

# Exploring Credit Data

Marlene Müller and Wolfgang Härdle

October 17, 2002

**Abstract:** Credit scoring methods aim to assess the default risk of a potential borrower. This involves typically the calculation of a credit score and the estimation of the probability of default.

One of the standard approaches is logistic discriminant analysis, also referred to as logit model. This model maps explanatory variables for the default risk to a credit score using a linear function. Nonlinearity can be included by using polynomial terms or piecewise linear functions. This may give however only a limited reflection of a truly nonlinear relationship. Moreover, an additional modeling step may be necessary to determine the optimal polynomial order or the optimal interval classification.

This paper presents semiparametric extensions of the logit model which directly allow for nonlinear relationships to be part of the explanatory variables. The technique is based on the theory generalized partial linear models. We illustrate the advantages of this approach using a consumer retail banking data set.

---

The research for this paper was supported by Sonderforschungsbereich 373 “Quantifikation und Simulation Ökonomischer Prozesse”, Humboldt-Universität zu Berlin (Germany). Address for correspondence: Dr. Marlene Müller, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany. email: marlene@wiwi.hu-berlin.de. We are grateful to Bernd Rönz, Humboldt-Universität zu Berlin, for his cooperation at an earlier stage of this project.

# 1 Introduction

Credit scoring methods aim to estimate the default risk of potential borrowers and to classify them into groups according to their default risk. This involves typically the calculation of a credit score and the estimation of the probability of default (PD).

From a statistical point of view, classification between risky and non-risky borrowers is first of all a discriminant analysis problem. Classical solutions to this are linear or quadratic discriminant analysis and — on a more advanced level — logistic discriminant analysis. All these methods are based on a score depending on explanatory variables. Typically, the score summarizes the explanatory variables in a predefined form (linear or quadratic). More complex nonlinear mappings can be considered by using polynomial terms or piecewise linear functions. This gives, however, only an imprecise reflection of a truly nonlinear relationship. Moreover, an additional modeling step is necessary to determine the optimal polynomial order or the optimal interval classification.

Recently developed methods allow for a flexible modeling via neural networks and classification trees, for applications see Arminger, Enache and Bonne (1997) and Henley and Hand (1996). Overviews on these methods for consumer credit risk can be found in Hand and Henley (1997) and Hand (2001). These nonparametric approaches do not restrict the possible nonlinear impact of explanatory variables. However, it is often hard to interpret the resulting relationships between the explanatory variables and the classification rule. This motivates our semiparametric approach.

We consider a modification of logistic discriminant analysis that allows for a more flexible handling of a subset of the explanatory variables. Our approach is based on generalized partial linear models which extend the “easy to interpret” structure of the logistic model by nonparametric components. A particularly interesting feature of logistic discriminant analysis (equivalently: fitting a logit model) is that simultaneously credit scores and PDs are estimated. This leads to a growing interest in the logit model for redesigning credit rating systems according to the requirements of the New Basel Capital Accord (“Basel II”, cf. Banking Committee on Banking Supervision, 2001).

The paper is organized as follows: Section 2 explains the data structure for cross-sectional credit samples and provides the notation of the data that we use throughout the paper. Section 3 recalls the important terms for logistic discriminant analysis (the logit model) and presents the results for our specific sample. Section 4 introduces the semiparametric extension of the logit model. We estimate here several specifications of this semiparametric model and compare the resulting fits to the estimated logit model. Finally, Section 5 discusses the estimated models with respect to performance criteria.

## 2 Data Structure

Before we describe the data that we use in the following, let us consider a typical example for a cross-sectional credit data set. Suppose we have a sample of customers that apply for a loan to buy a car. Assume further, that we have information if these customers paid their installments without problems ( $Y = 0$ ) or not ( $Y = 1$ ). For the sake of simplicity, we will call these two categories non-default and default in the following. Obviously, we have

now a default indicator  $Y$  and explanatory variables  $X = (X_1, \dots, X_p)$  for each member of the sample. Table 1 shows for illustration descriptive statistics on a subsample of the credit data used in Fahrmeir and Hamerle (1984) and Fahrmeir and Tutz (1994).

		Yes	No	(in %)	
$Y$	default	26.4	73.6		
$X_1$	previous loans OK	66.2	33.8		
$X_2$	employed	73.2	26.8		
		Min	Max	Mean	S.E.
$X_3$	duration (in months)	4	54	21.8	10.6
$X_4$	amount (in DM)	428	14179	3902.3	2621.9
$X_5$	age (in years)	19	75	34.2	10.8

Table 1: Example data: Sample on loans for cars.

Note that Table 1 reflects the usual structure of the explanatory variables in credit data sets: The variables may be of discrete (binary, categorical) or of continuous form. For the discrete data, a sufficiently complex representation is possible by using dummy variables. For the continuous variables, an appropriate way of including them into the score has to be found.

The data that we explore and analyze in the rest of this paper have been provided by the French bank Compagnie Bancaire. The used estimation sample consists of 6180 cases (clients) and 24 variables:

- Response variable  $Y$  (credit worthiness, binary, 1 denotes default). The number of faulty clients is relatively small (6%) which is typical for credit data.
- Metric explanatory variables  $X_2$  to  $X_9$ . All of them have (right) skewed distributions. Variables  $X_6$  is discrete with only five different realizations.  $X_8$  and  $X_9$  in particular have one realization which covers a majority of observations.
- Categorical explanatory variables  $X_{10}$  to  $X_{24}$ . Six of them are binary. The others have three to eleven categories (not ordered).

In addition to the estimation sample, the bank provided us with a validation data set of 1998 cases. Table 2 gives the number of non-defaults and defaults in the estimation and validation data sets. We refer to Müller and Rönz (2000) for additional details.

	Estimation data set	Validation data set
0 (non-defaults)	5808 (94%)	1891 (94.6%)
1 (defaults)	372 (6%)	107 (5.4%)
total	6180	1998

Table 2: Defaults and non-defaults in the French bank sample.

We now describe the variables in the estimation sample in more detail. The validation sample will be only used to evaluate the semiparametric models.

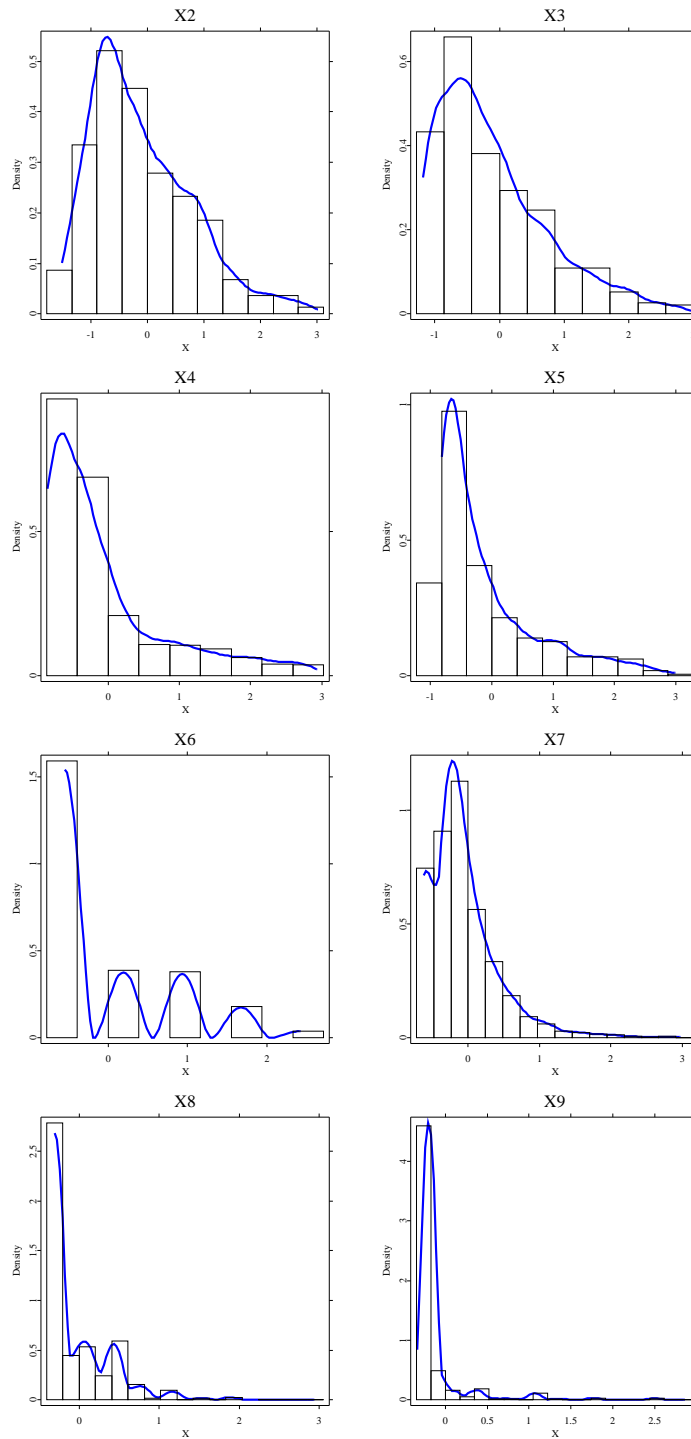


Figure 1: Histograms and kernel density estimates, variables X2 to X9.

We plot first the estimated probability density functions for the metric variables using histograms and kernel density estimates. For more statistical and numerical details on density estimation we refer to the monographs of Silverman (1986), Härdle (1991), or Scott (1992). Figure 1 shows the density estimates for the variables X2 to X9. For the kernel estimators we employed a rule-of-thumb bandwidths as smoothing parameter. From the figure we can conclude that the variables X6, X8 to X9 are of quasi-discrete

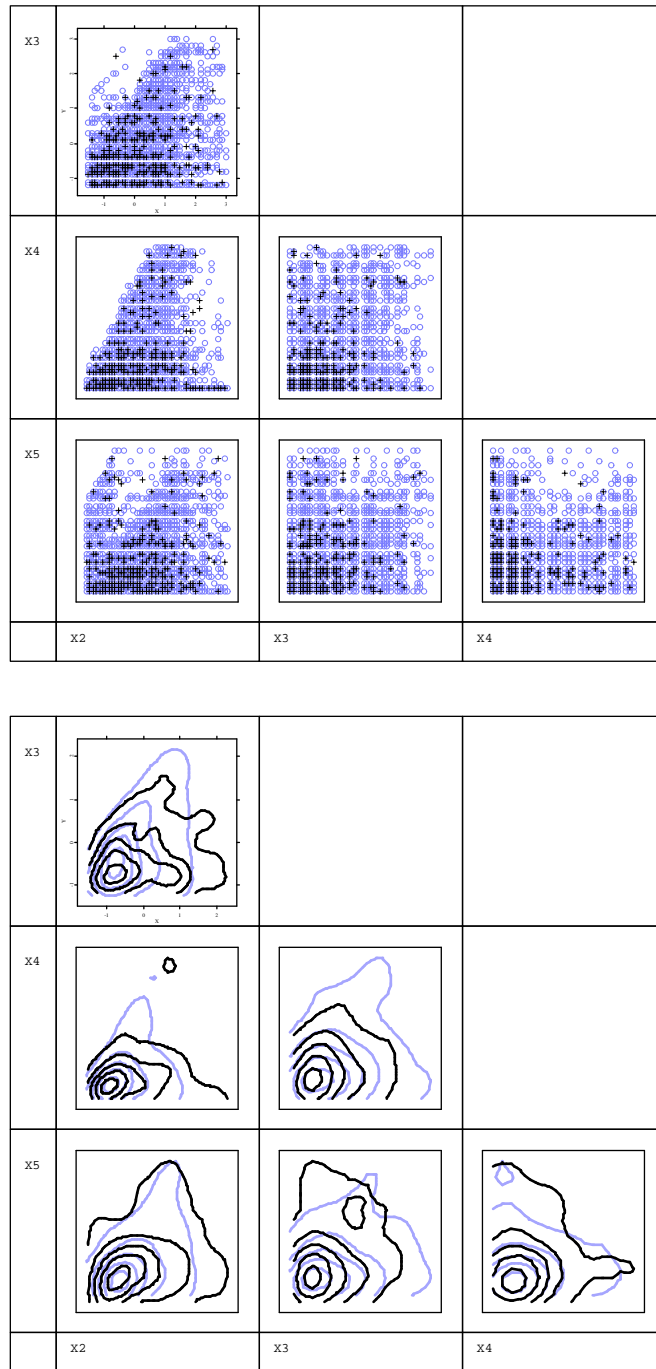


Figure 2: Scatterplots (upper display) and contour-plots (lower display), variables X2 to X5. Default observations and contours for default are emphasized in black.

structure. Since nonparametric components require continuous variation of the relevant variables, we will therefore concentrate on variables X2 to X5 and X7 for a nonparametric analysis.

As a second step in exploration we display bivariate plots for the variables X2 to X5. Figure 2 shows all bivariate scatterplots of X2 to X5. Due to the large number of non-defaults, it is difficult to capture their bivariate distribution. We therefore show bivariate

contours of the estimated densities for defaults and non-defaults. These density estimates are again kernel estimates using a rule-of-thumb bandwidth (Härdle, Müller, Sperlich and Werwatz, 2003; Scott, 1992). The figure shows that the assumptions of linear or quadratic discriminant analysis (circular or elliptical contours) are not fulfilled.

### 3 Logistic Credit Scoring

Logistic discriminant analysis assumes that the probability of belonging to the group of faulty clients is given by

$$P(Y = 1|X) = F \left( \sum_{j=2}^{24} \beta_j^\top X_j + \beta_0 \right), \quad (1)$$

where

$$F(u) = \frac{1}{1 + \exp(-u)}$$

is the logistic (cumulative) distribution function.  $X_j$  denotes the  $j$ -th variable itself if it is metric ( $j \in \{2, \dots, 9\}$ ) or a vector of dummies if it is categorical ( $j \in \{10, \dots, 24\}$ ). For all categorical variables we used the first category as reference.

Model (1) can be motivated as follows: Suppose that we know the true (negative) credit score which has the form

$$Y^* = v(X) - u$$

with  $v(\bullet)$  denoting a “regression” (or index) function and  $u$  an error term. We observe a default if the score  $Y^*$  is positive. (For practical purposes we consider higher score values to indicate higher risk of default.) Thus, our model is

$$Y = \begin{cases} 1 & \text{if } Y^* = v(X) - u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to (1) if  $u$  has a (standard) logistic distribution and

$$v(X) = \sum_{j=2}^{24} \beta_j^\top X_j + \beta_0 = \beta^\top X \quad (2)$$

holds. Modifications of the logit model usually concern the distributional assumptions (a Gaussian distribution of  $u$  leads to the probit model) or the assumptions on the index function  $v(\bullet)$ .

The logit model is estimated by maximum-likelihood (cf. McCullagh and Nelder, 1989). Table 3 shows the estimated coefficients  $\beta_j$ . We find that most of the variables contribute to the explanation of the response  $Y$ . As mentioned above, the modeling for the categorical variables is sufficiently complex due to their representation by dummy variables. For the metric variables we achieve different levels of significance. Variables X2, X3, X6, X8, and X9 have coefficients significantly different from zero. This means, their effect on the response is obviously well specified by considering them as a linear component in the index function  $v(\bullet)$ .

Variable	Coefficient	S.E.	<i>t</i> -value	Variable	Coefficient	S.E.	<i>t</i> -value
constant	<b>-2.605280</b>	0.5890	-4.42	X19#2	-0.086954	0.3082	-0.28
X2	<b>0.246641</b>	0.1047	2.35	X19#3	0.272517	0.2506	1.09
X3	<b>-0.417068</b>	0.0817	-5.10	X19#4	-0.253440	0.4244	-0.60
X4	-0.062019	0.0849	-0.73	X19#5	0.178965	0.3461	0.52
X5	-0.038428	0.0816	-0.47	X19#6	-0.174914	0.3619	-0.48
X6	<b>0.187872</b>	0.0907	2.07	X19#7	0.462114	0.3419	1.35
X7	-0.137850	0.1567	-0.88	X19#8	<b>-1.674337</b>	0.6378	-2.63
X8	<b>-0.789690</b>	0.1800	-4.39	X19#9	0.259195	0.4478	0.58
X9	<b>-1.214998</b>	0.3977	-3.06	X19#10	-0.051598	0.2812	-0.18
X10#2	-0.259297	0.1402	-1.85	X20#2	-0.224498	0.3093	-0.73
X11#2	<b>-0.811723</b>	0.1277	-6.36	X20#3	-0.147150	0.2269	-0.65
X12#2	-0.272002	0.1606	-1.69	X20#4	0.049020	0.1481	0.33
X13#2	0.239844	0.1332	1.80	X21#2	0.132399	0.3518	0.38
X14#2	-0.336682	0.2334	-1.44	X21#3	<b>0.397020</b>	0.1879	2.11
X15#2	<b>0.389509</b>	0.1935	2.01	X22#2	-0.338244	0.3170	-1.07
X15#3	0.332026	0.2362	1.41	X22#3	-0.211537	0.2760	-0.77
X15#4	<b>0.721355</b>	0.2580	2.80	X22#4	-0.026275	0.3479	-0.08
X15#5	0.492159	0.3305	1.49	X22#5	-0.230338	0.3462	-0.67
X15#6	<b>0.785610</b>	0.2258	3.48	X22#6	-0.244894	0.4859	-0.50
X16#2	<b>0.494780</b>	0.2480	2.00	X22#7	-0.021972	0.2959	-0.07
X16#3	-0.004237	0.2463	-0.02	X22#8	-0.009831	0.2802	-0.04
X16#4	0.315296	0.3006	1.05	X22#9	0.380940	0.2497	1.53
X16#5	-0.017512	0.2461	-0.07	X22#10	-1.699287	1.0450	-1.63
X16#6	0.198915	0.2575	0.77	X22#11	0.075720	0.2767	0.27
X17#2	-0.144418	0.2125	-0.68	X23#2	-0.000030	0.1727	-0.00
X17#3	<b>-1.070450</b>	0.2684	-3.99	X23#3	-0.255106	0.1989	-1.28
X17#4	-0.393934	0.2358	-1.67	X24#2	0.390693	0.2527	1.55
X17#5	<b>0.921013</b>	0.3223	2.86				
X17#6	<b>-1.027829</b>	0.1424	-7.22				
X18#2	0.165786	0.2715	0.61				
X18#3	0.415539	0.2193	1.89				
X18#4	<b>0.788624</b>	0.2145	3.68				
X18#5	<b>0.565867</b>	0.1944	2.91	df			6118
X18#6	0.463575	0.2399	1.93	Log-Lik.			-1199.6278
X18#7	<b>0.568302</b>	0.2579	2.20	Deviance			2399.2556

Table 3: Results of the logit estimation. Bold coefficients are significant at 5%.

For X4, X5, and X7 non-significant coefficients indicate that either these variables have no influence on the response or that their specification is insufficient. We will now investigate the latter conjecture. As a further graphical tool we use scatterplots for these explanatory variables. In contrast to linear regression, it is not useful to directly plot  $X_j$  vs.  $Y$ . However, if we assume model (1) as the underlying, the logits

$$\log \left( \frac{P(Y = 1|X)}{P(Y = 0|X)} \right)$$

should relate in a linear way to the explanatory variables  $X$ . We therefore divide the range of each of the variables X2 to X5 and X7 into intervals (classes) of similar length

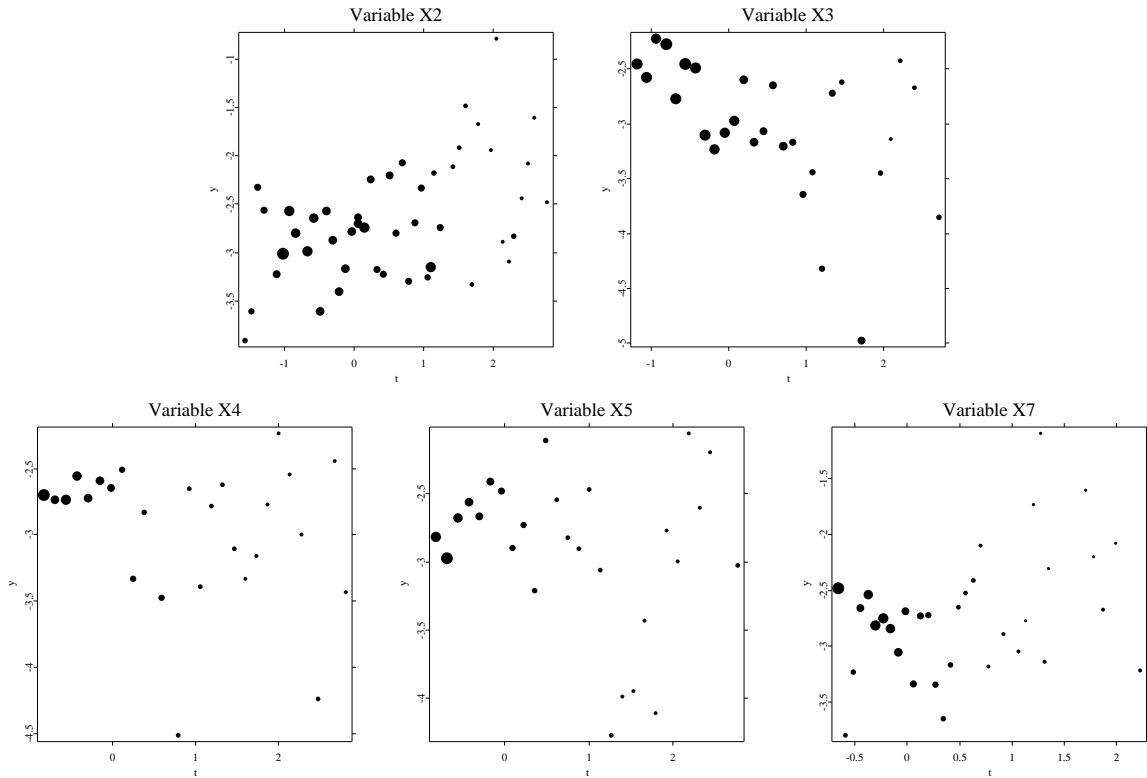


Figure 3: Marginal dependencies, variables X2 to X5, X7. Thicker bullets correspond to more observations in a class.

and estimate the logits in these intervals using the observed frequencies of  $Y = 0$  and  $Y = 1$ . The class centers are then plotted against the estimated logits. The resulting “scatterplots” are presented in Figure 3. It is obvious that the scatterplots for X2 and X3 follow a linear trend whereas for the other three variables a linear relationship is not obvious. The variables X4, X5 and X7 are hence the most interesting components for considering a nonlinear (nonparametric) modification of the index function (2).

## 4 Semiparametric Credit Scoring

The logit model (1) is a special case of the the generalized linear model (GLM, see McCullagh and Nelder, 1989) which is defined as

$$E(Y|X) = G(\beta^\top X)$$

with  $G(\bullet)$  denoting a “link” function. Since in our problem  $Y$  is binary, it holds

$$E(Y|X) = P(Y = 1|X).$$

Thus, the logit model is a GLM with the logistic distribution function  $F(\bullet)$  as link function. This property makes it easy to consider several extensions of the GLM which then hold automatically for the logit model.

The semiparametric modification that we consider here generalizes the linear argument (2) to a partial linear argument. Consider a vector of explanatory variables that splits up

into a vector  $X$  and a second vector  $T$ . The generalized partial linear model (GPLM)

$$E(Y|X, T) = G\{\beta^\top X + m(T)\}$$

allows us to describe the influence of the component  $T$  nonparametrically. As before, we assume  $G(\bullet)$  to be a known function (here the logistic link  $F$ ) and  $\beta$  to be an unknown parameter vector. In addition we have to estimate  $m(\bullet)$ , an unknown smooth function. The parametric component  $\beta$  and the nonparametric function  $m(\bullet)$  can be estimated in several ways, for a comparison of estimation algorithms and their numerical properties see Müller (2001). Details on the implementation of these estimators can be found in Müller (2000).

We consider the GPLM for several of the metric variables separately as well as for combinations of them. As mentioned earlier, we only consider variables X2 to X5 and X7 to be used within a nonparametric function because of the quasi-discrete structure of X6, X8 and X9. The semiparametric modification of the logit model takes the following form, as indicated here for the example of including X5 in a nonlinear way:

$$P(Y = 1|X) = F \left( \sum_{j=2, j \neq 5}^{24} \beta_j^\top X_j + m_5(X_5) \right).$$

A possible intercept is contained in the function  $m_5(\bullet)$ .

Table 4 contains the parametric coefficients for the parametric and semiparametric estimates for variables X2 to X9. Coefficients for X10 to X24 are estimated in each of the specified models, but are not listed here. The column headed by “logit” repeats the parametric logit estimates from Table 3.

	nonparametric in							
	logit	X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
constant	<b>-2.605</b>	-	-	-	-	-	-	-
X2	<b>0.247</b>	-	<b>0.243</b>	<b>0.241</b>	<b>0.243</b>	<b>0.233</b>	<b>0.228</b>	-
X3	<b>-0.417</b>	<b>-0.414</b>	-	<b>-0.414</b>	<b>-0.416</b>	<b>-0.417</b>	<b>-0.408</b>	<b>-0.399</b>
X4	-0.062	-0.052	-0.063	-	-0.065	-0.054	-	-
X5	-0.038	-0.051	-0.045	-0.034	-	-0.042	-	-
X6	<b>0.188</b>	<b>0.223</b>	<b>0.193</b>	<b>0.190</b>	<b>0.177</b>	<b>0.187</b>	0.176	<b>0.188</b>
X7	-0.138	-0.138	-0.142	-0.131	-0.146	-	-0.135	-0.128
X8	<b>-0.790</b>	<b>-0.777</b>	<b>-0.800</b>	<b>-0.786</b>	<b>-0.796</b>	<b>-0.793</b>	<b>-0.792</b>	<b>-0.796</b>
X9	<b>-1.215</b>	<b>-1.228</b>	<b>-1.213</b>	<b>-1.222</b>	<b>-1.216</b>	<b>-1.227</b>	<b>-1.214</b>	<b>-1.215</b>

Table 4: Parametric coefficients in parametric and semiparametric logit, variables X2 to X9. Bold values are significant at 5%.

It turns out, that all linear coefficients vary little over the different estimates. This holds as well for their significance. Variables X4, X5 and X7 are constantly insignificant over all estimates. The semiparametric logit model is estimated by semiparametric maximum-likelihood, a combination of maximizing a classical (parametric) likelihood for estimating  $\beta$  and a smoothed (local) likelihood for estimating the function  $m(\bullet)$ . The fitted curves for the nonparametric components according to Table 4 can be found in Figure 4 (separate

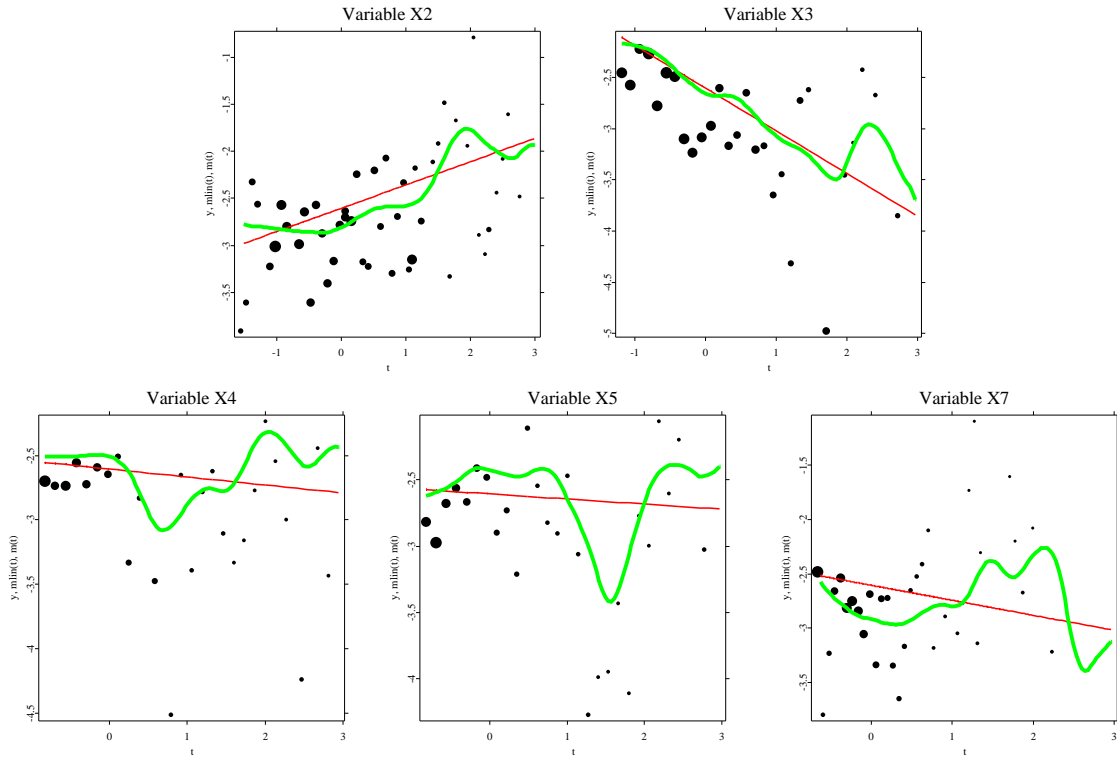


Figure 4: Estimated curves for variables X2 to X5 and X7. Parametric logit fits (thin dashed lines) and GPLM logit fits (thick solid curves).

nonparametric functions in X2 to X5 and X7) and Figure 5 (bivariate function in X4 and X5).

For the assessment of whether the semiparametric fit outperforms the parametric logit or not, we present the reported statistical characteristics in Table 5. The deviance is minus twice the estimated log-likelihood of the fitted model in our case. For the logit model, the degrees of freedom just denote

$$df = n - k$$

where  $n$  is the sample size and  $k$  the number of estimated parameters. In the semi-parametric case, a corresponding number of degrees of freedom can be approximated using the trace of the corresponding hat matrix. The deviance and the (approximate) degrees of freedom of the parametric and the semiparametric model can then be used to construct a likelihood ratio test to compare both models (Buja, Hastie and Tibshirani, 1989; Müller, 2001). The obtained significance levels from these tests are denoted by  $\alpha$ . Finally, we report pseudo  $R^2$  values in the style of McFaddens pseudo  $R^2$  values for the logit case (Greene, 1993, Sec. 21.4.2) representing an analog to the linear regression coefficient of determination.

It is obvious to see that in particular models containing variable X5 in the nonparametric part considerably decrease the deviance and increase the coefficient of determination  $R^2$ . Accordingly, the significance level for the test of parametric versus nonparametric modeling decreases.

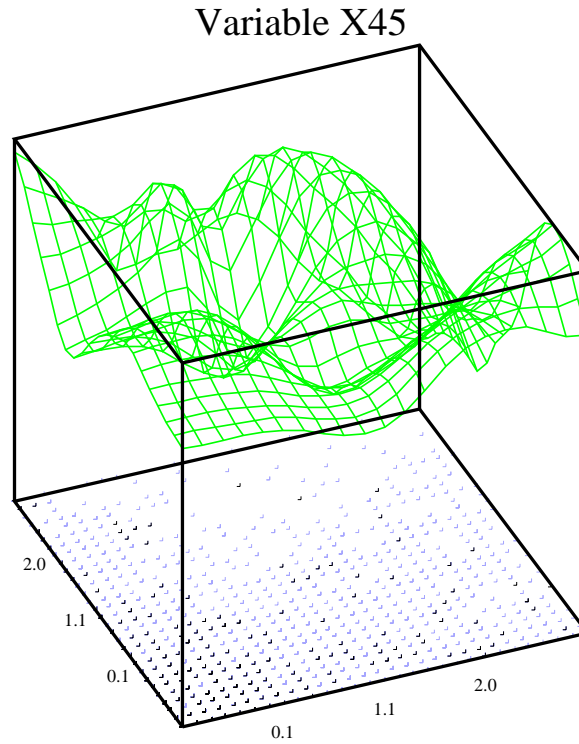


Figure 5: Bivariate nonparametric surface for variables X4, X5.

	logit	nonparametric in						
		X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
deviance	2399.26	2393.03	2395.19	2391.29	<b>2387.17</b>	<b>2388.63</b>	<b>2372.63</b>	<b>2372.43</b>
df	6118.00	6113.72	6113.57	6113.34	6113.41	6113.61	6103.82	6100.23
$\alpha$	–	0.210	0.458	0.133	<b>0.026</b>	<b>0.041</b>	<b>0.023</b>	<b>0.077</b>
AIC	2523.3	2525.6	2528.0	2524.6	2520.4	2521.4	2525.0	2533.0
Pseudo-R <sup>2</sup>	14.7%	14.9%	14.8%	15.0%	15.1%	15.1%	15.6%	15.6%

Table 5: Statistical characteristics in parametric and semiparametric logit fits. Bold values are significant at 10%. Estimation data set.

## 5 Evaluation of the Scores

For a credit rating system it is important that relevant explanatory variables are detected and enter the model in an optimal way. The semiparametric technique introduced above may help to find transformations of explanatory variables that improve the prediction of defaults.

How can different models (different scores) be compared? The easiest approach is to use misclassification rates. Suppose we have estimated the score  $S = S(X)$  for a potential

borrower. For example,  $S$  may denote

$$S = \sum_{j=2}^{24} \beta_j^\top X_j + \beta_0$$

in the parametric logit model and

$$S = \sum_{j=2, j \neq 5}^{24} \beta_j^\top X_j + m_5(X_5)$$

in the semiparametric logit model when fitting  $X_5$  nonparametrically. Typically one predicts

$$\hat{Y} = \begin{cases} 1 & F(S) > \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where the threshold  $\tau$  is taken as

$$\tau = 0.5.$$

Considering a range of  $\tau$ -values allows us to obtain a more detailed picture of the classification of different score values. Table 6 reports misclassified observations from the validation sample (of size 1998) at three different threshold values  $\tau$ .

threshold $\tau$	logit	nonparametric in						
		X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
0.25	129	133	129	136	130	128	132	130
non-default	41	44	40	49	40	40	46	40
default	88	89	89	87	90	80	86	90
0.5	111	110	111	111	110	108	111	110
non-default	5	5	5	5	5	2	5	4
default	106	105	106	106	105	106	106	106
0.75	107	107	107	107	107	107	107	107
non-default	0	0	0	0	0	0	0	0
default	107	107	107	107	107	107	107	107

Table 6: Misclassifications for  $\hat{Y} = 1$  (default) if  $F(S) \leq t$  and  $\hat{Y} = 0$  (non-default) if  $F(S) > t$ . Validation data set.

The Lorenz curve (cumulative accuracy profile, CAP) visualizes the accuracy of the score with respect to its predictive power for a default. Figure 6 shows the principle of the Lorenz curve. For both axes, sorted score values (from bad=high to good=low) are considered. The horizontal scale shows the percentages of observations above a certain value  $s$ , whereas the vertical axis shows percentages of faulty observations above this value  $s$ . Mathematically, the Lorenz curve is a plot of

$$P(S > s) \quad \text{versus} \quad P(S > s | Y = 1).$$

Typically, i.e., if the PD is a monotone increasing function of the score, the curve is concave and located above the diagonal. The diagonal can be interpreted as a “worst score”: If  $S$  and  $Y$  have no relation at all, then  $P(S > s) = P(S > s | Y = 1)$ . The “best score” does a perfect separation of defaults and non-defaults. This leads to the optimal

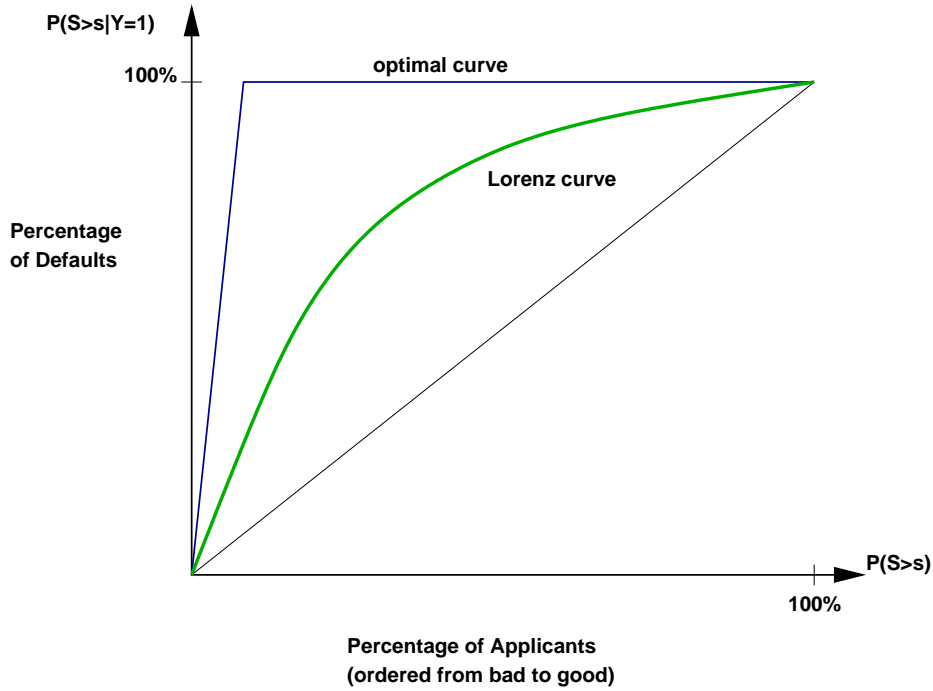


Figure 6: Principle of the Lorenz curve.

curve shown in Figure 6. A quantitative measure for the performance of a score is based on the area between the Lorenz curve and the diagonal. The Gini coefficient  $G$  denotes twice this area. To compare different scores, their accuracy ratios

$$AR = \frac{G}{G_{opt}},$$

i.e., the Gini coefficient  $G$  relative to the Gini coefficient of the optimal Lorenz curve can be used. Variants of the Lorenz curve are the receiver operating characteristic (ROC) curve (Hand and Henley, 1997) and the performance curve (Gourieroux and Jasiak, 2001, Ch. 4). See also Sobehart and Keenan (2001) for a relation between Lorenz curve and ROC and Keenan and Sobehart (1999) for a general overview on criteria for measuring the accuracy of credit scores.

Lorenz curves can also be used for assessing the impact of single variables. Table 7 shows the  $AR$  values for all metric explanatory variables on the estimation data set. We find that those variables which are highly significant in the logit fit (cf. Table 3) also achieve high accuracy ratios. (Note that we used appropriate  $+/-$  signs here for each variable, such that the maximal possible  $AR$  is reported.)

	nonparametric in							
	+X2	-X3	-X4	+X5	+X6	-X7	-X8	-X9
$AR$	0.076	0.168	0.043	0.023	0.024	0.052	0.165	0.107

Table 7: Accuracy ratios of variables X2 to X9. Estimation data set.

Let us have a closer look at the Lorenz curves for the three variables X4, X5, X7 which

had an obvious nonlinear effect in the score. Figure 7 shows the Lorenz curves and in comparison density estimates separately for defaults and non-defaults. In particular for X5 and X7 we see that the impact of these two variables on Y is non-monotonous: The Lorenz curve crosses the diagonal and the densities cross several times. This means that the nonlinear relationship in the index function  $v(\bullet)$  is as well reflected in the Lorenz curve (and vice versa).

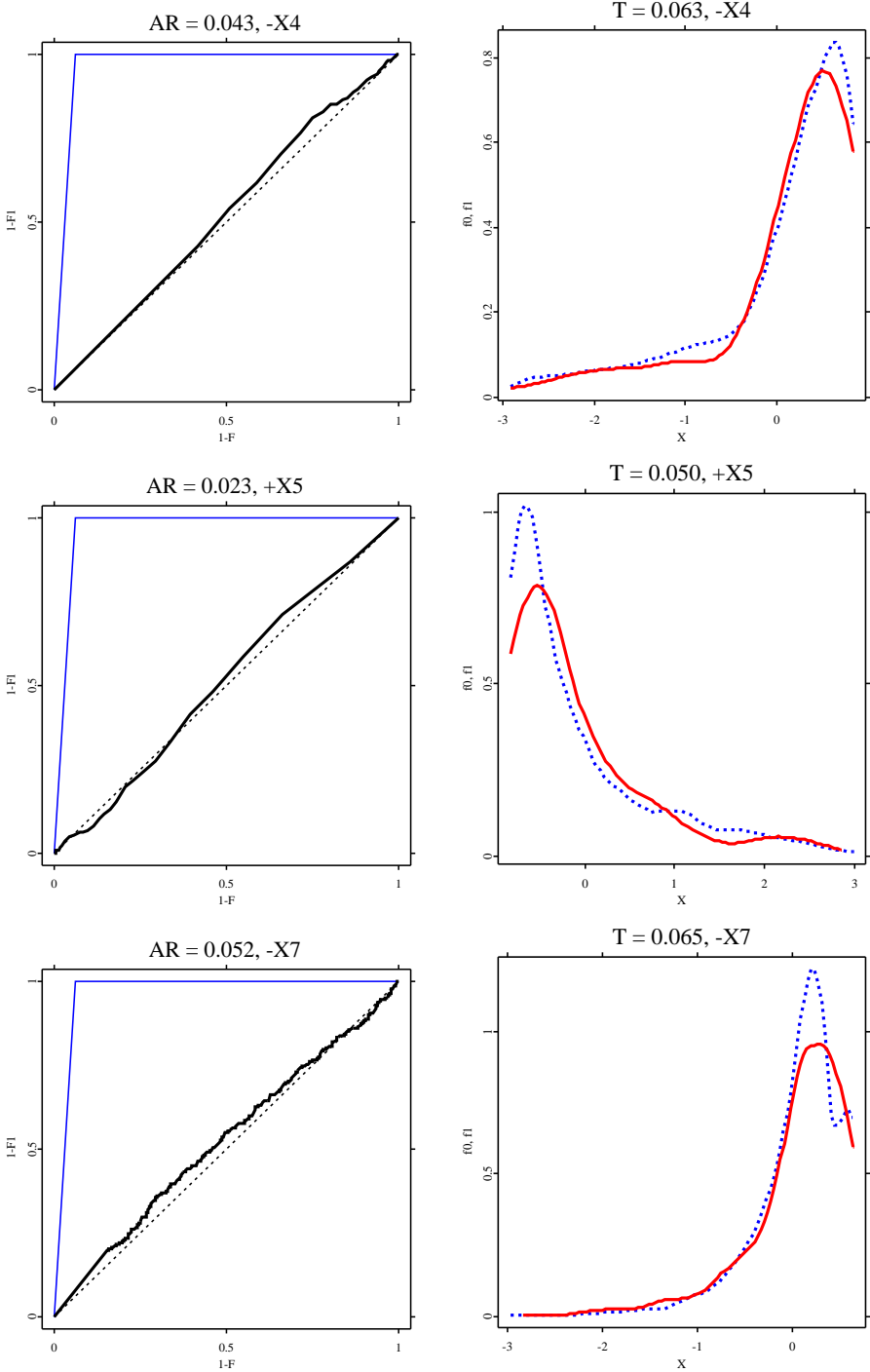


Figure 7: Lorenz curves (left) and density estimates (right, conditionally on default/non-default) for X4, X5, X7.

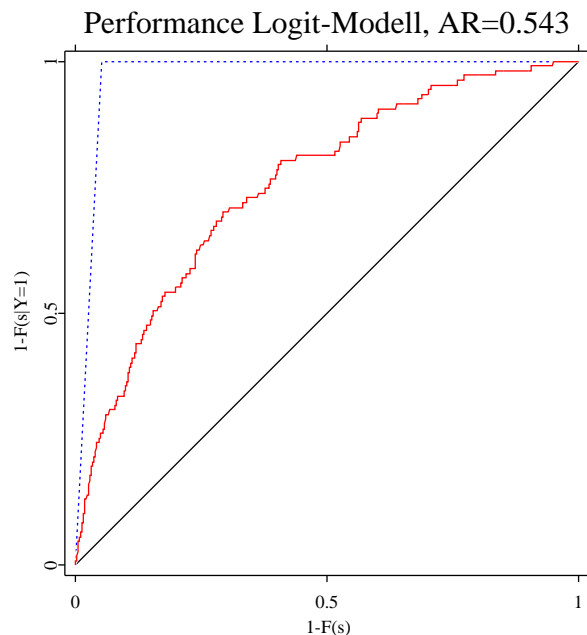


Figure 8: Lorenz curve for logit model (solid) and optimal curve (dashed).

Consider now the Lorenz curves and  $AR$  values for the fitted logit and semiparametric model. Figure 8 shows the result for the logit fit achieving an  $AR$  value of 0.543. Note that most of the performance of the score is contributed by the categorical variables. The continuous variables altogether explain only a small part of the default.

		nonparametric in						
	logit	X2	X3	X4	X5	X7	X4,X5	X2, X4,X5
$AR$	0.543	0.538	0.543	0.527	<b>0.556</b>	0.538	<b>0.548</b>	<b>0.552</b>

Table 8: Accuracy ratios in parametric and semiparametric logit. Bold values improve the logit fit. Validation data set.

Table 8 compares the  $AR$  performance of the parametric logit fit and the semiparametric logit fit obtained by separately including X2 to X5 nonparametrically. Indeed, the semiparametric model for the influence of X5 improves the performance with respect to the parametric model. The semiparametric models for the influence of X2 to X4 do not improve the performance with respect to the parametric model, though.

Figure 9 compares the performance of the parametric logit fit and the semiparametric logit fit obtained by jointly including X4, X5 nonparametrically. This performance curve improves versus nonparametrically fitting only X4, but shows less power versus fitting only X5. Hence, the improvement of using both variables jointly may be explained by the influence of X5 only.

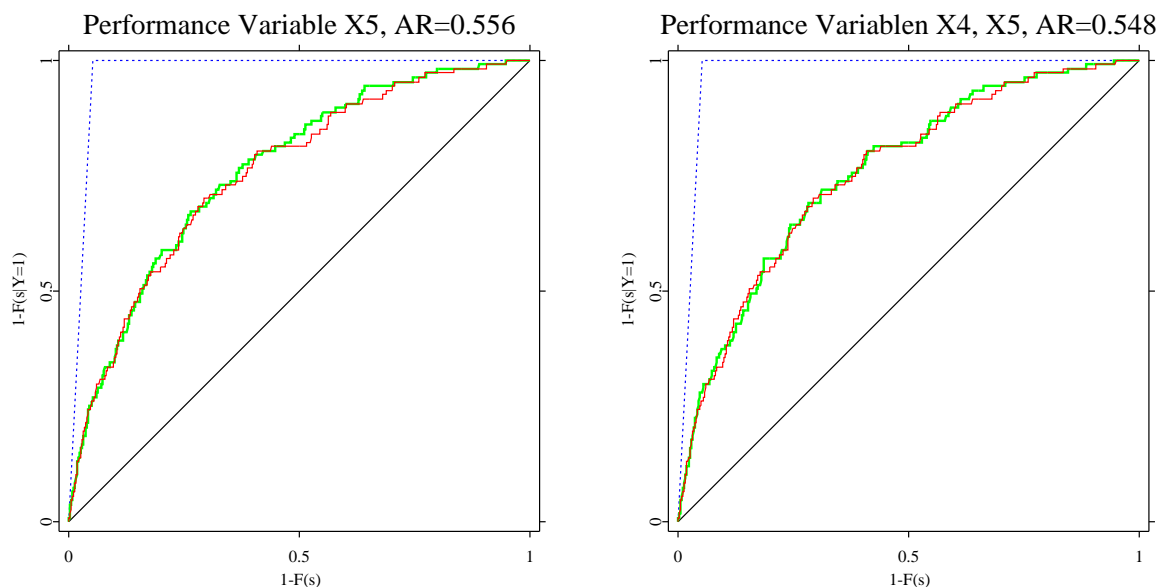


Figure 9: Performance curves with variables X5 (left) and with variables X4, X5 (right) jointly included nonparametrically. Validation data set.

## References

- Arminger, G., Enache, D. and Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks, *Computational Statistics, Special Issue: 10 Years AG GLM* **12**: 293–310.
- Banking Committee on Banking Supervision (2001). *The New Basel Capital Accord*, Bank for International Settlements, <http://www.bis.org>.
- Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**: 453–555.
- Fahrmeir, L. and Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.
- Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.
- Gourieroux, C. and Jasiak, J. (2001). *Econometric Analysis of Individual Risks*, Course Script, <http://dept.econ.yorku.ca/~jasiakj>.
- Greene, W. H. (1993). *Econometric Analysis*, 2 edn, Prentice Hall, Englewood Cliffs.
- Hand, D. J. (2001). Modelling consumer credit risk, *IMA Journal of Management mathematics* **12**: 139–155.
- Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society, Series A* **160**: 523–541.

- Härdle, W. (1991). *Smoothing Techniques. With Implementations in S*, Springer, New York.
- Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2003). *An Introduction to Non- and Semiparametric Models*, Springer, forthcoming.
- Henley, W. E. and Hand, D. J. (1996). A  $k$ -nearest-neighbor classifier for assessing consumer credit risk, *Statistician* **45**: 77–95.
- Keenan, S. and Sobehart, J. (1999). Performance measures for credit risk models, *Technical report*, Moody's Investor Service, Global Credit Research.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.
- Müller, M. (2000). Generalized partial linear models, in W. Härdle, Z. Hlávka and S. Klinka (eds), *XploRe Application Guide*, Springer and <http://www.xplo-re-stat.de>.
- Müller, M. (2001). Estimation and testing in generalized partial linear models – a comparative study, *Statistics & Computing* **11**: 299–309.
- Müller, M. and Rönz, B. (2000). Credit scoring using semiparametric methods, in J. Franke, W. Härdle and G. Stahl (eds), *Measuring Risk in Complex Stochastic Systems*, Springer.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, Chichester.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.
- Sobehart, J. and Keenan, S. (2001). Measuring default accurately, *Risk, Credit Risk Special Report* **14**(3): 31–33.