SFB 649 Discussion Paper 2007-035

Estimating Probabilities of Default With Support Vector Machines

Wolfgang Härdle*
Rouslan Moro**
Dorothea Schäfer***

* Humboldt-Universität zu Berlin, Germany
** Humboldt-Universität zu Berlin & DIW Berlin, Germany
*** DIW Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Estimating Probabilities of Default With Support Vector Machines ⋆

W. K. Härdle [a], R. A. Moro [b], D. Schäfer [c]

[a] *CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.*

[b] *German Institute for Economic Research, Königin-Luise-Straße 5, 14195 Berlin, Germany and CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.*

[c] *German Institute for Economic Research, Königin-Luise-Straße 5, 14195 Berlin, Germany.*

Abstract

This paper proposes a rating methodology that is based on a non-linear classification method, the support vector machine, and a non-parametric technique for mapping rating scores into probabilities of default. We give an introduction to underlying statistical models and represent the results of testing our approach on German Bundesbank data. In particular we discuss the selection of variables and give a comparison with more traditional approaches such as discriminant analysis and the logit regression. The results demonstrate that the SVM has clear advantages over these methods for all variables tested.

JEL classification: C14; G33; C45

Keywords: Bankruptcy; Company rating; Default probability; Support vector machines

1 Introduction

Banking throughout the world, both central and commercial, is based on credit or trust in the debtor's ability to fulfil his obligations. Facing increasing pressure from markets and regulators, banks build their trust to an ever increasing degree on statistical techniques for corporate bankruptcy prediction known as *rating* or *scoring*. Their main purpose is to estimate the financial situation of a company and, if possible, the probability that a company defaults on its obligations within a certain period.

Application of statistical models to corporate bankruptcy was made popular after the introduction of discriminant analysis (DA) by Altman (1968). Later the logit and probit models were suggested in Martin (1977) and Ohlson (1980). All these models belong to the class of Generalised Linear Models (GLM) and could also be interpreted using a latent (score) variable. Their core decision element is a linear score function (graphically represented as a hyperplane in a multidimensional space) separating successful and failing companies. The company score is computed as a value of that function. In the case of the probit and logit models the score is – via a link function – directly transformed into a probability of default (PD). The major disadvantage of these popular approaches is the enforced linearity of the score and, in the case of logit and probit models, the prespecified form of the link function (logit and Gaussian) between PDs and the linear combination of predictors.

In this paper we are introducing and evaluating a new way of assessing company's creditworthiness. The proposed rating methodology is based on the non-linear classification method, the support vector machine (SVM), and a non-parametric technique for mapping rating scores into probabilities of default (see the Appendix and Chapter 5). The SVM is based on the principle of a safe separation of solvent and insolvent companies in such a way that the distance between the classes is maximised while missclassifications are penalised proportionaly to the distance from their class. The method allows the use of

⋆ The authors gratefully acknowledge that the project is co-financed by the Stiftung Geld und Währung. We thank German Bundesbank for providing access to the unique database of the financial statements of German companies. The data analysis took place on the premises of the German Bundesbank in Frankfurt. The work of R. A. Moro was financially supported by the German Academic Exchange Service (DAAD) and German Bundesbank. This research was also supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk". We are grateful to Laura Auria and Ralf Körner of the German Bundesbank for their cooperation and valuable suggestions and to Wolfgang Bühler for the discussion of the paper at the European Finance Association meeting in Zürich in 2006.

Email addresses: haerdle@wiwi.hu-berlin.de (W. K. Härdle), rmoro@diw.de (R. A. Moro), dschaefer@diw.de (D. Schäfer).

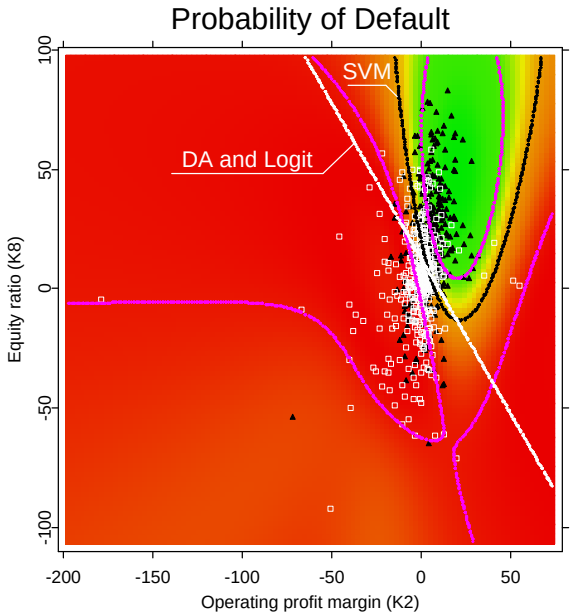


Figure 1. A classification example. The boundary between the classes of solvent (black triangles) and insolvent companies (white rectangles) was estimated using DA and logit regression (two indistinguishable lines) and an SVM (a non-linear curve).

kernel techniques (Hastie, Tibshirani, and Friedman (2001)) and, therefore, non-linear separating surfaces in contrast to DA, logit and probit models that rely on linear ones. Figure 1 illustrates the qualitative step forward that we are proposing in this paper. The straight line is the linear hyperplane separating solvent and insolvent companies based on DA or the logit model. The curved lines are the separation surface and the bounds calculated with the SVM technique. It is evident that the non-linear separation outperformes the linear one and translates into a better classification performance. Another important feature of the SVM is its automatic rather than manual surface shape identification.

We examine here empirically whether the adoption of SVMs leads to a more accurate prediction of default events than the use of DA and Logit/Probit models. Our study has potential implications for supervisory agencies, banks and firms: we illustrate that non-monotonicity and non-linearity in the data significantly influences accuracy. For supervisory agencies our assessments show the magnitude of the impact of simplified quantitative models on the PD estimation and, therefore, on capital requirements.

When following the DA, logit or probit approach we automatically impose (through a modelling bias) a monotonic relationship between financial and economic indicators and PDs. A typical example is the imposed monotonic

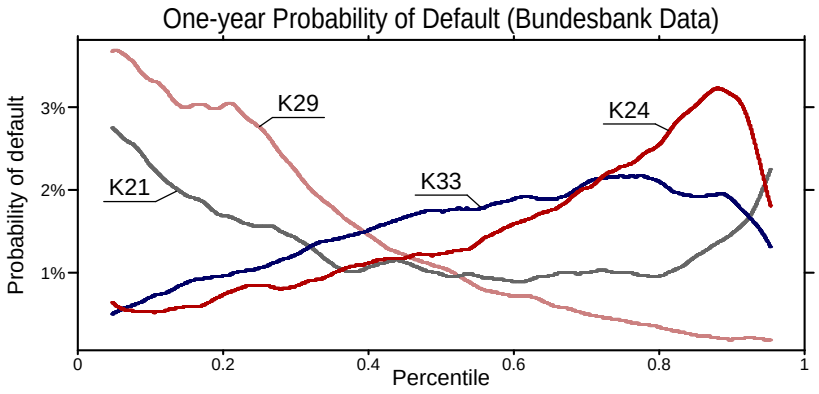


Figure 2. One year PDs evaluated for several financial ratios on the German Bundesbank data. The ratios are the net income change K21; net interest ratio K24; interest coverage ratio K29 and the logarithm of total assets K33.

decreasing relation as for the interest coverage ratio (Figure 2). However, in reality this dependence is often non-monotonic as for such important indicators as the company size or net income change. In the latter case companies that grow too fast or too slow have a higher probability of default. Non-linear dependencies in the data which are confirmed in the literature (Fernandes (2005), Manning (2004)) and are accounted for in the marketed models (Falkenstein, Boral, and Carty (2000)) are the reason for contemplating non-linear techniques as alternatives.

In order to be able to capture non-linearity, the score function – indicating the PD – must be flexible and based on very general criteria. The SVM is a non-linear statistical technique that in many applications, such as optical character recognition, medical diagnostics and electrical load forecasting, showed very good accuracy. It has as a solution a flexible classification function and is controlled by adjusting only few parameters. The SVM solution is stable, i.e. changes slowly in response to a slow change of the data, since the method is based on the convex optimisation problem (Tikhomirov (1996)). Its overall good performance and flexibility, eliminating the manual selection of the score function, make the SVM a suitable candidate for company rating (Härdle, Moro, and Schäfer (2005)).

The purpose of classification methods is to separate insolvent ($y = 1$) from solvent ($y = -1$) companies described with a d dimensional vector of characteristics x, usually financial ratios. Here we use $y \in \{-1, 1\}$ instead of the common $y \in \{0, 1\}$ notation since it is more convenient in the following formal expressions. The SVM does the separation of the two groups with the

maximum distance (margin) between them. The score for x is computed as

$$f(x) = \sum_{i=1}^{n} K(x_i, x)\alpha_i y_i + b. \tag{1}$$

In our case the kernel $K(x_i, x)$ is, up to a constant, a Gaussian density function, with $x - x_i$ as an argument, which measures the proximity of an observation x of an unknown class to the observation x_i whose class y_i is known. The closer x and x_i are, the larger is $K(x_i, x)$; therefore, the score $f(x)$ is primarily defined by the observations that are close to x. The factors α_i are the solution of an SVM (Lagrange multipliers) and have higher magnitudes for the observations at the boundary between the classes which are most relevant for classification. The mathematical details are described in the Appendix.

The rest of the paper proceeds as follows. Data and variable selection is presented in Sections 2 and 3. Section 4 outlines the comparative results. Then the non-parametric technique of estimating a single firm's PD is introduced. Finally, Section 6 concludes.

2 Data

For this study we use German Bundesbank data. The data are cross-sectional, since each firm enters the dataset only for several years. It covers the years 1987–2005 and contains around 500000 balance sheets and income statements for solvent and around 8000 statements for bankrupt firms. German Bundesbank condenses the balance sheet information for each firm into 33 financial predictors. We apply the Bundesbank ratios for comparison purposes. Table 1 presents the summary statistics for each predictor.

We have selected a homogenous sample spanning from 1992 to 1998. In 1991 German reunification and in 1999 the change in accounting procedure in the Bundesbank were the events that brought about a break in the data. The distribution of the data over the years for solvent and insolvent companies after cleaning the observations with missing variables is given in Table 2.

The last annual report of a company before it goes bankrupt receives the indicator $y = 1$, all the others $y = -1$. The last reporting date precedes bankruptcy by 0.5–3.5 years.

Not all predictors are equally relevant for the SVM as well as DA and Logit analysis. Moreover, since many predictors are highly correlated, even a small group of them already contains most classification information. Adding additional variables highly correlated with the previously included ones does not

Table 1
Summary Statistics. q_α is an α quantile. IQR is the interquartile range.

Var.	Name	Group	$q_{0.01}$	Median	$q_{0.99}$	IQR
K1	Pre-tax profit margin	Profitability	-26.9	2.3	78.5	5.9
K2	Operating profit margin	Profitability	-24.6	3.8	64.8	6.3
K3	Cash flow ratio	Liquidity	-22.6	5.0	120.7	9.4
K4	Capital recovery ratio	Liquidity	-24.4	11.0	85.1	17.1
K5	Debt cover	Liquidity	-42.0	17.1	507.8	34.8
K6	Days receivable	Activity	0.0	31.1	184.0	32.7
K7	Days payable	Activity	0.0	23.2	248.2	33.2
K8	Equity ratio	Financing	0.3	14.2	82.0	21.4
K9	Equity ratio (adj.)	Financing	0.5	19.3	86.0	26.2
K10	Random Variable	Test	-2.3	0.0	2.3	1.4
K11	Net income ratio	Profitability	-29.2	2.3	76.5	5.9
K12	Leverage ratio	Leverage	0.0	0.0	164.3	4.1
K13	Debt ratio	Liquidity	-54.8	1.0	80.5	21.6
K14	Liquidity ratio	Liquidity	0.0	2.0	47.9	7.1
K15	Liquidity 1	Liquidity	0.0	3.8	184.4	14.8
K16	Liquidity 2	Liquidity	2.7	63.5	503.2	58.3
K17	Liquidity 3	Liquidity	8.4	116.9	696.2	60.8
K18	Short term debt ratio	Financing	2.4	47.8	95.3	38.4
K19	Inventories ratio	Investment	0.0	28.0	83.3	34.3
K20	Fixed assets ownership r.	Leverage	1.1	60.6	3750.0	110.3
K21	Net income change	Growth	-50.6	3.9	165.6	20.1
K22	Own funds yield	Profitability	-510.5	32.7	1998.5	81.9
K23	Capital yield	Profitability	-16.7	8.4	63.1	11.0
K24	Net interest ratio	Cost struct.	-3.7	1.1	36.0	1.9
K25	Own funds/pension prov. r.	Financing	0.4	17.6	84.0	25.4
K26	Tangible asset growth	Growth	0.0	24.2	108.5	32.6
K27	Own funds/provisions ratio	Financing	1.7	24.7	89.6	30.0
K28	Tangible asset retirement	Growth	1.0	21.8	77.8	18.1
K29	Interest coverage ratio	Cost struct.	-1338.6	159.0	34350.0	563.2
K30	Cash flow ratio	Liquidity	-14.1	5.2	116.4	8.9
K31	Days of inventories	Activity	0.0	42.9	342.0	55.8
K32	Current liabilities ratio	Financing	0.3	58.4	98.5	48.4
K33	Log of total assets	Other	4.9	7.9	13.0	2.1

substantially increase available information but introduces additional noise reducing overall model performance. The identification of variables relevant for each model is the task of the variable selection procedure.

Table 2
The distribution of the data over the years for solvent and insolvent companies for the period 1992–1998 for the observations without missing variables.

Year	Solv.	Insolv. (%)	Total
1992	41626	621 (1.47%)	42247
1993	41202	691 (1.65%)	41893
1994	40814	622 (1.50%)	41436
1995	40869	586 (1.41%)	41455
1996	39011	564 (1.43%)	39575
1997	34814	658 (1.85%)	35472
1998	27903	646 (2.26%)	28549
Total	266239	4388 (1.62%)	270627

3 Variable Selection

Our judgements about model accuracy are based on widely accepted criteria: the accuracy ratio (AR), which will be used here as a criterion for model selection, and alpha and beta errors. AR is the ratio of the areas between (i) the cumulative default curves for the considered model and the random model and (ii) the ideal and the random model. An alpha error is the percentage of insolvent companies among solvent ones and a beta error is the percentage of solvent companies among insolvent ones. A classification method has a higher power if for a given alpha error it delivers a lower beta error. Higher ARs and lower alpha and beta errors indicate better model accuracy. The complementary to the sum of alpha and beta errors is the percentage of correctly classified out-of-sample observations or hit ratio (HR).

Since it is practically impossible to try all combinations of variables to choose one that yields overall the best AR, we need to apply a selection procedure. We will apply a backward variable selection procedure (BSP) and, in parallel, a forward selection procedure (FSP) for all three competitors: DA, logistic regression and SVM. The BSP starts with the full model which includes all variables. At the first step one of the variables is consecutively excluded and the AR of each reduced model is computed. The model that has the highest AR will be examined at the second step when one more variable is consecutively excluded and ARs are compared. The procedure continues until a univariate model is selected by reducing a bivariate model. The FSP starts with the selection of a univariate model and continues until all variables are included. At each step the variable is kept whose addition to the model produced the

Table 3
Variables included in the DA, Logit and SVM models that produced the highest ARs. "1" denotes a variable that was selected. The values in parenthesis are the median AR achieved for the model reported.

Model	Variables included in the model, K*																														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Backward selection																																	
DA (59.20)	.	.	.	.	1	1	1	.	1	.	1	1	.	1	1	1	1	.	.	.	.	.	.	.	.	.	.	1	1	.	1	.	.
Logit (59.16)	.	.	.	.	1	1	1	.	1	.	1	1	.	.	.	1	1	.	.	.	.	.	1	.	.	.	.	1	1	.	1	.	.
SVM (61.11)	.	.	.	1	.	1	1	.	1	.	.	1	.	.	1	.	.	.	.	.	1	.	.	.	1	1	.	1	1	.	1	1	1
Forward selection																																	
DA (59.32)	1	.	.	1	1	1	1	1	.	.	1	1	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1	.
Logit (59.05)	.	.	.	.	1	1	1	1	1	.	1	1	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	1	.	1	.	1
SVM (60.75)	.	.	.	.	1	1	1	.	1	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	1	1	.	.	1	.	1	1	1

highest AR.

The application of an FSP makes more sense when the number of variables included d is small. For example, if $d = 1$, the FSP selects the most accurate model, that is not true for $d >= 2$. The BSP selects the most accurate model if d is smaller by 1 than the number of variables available and is more suitable if expected d is large.

The comparison of models at each step is done on the basis of a robust measure of AR that is not sensitive to extreme values: median AR computed on bootstrapped data (Efron and Tibshirani (1993), Horowitz (2001)). We randomly select training and validation sets as subsamples of 400 solvent and 400 insolvent companies each. The relatively small size of the training and validation sets of 800 observations each is required by the bootstrap procedure. The two sets are not overlapping, i.e. do not contain common observations. The Monte Carlo experiment is repeated 100 times to compute one distribution of ARs. The performance of DA, Logit and SVM is summarised in Figure 3. The median for the SVM approach is for almost all models higher than for the alternative methods. The variables included in the selected models are reported in Table 3. In the BSP maximum AR is achieved for the SVM containing 14 variables.

The SVM model used for variable selection has the parameters $r = 5$ and $c = 10$ (see the Appendix) selected a priori without optimisation. The optimisation of r and c can further boost the SVM performance. Higher values of c and lower values of r correspond to more complex models. When a model becomes too complex, accuracy drops dramatically. The dependence of AR from r and c for the 14-variable model with the highest median AR is represented in Figure 4.

It should be noted that the standard normally distributed variable K10 does

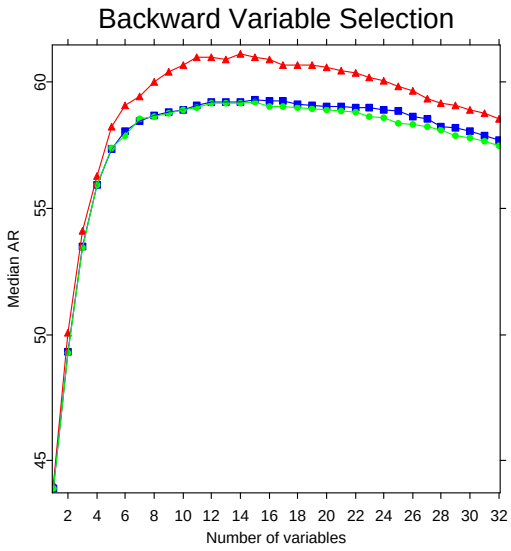


Figure 3. Median AR for DA (rectangles), Logit (circles) and SVM (triangles) for models with different numbers of predictors. At each step a model with the highest median AR is selected.

not contain any information and was artificially added to the data set for comparison purposes. It is already included into most models prior to step 20 out of total 32 steps. This means that the variables added to the model at the last steps are as redundant as K10.

The FSP does not have any clear advantage over the BSP. Since the BSP delivers a slightly higher median AR for two out of three models we will use the BSP selected models for further analysis.

4 Comparison of DA, Logistic Regression and SVM

Upon having chosen variables for each model we can compare their performance on the data from 1992–1998 and beyond that period. Since the selection procedure was done independently for DA, a logistic regression and SVM, we do not introduce any bias against or in favour of any model. The number of variables in each model will be different as indicated in Table 3.

The data used in the DA and Logit models were always processed as following: if $x < q_{inf}(x)$ then $x = q_{inf}(x)$ and if $x > q_{sup}(x)$ then $x = q_{sup}(x)$; $q_{inf}(x) = Median(x) - 1.5IQR(x)$ and $q_{sup}(x) = Median(x) + 1.5IQR(x)$. Thus, the DA and Logit procedures applied were *robust versions* not sensitive to outliers. Here IQR denotes the interquartile range.

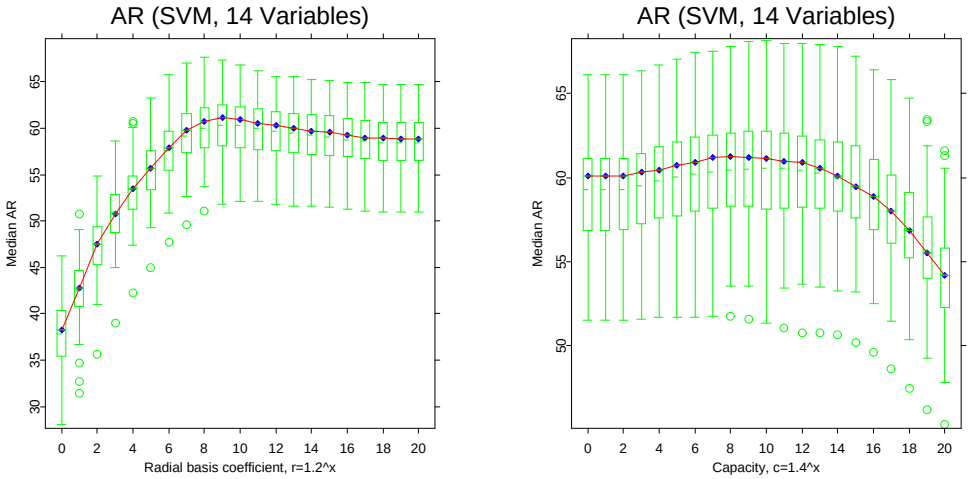


Figure 4. Left panel: the AR for different radial basis coefficients r. Capacity is fixed at $c = 10$. Right panel: the AR for different capacities c. The radial basis coefficient r is fixed at $r = 5$. The training and validation data sets are bootstrapped 100 times without overlapping from the data for 1992-1998. Each training and validation set contains 400 solvent and 400 insolvent companies.

Table 4
Forecasting accuracy improvement for each pair of models and the median AR for an SVM (the highest AR among the three models). 100 bootstrapped training and 100 bootstrapped testing samples are used. All figures are reported as percentage of the ideal AR (100%).

Training	Testing	SVM-DA	SVM-Logit	Logit-DA	AR (SVM)
1992	1995	0.99	0.96	0.01	61.79
1993	1996	3.12	3.05	0.10	60.73
1994	1997	3.56	3.62	-0.07	59.28
1995	1998	2.44	2.30	0.03	59.18
1992-1998	after 1998	2.20	1.74	0.11	58.57

Firstly, we will compare forecasting accuracy within 1992–1998. The data from 1992 will be used to forecast defaults in 1995, the data from 1993 to forecast defaults in 1996, etc. This gives a 3.5–6.5 year forecasting horizon. Secondly, the data from 1992–1998 will be used for forecasting defaults in 1999 and beyond. The latter test is performed completely out-of-sample and out-of-time. Since 1999 represents a break in the data when the reporting procedure at the Bundesbank was changed, forecasting beyond 1998 leads to lower accuracy.

When we switched from the testing design with 100 bootstrapped samples (Table 4) to training on the whole available data set without bootstrapping,

Table 5
Forecasting accuracy improvement for each pair of models and the AR estimated for an SVM (the highest AR among the three models). All data for the given years are used. All figures are reported as percentage of the ideal AR (100%).

Training	Testing	SVM-DA	SVM-Logit	Logit-DA	AR (SVM)
1992	1995	2.82	2.39	0.43	60.98
1993	1996	5.10	4.66	0.44	60.98
1994	1997	5.72	5.14	0.58	59.49
1995	1998	4.33	3.98	0.35	59.97
1992-1998	after 1998	5.04	4.03	1.01	59.86

as it will happen in practice (Table 5), we got an improvement in AR for the SVM. This is an indication that the risk to be non-representative is higher for small samples. However, both DA and Logit, compared to the SVM, perform substantially worse without bootstrapping that is due to the higher model risk associated with them.

When trained on the data for 1992 and tested on that for 1995 with a bootstrap procedure the SVM outperforms DA and Logit in 93% and 92% cases with a median improvement 2.44% and 2.30% and mean improvement 2.31% and 2.27% respectively, measured as percentage of the AR for the ideal model (Figure 5). The results for other years are very similar.

Figure 6 shows the comparison of DA, Logit and an SVM in terms of model power. Since the represented dependence is very noisy because of a small number of insolvencies in the sample, we applied a k-NN smoothing procedure with the window equal to $n/10$ or 1/10th of all observations in the sample. The training data are from 1995, testing data are from 1998. Two observations can be made. Firstly, an SVM has a higher power since its curve lies below those for DA and Logit. Secondly, many observations for the smallest alphas, more precisely 11%, when evaluated with an SVM lie in the area where no observations evaluated with DA or Logit are located. This means that an SVM in contrast to DA or Logit is able to locate the cluster of the companies with the lowest insolvency risk.

A higher power of the SVM and its ability to identify the most solvent companies avoiding unnecessary discrimination against them on a cautionary principle are particularly valuable features. Application of an SVM instead of DA or Logit will allow to issue more credit without increasing risk because of a better separation of solvent and insolvent companies.

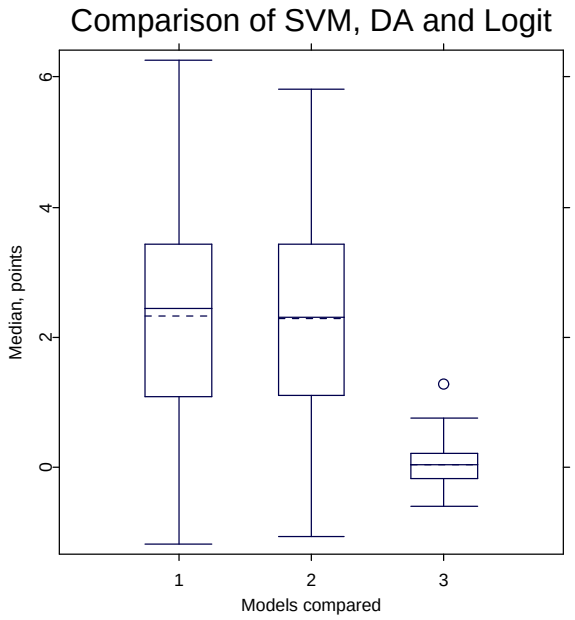


Figure 5. The improvement in AR of (i) SVM over DA, (ii) SVM over Logit and (iii) Logit over DA for the models with the highest median AR as they were selected by the BSP. The training data: 1995; testing data: 1998

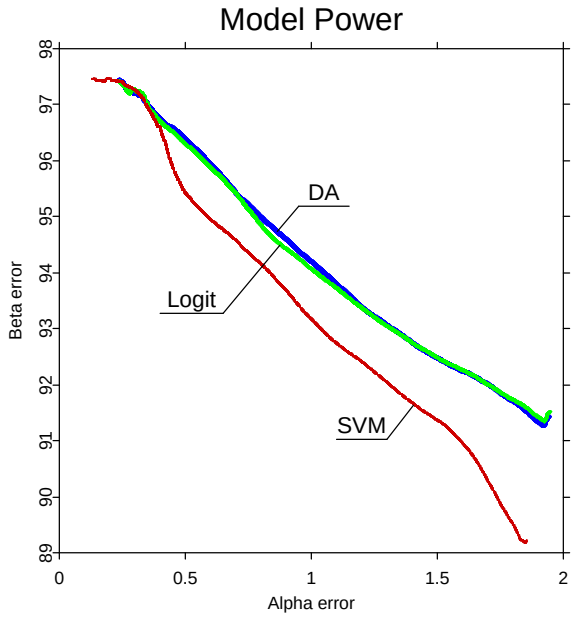


Figure 6. The power of a model: beta errors as a function of alpha errors. An SVM has a higher power than DA or Logit since it has smaller beta errors for the same alpha errors. Predictors were selected by the BSP. The training data: 1995; testing data: 1998.

5 Conversion of Scores into PDs

There is another way to look at a company score. It defines the distance between companies in terms of PD: the lower the difference in scores, the closer are companies. If a company has a higher score, it lies farther from successful companies and, therefore, its PD should be higher. This means that the dependence between scores and PDs is assumed to be monotonic. No further assumptions about the form of this dependence will be made in contrast to the already analysed Logit model with a prespecified functional form.

The conversion procedure consists of the estimation of PDs for the observations of the training set with a subsequent monotonisation (step one and two) and the computation of a PD for a new company (step three).

Step one is the estimation of PDs for the companies of the training set. This is done using standard smoothing techniques to preliminary evaluate PDs for all n observations of the training set:

$$\widetilde{PD}(z) = \frac{\sum_{i=1}^{n} w(z - z_i) I(y_i = 1)}{\sum_{i=1}^{n} w(z - z_i)}, \tag{2}$$

where $w(z - z_i) = \exp\{(z - z_i)^2 / 2h^2\}$. The rank of the i-th company $z_i = Rank\{f(x_i)\}$ varies from 1 to n depending on its score $f(x_i)$; the higher the score is, the higher is the rank. h is a bandwidth, in our case $h = 0.09n$. The smaller is the bandwidth, the smoother is $\widetilde{PD}(z)$. When $h \to 0$ no smoothing is performed and all $\widetilde{PD}(z_i)$, $i = 1, 2, \ldots, n$, will be either 1 or 0; when $h \to \infty$, all $\widetilde{PD}(z_i)$ will have the same value equal to the average probability of default for the training set.

Using the company rank z instead of the score $f(x)$ we obtain a k-NN smoother with Gaussian weights $\frac{w(z-z_i)}{\sum_{j=1}^{n} w(z-z_j)}$ which decay gradually as $|z - z_i|$ grows. This differs from the most commonly used k-NN smoother that relies on the uniform weights $\frac{1}{k} I(|z - z_i| < k/2 + 1)$.

The preliminary PDs evaluated at step one are not necessarily a monotonic function of the score. This is due to the fact that companies with close scores may have for different reasons a non-concordant binary survival indicator y. The monotonisation of $\widetilde{PD}(z_i)$, $i = 1, 2, \ldots, n$ is achieved at step two using the Pool Adjacent Violator (PAV) algorithm (Barlow, Bartholomew, Bremmer, and Brunk (1972)). Figure 7 illustrates the workings of the algorithm. The companies are ordered according to their rank and have here the indicator $y = 1$ for insolvent and $y = 0$ for solvent companies. The thin line denotes the PDs estimated using the k-NN method with uniform weights and $k = 3$. At the interval between the observations with rank 1 and 2 monotonicity is

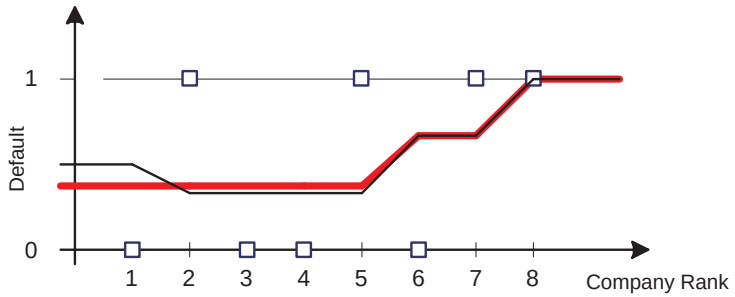


Figure 7. Monotonisation of PDs with the pool adjacent violator algorithm. The thin line denotes PDs estimated with the k-NN method with uniform weights and $k = 3$ before monotonisation and the bold line after monotonisation. Here $y = 1$ for insolvencies, $y = 0$ for solvent companies.

violated and is corrected with the PAV algorithm. The bold line shows PDs after monotonisation.

The PAV algorithm solves the following optimisation problem: given data $\{z_i, y_i\}_{i=1}^n$ with $z_1 \leq z_2 \leq \ldots \leq z_n$ find the monotonic increasing function $m(z_i)$, i.e. $m(z_1) \leq m(z_2) \leq \ldots \leq m(z_n)$ that minimises $\sum_{i=1}^n \{y_i - m(z_i)\}^2$. The solution to this problem is pooling (averaging) the adjacent observations that are violating monotonicity. The PAV acronym comes from this property. Mammen (1991) has shown that one can equivalently start with the PAV step and then smooth with a Nadaraya-Watson kernel estimator (Nadaraya (1964)).

As a result we obtain monotonised probabilities of default $PD(x_i)$ for the observations of the training set. A PD for any observation x of the testing set is computed by interpolating PDs for two adjacent, in terms of the score, observations from the training set. If the score for x lies beyond the range of the scores of the training set, then $PD(x)$ is set equal to the score of the first neighbouring observation of the training set. Figure 8 shows the PD and the cumulative PD (CPD) curve estimated on the binary data represented as circles. The CPD was evaluated as

$$CPD(z) = \frac{\sum_{i=1}^n I(y_i) I(z_i \leq z)}{\sum_{i=1}^n I(y_i)}.$$

Figure 9 represents PDs estimated with an SVM trained on the 1995 year data. The PDs for the rating classes, as they are denoted by Moody's, are reported in Table 6. Around 1800 companies or 6.30% of all companies in 1995 were classified as belonging to the class A2 or above with $PD \leq 0.095\%$. The securities of these companies can be used as a collateral for refinancing since they have PDs less than 0.1%, the threshold level set by the European Central Bank.

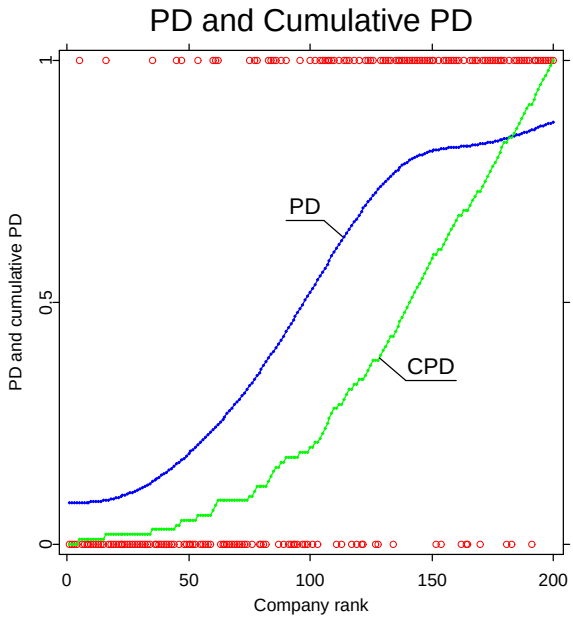


Figure 8. Smoothing and monotonisation of binary data ($y = 1$, 'default' or $y = 0$, 'non-default') represented as circles with a k-NN method and a pool adjacent violator (PAV) algorithm. The estimated PD equals, up to the scale, the first derivative of the cumulative PD.

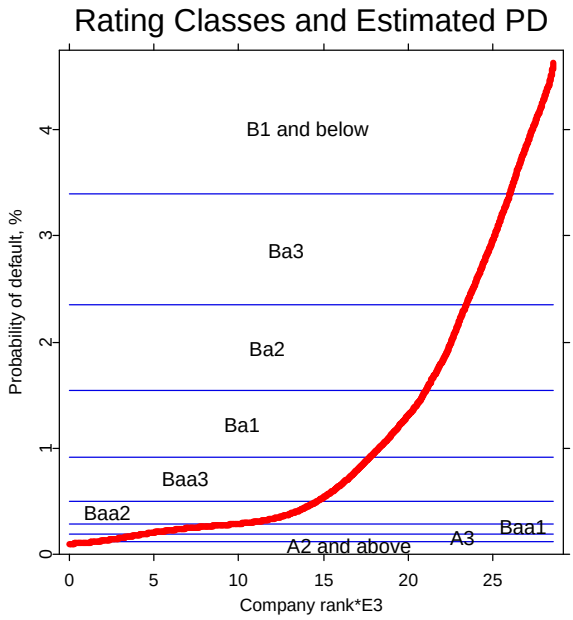


Figure 9. One year probabilities of default estimated with an SVM for 1995.

Table 6
One year PDs of the rating classes represented in Figure 9, the number and percentage of observations in each class for 1995. The total number of observations is 28549. The classes are denoted using the Moody's notation. The PDs of rating classes are reported as in Cantor, Emery, and Stumpp (2006).

Rating classes	PD, %	Number	Percentage
A2 and above	≤ 0.095	1799	6.30%
A3	0.150	2617	9.17%
Baa1	0.231	5126	17.96%
Baa2	0.331	5039	17.65%
Baa3	0.660	3191	11.18%
Ba1	1.171	3256	11.41%
Ba2	1.914	2373	8.31%
Ba3	2.783	2579	9.03%
B1 and below	≥ 4.002	2569	9.00%

6 Conclusion

In this paper we show that a rating model based on SVMs is dominating traditional linear parametric approaches such as DA and logistic regression. The forecasting accuracy improvement is significant already for small samples. We demonstrate how non-linear non-parametric techniques can be a basis for a rating model. The implementation of an SVM rating model and its extensive testing on the data of the German Bundesbank was performed. We believe that non-parametric techniques such as the SVM will become more commonplace in company rating since they better represent data, provide higher forecasting accuracy and allow to classify more companies as solvent without compromising stability.

7 Appendix

The SVM technique is based on margin maximisation between two data classes (Vapnik (1995)). The margin (Figure 10) is the distance between the hyperplanes bounding each class where in the hypothetical case of linearly perfectly separable data no observation may lie. Only those observations, so called support vectors, that lie on the margin boundaries (for linearly non-separable data also within or on the wrong side of the margin) determine the SVM solu-

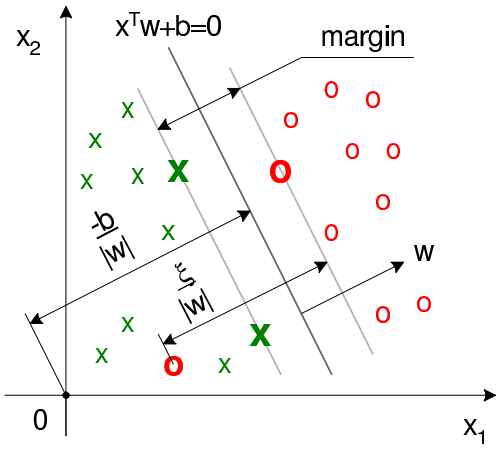


Figure 10. The separating hyperplane $x^\top w + b = 0$ and the margin in a linearly non-separable case. The observations marked with bold crosses and zeros are support vectors. The hyperplanes bounding the margin zone equidistant from the separating hyperplane are represented as $x^\top w + b = 1$ and $x^\top w + b = -1$.

tion. This is in the contrast to DA or logistic regression where all observations are used to derive the solution independently of their position relative to the opposite class.

To account for misclassifications the penalty ξ_i is introduced, which is related to the distance from the hyperplane bounding observations of the same class to observation i. If a misclassification occurs, $\xi_i > 0$. All observations satisfy the following two constraints:

$$y_i(x_i^\top w + b) \geq 1 - \xi_i, \tag{3}$$

$$\xi_i \geq 0. \tag{4}$$

For the canonical representation as in (3) the margin equals $2/\|w\|$. The convex objective function

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} C_i \xi_i.$$

is to be minimised under constraints (3) and (4). This leads to the primal problem

$$L_P = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} C_i \xi_i - \sum_{i=1}^{n} \alpha_i \{y_i \left(x_i^\top w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n} \mu_i \xi_i. \tag{5}$$

The parameters C_i are called capacity. They are related to the width of the margin zone. The smaller the C_i are, the bigger margins are possible. For a classical SVM $C_i = C$. In our case in order to control for the number of observations and dimensionality and to make an SVM suitable for the datasets

with any ratio of solvent and insolvent companies we compute C_i as

$$C_i = c\left\{\frac{I(y_i = 1)}{2n_+} + \frac{I(y_i = -1)}{2n_-}\right\}.$$

This compact representation allows to control the complexity of a linear SVM with only one parameter c.

By including the Karush-Kuhn-Tucker (KKT) first order optimality conditions (Gale, Kuhn, and Tucker (1951)) in (5) the dual Lagrangian L_D is derived

$$L_D = \frac{1}{2}w(\alpha)^\top w(\alpha) - \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\delta_i\alpha_i + \sum_{i=1}^{n}\gamma_i\left(\alpha_i - C_i\right) - \beta\sum_{i=1}^{n}\alpha_i y_i. \qquad (6)$$

α_i, δ_i, γ_i and β are Lagrange multipliers for all $i = 1, \ldots, n$. The function $w(\alpha)^\top w(\alpha)$ is a scalar product in some Hilbert space. For a linear SVM

$$w(\alpha)^\top w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j. \qquad (7)$$

When substituting the scalar product by the kernel function $K(x_i, x_j)$ a more general form is applicable:

$$w(\alpha)^\top w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j). \qquad (8)$$

The kernel function is a convenient way of mapping low dimensional data into a highly dimensional (often infinitely dimensional) space of features. It must satisfy the Mercer conditions (Mercer (1909)), i.e. be symmetric and semipositive definite or, in other words, represent a scalar product in some Hilbert space (Weyl (1928)).

Figure 11 shows a simple example of a mapping. The quadratic kernel function $K(x_i, x_j) = (x_i^\top x_j)^2$ maps two dimensional data into a three-dimensional space of features. The three features correspond to the three components of a quadratic form in two dimensions: $\tilde{x_1} = x_1^2$, $\tilde{x_2} = \sqrt{2}x_1x_2$ and $\tilde{x_3} = x_2^2$. The transformation is $\Psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$. By employing the kernel transformation higher order dependencies between variables are accounted for. The data separable in the data space only with a quadratic function will be separable in the feature space with a linear function. Thus, a non-linear SVM in the data space is equivalent to a linear SVM in the feature space. The number of features will grow fast with the dimension of the data d and the degree of the polynomial kernel.

Non-linear extensions of popular methods such as DA or logistic regression also exist when instead of original variables the transformed ones are used. Non-linear DA and logistic regression can be as powerful as SVM, however,

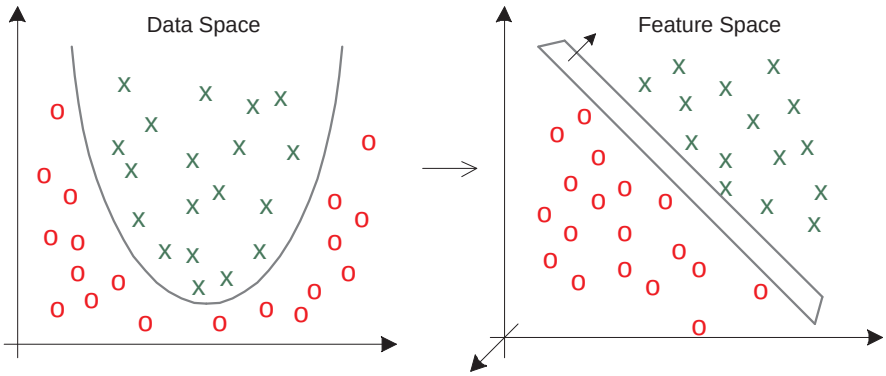


Figure 11. Mapping from a two-dimensional data space into a three-dimensional space of features $\mathbb{R}^2 \mapsto \mathbb{R}^3$.

require substantial experience from the operator for the choice of the transformations. SVM does this automatically on a robust theoretical basis with only the complexity parameter left to be optimised, which can be very easily accomplished automatically as well. In a sense an SVM tries a great number of transformations and selects without any supervision those that correspond most adequately to the data and chosen complexity.

In our study we applied an SVM with an anisotropic Gaussian or radial basis kernel

$$K(u,v) = \exp\left\{-(u-v)^\top r^{-2}\Sigma^{-1}(u-v)/2\right\}, \tag{9}$$

where r is a coefficient and Σ is a scaling matrix, which in our case is a variance-covariance matrix. The coefficient r is related to the complexity of classifying functions: the higher the r is, the lower is the complexity. If kernel functions allow for sufficiently rich feature spaces, the performance of SVMs with different kernels is comparable in terms of out-of-sample forecasting accuracy (Vapnik (1995)).

The company score is computed as:

$$f(x) = x^\top w + b, \tag{10}$$

where $w = \sum_{i=1}^n \alpha_i y_i x_i$ and $b = -\frac{1}{2}(x_+ + x_-)^\top w$; x_+ and x_- are any observations from the opposite classes for which constraint (3) becomes equality. By substituting the scalar product with a kernel function a non-linear score function is derived:

$$f(x) = \sum_{i=1}^n K(x_i, x)\alpha_i y_i + b, \tag{11}$$

where $b = -\frac{1}{2}\left\{\sum_{i=1}^n \alpha_i y_i K(x_i, x_+) + \sum_{i=1}^n \alpha_i y_i K(x_i, x_-)\right\}$; x_+ and x_- being any observations from the opposite classes for which $0 < \alpha < C$. The non-parametric score function (11) does not have a compact closed form representation. This may necessitate the use of graphical tools for its visualisation. Given the canonical representation $y_i f(x_i) = 1$ for the observations lying exactly on

the boundaries, the score of the separating function is $f(x) = 0$. Thus, SVM classifies a new firm x_k as solvent if $f(x_k) < 0$, and as insolvent if $f(x_k) > 0$. Note that the capacity c and the complexity term r are exogenous parameters to the model. c is the penalty weight of in-sample false classifications, r defines kernel complexity. Both values have to be fixed a priori.

References

Altman, E., September 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23 (4), 589–609.

Barlow, R. E., Bartholomew, J. M., Bremmer, J. M., Brunk, H. D., 1972. Statistical Inference Under Order Restrictions. John Wiley & Sons, New York, NY.

Cantor, R., Emery, K., Stumpp, P., 2006. Probability of default ratings and loss given default assessments for non-financial speculative-grade corporate obligors in the united states and canada.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, New York, NY.

Falkenstein, E., Boral, A., Carty, L., May 2000. Riskcalc for private companies: Moody's default model.

Fernandes, J. E., April 2005. Corporate credit risk modeling: Quantitative rating system and probability of default estimation. http://pwp.netcabo.pt/jed_fernandes/JEF_CorporateCreditRisk.pdf.

Gale, D., Kuhn, H. W., Tucker, A. W., 1951. Linear Programming and the Theory of Games, in Activity Analysis of Production and Allocation, T. C. Koopmans (ed.). John Wiley & Sons, New York, NY.

Härdle, W., Moro, R. A., Schäfer, D., 2005. Predicting Bankruptcy with Support Vector Machines in Statistical Tools in Finance, W. Härdle (ed.). Springer Verlag, Berlin.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer Verlag, Berlin.

Härdle, W., Simar, L., 2003. Applied Multivariate Statistical Analysis. Springer Verlag.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Verlag.

Horowitz, J. L., 2001. The Bootstrap. Vol. 5. Elsevier Science B. V.

Mammen, E., 1991. Estimating a smooth monotone regression function. Anals of Statistics 19, 724–740.

Manning, M. J., 2004. Exploring the relationship between credit spreads and default probabilities. Working Paper No. 225, Bank of England.

Martin, D., 1977. Early warning of bank failure: A logit regression approach. Journal of Banking and Finance 1, 249–276.

Mercer, J., 1909. Functions of positive and negative type and their connection

with the theory of integral equations. Philosophical Transactions of the Royal Society of London 209, 415–446.

Nadaraya, E. A., 1964. On estimating regression. Theory of Probability and its Applications 10, 186–190.

Ohlson, J., Spring 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 109–131.

Tikhomirov, V. M., January 1996. The evolution of methods of convex optimization. The American Mathematical Monthly 103 (1), 65–71.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York, NY.

Weyl, H., 1928. Gruppentheorie und Quantenmechanik. Hirzel, Leipzig.

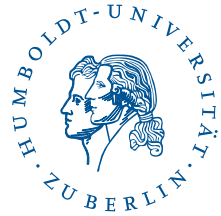
SFB 649 Discussion Paper Series 2007

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Trade Liberalisation, Process and Product Innovation, and Relative Skill Demand" by Sebastian Braun, January 2007.
002 "Robust Risk Management. Accounting for Nonstationarity and Heavy Tails" by Ying Chen and Vladimir Spokoiny, January 2007.
003 "Explaining Asset Prices with External Habits and Wage Rigidities in a DSGE Model." by Harald Uhlig, January 2007.
004 "Volatility and Causality in Asia Pacific Financial Markets" by Enzo Weber, January 2007.
005 "Quantile Sieve Estimates For Time Series" by Jürgen Franke, Jean-Pierre Stockis and Joseph Tadjuidje, February 2007.
006 "Real Origins of the Great Depression: Monopolistic Competition, Union Power, and the American Business Cycle in the 1920s" by Monique Ebell and Albrecht Ritschl, February 2007.
007 "Rules, Discretion or Reputation? Monetary Policies and the Efficiency of Financial Markets in Germany, 14th to 16th Centuries" by Oliver Volckart, February 2007.
008 "Sectoral Transformation, Turbulence, and Labour Market Dynamics in Germany" by Ronald Bachmann and Michael C. Burda, February 2007.
009 "Union Wage Compression in a Right-to-Manage Model" by Thorsten Vogel, February 2007.
010 "On σ−additive robust representation of convex risk measures for unbounded financial positions in the presence of uncertainty about the market model" by Volker Krätschmer, March 2007.
011 "Media Coverage and Macroeconomic Information Processing" by Alexandra Niessen, March 2007.
012 "Are Correlations Constant Over Time? Application of the CC-TRIG$_t$-test to Return Series from Different Asset Classes." by Matthias Fischer, March 2007.
013 "Uncertain Paternity, Mating Market Failure, and the Institution of Marriage" by Dirk Bethmann and Michael Kvasnicka, March 2007.
014 "What Happened to the Transatlantic Capital Market Relations?" by Enzo Weber, March 2007.
015 "Who Leads Financial Markets?" by Enzo Weber, April 2007.
016 "Fiscal Policy Rules in Practice" by Andreas Thams, April 2007.
017 "Empirical Pricing Kernels and Investor Preferences" by Kai Detlefsen, Wolfgang Härdle and Rouslan Moro, April 2007.
018 "Simultaneous Causality in International Trade" by Enzo Weber, April 2007.
019 "Regional and Outward Economic Integration in South-East Asia" by Enzo Weber, April 2007.
020 "Computational Statistics and Data Visualization" by Antony Unwin, Chun-houh Chen and Wolfgang Härdle, April 2007.
021 "Ideology Without Ideologists" by Lydia Mechtenberg, April 2007.
022 "A Generalized ARFIMA Process with Markov-Switching Fractional Differencing Parameter" by Wen-Jen Tsay and Wolfgang Härdle, April 2007.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".



023 "Time Series Modelling with Semiparametric Factor Dynamics" by Szymon Borak, Wolfgang Härdle, Enno Mammen and Byeong U. Park, April 2007.
024 "From Animal Baits to Investors' Preference: Estimating and Demixing of the Weight Function in Semiparametric Models for Biased Samples" by Ya'acov Ritov and Wolfgang Härdle, May 2007.
025 "Statistics of Risk Aversion" by Enzo Giacomini and Wolfgang Härdle, May 2007.
026 "Robust Optimal Control for a Consumption-Investment Problem" by Alexander Schied, May 2007.
027 "Long Memory Persistence in the Factor of Implied Volatility Dynamics" by Wolfgang Härdle and Julius Mungo, May 2007.
028 "Macroeconomic Policy in a Heterogeneous Monetary Union" by Oliver Grimm and Stefan Ried, May 2007.
029 "Comparison of Panel Cointegration Tests" by Deniz Dilan Karaman Örsal, May 2007.
030 "Robust Maximization of Consumption with Logarithmic Utility" by Daniel Hernández-Hernández and Alexander Schied, May 2007.
031 "Using Wiki to Build an E-learning System in Statistics in Arabic Language" by Taleb Ahmad, Wolfgang Härdle and Sigbert Klinke, May 2007.
032 "Visualization of Competitive Market Structure by Means of Choice Data" by Werner Kunz, May 2007.
033 "Does International Outsourcing Depress Union Wages? by Sebastian Braun and Juliane Scheffel, May 2007.
034 "A Note on the Effect of Outsourcing on Union Wages" by Sebastian Braun and Juliane Scheffel, May 2007.
035 "Estimating Probabilities of Default With Support Vector Machines" by Wolfgang Härdle, Rouslan Moro and Dorothea Schäfer, June 2007.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".