Estimating Probabilities of Default With Support Vector Machines

W. K. Härdle\textsuperscript{a}, R. A. Moro\textsuperscript{b}, D. Schäfer\textsuperscript{c}

\textsuperscript{a}CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

\textsuperscript{b}German Institute for Economic Research, Mohrenstraße 58, 10117 Berlin, Germany and CASE – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

\textsuperscript{c}German Institute for Economic Research, Mohrenstraße 58, 10117 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 Deutsche Bundesbank data. In particular we discuss the selection of variables and give a comparison with the logistic regression. The results demonstrate that the SVM has clear advantages over this method for all variables tested.

\textit{JEL classification:} C14; G33; C45

\textit{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). Similar to them is the hazard or survival analysis approach (Glennon and Nigro (2005)). 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 cumulative Gaussian) between PDs and the linear combination of predictors. For more details about rating models see Altman and Saunders (1998).

In this paper we are introducing an alternative way of assessing a 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 Section 5). Among previous works discussing the application of the

* The authors gratefully acknowledge that the project is co-financed by the Stiftung Geld und Währung. We thank Deutsche 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 Deutsche Bundesbank in Frankfurt. The work of R. A. Moro was financially supported by the German Academic Exchange Service (DAAD) and Deutsche 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 Deutsche 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 as well as to the discussants at the European Meeting of the Econometric Society in Vienna in 2006.

Email addresses: haerdle@wiwi.hu-berlin.de (W. K. Härdle), rmoro@diw.de (R. A. Moro), dschaefer@diw.de (D. Schäfer).
SVM to company rating are Martens, Baesens, van Gestel, and Vanthienen (2006) and Friedman (2002). The former gives a comparison of the SVM with a number of statistical methods on five different data sets, including one credit scoring data set. The authors, however, do not focus on the economic and financial application. The latter paper represents a brief manual on the results of implementing the Proximal SVM (PSVM, Fung and Mangasarian (2001)) for credit rating. In contrast to these two papers we give a comprehensive description of the rating technique, discuss the variable selection and the influence of the SVM parameters on the performance.

We focus on the cross-sectional analysis of the data as opposed to the time series approach (e.g. the Merton’s model, Merton (1974)). This is justified by the fact that time series data are seldomly available for small and medium size private companies that predominate in our sample. Moreover, assumptions of a stochastic process are prone to misspecification (Chen, Ju, Mazumdar, and Verma (2006)).

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 misclassifications are penalised. The method allows the use of kernel techniques (Hastie, Tibshirani, and Friedman (2001)) and, therefore, non-linear separating surfaces in contrast to classical DA, logit and probit models that rely on linear ones. The SVM can be considered as a generalised linear method. 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 outperforms 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 Logit model. The DA is a linear classification technique comparable to Logit, its performance is not higher than that of Logit and the classification functions are often indistinguishable (see Figure 1). This was the reason not to report the comparison results for DA and focus entirely on the comparison of the SVM and Logit. 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 Logit approach we automatically impose (through a mod-
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).

elling bias) a monotonic relationship between financial and economic indicators and PDs. A typical example is the monotonic decreasing relation for PD from the interest coverage ratio (Figure 2). This ratio represents the income from recurring operations as a percentage of interest expenses (see Appendix B) and indicates the ability of a firm to cover current costs of debt. The PD is the highest for the companies with the lowest ratios and the lowest for the companies with the highest debt coverage. However, this dependence is non-monotonic for such important indicators as the company size (K33) which is the logarithm of total assets, the net interest ratio (K24) representing net interest as a percentage of sales or the net income change (K21) which is computed in % relative to the net income in the previous year. K21 shows the profitability growth rate. In this case excessively high growth rates of the net income turn out to be as disadvantageous for a firm as low or negative growth rates. This is understandable because extremely high growth rates may not be sustainable and are likely to indicate a high volatility of income and, as a consequence, a higher PD. This result is completely in accordance with the Merton’s model (Merton (1974)), but we need a non-linear technique to discover such a dependence in the data. Non-monotonic dependencies in the data which were confirmed in the literature (Fernandes (2005), Manning (2004), Sjur, Hol, and van der Wijst (2007)) and are accounted for in the marketed models (Falkenstein, Boral, and Carty (2000)) are the reason for contemplating non-linear
One-year Probability of Default (Bundesbank Data)

Figure 2. One year PDs evaluated for several financial ratios on the Deutsche 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.

techniques as alternatives. Moreover, even if the general form of the dependence of PD from financial ratios is known, e.g. it is monotonously decreasing, the question is still open about the functional form of that dependence which can be highly non-linear.

In order to be able to capture non-linearity and non-monotonicity, the score function that can later be translated into 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 a 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 separates the two groups with the maximum distance (margin) between them. The score for \( x \) is computed as

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

In our case the kernel \( K(x, x_i) \) 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 $K(x, x_i)$; therefore, the score $f(x)$ is primarily defined by the observations that are close to $x$. The $n$ factors $\alpha_i$ (Lagrange multipliers) are the free parameters which are the solution of an SVM optimisation problem and have higher magnitudes for the observations at the boundary between the classes which are most relevant for classification. The Gaussian kernel (21) relies on two function complexity parameters that have to be set a priori, $r$ and $c$. The former is the radial basis coefficient that determines the minimum size of the data cluster or other data features that can be replicated with a kernel. The latter is the capacity coefficient determining the relative importance of in-sample misclassifications vs. the generalisation ability. See Appendix A for more details.

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 Deutsche Bundesbank data. Most firms enter the dataset only for few 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. Deutsche Bundesbank condenses the balance sheet information for each firm into 33 financial predictors. Some of those ratios are similar to regular ones reported in the literature, some are specific to German companies such as K25 (own funds and pension provisions ratio). Table 1 presents the summary statistics for each predictor.

We 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 $y = 1$ and its reports from previous years are excluded from analysis to avoid the confusion with very different times to default. For the rest (solvent) companies $y = -1$. The bankruptcy follows 0.5–3.5 years after the last reporting date.

Not all predictors are equally relevant for the SVM as well as Logit analy-
Table 1
Summary Statistics. $q_\alpha$ is an $\alpha$ quantile. IQR is the interquartile range.

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

sis. 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 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.

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

3 Variable Selection and Model Validation

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 or the error of type I is the percentage of insolvent companies among those classified as solvent, and a beta error or the error of type II is the percentage of solvent companies among those classified as insolvent. A classification method has a higher power if for a given alpha error it delivers a lower beta error.

The AR is the preferred accuracy indicator since it is invariant to any strictly monotonous transformation of the score and thus better reflects the pure performance of the classification model unaffected by any subsequent manipulations with the score. $\alpha$ and $\beta$ errors are, in contrast, sensitive to the threshold score value separating solvent from insolvent companies. The comparison of models 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 800 companies (400 solvent and 400 insolvent ones). An SVM and a Logit are trained on the training subsample and their accuracy is evaluated on the validation subsample. The relatively small size of training and validation sets is required by the bootstrap procedure to reduce the bias of the AR estimator.
Both subsets are extracted randomly from the pool of data for 1992–1998 without overlapping, i.e. they do not contain common observations, therefore, the validation is performed out-of-sample. This is essential since the SVM can perform in-sample classification without errors. The period 1992–1998 was selected because it contains relatively homogeneous data between the reunification of Germany in 1991 and the change of the accounting procedure of the Bundesbank in 1999.

One estimation of the AR is made in one cycle of the Monte Carlo experiment which contains 100 cycles or repetitions, i.e. we obtain 100 estimates of AR from which the distribution and the median AR can be derived. At each cycle we also compute the improvement in AR of SVM over Logit in order to derive the median improvement.

Since it is practically impossible to try all combinations of variables in order to choose the one that yields overall the best median AR, we need to apply a selection procedure. We will apply a well established backward variable selection procedure (BSP) and, in parallel, a forward selection procedure (FSP) for the two competitors: a logistic regression and SVM. The BSP starts with the full model which includes all variables from K1 to K33. At the first step one of the variables is consecutively excluded and the median AR of each reduced model is computed on bootstrapped subsamples. The model that has the highest median AR will be examined at the second step when one more variable is consecutively excluded and median 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 resulted in the highest median AR compared to other models with the same number of variables.

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 the expected $d$ is large.

The variables which provide the highest median AR for SVM (14 variables in BSP and 12 variables in FSP) and Logit (12 variables for BSP and FSP) are reported in Table 3. Figure 3 shows the median ARs for univariate, bivariate, trivariate, etc. SVM and Logit models. The variables are selected with the BSP for each model independently and may vary. The median for the SVM approach is for all models higher than for the alternative method, Logit.

The SVM model used for variable selection has the parameters $r = 5$ and $c = 10$ (see Appendix A) chosen a priori without optimisation. $r$ is the radial
Table 3

Variables included in the 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. In the BSP (FSP) an SVM selects 14 (12) variables, while Logit selects 12 (12) variables.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables included in the model, K*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>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</td>
</tr>
<tr>
<td>Backward selection</td>
<td></td>
</tr>
<tr>
<td>Logit (59.16)</td>
<td>. . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>SVM (61.11)</td>
<td>. . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Forward selection</td>
<td></td>
</tr>
<tr>
<td>Logit (59.05)</td>
<td>. . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>SVM (60.75)</td>
<td>. . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>

Figure 3. Median AR for the Logit and SVM models with a different number of predictors. At each step a model with the highest median AR is selected.

basis coefficient of the Gaussian kernel and $c$ is the capacity coefficient. The optimisation of $r$ and $c$ can further boost the SVM performance. Both control the model complexity. Higher values of $c$ and lower values of $r$ correspond to more complex models. When a model becomes too complex, accuracy drops dramatically. This is called overfitting and happens for excessively small values of $r$ and large values of $c$ (see Figure 4). Here the dependence of AR from $r$ and $c$ is illustrated for the SVM model with the 14 variables selected with the BSP (see Table 3). For moderate values of $r$ and $c$ the accuracy reaches its maximum, while for high $r$ and small $c$ the accuracy is lower. High values of $r$ correspond to near linear separating surfaces.
Figure 4. Left panel: the AR for different radial basis coefficients \( r \). The capacity coefficient is fixed at \( c = 10 \). Right panel: the AR for different capacity coefficients \( c \). The radial basis coefficient \( r \) is fixed at \( r = 5 \). The training and validation data sets are bootstrapped 100 times without overlapping from the homogeneous data for 1992–1998. Each training and validation set contains 800 observations (400 solvent and 400 insolvent companies).

When using an SVM a care should be taken to avoid too high complexities (overfitting) which deteriorate the generalisation ability of the SVM, i.e. the ability to perform classification on the unknown validation sample. The search of the \( r \) and \( c \) that deliver the highest median AR is nevertheless a simple two dimensional optimisation problem that can be solved by alternatively adjusting \( r \) and \( c \). In most cases an approximate solution with only few iteration steps guarantees an accuracy superior to that of the Logit model.

It is common to apply some transformation to the variables that display a non-linear dependence from PDs before using them with linear methods. By applying the transformation, it is argued, the assumptions of the Logit model are fulfilled and it can produce very good classification results comparable to more complex methods such as the SVM. In this case, however, humans will take over the job of statistical methods as they have to estimate suitable transformations. The SVM can estimate those non-linear transformations and, hence, the dependence of the output from the input automatically. Only the desired complexity of the SVM or, in other words, of the transformations has to be set a priori.

To make a fair comparison, we nevertheless transformed the data as it is common in practice, namely capped the outliers: 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 Logit procedure applied is a ro-
\textit{bust} Logit not sensitive to outliers. Here $IQR$ denotes the interquartile range. This variable transformation is actually used in the Deutsche Bundesbank to prepare the data.

To illustrate the significance of the selected variables a random standard normally distributed variable $K10$ was added which does not contain any information about bankruptcies. If used alone it has AR distributed around 0. Not surprisingly it is never included in the models with few variables. However, $K10$ is already included into most models after step 20 out of total 32 steps of the selection procedure. This indicates that the additional variables added at the last steps of the selection procedure do not contain new information compared to the already included ones and are as redundant as $K10$. A similar conclusion can be made by observing the falling prediction accuracy for the models with more than around 14 variables (see Figure 3). Essentially, new variables highly dependent on previous ones only inject noise without adding new information that leads to lower prediction accuracy.

The FSP does not have any advantage over the BSP, therefore the BSP was used as a variable selection procedure.

4 Comparison of Logistic Regression and SVM

Upon having chosen variables for each model we can compare model performance on the data from 1992–1998 and beyond that period. Since the selection procedure was done independently for a logistic regression and SVM, we do not introduce a bias against or in favour of any model. The number of variables in each model is different. The BSP results in 14 selected variables for an SVM and 12 variables for Logit (Table 3). Additionally, we are also reporting the comparison results for the Logit and SVM on the same set of variables to exclude the effects of the variable selection procedure.

A classification procedure consists of two steps: (i) model training and validation and (ii) the estimation of company classes. Training is the estimation of the free parameters of the model, in the case of the SVM, Lagrange multipliers. Validation is required to select variables and calibrate the model parameters fixed a priori, in the case of the SVM, $c$ and $r$. In our study we used the fixed values $c = 10$ and $r = 5$ but illustrated the effects of varying $c$ and $r$ in Figure 4. After training and validation the model is ready for classification. The classification accuracy can be assessed on the testing data which beside the company characteristics $x$ also contain the classes $y$, ‘solvent’ or ‘insolvent’.

In our tests we mimic the classification procedure applied in practice using historical data from 1992–1998 for training and validation and the data beyond
Table 4
Forecasting accuracy improvement for an SVM vs. Logit and the median AR for an SVM. 100 pairs of bootstrapped training testing samples are used. All figures are reported as a percentage of the ideal AR (100%).

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>SVM-Logit</th>
<th>AR (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>1995</td>
<td>0.96</td>
<td>61.79</td>
</tr>
<tr>
<td>1993</td>
<td>1996</td>
<td>3.05</td>
<td>60.73</td>
</tr>
<tr>
<td>1994</td>
<td>1997</td>
<td>3.62</td>
<td>59.28</td>
</tr>
<tr>
<td>1995</td>
<td>1998</td>
<td>2.30</td>
<td>59.18</td>
</tr>
<tr>
<td>1992-1998 after 1998</td>
<td>1.74</td>
<td>58.57</td>
<td></td>
</tr>
</tbody>
</table>

Three alternative procedures are used. The first one is training and testing on 100 pairs of the bootstrapped training and testing subsamples. This allows the estimation of the distribution of the AR and, hence, the median improvement of the AR for an SVM vs. Logit (Table 4). When trained on the data for 1992 and tested on that for 1995 with a bootstrap procedure the SVM outperforms Logit in 92% cases with a median improvement 2.30% and mean improvement 2.27%, measured as percentage of the AR for the ideal model. The results for other years are very similar.

Secondly, we repeated the comparison of the SVM vs. Logit for the identical set of nine variables, K06, K07, K09, K12, K29, K31, K05, K25 and K33, which were selected simultaneously both by the SVM and Logit either with
Table 5
Forecasting accuracy improvement for an SVM vs. Logit on common 9 variables (K06, K07, K09, K12, K29, K31, K05, K25 and K33) and the median AR for an SVM. 100 bootstrapped pairs of training and testing samples are used. All figures are reported as a percentage of the ideal AR (100%).

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>SVM-Logit</th>
<th>AR (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>1995</td>
<td>0.95</td>
<td>61.31</td>
</tr>
<tr>
<td>1993</td>
<td>1996</td>
<td>2.62</td>
<td>59.67</td>
</tr>
<tr>
<td>1994</td>
<td>1997</td>
<td>2.65</td>
<td>57.39</td>
</tr>
<tr>
<td>1995</td>
<td>1998</td>
<td>1.29</td>
<td>58.35</td>
</tr>
<tr>
<td>1992-1998 after 1998</td>
<td>0.88</td>
<td>58.18</td>
<td></td>
</tr>
</tbody>
</table>

Table 6
Forecasting accuracy improvement for an SVM vs. Logit and the AR estimated for an SVM. All data for the given years are used. All figures are reported as a percentage of the ideal AR (100%).

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
<th>SVM-Logit</th>
<th>AR (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>1995</td>
<td>2.39</td>
<td>60.98</td>
</tr>
<tr>
<td>1993</td>
<td>1996</td>
<td>4.66</td>
<td>60.98</td>
</tr>
<tr>
<td>1994</td>
<td>1997</td>
<td>5.14</td>
<td>59.49</td>
</tr>
<tr>
<td>1995</td>
<td>1998</td>
<td>3.98</td>
<td>59.97</td>
</tr>
</tbody>
</table>

the BSP or FSP (see Table 3). On these identical variables the SVM is still better than Logit but the improvement is slightly lower (Table 5). This can be attributed to the fact that the SVM can make a better use of the variables with a non-monotonic or non-linear relation with PDs which may be overlooked by Logit.

Thirdly, all observations for the period were used (Figure 6). When we switched from the testing design with 100 bootstrapped samples (Table 4) to training on the whole available data set without bootstrapping, as it will happen in practice (Table 6), we got an improvement in AR for the SVM. This is an indication that the risk of getting a non-representative sample is smaller for large samples. However, Logit, compared to the SVM, performs substantially worse without bootstrapping that is due to the higher model risk associated with it.
Figure 5 shows the comparison of an SVM and Logit in terms of $\alpha$ and $\beta$ errors. Such a plot illustrates model power. Since the represented dependence is very noisy because of a small number of insolvencies in the sample and the dependence of the $\alpha$ and $\beta$ errors (errors of type I and II) from the threshold score value separating solvent and insolvent companies, we applied a $k$-NN smoothing procedure with a window equal to $n/10$ or $1/10$th of all observations in the sample. The training data are from 1995, the testing data are from 1998. All observations for these years were used. Two conclusions can be made. Firstly, an SVM has a higher power since its curve lies below those for Logit. Secondly, many observations for the smallest $\alpha$ or type I error, more precisely 11%, when evaluated with an SVM lie in the area where no observations evaluated with Logit are located. This means that an SVM in contrast to Logit is able to locate the cluster of the companies with the lowest insolvency risk. In practice many of those would be the investment grade firms whose securities can be used in refinancing operations.

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 Logit would allow to issue more credit without increasing risk because of a better separation between solvent and insolvent companies.

5 Conversion of Scores into PDs

The conversion of rating scores into PDs provides us with a link to the existing rating classes reported by rating agencies such as Moody’s and S&P. Figure 8 illustrates this objective. In the Logistic model a sigmoid function is applied to estimate PD assuming the logistic distribution of the latent variable. However, such an assumption is often not compatible with reality. The company score, as it is computed by the SVM, 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 transformation from the score to PD.

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
Figure 5. The power of a model: beta errors as a function of alpha errors. An SVM has a higher power than 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, all observations from these years are used. An alpha error a.k.a. the error of type I is the percentage of insolvent companies among those classified as solvent, and a beta error or the error of type II is the percentage of solvent companies among those classified as insolvent.

is done using standard smoothing techniques in order to preliminary evaluate PDs for all \( i = 1, 2, \ldots, n \) observations of the training set:

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

where \( w(z - z_i) = \exp\left\{ \frac{(z - z_i)^2}{2h^2} \right\} \). The rank of the \( i \)-th company \( z_i = \text{Rank}\{ f(x_i) \} \) can be 1, 2, 3, \ldots up 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 \( PD(z) \). When \( h \to 0 \) no smoothing is performed and all \( \hat{PD}(z_i), i = 1, 2, \ldots, n, \) will be either 1 or 0; when \( h \to \infty \), all \( \hat{PD}(z_i) \) 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 \( \sum_{j=1}^{w(z - z_i)} w(z - z_j) \) which decay gradually as \( |z - z_i| \) grows. This differs from the most commonly used \( k \)-NN smoother that relies on the uniform weights \( \frac{1}{k} I(|z - z_i| < \frac{k}{2} + 1) \).
Figure 6. 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.

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 \( \tilde{PD}(z_i), i = 1, 2, \ldots, n \) is achieved at step two using the Pool Adjacent Violator (PAV) algorithm (Barlow, Bartholomew, Bremmer, and Brunk (1972)). Figure 6 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 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 7 shows the PD and the cumulative PD (CPD) curve estimated on the binary data represented as circles. The CPD was evaluated as
Figure 7. 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.

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

The first smoothing step can be dropped and only PAV algorithm applied. Although this produces gives asymptotically the same results, the PD estimates for the companies with the extreme scores can equal 100% in this case. This requires a correction, e.g. as a preliminary smoothing done at step one.

Figure 8 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 7. 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 by the European central banks as a collateral in their refinancing operations since they have PD less than 0.1%, the threshold level set by the European Central Bank.
Figure 8. One year probabilities of default estimated with an SVM for 1995.

Table 7
One year PDs of the rating classes represented in Figure 8, 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).

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

The classical Support Vector Machine applied in this paper is a statistical technique for binary classification that is based on the Tikhonov regularisation principle (Tikhonov (1963), Tikhonov and Arsenin (1977)). It is based on linear classifiers that simultaneously maximise the margin or the distance between the classes and minimise empirical risk related to misclassifications on a given data set (Vapnik (1995)).

Figure 9 illustrates the maximum margin classification for linearly separable and non-separable data in a two-dimensional case. The separating function generated by a linear SVM is

$$x^\top w + b = 0.$$  \hspace{1cm} (3)

Such a classification rule makes an SVM similar to Logit. $x_i$ is a $d \times 1$ vector of the characteristics of firm $i$, e.g. financial ratios described in Appendix B, whereas $d$ is the number of characteristics or variables used. $w$ is a $d \times 1$ vector of weights which determine the slope of the separating function. The scalar $b$ is a location parameter or a threshold.

The margin is the empirically estimated distance between the opposite classes of observations. In Figure 9 it is shown as the distance between the margin boundaries – the parallel lines located symmetrically on both sides of the separating function. In a perfectly separable case such as in Figure 9, left panel, no observations may lie in the margin zone and all observations must satisfy the constraints:
Figure 9. The separating hyperplane \( \mathbf{x}^\top \mathbf{w} + b = 0 \) and the margin in a linearly separable (left) and non-separable (right) case. Crosses denote solvent companies, zeros are the insolvent ones. The hyperplanes bounding the margin zone equidistant from the separating hyperplane are represented as \( \mathbf{x}^\top \mathbf{w} + b = 1 \) and \( \mathbf{x}^\top \mathbf{w} + b = -1 \). The misclassification penalty in the non-separable case is proportional to the distance \( \xi/\|\mathbf{w}\| \).

\[
\begin{align*}
x_i^\top \mathbf{w} + b &= 1 \quad \text{for } y_i = 1, \\
x_i^\top \mathbf{w} + b &\leq -1 \quad \text{for } y_i = -1.
\end{align*}
\]

The constraints ensure that the observations of the opposite classes lie on the opposite sides from the margin gap.

Misclassifications may occur if data are linearly non-separable as in Figure 9, right panel. Here the bold zero on the left-hand side of the separating line shows a solvent company that is classified as insolvent. SVM adjusts the weights \( \mathbf{w} \) and the location parameter \( b \) in such a way that the margin is maximised and the sum of misclassification errors \( \xi_i \) is minimised. \( \xi_i \geq 0 \) is also called a slack variable and is introduced to (4) and (5) to ensure that these constraints are satisfied. For any observation \( x_i \) the modified constraints must hold:

\[
\begin{align*}
x_i^\top \mathbf{w} + b &\geq 1 - \xi_i \quad \text{for } y_i = 1, \\
x_i^\top \mathbf{w} + b &\leq -1 + \xi_i \quad \text{for } y_i = -1.
\end{align*}
\]

For the representation (4) – (7) when 1 appears on the right hand side the margin equals \( 2/\|\mathbf{w}\| \). Here \( \|\mathbf{w}\| \) is the Euclidean norm or the length of vector \( \mathbf{w} \).

Only the observations lying on the margin boundaries or on the wrong side of the margin determine the SVM solution. These observations are marked with bold crosses and zeros. They are called support vectors, hence the name of the
method. This contrasts to Logit where all observations are used to derive the solution.

The primal minimisation problem to be solved is convex and has a unique solution:

\[
\min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} C_i \frac{\xi_i}{\|w\|}
\]

\[
\text{s.t. } y_i(x_i^\top w + b) \geq 1 - \xi_i,
\]

\[
\xi_i \geq 0.
\]

(8)  

(9)  

(10)

Here (6) and (7) are rewritten as one constraint. It is easier to see that the problem is convex if we rewrite the optimised functional in (8) as \( \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} C_i \xi_i \). The first term is the inverse margin, which equals \( 2/\|w\| \). By minimising this term we maximise the margin. The second term is a sum of weighted errors that are measured as a distance to a misclassified observation \( i \) from the boundary of its class \( \xi_i/\|w\| \). The parameters \( C_i \)'s which are called capacity represent the penalty weights of in-sample false classifications for each company observation \( i \). The SVM will give priority to the correct classification of the companies with higher \( C_i \)'s. Capacity is related to the width of the margin zone. Smaller \( C_i \)'s are associated with bigger margins. In our case \( C_i \) are set equal for the same class. In order to make SVMs comparable for a different number of observations and various ratios between solvent and insolvent companies we compute \( C_i \)'s as \( c/2n_+ \) for insolvent and \( c/2n_- \) for solvent companies. Here \( n_+ \), and \( n_- \) are the numbers of insolvent and solvent companies in the training set, \( c \) is the coefficient that is used to control the capacity of the SVM. In contrast to \( C_i \) it is invariant of the number of observations in the training data set and provides a convenient basis for comparing SVMs. This formulation implies that in a sample with mostly solvent companies, misclassifications of insolvent companies are given a higher weight. If the number of solvent and insolvent companies is the same, then \( C_i = c/n \).

The primal problem (8) – (10) rewritten in a Lagrangian formulation is

\[
\min_{w,b,\xi} \max_{\alpha,\mu} L_P = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} C_i \xi_i - \sum_{i=1}^{n} \alpha_i \{y_i(x_i^\top w + b) - 1 + \xi_i\} - \sum_{i=1}^{n} \xi_i \mu_i.
\]

The Karush-Kuhn-Tucker Conditions or first order optimality conditions are:
\[ \frac{\partial L_P}{\partial w_k} = 0 \iff w_k = \sum_{i=1}^{n} \alpha_i y_i x_{ik}, \quad k = 1, \ldots, d, \quad (11) \]

\[ \frac{\partial L_P}{\partial b} = 0 \iff \sum_{i=1}^{n} \alpha_i y_i = 0, \quad (12) \]

\[ \frac{\partial L_P}{\partial \xi_i} = 0 \iff C_i - \alpha_i - \mu_i = 0, \quad (13) \]

\[ \alpha_i \{ y_i (x_i^\top w + b) - 1 + \xi_i \} = 0, \]

\[ \mu_i \xi_i = 0, \]

\[ y_i (x_i^\top w + b) - 1 + \xi_i \geq 0, \]

\[ \alpha_i \geq 0, \]

\[ \mu_i \geq 0, \]

\[ \xi_i \geq 0, \]

where \( x_{ik} \) is the \( k \)-th characteristic of company \( i \). The dual problem is equivalent to the primal one since the minimised function is convex (Gale, Kuhn, and Tucker (1951)). By substituting equations (11) – (13) into the primal Lagrangian we derive the dual problem:

\[
\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j, \quad (14) \\
\text{s.t.} \quad 0 \leq \alpha_i \leq C_i, \quad \sum_{i=1}^{n} \alpha_i y_i = 0.
\]

The \( n \) Lagrange multipliers \( \alpha_i \) are the free parameters to be estimated. They represent the weights with which each observation influences the solution (see (16) and (20)). Those observations have higher weights which are harder to classify, i.e. which lie closer to the margin zone. On the contrary, the coefficients in the logistic regression are the weights assigned to each variable and can not be directly compared to Lagrange multipliers. Problem (14) can be equivalently expressed in a matrix notation:

\[
\max_{\alpha} \iota^\top \alpha - \alpha^\top H \alpha, \quad (15) \\
\text{s.t.} \quad 0 \leq \alpha \leq C, \quad y^\top \alpha = 0.
\]

Here \( \alpha \) is a vector of Lagrange multipliers, \( \iota \) is a vector of 1’s, \( y \) is a vector of company classes, +1 for solvent or −1 for insolvent ones and \( C \) here is a vector of the coefficients \( C_i \); all vectors are of the size \( n \times 1 \). The \( n \) components of the vector \( \alpha \) are obtained as the solution of the constrained maximisation problem (15). The \( i,j \)'th element of the matrix \( H \) is
\[ h_{ij} = y_i y_j x_i^\top x_j = y_i y_j \sum_{k=1}^{d} x_{ik} x_{jk}. \]

The reader who desires to construct an SVM independently may find the problem formulation in the matrix notation (15) more convenient. The SVM problem is a classical quadratic optimisation problem (Fletcher (1987)) that can be solved with numerous software packages such as Matlab (routines minq or minqdef) or using algorithms specifically developed for the SVM such as the Sequential Minimal Optimisation (SMO) (Platt (1998)).

Equation (11) of the KKT optimality conditions determines the weights \( w_k, k = 1, \ldots, d \) for the \( k \)-th characteristic of a company. By substituting (11) into (3) we derive the classification rule:

\[
f(x) = x^\top w + b = x^\top \sum_{i=1}^{n} \alpha_i y_i x_i + b = \sum_{i=1}^{n} \alpha_i y_i x_i^\top x + b
\]

\[
\begin{cases}
  f(x) < 0 & \Rightarrow x \text{ is solvent,} \\
  f(x) \geq 0 & \Rightarrow x \text{ is insolvent.}
\end{cases}
\]

To derive the coefficient \( b \) we will use the fact that the separating hyperplane \( f(x) = 0 \) (see Figure 9) lies equidistant from the hyperplanes bounding the classes:

\[
x_+^\top w + b = 1 \quad \text{for} \quad y_+ = 1, \quad (17)
\]

\[
x_-^\top w + b = -1 \quad \text{for} \quad y_- = -1, \quad (18)
\]

where \( x_+ \) is any support vector that lies on or ‘supports’ the hyperplane for \( y = 1 \) and \( x_- \) is any support vector that lies on the hyperplane for \( y = -1 \). Both \( x_+ \) and \( x_- \) have dimensions \( d \times 1 \). By summing (17) and (18) we derive:

\[
b = \frac{1}{2} \left( x_+^\top + x_-^\top \right) w = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i \left( x_+^\top + x_-^\top \right) x_i.
\]

To reduce numerical errors when training the SVM it is desirable to use averages over all \( x_+ \) and \( x_- \) instead of two arbitrary chosen support vectors.

Note that the classification rule (16) depends only on the scalar product \( x^\top x_i \), not on the original \( x \) and \( x_i \). This makes possible a ‘kernel trick’, i.e. an implicit mapping of low dimensional data into a highly dimensional Hilbert feature space and performing a linear classification there, e.g. with an SVM. A kernel transformation corresponds to (i) performing a variable transformation and (ii) taking a scalar product of transformed variables.
In practice $x^\top x_i$ in the SVM formulation (14) is replaced with a kernel function $K(x, x_i)$ which represents a scalar product in a feature space (Weyl (1928)). Then the elements of the matrix $H$ in (15) are $h_{ij} = y_i y_j K(x_i, x_j)$. A kernel function must satisfy the Mercer conditions (Mercer (1909)), i.e. be symmetric and semipositive definite as a scalar product. It can map data into infinitely dimensional spaces as in the case with Gaussian kernels. The number of Lagrange multipliers $\alpha_i$ – parameters to be estimated – is $n$ and can be large for large data sets. However, by selecting a small $C_i$’s and, hence, a narrow interval $[0, C_i]$ in which $\alpha$ may vary we can avoid overfitting and extremely high complexities of the SVM classifier.

Figure 10 shows a simple mapping example. The quadratic kernel function

$$K(x, x_i) = (x^\top x_i)^2$$

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: $\tilde{x}_1 = x_1^2$, $\tilde{x}_2 = \sqrt{2} x_1 x_2$, and $\tilde{x}_3 = x_2^2$. The transformation from a two dimensional data space into a three dimensional feature space is $\Psi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)^\top$. However, we do not need to know the transformation $\Psi$ explicitly and can equivalently apply the kernel $K(x_1, x_2) = \Psi(x_1, x_2)^\top \Psi(x_1, x_2)$ to represent quadratic dependencies between input variables. The data separable in the data space of $x_1$ and $x_2$ only with a quadratic function will be separable in the feature space of $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ 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 is growing fast with the dimension of the data $d$ and the degree of the polynomial kernel making a direct data transformation not feasible and the advantages of the data transformation via a kernel obvious.

By substituting the scalar product in (16) with a kernel function a non-linear score function $f$ is derived:

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

where, by analogy with (19):

$$b = -\frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i \{K(x_+, x_i) + K(x_-, x_i)\}.$$

The non-parametric score function (20) does not have a compact closed form representation.
In our study we applied an SVM with an anisotropic Gaussian or radial basis kernel

\[ K(x, x_i) = \exp \left\{ -\frac{(x - x_i) \top r^{-2}\Sigma^{-1}(x - x_i)}{2} \right\}, \tag{21} \]

where \( r \) is a coefficient and \( \Sigma \) is a scaling matrix, which in our case is a variance-covariance matrix of the training characteristics \( x \). The \( k_1, k_2 \)-th element of the matrix is:

\[
\sigma_{k_1,k_2} = \frac{1}{n} \sum_{i=1}^{n} \left( x_{i,k_1} - \frac{1}{n} \sum_{j=1}^{n} x_{j,k_1} \right) \left( x_{i,k_2} - \frac{1}{n} \sum_{j=1}^{n} x_{j,k_2} \right).
\]

Here \( \sigma_{k_1,k_2} \) is the covariance between two financial ratios \( x_{k_1} \) and \( x_{k_2} \), e.g. the operating profit margin (K2) and the equity ratio (K8) as in Figure 1. \( \Sigma \) is used to bring all variables to the same scale and exclude the excessive influence of the variables with high variance. The ability to use differently scaled data explains the term ‘anisotropic’ in the kernel name. Before computing \( \Sigma \) and training an SVM the outliers should be processed, e.g. capped. 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)). Note that only the capacity \( C_i \) and the complexity coefficient \( r \) are to be set a priori. The Lagrange multiplies are the free parameters that are computed when training the SVM.

The SVM has a substantial advantage in comparison to the logistic regression with transformed variables, namely, it does not require to specify the transformation but estimates it from a broad range of possible ones defined implicitly by the kernel function type and the SVM capacity coefficient. This advantage
of the SVM is fully revealed when data are new or the relevance of well known transformations must be tested.

8 Appendix B

K1 pre-tax profit margin (profitability): earnings before taxes as a percentage of sales.
K2 operating profit margin (profitability): operating income as a percentage of sales.
K3 cash flow ratio (liquidity): cash flow as a percentage of net sales.
K4 capital recovery ratio (liquidity): net receipts as a percentage of the invested capital.
K5 debt cover (liquidity): net receipts as a percentage of accounts payable minus cash and cash equivalents.
K6 days receivable (activity): accounts receivable in proportion to net sales times 360 days.
K7 days payable (activity): accounts payable in proportion to net sales times 360 days.
K8 equity ratio (financing): own capital as a percentage of total capital.
K9 adjusted equity ratio (financing): own and shareholder’s capital as a percentage of total capital.
K10 random variable (test variable): a standard normally distributed variable.
K11 net income ratio (profitability): income excluding extraordinary items as a percentage of sales.
K12 leverage ratio (leverage): guarantees and other commitments as a percentage of total assets.
K13 debt ratio (liquidity): financial needs for production and investment as a percentage of total assets.
K14 liquidity ratio (liquidity): liquid assets as a percentage of total assets.
K15 liquidity 1 (liquidity): liquid assets as a percentage of the short term debt.
K16 liquidity 2 (liquidity): liquid assets and accounts receivable as a percentage of the short term debt.
K17 liquidity 3 (liquidity): current assets as a percentage of the short term debt.
K18 short term debt ratio (financing): short term debt as a percentage of total assets.
K19 inventories ratio (investment): inventories as a percentage of total assets.
K20 fixed assets ownership ratio (leverage): own capital as a percentage of fixed assets.
K21 net income change (growth): net income change in % in comparison to the previous year.
K22 own funds yield (profitability): earnings before taxes as a percentage of own capital.
K23 capital yield (profitability): earnings before interest expenses and taxes as a percentage of total assets.
K24 net interest ratio (cost structure): net interest as a percentage of sales.
K25 own funds and pension provisions ratio (financing): own capital and pension provisions as a percentage of total assets.
K26 tangible asset growth (growth): the addition of tangible assets in %.
K27 own funds and provisions ratio (financing): own capital and provisions as a percentage of total assets.
K28 tangible asset retirement (growth): retirement of tangible assets in %.
K29 interest coverage ratio (cost structure): income from recurring operations as a percentage of interest expenses.
K30 cash flow ratio (liquidity): earnings after depreciation and amortisation, taxes and provisions as a percentage of sales.
K31 days of inventories (activity): inventories in proportion to sales times 360 days.
K32 current liabilities ratio (financing): short term accounts payable as a percentage of short and long term debt.
K33 logarithm of total assets (size).

References


support vector machines. technical report msr-tr-98-14, Microsoft Research.