A Latent Variable Approach to Validate Credit Rating Systems

Kurt Hornik[a], Rainer Jankowitsch[b], Christoph Leitner[a], Manuel Lingo[bc],
Stefan Pichler[b], Gerhard Winkler[bc]

September 2008

[a] Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Vienna, Austria

[b] Department of Finance and Accounting, Wirtschaftsuniversität Wien, Heiligenstädter Strasse 46-48, A-1090 Vienna, Austria

[c] Oesterreichische Nationalbank, Otto Wagner Platz 3, A-1090 Vienna, Austria

Abstract

We suggest a new parametric framework to assess the accuracy of estimated default probabilities (PDs). Whereas the traditional methods to validate credit rating systems focus primarily on the discriminatory power, recent advances in credit risk management and banking regulation called forth the need for accurate estimates of model parameters. Thus, the focus of rating validation has shifted to the accuracy of PD estimates. In this paper we introduce bank and obligor specific error terms which are additive to a suitably transformed latent variable, which in our approach has the interpretation of the "true", unobservable PD of the underlying obligor. Provided with rating information from different sources (e.g., banks, rating tools, or rating agencies) and an appropriate model for the distribution of the latent PDs the parameters of the errors distribution can be obtained. In a specific application we assume the error terms to be jointly normal and estimate this model for a data set from the Austrian Credit Register where we observe PD estimates for 2090 obligors provided by 13 banks.

Keywords: Rating validation, latent trait, latent variable, probability of default.

JEL classification: G21, C33.

1 Introduction

The validation of credit rating systems has become an important field of research in the last years (see, e.g., Krahnen and Weber, 2001; Crouhy et al., 2001, for a discussion of the latest developments in credit rating systems). Recent advances in credit risk management and banking regulation called forth the need for accurate estimates of model parameters. According to the new regulatory Basel II framework which is currently also implemented in many national legislations, banks have several incentives to make use of internal rating systems to estimate risk parameters which are the essential input to calculate their regulatory capital requirements (Bank for International Settlements, 2004). One of the most important risk parameter is the probability of default (PD) which is defined to measure the likelihood of the occurrence of a default event for a certain obligor over a one year horizon. Modern credit risk management is crucially based on the risk-adjusted pricing of loans and other credit-risk contingent claims which again heavily relies on a valid and accurate PD estimation methodology. The accuracy of PD estimates is of particular importance for virtually all pricing models for structured credit derivatives. While some pricing models need accurate measures of the average PD of a bond or loan portfolio as an input, some more advanced models require the distribution of individual PDs of such a portfolio which imposes even higher challenges on the validity of the estimation models.

In this paper we introduce a new framework to assess the accuracy of PD estimates. In contrast to *backtesting* methods where ex ante PD estimates are compared to ex post realizations of actual defaults our model is based on the existence of contemporaneous PD estimates for the same obligor provided by different rating sources. Such a *benchmarking* approach is particularly helpful when sufficient default observations are not available or competing rating systems are to be compared (for a discussion of benchmarking approaches see Bank for International Settlements, 2005; Hornik et al., 2007).

Traditional methods to validate credit rating systems focus primarily on the discriminatory power, i.e., the ability to ex ante distinguish between defaulting and non-defaulting obligors. Among the best-known methods of this type are analyses based on the Accuracy Ratio and the Receiver Operating Characteristic (e.g., Bank for International Settlements, 2005). Unfortunately, these methods do not provide any conclusive information about the accuracy of the PD estimates. Consider a hypothetical rating system which systematically underestimates the true PDs by one half. Obviously, such a rating system, though remarkably inaccurate, has maximal theoretical discriminatory power. On the other hand, in a population of obligors with identical PDs even a perfectly accurate PD estimation will show zero discriminatory power. Recent studies also deal with the shortcomings of the use of these concepts for rating validation (e.g., Lingo and Winkler, 2008). As pointed out by the regulatory bodies (Bank for International Settlements, 2005), validation methods have to aim at directly assessing the *calibration quality*, i.e., the accuracy and reliability of PD estimates (e.g., Stein, 2002).

In our model the "true" PD is taken as a latent variable. Rating outcomes from different sources (e.g., banks, rating tools, or rating agencies) are treated as noisy observations of this latent variable. A parametric specification of such a model has to include the following components: (i) The distribution of the latent PDs, (ii) the formal relation how the noise or error terms are linked to the latent PDs, and (iii) the distribution of the error terms. In general, provided with sufficiently many observations of PD estimates for a set of obligors stemming from different sources such a model can be estimated via maximum likelihood. The key assumption of our model is that the error terms are related additively to the latent PDs transformed from the unit interval to the real axis via a suitable *link function* (e.g., the probit function which is discussed in this paper and used in the empirical example). The means and (co)variances of these distributions are the key outcome of the model.

The mean parameters indicate the rating *bias*, i.e., the expected shift of PD estimates of a certain rating system compared to the average PD across all rating sources. The variance parameters reflect the general size of undirected estimation errors, i.e., they reflect the *precision* of the rating system. Finally, the covariances convey information about potential error dependencies across rating systems.

The proposed model can also be used to compute *consensus* PDs for all obligors, which particularly in the case of only partially available rating information is non trivial (e.g., Cook et al., 1986, 2007). To the best of our knowledge there is no viable methodology available to obtain consensus ratings for these kind of data sets. The estimation of consensus PDs is thus one of the major contributions of this paper. Deviations between observed PD estimates and consensus PDs can be analyzed on an atomistic case by case basis. Further, these differences can be aggregated on rating source or obligor levels, which can aid in detecting potential systematic patterns or anomalies in rating behavior.

Our framework is applicable in the context of banking supervision and development of rating models. Banking supervisors are interested in the parameters of the error terms to assess the calibration quality of the internal rating system of a supervised bank. Supervisors might also be interested in the consensus PD estimates for analyzing financial stability of a banking system (see Elsinger et al., 2006). Finally, developers of rating systems such as banks and rating agencies have a natural interest in comparing their outcomes to peers at different stages of the development. Note, however, that as any benchmarking method our model does only provide information about the relative rather than the absolute rating errors since actual default information is not incorporated.

We estimate the proposed model for a multi-rater panel provided by the Austrian central bank where we observe PD estimates for 2090 obligors provided by 13 banks. We employ standard maximum likelihood techniques to estimate the parameters of the distributions of the errors and latent PDs. Our empirical example shows that our framework is suitable to identify bank specific regularities with respect to grouping variables, like industry affiliation, legal form and exposure size, and to conduct an outlier analysis of the estimated rating errors of individual banks.

The model presented in this paper is related to other studies on benchmarking credit rating systems (e.g., Hornik et al., 2007; Stein, 2002; Carey, 2001). In contrast to these contributions, our model explicitly proposes a probabilistic framework which reflects the stochastic nature of the true PDs and the rating errors incurred in the process of PD estimation. This allows to directly estimate parameters reflecting the calibration quality of rating systems and derive consensus PD estimates consistent with the suggested framework.

This paper is organized as follows. Section 2 motivates and formally describes the proposed latent variable model for PD estimation. Section 3 contains a description of the data set and the estimation methodology, and a presentation of the empirical results. Section 4 summarizes the results and indicates possible directions of future research.

2 A Latent Variable Model for PD Estimation

The basic assumption of our model is that raters cannot observe the "true" PD of the obligor. This assumption is justified by the general informational asymmetry between firm owners and debt holders which constitutes the cornerstone of modern corporate finance (e.g., Leland and Pyle, 1977; Berk and DeMarzo, 2007). This asymmetry can be due to limited access to the existing information, such as incomplete accounting information (Duffie and Lando, 2001), or delayed observations of the driving risk factors (Guo et al., 2008). Generally, the modeler is entirely free in the specification of the functional form how the estimation error is related to the latent PD.

The motivation for the specification used in this paper builds on the main concept of *structural* or *firm value* models (e.g., Merton, 1974; Lando, 2004). The standard models assume the firm value to be the only driving factor of credit risk and to follow a Geometric Brownian Motion. As a consequence, an important stylized property of these models is that the probit of the PD is linear in the natural log of the firm value. To fix notations, let V_i be the log asset value of firm i and PD_i its probability of default. The basic model of Merton (1974) can be written as

$$PD_i = \Phi(-DD_i), \qquad DD_i = a_i + b_i V_i,$$

where Φ is the distribution function of the standard normal distribution, a_i and b_i are constants independent of V_i, and DD_i is the so-called *distance to default* of firm i (Crosbie and Bohn, 2003; Bharath and Shumway, 2008). Along the lines of Duffie and Lando (2001), we assume that the error in the observation of the firm value is normal and additive to the log of the firm value. This assumption can also be justified by the wide-spread use of structural models for PD estimation in the banking industry which have gained additional importance by the introduction of the Basel II supervisory framework. The most prominent industry model was developed by Moody's KMV (Crosbie and Bohn, 2003) and is used in several extensions and modifications by many financial institutions. In this class of models the distance to default is derived from stock market and accounting data and used as the key input to the PD estimation. It thus seems natural to assume that erroneous observations and incomplete information lead to normally distributed errors which are additive to the distance to default.

The "estimate" PD_{ij} as derived by bank j for the true PD of firm i is thus of the form

$$PD_{ij} = \Phi(-DD_i + \epsilon_{ij})$$

where ϵ_{ij} is the corresponding error (which depends on j in particular as raters might have access to different information sets for firm i). Equivalently,

$$\Phi^{-1}(PD_{ij}) = -DD_i + \epsilon_{ij} = \Phi^{-1}(PD_i) + \epsilon_{ij}$$

A second important class of industry models estimate PDs based on a probit (or logit) regression of observed default indicators on a set of risk characteristics of the firm (e.g., Blume et al., 1998; Nickell et al., 2000). A linear combination of the risk characteristics of the firm where the resulting regression coefficients are used as weights constitutes the *rating score* of the firm. The PD estimate is then obtained by transforming the score to the unit interval by the corresponding transformation. In the case of the probit transformation this approach is fully consistent with our framework, because the error term in the regression is normal and additive to the score.

Consolidating the Merton type and the regression approaches, we are led to considering a general class of models relating the raters' estimated (observed) PDs to the (unobservable) true PDs in the form

$$\ell(PD_{ij}) = \ell(PD_i) + \epsilon_{ij}.$$

for a suitable (strictly monotonically increasing) *link function* ℓ mapping the $(0, 1)$ PD scale to $(-\infty, +\infty)$. Via ℓ, the PDs are mapped to corresponding scores. On the score scale, the rating errors are modeled additively. Let $S_{ij} = \ell(PD_{ij})$ and $S_i = \ell(PD_i)$ denote the observed and latent scores, respectively. For Merton type models, $\ell = \Phi^{-1}$ and $S_i = -DD_i$. Writing μ_{ij} and σ_{ij} denote the mean and standard deviation of ϵ_{ij}, respectively, the above latent trait model can be written as

$$S_{ij} = S_i + \mu_{ij} + \sigma_{ij} Z_{ij} \tag{1}$$

where the standardized rating errors $Z_{ij} = (\epsilon_{ij} - \mu_{ij})/\sigma_{ij}$ have mean zero and unit variance. We prefer to think of PD_i and hence the corresponding scores S_i as drawn randomly from an underlying obligor population, and assume that rating errors are independent of the true PDs and that true PDs and rating errors are i.i.d. across obligors. (Of course, these assumptions could be relaxed if more involved probabilistic specifications are desired.) We thus obtain a *mixed-effects model* (e.g., Pinheiro and Bates, 2000) for the observed S_{ij} where the latent true PD scores enter as random effects. We refer to this model as the *latent trait model* for the multi-rater panel of PD estimates.

Given parametric models for the μ_{ij} and σ_{ij} and the distributions of true PD scores and standardized rating errors, the latent trait model can be estimated using e.g. marginal maximum likelihood (provided that the marginal distributions of the S_{ij}, i.e., the convolutions of the true PD score and rating error distributions, can be computed well enough), or Bayesian techniques.

A very simple parametric specification of the bias/variance structure of the rating errors is $\mu_{ij} = \mu_j$ and $\sigma_{ij} = \sigma_j$, in which case the rating errors would be independent of the obligors and their characteristics (in particular, their creditworthiness itself). We suggest to employ flexible models of the form

$$\mu_{ij} = \mu_{g(i),j}, \qquad \sigma_{ij} = \sigma_{g(i),j},$$

where $g(i) \in \{1, \dots, G\}$ is the group of obligor i, for a suitable grouping of obligors relative to which raters exhibit homogeneous rating error characteristics. Note that we use μ for the means of the rating errors and the respective model parameters, with $\mu_{ij} = \mathbb{E}(\epsilon_{ij})$ and $\mu_{g,j}$ the parameter for group g and rater j. The σ notation is analogous. Such groups can e.g. be defined by industry, obligor "type", size, and legal form, or combinations thereof. The importance of accounting for industry group effects in the analysis of creditworthiness patterns has e.g. been emphasized in Crouhy et al. (2001). As

$$\mathbb{E}(S_{ij}) = \mathbb{E}(S_i) + \mu_{g(i),j},$$

conditions relating the rating biases to the mean observed PD scores are required to ensure identifiability. More generally, note that common random effects in the rating errors cannot be separated from the true PD scores. A natural condition consistent with the interpretation as rating *bias* relative to an underlying unbiased truth is $\sum_j \mu_{g,j} = 0$, i.e., that the raters' average rating bias within each obligor group is zero. With this identifiability constraint, the marginal group effects in the means are absorbed into the means of the latent scores, for which we shall employ the basic model $\mathbb{E}(S_i) = \nu_{g(i)}$ consistent with the identifiability constraint. Possible models for the variances of the PD scores include $\mathrm{var}(S_i) = \tau^2$ (constant for all obligors) or $\mathrm{var}(S_i) = \tau^2_{g(i)}$ (constant for all obligors in the same group).

With these specifications, the consensus score for obligor i is given by $\hat{S}_i$, the estimated random effect in the fitted mixed-effects model. Consensus PD estimates are readily obtained by transforming the consensus scores back to the PD scale using the inverse ℓ^{-1} of the link function, i.e., as $\widehat{PD}_i = \ell^{-1}(\hat{S}_i)$. Finally, *residuals* are the part of the observations unexplained by the model and given by $S_{ij} - \hat{S}_i - \hat{\mu}_{ij}$, the observed scores minus the estimated random and fixed effects.

In what follows, we assume that true PD scores and rating errors have a (multivariate) normal distribution; possible extensions are discussed in Section 4. With these assumptions, the latent trait model can be estimated using standard software for mixed-effects models, provided that these allow for sufficiently flexible specifications of the error covariance structure.

3 Empirical Analysis

3.1 Data Description

For the empirical example we employ a data set on rating information provided by Oesterreichische Nationalbank, the Austrian central bank. The data contain rating information (one-year PD estimates) from 13 major Austrian banks on 2090 obligors in September 2007 and cover a significant share of the Austrian credit market. For each obligor, at least two PD estimates are available. The number of *co-ratings* (occurrences of ratings of a single obligor by two different banks) is 5460. In addition to the PD estimates we have cross-sectional information about the obligors, like legal form, industry affiliation and outstanding exposure. Table 1 reports descriptive statistics of the data set.

Note that even the smallest bank has at least 70 obligors in common with one or more of the other institutions. Apart from looking at the number of co-ratings on a bank level, we also compute the number of co-ratings on an obligor level. The median number of these co-ratings is 2, suggesting that most obligors have business relations to only a small number of banks.

For a deeper analysis we group all obligors by their industry affiliation and their legal form. Based on the NACE codes (European Commission, 2008) we classify obligors to nine main industries. Table 2 shows the distribution of the obligors across the industries.

	Min.	Median	Max.	Mean
Number of obligors per bank	70	182	1700	420
Number of ratings per obligor	2	2	11	2.5
Size of banks measured by their total assets in Euro billions	1.0	8.7	128.5	11.8

Table 1: Descriptive statistics of the characteristics of the rating information and the 13 Austrian banks in the data set.

Label	Industry	No. of co-ratings	No. of co-ratings in %
Manufac	Manufacturing	938	17.2
Energy	Energy & Environment	180	3.3
Constr	Construction	184	3.4
Trading	Trading	641	11.7
Finance	Financial Intermediation	1737	31.8
RealEst	Real Estate & Renting	754	13.8
Public	Public Sector	344	6.3
Service	Service	435	8.0
Private	Private Individuals	247	4.5
Total		5460	100.0

Table 2: Distribution of the co-ratings of the 13 Austrian banks across industries.

Table 2 shows that the total numbers of co-ratings (5460) is not uniformly distributed across the nine industries, ranging from 180 co-ratings in Energy & Environment ("Energy") to 1737 co-ratings in Financial Intermediation ("Finance"). With 13 banks and 9 industries there are 117 possible sub-portfolios to be analyzed. However in 17 of these there are no observations.

In addition the obligors can be grouped with aspect to their legal form yielding that 79.6% of the obligors are limited companies, 12.2% unlimited companies, and 8.2% are private individuals.

Finally, we use information on the banks' relative exposures against each obligor. Relative exposure is measured as the outstanding amount against the obligor expressed as a fraction of the total volume of outstanding loans of a specific bank and serves as a rough indicator for the size of the obligor.

3.2 Results

This section describes the model selection process and the empirical results. The general model class allows for many competing model specifications based on the data structure (see Section 2). Our selection process yields a model using the industry as grouping variable. We present bank and industry specific analyses based on rating biases and error variances derived from this specification. Furthermore, we analyze the errors from the consensus rating of this model on the obligor level based on exposure and legal form.

3.2.1 Model Selection and Parameter Estimates

Equation 1 in Section 2 describes a very general model class allowing for a variety of specifications for the means and variances of the normal distributions of the latent PD scores and rating errors. We use parametric models which group obligors based on the available co-variates (for the data set at hand, industry affiliation, legal form and exposure). For the error distributions, bank effects as well as group/bank interaction terms are considered. We also investigate models allowing for general correlation patterns between rating errors. For each fitted model, we compute the Akaike Information Criterion (AIC) and

Bayesian Information Criterion (BIC, also known as Schwarz's Bayesian criterion). The best model is then selected based on these criteria.

For all calculations we use the R system (version 2.6.1) for statistical computing (R Development Core Team, 2007). The parameters of the mixed-effects models are estimated via maximum likelihood (e.g., Pinheiro and Bates, 2000) using the R package nlme (Pinheiro et al., 2007). The best model found by the model selection procedure uses industry affiliation as the sole grouping variable and a single variance parameter PD score, and is given by

$$S_{ij} = S_i + \mu_{g(i),j} + \sigma_{g(i),j} Z_{ij}, \qquad S_i \sim N(\nu_{g(i)}, \tau^2), \tag{2}$$

where $\mu_{g,j}$ is the rating bias to the mean PD score of bank j for obligors in industry g, $\sigma_{g,j}$ is the standard deviation of the rating error of bank j for obligors in industry g, and ν_g is the mean PD score in industry g. This model forms the basis for further analysis. Note that the μ and σ parameters are unestimable for industry/bank combinations with no observations.

We begin our analysis of the estimation results by showing the parameters describing the distribution of the true latent scores. These are the industry specific means ν_g and the standard deviation τ. For ease of interpretation we additionally show the images under the inverse link function of the mean PD scores for each industry and the respective one standard deviation intervals (see Table 3).

	Industry								
	Manufac	Energy	Constr	Trading	Finance	RealEst	Public	Service	Private
ν_g	−2.542	−2.993	−2.448	−2.375	−3.256	−2.474	−3.330	−2.517	−2.296
$\Phi(\nu_g)$	55.1	13.8	71.8	87.7	5.6	66.8	4.3	59.2	108.4
$\Phi(\nu_g - \tau)$	17.7	3.8	23.8	29.8	1.5	22.6	1.1	19.8	40.0
$\Phi(\nu_g + \tau)$	151.1	44.2	190.9	227.5	19.2	174.7	15.1	157.0	267.4

Table 3: Industry specific means ν_g and PD intervals measured in basis points (10^{-4}). Intervals are obtained by applying the probit transformation to $\nu_g \pm \tau$.

We infer from Table 3 that on an aggregate level the portfolio of our sample banks might exhibit important differences in average credit quality across industries ranging from 4.3 basis points (bp; one bp corresponds to 10^{-4}) measured in terms of PDs for public obligors to 108.4bp for private obligors. Furthermore, a standard deviation τ of 0.357 on the score level yields intervals of different width on the PD scale depending on the industry specific ν_g. E.g., for public obligors we obtain an interval ranging from 1.1bp to 15.1bp whereas it is much wider among private obligors spanning from 40.0bp to 267.4bp.

	Bank												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Manufac	−0.214	NA	−0.313	−0.002	NA	0.120	0.097	−0.053	0.042	−0.138	−0.011	NA	0.472
Energy	−0.301	−0.148	−0.191	0.166	NA	NA	0.168	0.282	−0.063	−0.113	0.200	NA	NA
Constr	−0.211	−0.026	−0.123	0.053	NA	NA	NA	−0.113	−0.058	−0.012	−0.117	NA	0.607
Trading	−0.253	NA	−0.192	0.071	0.156	0.129	−0.122	0.176	−0.173	−0.082	0.048	−0.163	0.403
Finance	0.029	0.068	0.434	−0.259	−0.267	0.154	−0.259	−0.259	0.171	0.304	−0.259	−0.259	0.402
RealEst	−0.249	−0.145	−0.337	−0.004	0.234	0.008	0.080	0.094	−0.008	−0.251	0.013	0.090	0.474
Public	−0.084	0.148	0.297	−0.219	NA	0.130	−0.224	−0.216	0.162	0.048	NA	−0.236	0.194
Service	−0.214	−0.129	−0.316	−0.019	0.390	−0.156	0.068	0.221	−0.099	−0.132	0.023	NA	0.361
Private	−0.197	−0.085	−0.617	0.240	0.032	−0.029	0.285	NA	−0.195	0.169	0.193	0.202	NA

Table 4: Rating bias $\mu_{g,j}$ for bank/industry combinations of the 13 Austrian banks. In case of no observations the bias cannot be estimated, hence the NAs.

Analyzing the rating biases, Table 4 shows the bank specific bias estimates ($\mu_{g,j}$). A number of conclusions can be drawn from this analysis. First one might want to interpret the results from an aggregate bank specific perspective. Evidently, in columns 1 and 3 most parameter estimates show a negative sign, whereas the opposite holds for column 13 with all estimates being positive. Banks 1 and 3 thus seem to be too optimistic in their credit assessment whereas bank 13 might exhibit an extremely conservative rating behavior.

We also employ so-called relationship plots to visualize the estimated rating bias and error variance parameters. These plots use the strucplot framework according to Meyer et al. (2006) and the corresponding strucplot function of the R package vcd (Meyer et al., 2007) to visualize the relationship between measurements of a quantitative variable (here: estimated model parameters) and the interaction of two qualitative factors (here: industry/bank combinations). Each combination of factor levels is represented by a rectangular cell shaded by gray values representing the corresponding measurement values.

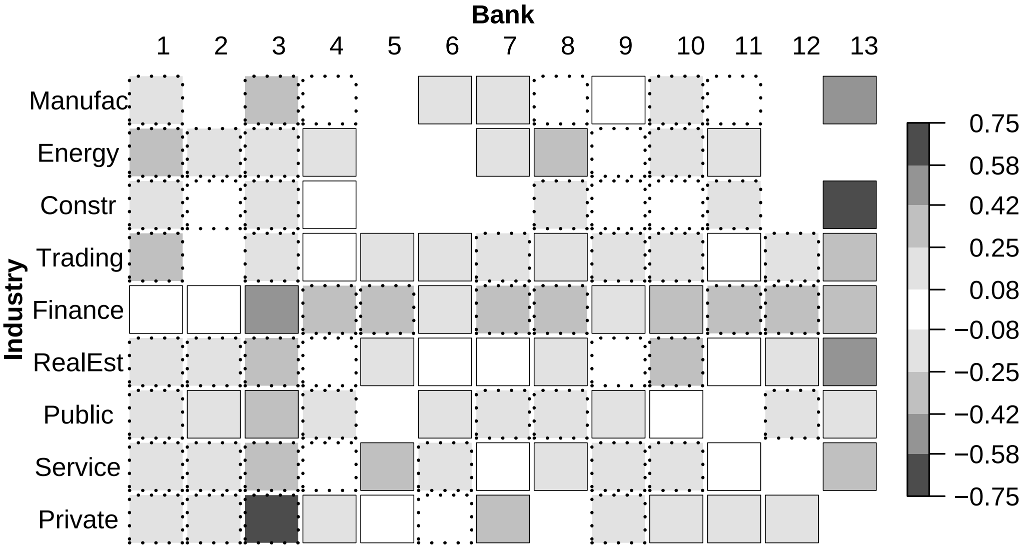

Figure 1: Rating bias for bank/industry combinations $\mu_{g,j}$ of the 13 Austrian banks. A dark cell represents a high absolute value, whereas a light cell represents a very low absolute value. If the underlying value is negative, the border of the cell is dotted instead of lined. In case of no observations the bias cannot be estimated, hence the missing cells.

We can also use the industries to better understand the general tendency of identified potential "outlier banks". Bank 1 rather generally overestimates credit quality relative to the other banks (particularly in the Energy, Trading, and Real Estate industries). Bank 3 possibly overestimates the credit quality for private individuals. Conversely, for bank 13 we note that it potentially acts over-cautiously in the Construction industry.

One of the key strengths of our framework is its ability to estimate the standard deviations of the bank specific errors and thus measure the precision of the respective rating systems. Table 5 contains the results and Figure 2 shows the corresponding relationship plot. The values range from 0.006 for bank 8 for public obligors to a maximum of 0.612 observed for bank 12 for private individuals indicating that the rating tool employed by this bank might be inappropriate for this industry. From a bank-wide perspective Figure 2 furthermore suggests that the level of precision is particularly low for bank 3. We observe the highest standard deviation levels for the Real Estate industry, indicating that the banks have particular difficulties in accurately assessing the credit quality and suggesting that informational asymmetries are

	Bank												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Manufac	0.447	NA	0.394	0.270	NA	0.407	0.098	0.033	0.140	0.249	0.136	NA	0.383
Energy	0.152	0.095	0.401	0.096	NA	NA	0.248	0.355	0.038	0.355	0.029	NA	NA
Constr	0.039	0.149	0.306	0.199	NA	NA	NA	0.252	0.284	0.256	0.155	NA	0.343
Trading	0.206	NA	0.521	0.175	0.411	0.054	0.212	0.109	0.160	0.296	0.221	0.028	0.215
Finance	0.328	0.166	0.412	0.012	0.041	0.446	0.018	0.020	0.337	0.380	0.020	0.026	0.360
RealEst	0.503	0.530	0.462	0.181	0.319	0.245	0.183	0.282	0.344	0.329	0.163	0.282	0.414
Public	0.221	0.204	0.331	0.036	NA	0.271	0.035	0.006	0.014	0.297	NA	0.048	0.242
Service	0.349	0.346	0.397	0.238	0.051	0.247	0.169	0.522	0.210	0.336	0.233	NA	0.149
Private	0.041	0.287	0.160	0.386	0.386	0.259	0.332	NA	0.133	0.428	0.283	0.612	NA

Table 5: Standard deviations $\sigma_{g,j}$ of the rating errors for bank/industry combinations of the 13 Austrian banks. In case of no observations the standard deviation cannot be estimated, hence the NAs.

rather pronounced in this industry.

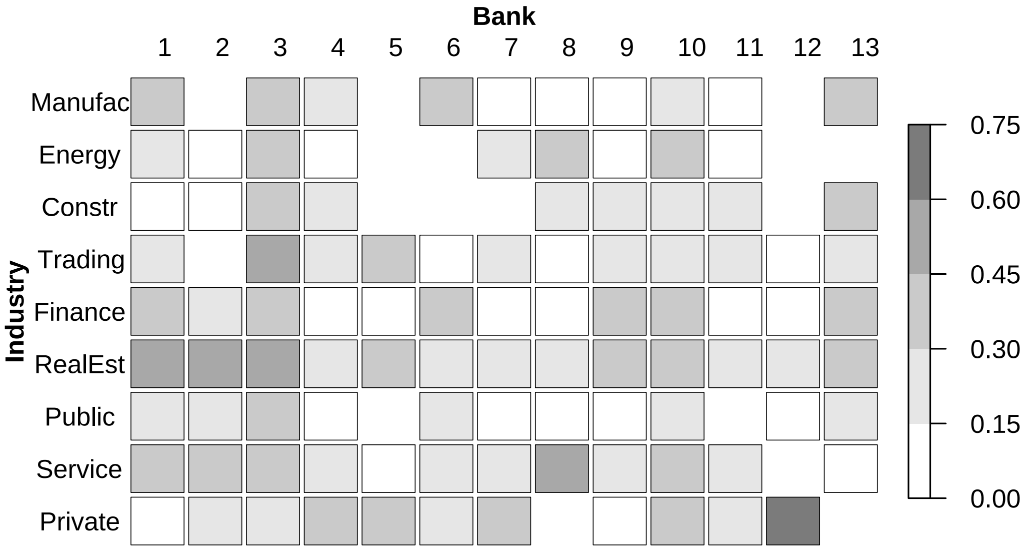

Figure 2: Standard deviations $\sigma_{g,j}$ of the rating errors for bank/industry combinations of the 13 Austrian banks. A dark cell represents a high absolute value, whereas a light cell represents a very low absolute value. In case of no observations the deviation can't be estimated, hence the missing cells.

3.2.2 Consensus Rating and Residual Analysis

The calibration results of the model presented in the previous section allows us to estimate a consensus rating for each obligor (see Section 2). The consensus rating itself can be used for many applications where deviations of individual raters or ratings from an aggregated rating is of interest, e.g., in banking supervision. In this section, the consensus ratings are used to calculate residuals for each rating which allows for a deeper economic analysis. The two variables not employed for grouping in the model specification, i.e., legal form and exposure, are used in this section to illustrate possible further analysis.

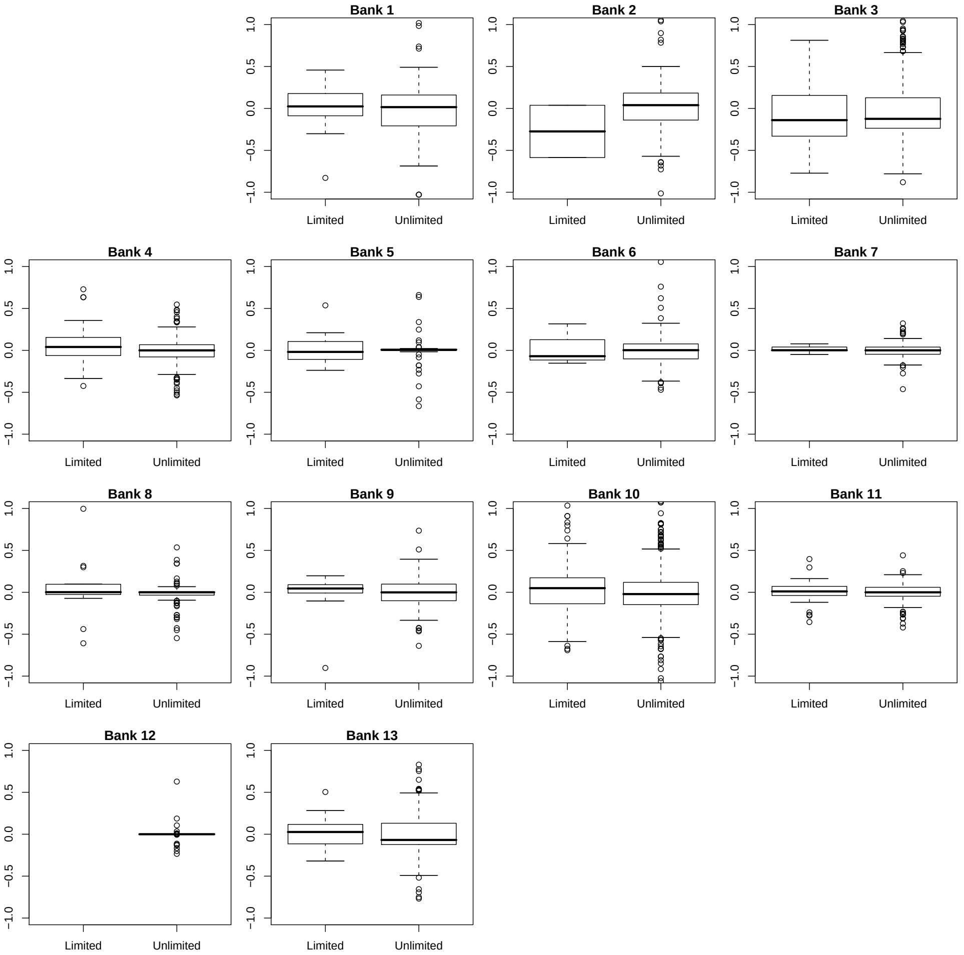

Figure 3: Residual analysis for all 13 banks across the legal forms: limited and unlimited companies.

Figure 3 presents the residuals for each bank for corporate obligors with limited and unlimited liability. Based on this representation we analyze the locations and dispersions of the residuals on the bank level. In general, the medians of the residuals are small in absolute terms and we find no structural effect over all banks, i.e., there is no general difference in terms of the median residuals between limited and unlimited liability obligors. On a bank specific level we find residual medians that markedly differ between the two legal forms, e.g., for banks 2 and 13. Bank 2 assigns favorable ratings for limited liability obligors and vice versa for bank 13. We also see no systematic difference in residual dispersion between limited and unlimited corporates, but note that individual banks exhibit rather marked levels of residual dispersion for a specific legal form. Such results could be particularly interesting for supervisors, as it might allow to identify problem areas of individual banks.

As a second illustrative example, we analyze the residuals with respect to the relative exposure size of the obligors. The relative exposure shows the importance of an obligor for the bank. Thus we are particularly interested to detect obligors with large relative exposures and too favorable ratings. Figure 4 shows residuals against relative exposures for two selected banks (bank 13 and bank 8). Bank 13 rates two obligors (marked in Figure 4) with high relative exposures (more than 2.5% of the total exposure) rather too favorable relative to the market consensus. For bank 8, however, we cannot find comparable outliers. Such an outlier analysis is very important, as it might help supervisors to identify problem loans which have a significant size within a bank's credit portfolio.

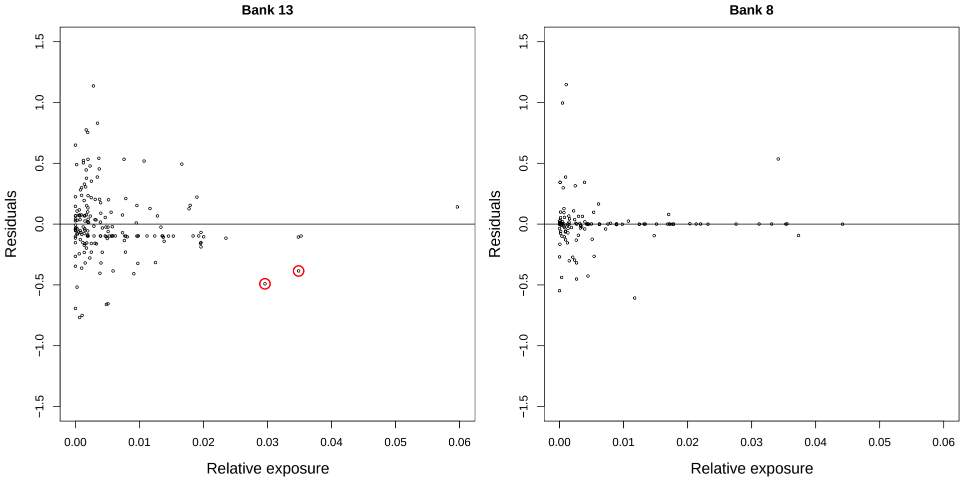

Figure 4: Residual analysis for two banks (bank 13 and bank 8) across the relative exposure.

4 Discussion

In this paper we proposed a new probabilistic framework for credit rating model validation in a multi-rater setup, i.e., in situations where PD estimates from different sources for the same obligors are available. In our model the unobservable true PD of all obligors is treated as a latent variable and raters obtain only noisy observations of the latent true PD. In the general framework three ingredients are to be specified: The distribution of the latent PD, the distribution of the error terms, and a suitable link function which transforms the PD to the real axis such that the error terms are additive. Building on the theory of structural credit risk models we propose a specification with normal error terms which uses the probit as a link function. In an empirical example we estimate suitable parametrizations of this model and present results on parameter estimates and possible economic interpretation. Our framework has a variety of potential applications: Banking supervisors might be interested in parameters of the error terms to assess the calibration quality of bank internal rating systems. Developers of rating systems such as banks and rating agencies have a natural interest in benchmarking their rating outcomes to competing models.

Results of benchmarking analyses always need to be assessed relative to the representativity of the available data panel. In our case, it obviously needs to be ensured that raters in the panel follow mutually distinct rating procedures. Data in the panel must be measurements of the *same* underlying entity. In the application framework of this paper, banks must supply ratings for the same underlying notion of obligor creditworthiness, which in our case are one-year issuer-specific point-in-time probabilities of default. If different notions are used, e.g. when trying to incorporate ordinal ratings of creditworthiness such as those provided by the major rating agencies, one could attempt to map these to their one-year PD equivalents. In this case, rating errors will include the corresponding mapping errors.

The suggested framework for modeling multi-rater panels of PD estimates is very general and allows for a variety of possible enhancements. We already indicated the possibility of including additional terms in the parametric specifications of the means and variances of the PD scores and rating errors, or allowing for correlations of rating errors across raters (again, note that a common error "factor" is indistinguishable from the latent PD score). In addition, one could aim at employing more flexible models for the distributions of the PD scores or rating errors, e.g., via suitable mixtures of normals. One could also try to model potential censoring effects mandated by regulatory frameworks. E.g., Paragraph 331 of Bank for International Settlements (2004) states that “the PD for retail exposures is the greater of the one-year PD associated with the internal borrower grade to which the pool of retail exposures is assigned or 0.03%”, suggesting to enhance Equation 1 along the lines of $S_{ij} = \max(S_i + \epsilon_{ij}, c_i)$ with *known* obligor-specific cutoffs c_i. One should note, however, that in many applications co-rating patterns are rather sparse, limiting the flexibility of statistical models which can be inferred from available data. Finally, one could think of extending the cross-sectional setup to a dynamic framework where the PD estimates are also observed at different points in time and hence the latent PDs and the error terms have to be modeled by suitable stochastic processes. Such a framework would allow forecasting future PDs as well as a lead-lag analysis across different raters. We extend to explore these possible enhancements in our future research.

References

Bank for International Settlements. International convergence of capital measurement and capital standards: A revised framework, 2004. URL `http://www.bis.org/publ/bcbs107.htm`.

Bank for International Settlements. Studies on the validation of internal rating systems (revised), 2005. URL `http://www.bis.org/publ/bcbs_wp14.pdf`.

J. Berk and P. DeMarzo. *Corporate Finance*. Statistics and Computing. Pearson International Edition, Boston, USA, 2007. ISBN 0-321-41680-5.

S. T. Bharath and T. Shumway. Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3):1339–1369, 2008.

M. Blume, F. Lim, and A. MacKinlay. The declining credit quality of US corporate debt: Myth or reality. *Journal of Finance*, 53:1953–1413, 1998.

M. Carey. Some evidence on the consistency of banks' internal credit ratings. *Federal Reserve Board Working Paper*, 2001.

W. D. Cook, M. Kress, and L. M. Seiford. Information and preference in partial orders: A bimatrix representation. *Psychometrika*, 51:197–207, 1986.

W. D. Cook, B. Golany, M. Penn, and T. Raviv. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research*, 34:954–965, 2007.

P. Crosbie and J. Bohn. Modeling default risk. *Moody's KMV*, December 2003.

M. Crouhy, D. Galai, and R. Mark. Prototype risk-rating system. *Journal of Banking and Finance*, 25:47–95, 2001.

D. Duffie and D. Lando. Term structures of credit spreads with incomplete information. *Econometrica*, 69: 633–664, 2001.

H. Elsinger, A. Lehar, and M. Summer. Risk assessment for banking systems. *Management Science*, 52(9): 1301–1314, 2006.

European Commission. Statistical classification of economic activities in the european community, 2008. URL `http://ec.europa.eu/environment/emas/documents/nace_en.htm`.

X. Guo, R. A. Jarrow, and Y. Zeng. Credit risk models with incomplete information. *Mathematics of Operations Research*, 2008. URL `http://www.ieor.berkeley.edu/~xinguo/papers/II.C.2.pdf`. To appear.

K. Hornik, R. Jankowitsch, M. Lingo, S. Pichler, and G. Winkler. Validation of credit rating systems using multi-rater information. *Journal of Credit Risk*, 3:3–29, 2007.

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

D. Lando. *Credit Risk Modeling*. Princeton University Press, Princeton, NJ, 1st edition, 2004.

H. E. Leland and D. H. Pyle. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance*, 32:371–387, 1977.

M. Lingo and G. Winkler. Discriminatory power: An obsolete validation criterion? *Journal of Risk Model Validation*, 2(1):1–27, 2008.

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

D. Meyer, A. Zeileis, and K. Hornik. The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17(3):1–48, 2006. URL `http://www.jstatsoft.org/v17/i03/`.

D. Meyer, A. Zeileis, and K. Hornik. *vcd: Visualizing Categorical Data*, 2007. URL `http://CRAN.R-project.org/package=vcd`. R package version 1.0-6.

P. Nickell, W. Perraudin, and S. Varotto. Stability of rating transitions. *Journal of Banking and Finance*, 24: 203–227, 2000.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag, New York, USA, 2000. ISBN 0-387-98957-9.

J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. *nlme: Linear and Nonlinear Mixed-Effects Models*, 2007. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-86.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

R. Stein. Benchmarking default prediction models: Pitfalls and remedies in model validation. Technical Report #020305, Moody's KMV, 2002. URL `http://riskcalc.moodysrms.com/us/research/crm/Validation_Tech_Report_020305.pdf`.