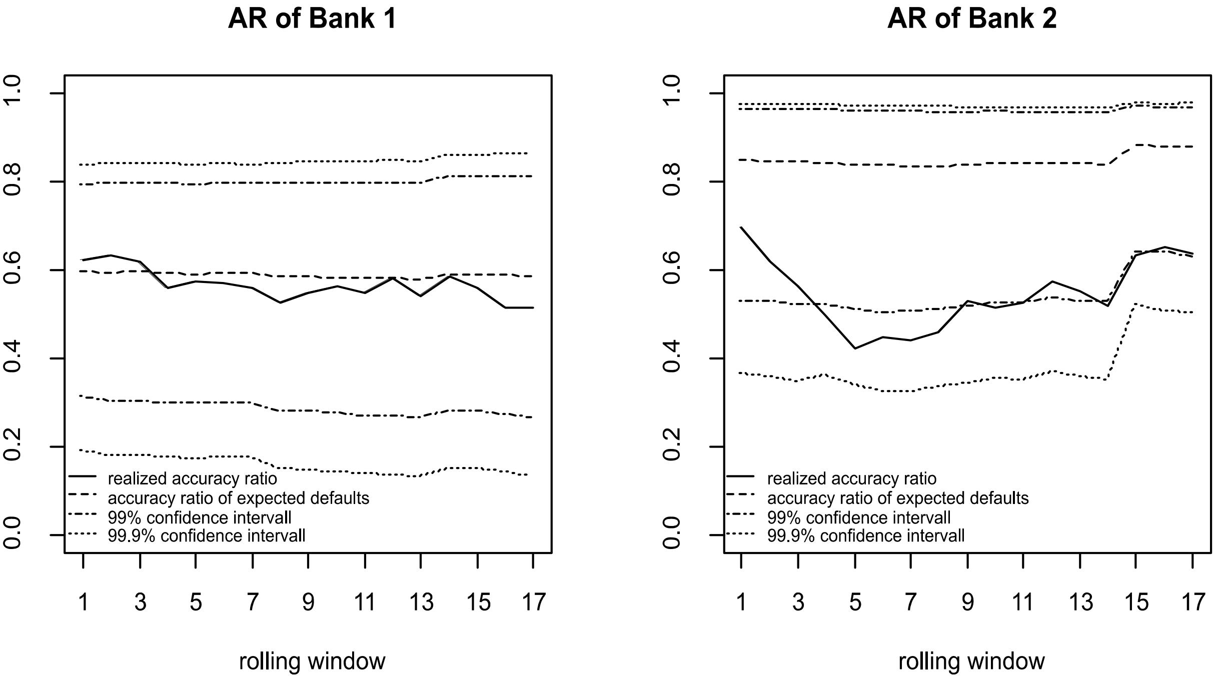
Figure 8 shows the realized AR as well as the 99% and 99.9% confidence intervals for both banks.

Figure 8: AR analysis of Bank 1 and Bank 2 assuming independence



Considering only the realized ARs one might assume that the rating systems of bank 1 and 2 are of similar quality since the realized ARs are similar in size. However, in contrast to bank 1 the realized ARs of bank 2 are never within the confidence intervals. This may indicate that the rating system of bank 2 is unsatisfactorily calibrated. On both, a 99% and a 99.9% confidence level we can reject the hypothesis that the ex-ante probabilities of default reported by bank 2 are the true probabilities of default. However, as shown in section 3.4, correlation increases the variance of the AR-distribution thus widening the confidence intervals. The realized ARs of bank 2 might therefore very well be within the confidence intervals, if correlation was present. Since we lack data on default correlation, we shall – analogously to our previous examples – assume that the intra-class correlation is equivalent to correlation as defined by the Basel II framework. By repeating our analysis, this time accounting for correlation, one obtains the confidence intervals shown in Figure 9.

Figure 9: AR analysis of Bank 1 and Bank 2 accounting for correlation



The effect of intraclass correlation is quite substantial for bank 1. Especially the lower bound of the interval decreases substantially. As expected, the realized ARs of bank 1 remain within the confidence intervals. Regarding bank 2, the effect of correlation also widens the confidence intervals, affecting primarily the lower boundary. The realized ARs of bank 2 are now within the 99.9% confidence interval but still outside or close to the 99% confidence interval. This suggests that the realized ARs of bank 2 are highly improbable even when considering intraclass correlation. We hence conclude that bank 2 may have to recalibrate its rating systems.

5. Conclusion

One of the most valued features of a good rating system is classically its high discriminatory power. It refers to the systems ability to distinguish, ex-ante, between subsequently defaulting and non-defaulting obligors. With the default event being random, however, most rating systems aim at estimating probabilities of default. It follows that there exists a true probability of default for any time horizon at any point in time for each and every obligor. Against this background, we have reconsidered common measures of discriminatory power in a probabilistic framework. We show that these measures depend on the distribution of the rating scores and the probabilities of default associated with the distinct ratings. As an important first result we are thus able to verify the proposition of Hamerle et al., 2003, that the measures are portfolio dependent. As a second result we can show that the measures are stochastic themselves. Depending on the distribution of the true ex-ante probabilities of default, the quality of each and every portfolio thus determines the shape of its distribution for both of the proposed measures, AR and AUROC.

In a validation context, the AR and AUROC measures, however, are calculated using a posteriori probabilities of default, i.e. realized default frequencies. Hence, the proposed measures of discriminatory power may be employed to test the calibration quality of a rating system. To this end we develop a formal testing procedure. Like any other test of calibration quality our procedure has two prerequisites:

- information on the ex-ante probabilities of default assigned to the different rating grades,
- and the realized default frequencies for each and every rating grade

Our test involves the comparison of the realized value of AR or AUROC to the (analytically derivable) desired level of confidence computed. This level is computed under the Null Hypothesis that the reported ex-ante probabilities of default are the true probabilities of default. Hence the test procedure is based on the idea that a significant deviation, of the realized value of AR or AUROC, from its expected value, indicates that the ex-ante probabilities of default cannot be the true ones and are thus miscalibrated. The proposed testing procedure therefore also makes the portfolio dependent AR and AUROC measures comparable across different portfolios. We discuss the power of the suggested proposed procedure when compared to alternative standard test procedures that allow for a joint testing of the calibration quality of different rating grades. Furthermore, we analyse the impact of default correlation on the distribution of the AR and AUROC measures. Finally, we empirically investigate our procedure's merits and pitfalls.

Our findings result in the following consequences: From the perspective of accuracy, it is essential to maximize granularity and optimise calibration quality. It thus appears wise from a supervisor's perspective to demand both, a certain minimum granularity as envisaged in the new Basel II framework (see paragraph 404 in Basel Committee on Banking Supervision, 2005a,) and a high calibration quality. That said, we demonstrate in this paper that maximum granularity and maximum calibration quality, relating to a situation in which each and every obligor is assigned its own true ex-ante probability of default, ensure automatically that a rating system will be as discriminative as it can be in terms of the proposed measures of discriminatory power. In a validation context, the proposed measures of discriminatory power are thus only useful in so far as they convey information on a rating system's calibration quality. It is hence not valid to treat discriminatory power as a concept different from calibration quality. Furthermore, a separate and explicit testing of discriminatory power for its own sake is redundant in a validation context. These crucial insights should be accounted for by rating agencies, banks and their supervisor in their validation tasks.

APPENDIX 1:

Along the lines of Engelmann et al. (2003) we show the relation of AR and AUROC. Starting off with the definition of the AR as given (7), we obtain the relation to the AUROC given in formula (9) of the paper as follows:

$$
\begin{aligned}
AR_{RS} &= \frac{\int_0^1 F_d(C)dF(C) - 0.5}{0.5 \cdot P(Y=0)} \\
&= \frac{P(Y=1)\cdot \int_0^1 F_d(C)dF_d(C) + P(Y=0)\cdot \int_0^1 F_d(C)dF_s(C) - 0.5}{0.5 \cdot P(Y=0)} \\
&= \frac{P(Y=1)\cdot 0.5 + P(Y=0)\cdot \int_0^1 F_d(C)dF_s(C) - 0.5}{0.5 \cdot P(Y=0)} \\
&= 2\cdot\left(\int_0^1 F_d(C)dF_s(C) - 0.5\right) \\
&= 2\cdot AUROC_{RS} - 1
\end{aligned}
$$

APPENDIX 2:

To prove that AR and AUROC are portfolio dependent as first discovered by Hamerle et al. (2003) and furthermore stochastic as noted by Blochwitz et al. (2005) we show drawing on the above mentioned works that both measures depend on the distribution of the rating scores and the probabilities of default associated with the rating scores.

In equation (10) the AR was defined as:

$$AR_{RS} = 2 \cdot \left(\int_{-\infty}^{+\infty} F_d(C) \cdot f_s(C) dC - 0.5 \right)$$

Using the definition of the hit rate $HR(C) = F_d(C) = P(R \leq C \mid Y = 1)$ as given in (4) and the definition of the false alarm rate $FAR(C) = F_s(C) = P(R \leq C \mid Y = 0)$ as given in (5) we may transform the equation above into:

$$\begin{aligned} AR_{RS} &= 2 \cdot \left(\int_{-\infty}^{+\infty} F_d(C) \cdot f_s(C) dC - 0.5 \right) \\ &= \int_{-\infty}^{+\infty} F_d(C) \cdot f_s(C) dC + \int_{-\infty}^{+\infty} F_d(C) \cdot f_s(C) dC - 1 \\ &= \int_{-\infty}^{+\infty} F_d(C) \cdot f_s(C) dC - \int_{-\infty}^{+\infty} (1 - F_d(C)) \cdot f_s(C) dC = \\ &= \int_{-\infty}^{+\infty} F(C|Y=1) \cdot f(C|Y=0) dC - \int_{-\infty}^{+\infty} [1 - F(C|Y=1)] \cdot f(C|Y=0) dC \end{aligned}$$

Based on the laws for conditional probability and the Bayes Theorem we may then transform this equation into:

$$\begin{aligned}
AR_{RS} &= \int_{-\infty}^{+\infty} F\left(C \middle| Y=1\right)\cdot f\left(C \middle| Y=0\right) dC - \int_{-\infty}^{+\infty} \left[1 - F\left(C \middle| Y=1\right)\right]\cdot f\left(C \middle| Y=0\right) dC \\
&= \int_{-\infty}^{+\infty} \frac{F\left(C, Y=1\right)}{P\left(Y=1\right)}\cdot \frac{f\left(C, Y=0\right)}{P(Y=0)} dC - \int_{-\infty}^{+\infty} \left(\frac{P\left(Y=1\right)}{P\left(Y=1\right)} - \frac{F\left(C, Y=1\right)}{P\left(Y=1\right)}\right)\cdot \frac{f\left(C, Y=0\right)}{P(Y=0)} dC \\
&= \frac{1}{P\left(Y=1\right)\cdot P(Y=0)}\cdot \left\langle \begin{aligned} &\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{C_1} f\left(C_2, Y=1\right) dC_2\right)\cdot f\left(C_1, Y=0\right) dC_1 - \\ &-\left[P\left(Y=1\right)\cdot \int_{-\infty}^{+\infty} f\left(C, Y=0\right) dC - \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{C_1} f\left(C_2, Y=1\right) dC_2\right)\cdot f\left(C_1, Y=0\right) dC_1\right] \end{aligned} \right\rangle \\
&= \frac{1}{P\left(Y=1\right)\cdot P(Y=0)}\cdot \left\langle 2\cdot\left[\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{C_1} f\left(C_2, Y=1\right) dC_2\right)\cdot f\left(C_1, Y=0\right) dC_1\right] - P\left(Y=1\right)\cdot P\left(Y=0\right)\right\rangle \\
&= 2\cdot\left\langle \frac{1}{P\left(Y=1\right)\cdot P(Y=0)}\cdot\left[\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{C_1} f\left(C_2, Y=1\right) dC_2\right)\cdot f\left(C_1, Y=0\right) dC_1\right]\right\rangle - 1 \\
&= 2\cdot\left\langle \frac{1}{P\left(Y=1\right)\cdot\left[1-P\left(Y=1\right)\right]}\cdot\left[\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{C_1} f\left(Y=1 \middle| C_2\right)\cdot f\left(C_2\right) dC_2\right)\cdot f\left(Y=0 \middle| C_1\right)\cdot f\left(C_1\right) dC_1\right]\right\rangle - 1
\end{aligned}$$

The last line corresponds to formula (10) of the paper and reveals the dependence of the one-dimensional summary measures of a rating system's discriminatory power on the probabilities of default and the distribution of the rating scores.

References

E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankcurptcy. *Journal of Finance*. 23:589-609, 1968.

E. I. Altman. The success of business failure prediction models. An international survey. *Journal of Banking and Finance*. 8:171-198, 1984.

A. Agresti. *Categorical data analysis*. 2. ed. New York: Wiley-interscience, 2002.

D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*. 12:387-415, 1975.

Basel Committee on Banking Supervision. Credit Ratings and Complementary Sources of Quality Information, 2000.

Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: A revised framework, 2005a.

Basel Committee on Banking Supervision. Studies on the validation of internal rating systems (revised), 2005b.

Basel Committee on Banking Supervision. Sound credit risk assessment and valuation for loans, 2006.

S. Blochwitz, A. Hamerle, S. Hohl, R. Rauhmeier and D. Rösch. Myth and Reality of Discriminatory Power for Rating Systems. *Wilmott Magazine:* 2-6, 2005.

A. Blöchinger and M. Leippold. Economic benefit of powerful credit scoring. *Journal of Banking and Finance*. 30:851-872, 2006.

C. Bluhm, L. Overbeck and C. Wagner. *An Introduction to Credit Risk Modelling*. Florida: Chapman&Hall, 2003.

F. Coppens, F. Gonzalez and G. Winkler. The performance of credit rating systems in the assessment of collateral used in Eurosystem monetary policy operations. *ECB Occasional Paper 65*. Frankfurt: European Central Bank, 2007.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 44:837-845, 1988.

A. I. Dimitras, S. H. Zanakis and C. Zopounidis. A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*. 90:487-513, 1996.

B. Engelmann, E. Hayden, and D. Tasche. Testing rating accuracy. *Risk*. 16:82-86, 2003.

T. Fawcett. ROC Graphs: Notes and practical considerations for data mining researchers. *Technical Report HPL-2003-4*, HP Labs, 2003.

A. Hamerle, R. Rauhmeier, and D. Rösch. Uses and misuses of measures for credit rating accuracy, Working Paper. Universität Regensburg, 2003.

D. Hand. *Construction of Assessment of Classification Rules*. New York: Wiley Series in Probability and Statistics, 1997.

D. Hosmer and S. Lemeshow. A goodness of test for the multiple logistic regression. *Communication in Statistics*. A10:1043-1069, 1980.

D. Hosmer, S. Lemeshow and J. Klar. Goodness of fit testing for multiple logistics regression analysis when the estimated probabilities are small. *Biometrical Journal*. 30: 991-924, 1988.

J. P. Krahnen and M. Weber. Generally accepted rating principles: A primer. *Journal of Banking and Finance*. 25:3-23, 2001.

R.C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*. 29:449–470, 1974

J. P. Morgan. *CreditMetrics™ – Technical Document*. New York, 1997.

P. J. Schoenbucher. Factor Models for Portfolio Credit Risk, Working Paper. Bonn University, 2000.

J. R. Sobehart, S. C. Keenan, and R. M. Stein. Benchmarking quantitative default risk models: a validation methodology. Moody's Investors Service. Global Credit Research, 2000.

J. R. Sobehart and S. C. Keenan. Measuring default accurately. *Risk*. 14:31-33, 2001.

D. Spiegelhalter. Probabilistic prediction in patient management and clinical trails. *Statistics in Medicine*. 5:421-433, 1986.

R. M. Stein. Benchmarking default prediction models: pitfalls and remedies in model validation. *Journal of Risk Model Validation.* 1:77-113, 2007.

R. M. Stein and F. Jordao. What is a more powerful model worth? Working Paper. Moody's KMV, 2003.

D. Tasche. A traffic lights approach to PD validation, Working Paper. Deutsche Bundesbank, Frankfurt, 2003.