SFB 649 Discussion Paper 2008-005

The Default Risk of Firms Examined with Smooth Support Vector Machines

Wolfgang Härdle*
Yuh-Jye Lee**
Dorothea Schäfer***
Yi-Ren Yeh**

* Humboldt-Universität zu Berlin, Germany
** National University of Science and Technology Taipeh, Taiwan
*** 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

The Default Risk of Firms Examined with Smooth Support Vector Machines*

Wolfgang Härdle†
Yuh-Jye Lee‡
Dorothea Schäfer§
Yi-Ren Yeh¶

December 3, 2007

Abstract

In the era of Basel II a powerful tool for bankruptcy prognosis is vital for banks. The tool must be precise but also easily adaptable to the bank's objections regarding the relation of false acceptances (Type I error) and false rejections (Type II error). We explore the suitability of Smooth Support Vector Machines (SSVM), and investigate how important factors such as selection of appropriate accounting ratios (predictors), length of training period and structure of the training sample influence the precision of prediction. Furthermore we show that oversampling can be employed to gear the tradeoff between error types. Finally, we illustrate graphically how different variants of SSVM can be used jointly to support the decision task of loan officers.
Keywords: Insolvency Prognosis, SVMs, Statistical Learning Theory, Non-parametric Classification

JEL: G30; C14; G33; C45

*We thank SFB 649 at Humboldt University of Berlin for providing the data.

†CASE – Center for Applied Statistics and Economics, Humboldt–Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, haerdle@wiwi.hu-berlin.de

‡National Taiwan University of Science and Technology, Department of Computer Science Information Engineering, Taipei 106, Taiwan, yuh-jye@mail.ntust.edu.tw

§German Institute for Economic Research (DIW)– Berlin, Mohrenstrasse 58, 10117 Berlin, Phone +49-30 89789-162, Fax +49-30 89789-104 Email: dschaefer@diw.de.

¶National Taiwan University of Science and Technology, Department of Computer Science Information Engineering, Taipei 106, Taiwan, D9515009@mail.ntust.edu.tw

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

1 Introduction

Default prediction is at the core of credit risk management and has therefore always obtained special attention. It has become even more important since the Basel Committee on Banking Supervision (Basel - II) established borrowers' rating as the crucial criterion for minimum capital requirements of banks. The methods for generating rating figures have developed significantly over the last 10 years (Krahnen and Weber, 2001). The rational behind the increased sophistication in predicting the borrowers default risk is the aim of banks to minimize their cost of capital.

In this paper we intend to contribute to increased sophistication by exploring the predicting power of Smooth Support Vector Machines (SSVM). SSVM is a variant of the basic SVM. The working principle of SVMs in general is described very easily. Imagine a bunch of observations in distinct classes such as balance sheet data from solvent and insolvent companies. Assume that the observations are such that they can not be separated by a linear function. Rather than fitting nonlinear curves to the data, SVMs handle this problem by using a specific transformation function, the kernel function, that maps the data from the original space into a higher dimensional space where a hyperplane can do the separation linearly. The constrained optimization calculus of SVM gives a unique optimal separating hyperplane and adjusts it in such a way that the elements of distinct classes possess the largest distance to the hyperplane. By re-transforming the separating hyperplane into the original space of variables, the typical non-linear separating function emerges (Vapnik, 1995). The main difference of SSVMs and SVMs is the following. With SSVMs the transition from the class of solvent to the class of insolvent companies occurs in a smooth way while the basic SVMs employ for separation a strictly defined cut off value.

Our aim is threefold when using SSVM. First, we examine the power of SSVM in predicting firms' defaults, second, we investigate how important factors, that are exogenous to the model such as selecting the appropriate set of accounting ratios, length of training period and structure of the training sample, influence the precision, and third, we explore how oversampling and downsampling affects the tradeoff between Type I and Type II errors. In addition, we illustrate graphically how loan officers can benefit from considering jointly the prediction results of different SSVM-variants.

There are basically three distinct approaches to predict the risk of default: option theory-based approaches, parametric models and non-parametric methods. While the first class relies on the rule of no arbitrage the latter both are based purely on statistic principles. The popular Merton (1974) model treats the firm's equity as the underlying asset of a call option held by shareholders. In case of insolvency shareholders deny exercising. The probability of default is derived from an adapted Black-Scholes formula. Later, several authors – e.g. Longstaff and Schwartz (1995), Mella-Barral and Perraudin (1997), Leland and Toft (1996) and Zhou (2001), to name only a few – propose variations to ease the strict assumptions on the structure of the data, imposed by the Merton model. These approaches are frequently denoted as structural models. However, the most challenging requirement is the knowledge of market values of debt and equity. This precondition is a severe obstacle for adequately using the Merton model as it is satisfied only in a small minority of cases.

Parametric statistical models can be applied to any type of data, whether they are market or book based. The first model introduced was discriminant analysis (DA) for univariate (Beaver, 1966) and multivariate models

(Altman, 1968). After DA usage of the logit and probit approach for predicting default were proposed in Martin (1977) and Ohlson (1980). These approaches rely on an a priori assumed functional dependence between risk of default and predictor. DA requires a linear functional dependence, or a preshaped polynomial functional dependence in advanced versions. Logit and probit tools work with monotonic relationships between default event and predictors such as accounting ratios. However, such restrictions often fail to meet the reality of observed data. This fact makes it clear, that there is a need for an approach that, in contrast to conventional methods, relaxes the requirements on data and/or lower the dependence on heuristics. Nonlinear classification methods such as Support Vector Machines (SVMs) or neural networks are strong candidates to meet these demands as they go beyond conventional discrimination methods. Tam and Kiang (1992) and Altman, Marco and Varetto (1994) focus on neural networks. In contrast, we concentrate on SVMs exclusively.

SVM is a relatively new technique and builds on the principles of statistical learning theory. It is easier to handle compared to neural networks. Furthermore, SVMs have a wider scope of application as the class of SVM models includes neural networks (Schölkopf and Smola, 2002). The power of the SVM-technology becomes evident in a situation as depicted in Figure 1 where operating profit margin and equity ratio are used as explanatory variables. A separating function similar to a parabola (in dark blue) appears in the $n = 2$-dimensional space. The accompanying pink lines represent the margin boundaries whose shape and location determine the distance of elements from the separating function. In contrast, the Logit approach and discriminant analysis (DA) yield the (white) linear separating function.

Selecting the best accounting ratios for executing the task of predicting

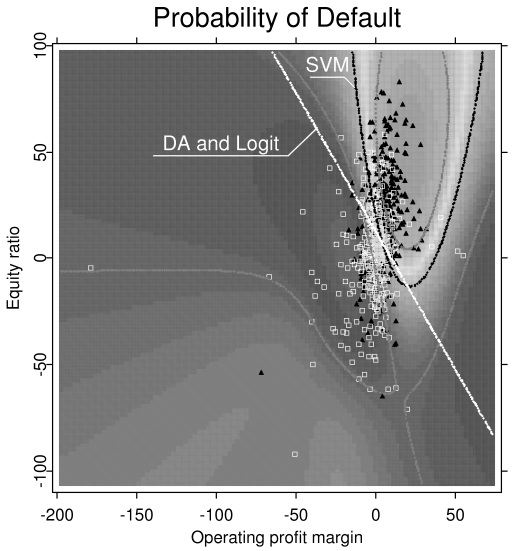

Figure 1: SVM-separating function (dark blue) with margin in a two-dimensional space.

is an important issue in practice but has not received appropriate attention in research. We address this issue of how important the chosen set of predictors is for the outcome. For this purpose we explore the prediction potential of SSVM within a two step approach. First, we derive alternative sets of accounting ratios that are used as predictors. The benchmark set comes from Chen, Härdle and Moro (2006). A second set is defined by a 1-norm SVM, and the third sets is based on the principle of adding only those variables that contain the most contrary information with respect to an a priori chosen initial set. We call the latter procedure the incremental forward selection of variables. As a result we are working with three variants of SSVM. In the second step then, these variants are compared with respect to their prediction power.

The analysis is built on 30 accounting ratios of 20,000 solvent and 1,000 insolvent German firms. Our finding show that the SSVM-types have an overall good performance with hit ratios ranging from 59.8 % to 74.1 % (mean). The SSVM based on predictors selected by the 1-norm SVM clearly outperforms the SSVM on the basis of incremental forward selection. It is also found that oversampling influences the trade off between Type I and Type II errors. Thus, oversampling can be used to make the relation of the two error types an issue of bank policy.

The rest of the paper is organized as follows. The following sections describe the data and the SVM methodology. In Section 3 and 4 the variable selection technique, the estimation procedure and the findings are explained. Section 5 illustrates some results. Section 6 concludes.

2 Data and measures of accuracy

In this study of the potential virtues of SVMs in insolvency prognosis the CreditReform database is employed. The database consists of 20,000 financially and ecomically solvent and 1,000 insolvent German firms observed once in the period of 1997 to 2002. Although the firms were randomly selected, accounting information dates most frequently in 2001 and 2002. Approximately 50% of the observations are coming from this period. The industry distribution of the insolvent firms is as follows: construction 39.7%, manufacturing 25.7%, wholesale & retail trade 20.1%, real estate 9.4% and others 5.1%. The latter includes businesses in agriculture, mining, electricity, gas and water supply, transport and communication, financial intermediation social service activities and hotels and restaurants. The 20000 solvent companies belong to manufacturing (27.4%), wholesale & retail trade (24.8%), real estate (16.9%), construction (13.9%) and the others (17.1%). There is only low coincidence between the insolvent group of “others” and the solvent

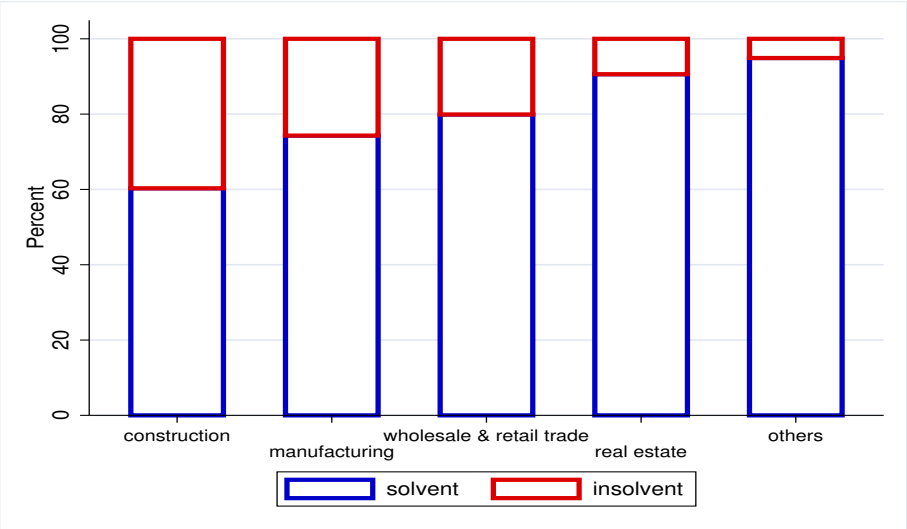

Figure 2: Portions of solvent and insolvent firms per industry

one. The latter comprises many firms of additional industries such as publication administration and defense, education and health. Figure 2 shows the portions of solvent and insolvent firm per industry. A set of balance sheet and income statement items describes each company. The ones we use for further analysis are described below:

- AD (Amortization and Depreciation)
- AP (Accounts Payable)
- AR (Account Receivable)
- CA (Current Assets)
- CASH (Cash and Cash Equivalents)
- CL (Current Liabilities)
- DEBT (Debt)

- EBIT (Earnings before Interest and Tax)
- EQUITY (Equity)
- IDINV (Growth of Inventories)
- IDL (Growth of Liabilities)
- INTE (Interest Expense)
- INV (Inventories)
- ITGA (Intangible Assets)
- LB (Lands and Buildings)
- NI (Net Income)
- OI (Operating Income)
- QA (Quick Assets)
- SALE (Sales)
- TA (Total Assets)
- TL (Total Liabilities)
- WC (Working Capital (=CA-CL))

Firms appear in the database several times in different years. However, each financial information from a particular year is treated as a single observation. The data of the insolvent firms are collected two years prior to insolvency. The firm size is measured by total assets. We construct 28 ratios to condense the balance sheet information (see Table 1). However, before dealing with the Creditreform dataset, some firms whose behavior is very different from other ones are filtered out in order to make the dataset more compact. The data preprocessing procedure is described as follows:

Table 1: The Definitions of Accounting Ratios used in the analysis

Variable	Ratio	Indicator for
X1	NI/TA	Profitability
X2	NI/SALE	Profitability
X3	OI/TAS	Profitability
X4	OI/SALE	Profitability
X5	EBIT/TA	Profitability
X6	(EBIT+AD)/TA	Profitability
X7	EBIT/SALE	Profitability
X8	EQUITY/TA	Leverage
X9	(EQUITY-ITGA)/	Leverage
	(TA-ITGA-CASH-LB)	Leverage
X10	CL/TA	Leverage
X11	(CL-CASH)/TA	Leverage
X12	TL/TA	Leverage
X13	DEBT/TA	Leverage
X14	EBIT/INTE	Leverage
X15	CASH/TA	Liquidity
X16	CASH/CL	Liquidity
X17	QA/CL	Liquidity
X18	CA/CL	Liquidity
X19	WC/TA	Liquidity
X20	CL/TL	Liquidity
X21	TA/SALE	Activity
X22	INV/SALE	Activity
X23	AR/SALE	Activity
X24	AP/SALE	Activity
X25	Log(TA)	Size
X26	IDINV/INV	Growth
X27	IDL/TL	Growth
X28	IDCASH/CASH	Growth

1. We exclude those firms whose total asset sizes are not in the range of 10^5 to 10^7 EUR and the year of 1996 (remaining insolvent: 967; solvent: 15,834).

2. In order to compute the accounting ratios AP/SALE, OI/TA, TL/TA, CASH/TA, IDINV/INV, INV/SALE, EBIT/TA and NI/SALE, we remove the firms with zero denominators (remaining insolvent: 816; solvent 11,005).

3. We drop outliers, that is, in the insolvent class the firms with the extreme values of financial indices will be removed (remaining insolvent: 811; solvent: 10468).

After this preprocessing, there are 11,279 firms in the dataset, including 811 insolvent and 10,468 solvent firms. In the following analysis, we focus on the revised dataset.

The performance of the SSVMs is evaluated on the basis of three measures of accuracy: Type I error rate (in %), Type II error rate (in %) and total error rate (in %). The hit ratio is 100-total error rate (in %). Type I error is the ratio of the number of predicting falsely insolvent companies to the number of insolvent companies. Similarly, the Type II error is the ratio of the number of predicting falsely solvent companies to the number of solvent companies. The error-types are defined as follows

- Type I error rate = FN/(FN+TP)×100%,
- Type II error rate= FP/(FP+TN)×100%,
- Total error rate =(FN+FP)/(TP+TN+FP+FN)×100%,

where

True Positive (TP):	Predict insolvent firms as insolvent ones
False Positive (FP):	Predict solvent firms as insolvent ones
True Negative (TN):	Predict solvent firms as solvent ones
False Negative (FN):	Predict insolvent firms as solvent ones.

Table 2 explains the terms used in the definition of error rates.

Table 2: Matrix for possibilities of prediction

		Predicted class	
		Positive	Negative
Actual	Positive	*True Positive* (TP)	*False Negative* (FN)
Class	Negative	*False Positive* (FP)	*True Negative* (TN)

3 SVM-Methodology

In recent years, the so-called support vector machine (SVM) which has its roots in the theory of statistical learning (Burges, 1998; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995) has become one of the most successful learning algorithms for classification as well as for regression (Drucker, Burges, Kaufman, Smola and Vapnik, 1997; Mangasarian and Musicant, 2000; Smola and Schölkopf, 2004; Lee, Hsieh and Huang, 2005). Some features of SVM make it particularly attractive for predicting the default risk of firms. SVMs are a non-parametric technique that learn the separating function from the data, they are based on a sound theoretical concept, do not require a particular distribution of the data, and deliver an optimal solution for the expected loss from misclassification. SVMs estimate the separating hyperplane between defaulting and non-defaulting firms under the constraint of a maximal margin between the two classes, see Vapnik (1995) and Schölkopf and Smola (2002).

SVMs can be formulated differently. However, in all variants either a constrained minimization problem or an unconstrained minimization problem is solved. The objective function in these optimization problems basically consists of two parts: a misclassification penalty part which stands for *model bias* and a regularization part which controls the *model variance*. We briefly introduce three different models: the smooth support vector machine (SSVM) (Lee and Mangasarian, 2001), the smooth support vector machine with reduced kernel technique (RSVM) and the 1-norm SVM. The SSVM will be used for classification and the 1-norm SVM will be employed to variable selection. The RSVM is applied for oversampling in order to mitigate the computational burden due to increasing the number of instances in the training sample.

3.1 Smooth support vector machine

The aim of SVM is to find the separating hyperplane with the largest margin from the training data. This hyperplane is "optimal" in the sense of statistical learning: it strikes a balance between overfitting and underfitting. Overfitting means that the classification boundary is too curved and therefore has less ability to classify unseen data correctly. Underfitting on the other hand, gives a too simple classification boundary and leaves too many misclassified observations (Vapnik, 1995). Given a training dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input data and $y_i \in \{-1, 1\}$ is the corresponding class label, striking the balance via an optimal hyperplane can be achieved by a two step procedure. First, a convex optimization problem is solved and, second, the kernel trick is applied. Below we describe how both steps lead to the SSVM we are aiming at. For reasons of reference we start the description by the convex optimization

problem of a conventional SVM:

$$\begin{array}{ll} \min\limits_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} & C\sum_{i=1}^{n}\xi_i+\frac{1}{2}\|w\|_2^2 \\ \text{s.t.} & y_i(w^\top \mathbf{x}_i+b)+\xi_i\geq 1 \\ & \xi_i\geq 0, \quad \text{for } i=1,2,\ldots,n, \end{array} \tag{1}$$

has to be solved where C is a positive parameter controlling the tradeoff between the training error (model bias) and the part of maximizing the margin (model variance) that is achieved by minimizing $\|w\|_2^2$. In contrast to the basic SVM of (1), a smooth support vector machine (SSVM) minimizes the square of the slack vector ξ with weight $\frac{C}{2}$. In addition, SSVM appends the term $\frac{b^2}{2}$ to the objective to be minimized results in the following minimization problem:

$$\begin{array}{ll} \min\limits_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} & \frac{C}{2}\sum_{i=1}^{n}\xi_i^2+\frac{1}{2}(\|w\|_2^2+b^2) \\ \text{s.t.} & y_i(w^\top \mathbf{x}_i+b)+\xi_i\geq 1 \\ & \xi_i\geq 0, \quad \text{for } i=1,2,\ldots,n. \end{array} \tag{2}$$

At a solution of (2), ξ is given by $\xi_i=\{1-y_i(w^\top \mathbf{x}_i+b)\}_+$ for all i where the *plus* function x_+ is defined as $x_+=\max\{0,x\}$. Thus, we can replace ξ_i in (2) by $\{1-y_i(w^\top \mathbf{x}_i+b)\}_+$. This will convert the problem (2) into an unconstrained minimization problem as follows:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n}\{1-y_i(w^\top \mathbf{x}_i+b)\}_+^2+\frac{1}{2}(\|w\|_2^2+b^2). \tag{3}$$

This formulation reduces the number of variables from $d+1+n$ to $d+1$. However, the objective function to be minimized is not twice differentiable which precludes the use of a fast Newton method. In SSVM, the plus function x_+ is approximated by a smooth $p-function$, $p(x,\alpha)=x+\frac{1}{\alpha}\log(1+e^{-\alpha x}), \alpha>0$. By replacing the plus function with a very accurate smooth approximation p-function gives the smooth support vector machine formulation:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n}p(\{1-y_i(w^\top \mathbf{x}_i+b)\},\alpha)^2+\frac{1}{2}(\|w\|_2^2+b^2), \tag{4}$$

where $\alpha > 0$ is the smooth parameter. The objective function in problem (4) is strongly convex and infinitely differentiable. Hence, it has a unique solution and can be solved by using a fast Newton-Armijo algorithm. In the second step, this formulation can be extended to the nonlinear SVM by using the kernel trick as follows:

$$\min_{(u,b)\in\mathbb{R}^{n+1}} \frac{C}{2}\sum_{i=1}^{n} p([1-y_i\{\sum_{j=1}^{n} u_j K(\mathbf{x}_i,\mathbf{x}_j)+b\}],\alpha)^2 + \frac{1}{2}(\|u\|_2^2+b^2), \quad (5)$$

where $K(\mathbf{x}_i,\mathbf{x}_j)$ is a kernel function. This kernel function represents the inner product of $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ where ϕ is a certain mapping from input space $\mathbb{R}^d$ to a feature space $\mathcal{F}$. We do not need to know the mapping ϕ explicitly. This is the so-called kernel trick. The nonlinear SSVM classifier can be expressed in matrix form as follows:

$$\sum_{u_j\neq 0} u_j K(A_j^\top,\mathbf{x}) + b = K(\mathbf{x},A^\top)u + b \quad (6)$$

where $A = [\mathbf{x}_1^\top;\cdots;\mathbf{x}_n^\top]$ and $A_j = \mathbf{x}_j^\top$

3.2 Reduced Support Vector Machine

In large scale problems, the full kernel matrix will be very large so it may not be appropriate to use the full kernel matrix when dealing with (5). In order to avoid facing such a big full kernel matrix, we brought in the reduced kernel technique (Lee and Huang, 2007). The key idea of reduced kernel technique is randomly selecting a portion of data as to generate a thin rectangular kernel matrix. Then it uses this much smaller rectangular kernel matrix to replace the full kernel matrix. In the process of replacing the full kernel matrix by a reduced kernel, we use the Nyström approximation (Smola and Schölkopf, 2000) for the full kernel matrix:

$$K(A,A^\top) \approx K(A,\tilde{A}^\top)K(\tilde{A},\tilde{A}^\top)^{-1}K(\tilde{A},A^\top), \quad (7)$$

where $K(A, A^\top) = K_{n\times n}$, $\tilde{A}_{\tilde{n}\times d}$ is a subset of A and $K(A, \tilde{A}) = \tilde{K}_{n\times\tilde{n}}$ is a reduced kernel. Thus, we have

$$K(A, A^\top)u \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}^\top, A)u = K(A, \tilde{A}^\top)\tilde{u}. \quad (8)$$

where $\tilde{u} \in \mathbb{R}^{\tilde{n}}$ is is an approximated solution of u via the reduced kernel technique. The reduced kernel method constructs a compressed model and cuts down the computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(\tilde{n}^3)$. It has been shown that the solution of reduced kernel matrix approximates the solution of full kernel matrix well. The SVM with the reduced kernel is called RSVM.

3.3 1-norm Support Vector Machine

The 1-norm support vector machine replaces the regularization term $\|w\|_2^2$ in (1) with the ℓ_1-norm of w. The ℓ_1-norm regularization term is also called the LASSO penalty (Tibshirani, 1996). It tends to shrink the coefficients w's towards zeros in particular for those coefficients corresponding to redundant noise features (Zhu, Rosset, Hastie and Tibshirani, 2003; Williams and Seeger, 2001). This nice feature will lead to a way to select the important ratios in our prediction model. The formulation of 1-norm SVM is described as follows:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \quad & C\sum_{i=1}^n \xi_i + \|w\|_1 \\ \text{s.t.} \quad & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n. \end{aligned} \quad (9)$$

The objective function of (9) is a piecewise linear convex function. We can reformulate it as the following linear programming problem:

$$\begin{aligned} \min_{(w,s,b,\xi)\in\mathbb{R}^{d+d+1+n}} \quad & C\sum_{i=1}^n \xi_i + \sum_{j=1}^n s_j \\ \text{s.t.} \quad & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & -s_j \leq w_j \leq s_j, \text{ for } j = 1, 2, \ldots, d, \\ & \xi_i \geq 0, \text{ for } i = 1, 2, \ldots, n, \end{aligned} \quad (10)$$

where s_j is the upper bound of the absolute value of w_j. At the optimal solution of (10) the sum of s_j is equal to $\|w\|_1$.

The 1-norm SVM can generate a very sparse solution w and lead to a parsimonious model. In a linear SVM classifier, solution sparsity means that the separating function $f(\mathbf{x}) = w^\top \mathbf{x} + b$ depends on very few input attributes. This characteristic can significantly suppress the number of the nonzero coefficients w's, especially when there are many redundant noise features (Fung and Mangasarian, 2004; Zhu et al., 2003). Therefore the 1-norm SVM can be a very promising tool for the variable selection tasks. We will use it to choose the important financial indices for our bankruptcy prognosis model.

4 Selection of Accounting ratios

In principle any possible combination of accounting ratios could be used as explanatory variables in a bankruptcy prognosis model. Therefore, appropriate performance measures are needed to gear the process of selecting the ratios with the highest separating power. In Chen et al. (2006) Accuracy Ratio (AR) and Conditional Information Entropy Ratio (CIER) determine the selection procedure's outcome. It turned out that the ratio "accounts payable divided by sales", X24 (AP/SALE), has the best performance values for a univariate SVM model. The second selected variable was the one combined with X24 that had the best performance of a bivariate SVM model. This is the analogue of forward selection in linear regression modeling. If one keeps on adding new variables one typically observes a declining change in improvement. This was also the case in that work where the performance indicators started to decrease after the model included eight variables. The described selection procedure is quiet lengthy, since there are at least 216 accounting ratio combinations to be considered. We will not employ the procedure here but use the chosen set of 8 variables as the benchmark set V1. Table 3 presents V1 in the first column.

We propose two different approaches for variable selection that will simplify the selection procedure. The first one is based on 1-norm SVM introduced in the section 3.2. The SVM was applied to the period from 1997 through 1999. We selected the variables according to the size of the absolute values of the coefficients w from the solution of the 1-norm SVM. Table 3 displays the eight selected variables as V2. We obtain 8 variables out of 28. Note that five variables, X2, X3, X5, X15 and X24 are also in the benchmark set V1.

The second variable selection scheme is incremental forward variable selection. The intuition behind this scheme is that a new variable will be added into the already selected set if it will bring in the most extra information. We measure the extra information for an accounting ratio using the distance between this new ratio vector and the space spanned by the current selected ratio subset. This distance can be computed by solving a least squares problem. The ratio with the farthest distance will be added into the selected accounting ratio set. We repeat this procedure until a certain stopping criteria is satisfied. V1 is used as the initial selected accounting ratio set. Then we follow the procedure to select 7 extra more accounting ratios. These 7 ratios are different from V1, and are called the variable set V3. We will use these three variable sets for the further data analysis in the coming section.

5 Experiments Setting and Simulation Results

In this section we present our experimental setting and results. We compare the performance of three sets of accounting ratios, V1, V2 and V3, in our SSVM-based insolvency prognosis model. The performance is measured by Type I error rate, Type II error rate and total error rate. Fortunately, in reality, there is only a small portion of companies insolvent compared to

Table 3: Selected variables

Variable	Definition	V1	V2	V3
X2	NI/SALE	x	x	
X3	OI/TAS	x	x	
X5	EBIT/TA	x	x	
X6	(EBIT+AD)/TA		x	x
X8	EQUITY/TA		x	x
X10	CL/TA			x
X11	(CL-CASH)/TA			x
X12	TL/TA	x		
X13	DEBT/TA			x
X15	CASH/TA	x	x	
X19	WC/TA			x
X20	CL/TL			x
X22	INV/SALE	x		
X23	AR/SALE		x	
X24	AP/SALE	x	x	
X26	IDINV/INV	x		

the number of solvent companies. Due to the small share in a sample that reflects reality, a simple classification such as Naïve Bayes or a decision tree tends to classify every company as solvent. That is accepting all companies' loan applications. This will lead to a very high Type I error rate while the total error rate and the Type II error rate are very small. Such kind of models is useless in practice.

Our cleaned data set consists of around 10% of insolvent companies. Thus, the sample is fairly unbalanced although the share of insolvent companies is higher than in reality. In order to deal with this problem, insolvency prognosis models start usually off with more balanced training and testing samples than reality provides. For example Härdle, Moro and Schäfer (2007) employ a down-sampling strategy and work with balanced

(50%/50%)-samples. The chosen bootstrap procedure repeatedly randomly selects a fixed number of insolvent firms from the training set and adds the same number of randomly selected solvent firms. However, in this paper, we adopt a an over-sampling strategy, to balance the size between the solvent and the insolvent firms, and refer to the down-sampling procedure primarily for reasons of reference.

Over-sampling duplicates the number of the insolvent firms a certain times. In this experiment, we duplicate in each scenario the number of insolvent firms as many times as necessary for reaching a balanced sample. Note that in our over-sampling scheme every solvent and insolvent company's information is utilized. This increases the computational burden due to increasing the number of training instances. We employ the reduced kernel technique introduced in section 3.2 to mediate this problem.

All classifiers we need in these experiments are nonlinear SSVM with the Gaussian kernel which is defined as:

$$K(\mathbf{x}, \mathbf{z}) = e^{-\gamma||\mathbf{x}-\mathbf{z}||_2^2},$$

where γ is the width parameter. In nonlinear SSVM, we need to determine two parameters, the penalty term C and γ. The 2-D grid search will consume a lot of time. In order to cut down the search time, we adopt the uniform design model selection method (Huang, Lee, Lin and Huang, 2007) to search an appropriate pair of parameters.

5.1 Performance of SSVM

We conduct the experiments in a scenario in which we always train the machine from the data in hand and then use the trained SVM to predict the next year's cases. This strategy simulates the real task of prediction which

Table 4: The scenario of our experiments

Scenario Scenario	Observation period of Training Set	Observation period of Testing Set
S1	1997	1998
S2	1997-1998	1999
S3	1997-1999	2000
S4	1997-2000	2001
S5	1997-2001	2002

binds the analyst to use past data for forecasting future outcomes. The experimental setting is described in Table 4. We perform these experiments for the three variable sets, V1 to V3, repeat them 30 times and compare in each experiment the over-sampling and the down-sampling scheme.

In Table 5 and Table 6 we report the results for the oversampling and downsampling strategy respectively. We give mean and standard deviation of Type I error rates and Type II error rates, and the mean and standard deviations of total error rates (misclassification rates). The randomness is very obvious in the down-sampling scheme (see Table 6). Each time we only choose negative instances with the same size of the whole positive instances. The observed randomness in our over-sampling scheme (Table 5) is due to applying the reduced kernel technique for solving the problem. We use the training set in the down-sampling scheme as the reduced set. That is, we use the whole insolvent instances and the equal size of solvents instances as our reduced set in generating the reduced kernel. Then we duplicate the insolvent part of kernel matrix to balance the size of insolvent and solvent firms.

Both Tables reveal that different variable schemes produce dissimilar results with respect to both precision and deviation of predicting. Variable

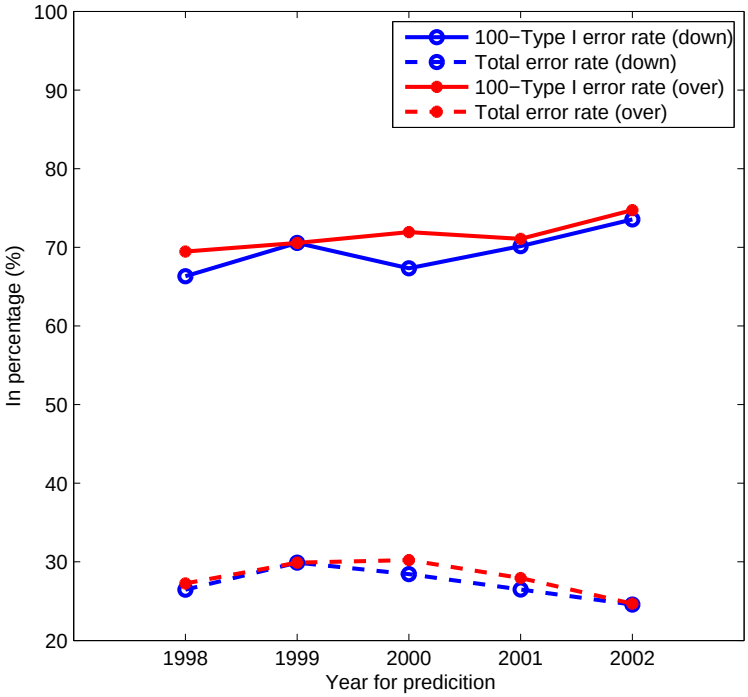

Figure 3: Learning curve for variables set V2

set V3 is clearly outperformed by both sets V1 and V2. The inferiority of V3 arises in both, the down-sampling and the over-sampling scenario. The over-sampling scheme shows better results in the Type I error rate (red line above, see Figure 3) but has slightly bigger total error rates (dashed red line below). It is also obvious that in almost all models a longer training period works in favor of the accuracy of prediction. The learning curve over the time frame the training sample covers shows an upward tendency for the number (100 - Type 1 error rate) although there is a disturbance for the forecast of the year 2000 that is based on training samples that cover 1998 till 1999. The total error rate goes down for both sampling strategies if the

Table 5: The results in percentage (%) of over-sampling for three variable sets

Set of accounting ratios	Scenario	Type I Error Rate		Type II Error Rate		Total Error Rate	
		mean	std	mean	std	mean	std
V1	S1	32.72	0.70	26.59	0.19	27.12	0.18
	S2	31.58	0.30	29.37	0.07	29.59	0.07
	S3	27.36	0.55	27.61	0.20	27.59	0.18
	S4	30.56	0.57	25.47	0.14	25.79	0.12
	S5	25.68	0.23	22.66	0.12	22.80	0.11
V2	S1	30.55	0.50	26.94	0.07	27.25	0.08
	S2	30.46	0.35	30.72	0.15	30.69	0.15
	S3	28.06	0.01	30.40	0.12	30.23	0.11
	S4	28.92	0.49	27.88	0.13	27.94	0.11
	S5	25.28	0.09	24.66	0.15	24.68	0.16
V3	S1	25.68	0.81	39.40	0.21	38.22	0.16
	S2	17.19	0.23	42.09	0.17	39.56	0.17
	S3	28.30	0.34	41.14	0.11	40.20	0.10
	S4	20.23	0.52	39.46	0.16	38.26	0.15
	S5	28.13	0.26	35.54	0.11	35.21	0.10

training period covers at least three years. One more thing worth pointing out here is that over-sampling schemes have much smaller standard deviations both in Type I error rate and total error rate.

In order to investigate the effect of the over-sampling versus the down-sampling scheme we follow the setting as above but use the V2 variable set. For each training-testing pair, we do over-sampling for positive instances from 6 to 15 times. We show the trend and effect in Figure 4. It is easy to find out that the Type I (II) error rate decreases (increases) as the over-sampling times increases. This feature implies that the machine would have a tendency of classifying all companies as solvent if the training sample had realistic shares of insolvent and solvent companies. Such behavior would

Table 6: The results in percentage (%) of down-sampling for three variable sets

Set of accounting ratios	Scenario	Type I Error Rate		Type II Error Rate		Total Error Rate	
		mean	std	mean	std	mean	std
V1	S1	33.25	3.67	27.85	2.09	28.32	1.73
	S2	31.75	2.00	28.79	1.14	29.09	1.00
	S3	30.86	1.52	26.79	1.20	27.09	1.09
	S4	31.15	1.79	24.76	1.13	25.15	1.00
	S5	27.85	2.27	22.44	0.98	22.68	0.89
V2	S1	33.68	3.44	25.78	2.53	26.46	2.05
	S2	29.45	2.18	29.95	1.69	29.90	1.39
	S3	32.66	2.50	28.10	1.45	28.43	1.19
	S4	29.86	1.85	26.25	0.98	26.47	0.91
	S5	26.46	2.33	24.48	1.23	24.56	1.15
V3	S1	30.20	5.21	37.28	2.80	36.67	2.29
	S2	19.98	2.86	41.16	1.68	39.01	1.40
	S3	30.14	1.79	38.64	1.10	38.02	0.94
	S4	23.90	2.18	36.95	1.59	36.14	1.42
	S5	29.37	1.28	34.48	1.13	34.26	1.04

produce a Type 1 error rate of 100 %. The more balanced the sample is the higher is the penalty for classifying insolvent companies as solvent. This fact is illustrated in Figure 4 by the decreasing curve with respect to the number of duplications of insolvent companies.

Often banks favor a strategy that allows them to minimize the Type II errors for a given number of Type I errors. The impact of over-sampling on the trade off between the two types of errors - shown in Figure 4 - implies that the number of over-sampling times is a strategic variable in training the machine. This number can be determined by the bank's aim regarding the relation of Type I and Type II errors.

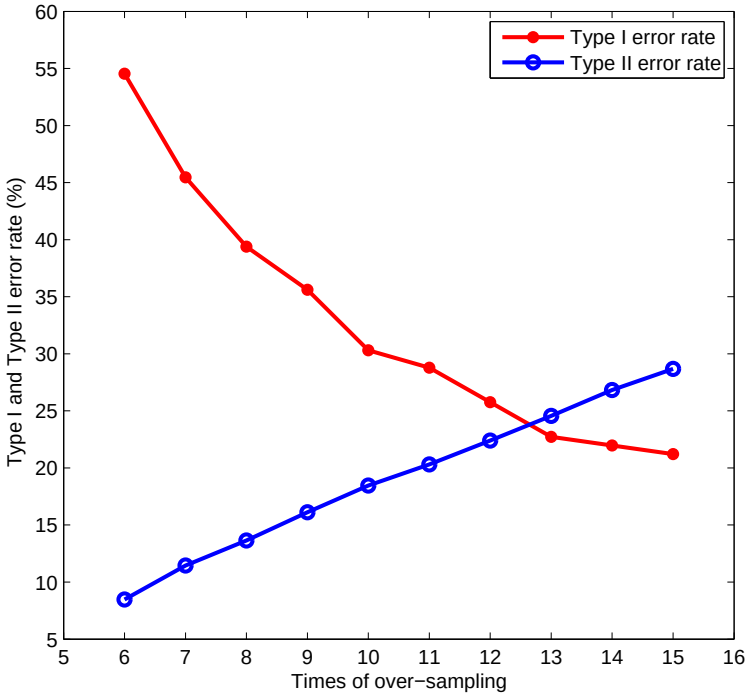

Figure 4: The effect of over-sampling on Type I and Type II error rates for scenario S5 and variables set V2

5.2 More Data visualization

Each SSVM-model has its own output value. We use these output to construct 2-D coordinate systems. Figure 5 shows an example for scenario S5 where the scores of the V2 model (V1 model) is represented by the vertical (horizontal) line. A positive (negative) value indicates predicted solvency (insolvency). We then map all solvent firms in the testing set onto the coordinate systems. There are 132 insolvent firms and 2866 solvent firms in this testing set. We also randomly choose the same amount of insolvent firms from the testing set as well. The plus points in the lower left quad-

rant and the circle points in the upper right quadrant show the number of Type I errors and Type II errors respectively in both models. Plus points in the upper right quadrant and circle points in the lower left quadrant reflect those firms that are predicted correctly by both models. Circles and plus points in the lower right quadrant (upper left quadrant) represent conflicting prognoses. We also report the number of insolvent firms and the number of solvent firms in each quadrant of Figure 5. In Figure 5, you can image the two different insolvency prognosis models generated by V1 and V2 respectively as different experts. Use their output values for each instance to plot. It provides a visualization tool and help bank officer to make the decision. That is, the proposed visualization scheme could be used to support loan officers in their final decision about accepting or rejecting the application of the client. If the application has been classified as solvent, or insolvent, by alternative machines, most likely the prognosis meets reality (the plus points in the upper right quadrant and the circle points in the lower left quadrant). Opposing forecasts, however, should be taken as a hint to evaluate this firm more thoroughly, for example by employing an expert team, or even by using a third machine.

6 Conclusion

In this paper we apply different variants of SVM to a unique dataset of German solvent and insolvent firms. We use an a-priori given set of predictors as benchmark, and suggest two further variable selection procedures, the first procedure uses a 1-norm SVM and the second, incremental way selects consecutively the variable that is the farthest one from the column space of current variable set. Given the three SSVM based on distinct variable sets, the relative performance of the types of smooth support vector machines is tested. The performance is measured by error rates. The two sets of variables selected by our own lead to a dissimilar performance SSVM with

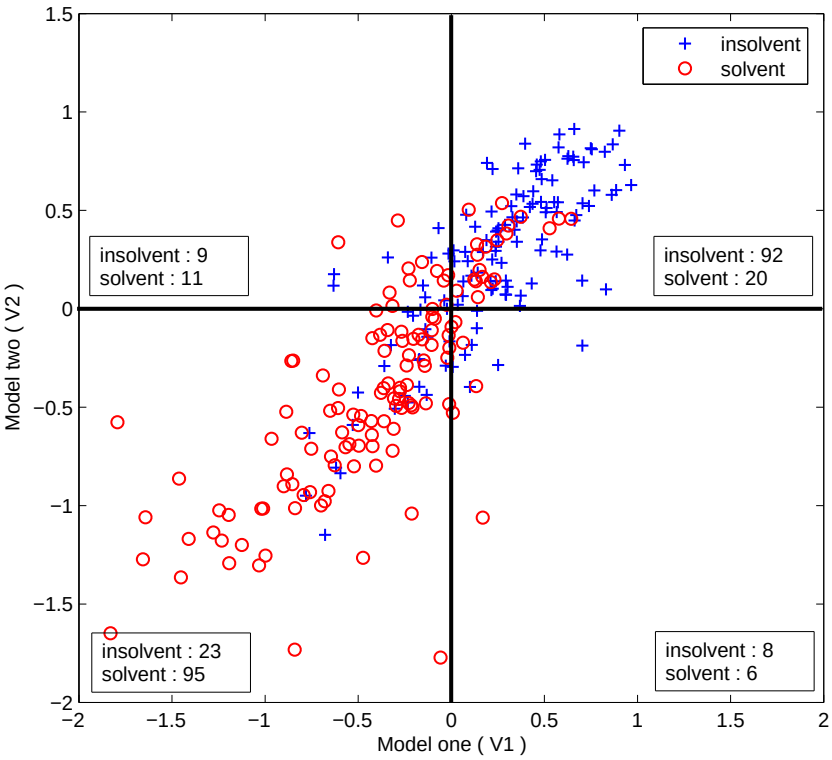

Figure 5: Data visualization via model one (generated by V1) and model two (generated by V2) in scenario S5

respect of prediction accuracy. The selection of variables by the 1-norm SVM clearly outperforms the incremental selection scheme. This finding hints at some superiority of SVMs for the variable selection procedures but further research is clearly necessary in this respect. The training period makes a clear difference, though. Results improve considerably if more years of observation were used in training the machine. Moreover the over-sampling scheme works very well in dealing with unbalanced datasets. It provides flexibility to control the trade-off between the Type I and Type II errors.

The results generated are very stable in term of small deviations of Type I error and total error rates.

References

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

Altman, E., Marco, G. and Varetto, F. (1994), 'Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience)', *Journal of Banking and Finance* (18), 505–529.

Beaver, W. (1966), 'Financial ratios as predictors of failures. empirical research in accounting: Selected studies', *Journal of Accounting Research* **4**, 71–111.

Burges, C. J. C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Chen, S., Härdle, W. and Moro, R. A. (2006), 'Estimation of default probabilities with support vector machines', *SFB 649 Discussion Paper 2006-077* .

Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. (1997), Support vector regression machines, *in* M. C. Mozer, M. I. Jordan and T. Petsche, eds, 'Advances in Neural Information Processing Systems -9-', MIT Press, Cambridge, MA, 155-161.

Fung, G. and Mangasarian, O. L. (2004), 'A feature selection Newton method for support vector machine classification', *Computational optimization and applications* **28**(2), 185–202.

Härdle, W., Moro, R. A. and Schäfer, D. (2007), 'Estimating probabilities of default with support vector machines', *SFB 649 Discussion Paper 2007-035* .

Huang, C. M., Lee, Y. J., Lin, D. K. J. and Huang, S. Y. (2007), 'Model selection for support vector machines via uniform design', *A special issue on Machine Learning and Robust Data Mining of Computational Statistics and Data Analysis, to appear* .

Krahnen, J. P. and Weber, M. (2001), 'Generally accepted rating principles: A primer', *Journal of Banking and Finance* **25**(1), 3–23.

Lee, Y.-J., Hsieh, W.-F. and Huang, C.-M. (2005), 'ϵ-SSVR: A smooth vector machine for ϵ-insensitive regression', *IEEE Transactions on Knowledge and Data Engineering* **17**, 678–685.

Lee, Y. J. and Huang, S. Y. (2007), 'Reduced support vector machines: A statistical theory', *IEEE Transactions on Neural Networks* **18**, 1–13.

Lee, Y. J. and Mangasarian, O. L. (2001), 'SSVM: A smooth support vector machine', *Computational Optimization and Applications* **20**, 5–22.

Leland, H. and Toft, K. (1996), 'Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads', *Journal of Finance* (51), 987–1019.

Longstaff, F. A. and Schwartz, E. S. (1995), 'A simple approach to valuating risky fixed and floating rate debt', *Journal of Finance* (50), 789–819.

Mangasarian, O. L. and Musicant, D. R. (2000), 'Robust linear and support vector regression', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(9), 950–955.

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

Mella-Barral, P. and Perraudin, W. (1997), 'Strategic debt service', *Journal of Finance* (52), 531–556.

Merton, R. (1974), 'On the pricing of corporate debt: The risk structure of interest rates', *The Journal of Finance* **29**(2), 449–470.

Ohlson, J. (1980), 'Financial ratios and the probabilistic prediction of bankruptcy', *Journal of Accounting Research* **18**(1), 109–131.

Schölkopf, B. and Smola, A. J. (2002), *Learning with Kernels*, MIT Press.

Smola, A. and Schölkopf, B. (2000), Sparse greedy matrix approximation for machine learning, *in* 'in Proc. 17th Int. Conf. Mcah. Learn.', San Francisco, CA, 911-918.

Smola, A. and Schölkopf, B. (2004), 'A tutorial on support vector regression', *Statistics and Computing* **14**, 199–222.

Tam, K. and Kiang, M. (1992), 'Managerial application of neural networks: the case of bank failure prediction', *Management Science* **38**(7), 926–947.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society* **58**(1), 267–288.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer, New York.

Williams, C. K. I. and Seeger, M. (2001), 'Using the Nyström method to speed up kernel machines', *Advances in Neural Information Processing Systems* **13**, 682–688.

Zhou, C. (2001), 'The term structure of credit spreads with jump risk', *Journal of Banking and Finance* (25), 2015–2040.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003), 1-norm support vector machines, *in* 'Advances in Neural Information Processing Systems 07'.

SFB 649 Discussion Paper Series 2008

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

001 "Testing Monotonicity of Pricing Kernels" by Yuri Golubev, Wolfgang Härdle and Roman Timonfeev, January 2008.
002 "Adaptive pointwise estimation in time-inhomogeneous time-series models" by Pavel Cizek, Wolfgang Härdle and Vladimir Spokoiny, January 2008.
003 "The Bayesian Additive Classification Tree Applied to Credit Risk Modelling" by Junni L. Zhang and Wolfgang Härdle, January 2008.
004 "Independent Component Analysis Via Copula Techniques" by Ray-Bing Chen, Meihui Guo, Wolfgang Härdle and Shih-Feng Huang, January 2008.
005 "The Default Risk of Firms Examined with Smooth Support Vector Machines" by Wolfgang Härdle, Yuh-Jye Lee, Dorothea Schäfer and Yi-Ren Yeh, January 2008.

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".