A review of some semiparametric regression models with application to scoring

Jean-Loïc Berthet[1] and Valentin Patilea[2]

[1] ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: jean-loic.berthet@ensai.fr)

[2] CREST-ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: patilea@ensai.fr)

Abstract. Some semiparametric models for binary response data are reviewed: single-index models, generalized partially linear models, generalized partially linear models single-index models and multiple-index models. All these models can be seen as extensions of the classical logistic regression. We test and compare these models using data on bankruptcy of French companies and data from credit business.

Keywords: Scoring, semiparametric regression, iterative methods, single and multiple-index, bandwidth choice.

1 Introduction

Classification techniques are used in many statistical applications. The objective of any classification model is to classify individuals in two or more groups based on a predicted outcome associated with each individual. Here, we are interested in statistical models classifying individuals in two groups: 'good' (or 'not default') and 'bad' (or 'default') individuals. Such models can be applied in banking and credit control, marketing, medicine, *etc.* The classification rule for an individual must be based on the information about the individual at the time of the decision. This information is contained in a vector of explanatory variables (factors, indicators, characteristics, ...) $\mathbf{X} = (X_1, ..., X_p)^\top$. Usually, the available information for an individual is synthesized into a single value usually called the *score*. The score aims to reflect the probability that the individual will 'not default'.

Various parametric and nonparametric methods can be used to solve classification problems (see, e.g., [Hand and Henley, 1997] for a review). Discriminant analysis, linear regression and logistic regression are the standard parametric techniques, while $k-$nearest neighbors, classification trees, neu-

ral networks and, more recently, support vector machine are some common nonparametric (distribution-free) procedures.

The simple, user friendly and easily interpretable character of the parametric regression models make them the most popular classification techniques in many application fields. The nonparametric methods, unlike the parametric methods, make no (or mild) assumption about the distribution of the observations and are therefore attractive when data on hand does not meet strict statistical assumptions. The price of this flexibility can be high, however. First, estimation precision decreases rapidly as the dimension of $\mathbf{X}$, the vector of explanatory variables, increases. This is the so-called curse of dimensionality. A second problem with nonparametric methods is that the results can be difficult to display, communicate, and interpret when $\mathbf{X}$ is multidimensional. A further problem with nonparametric methods is the difficulty to extrapolate the prediction to individuals with characteristics that are very different from the characteristics of the individuals that served for estimation.

The semiparametric methods represent an appealing compromise for constructing statistical models. By making assumptions that are of intermediate strength between the parametric and nonparametric approaches, the semiparametric models reduce the risk of misspecification relative to a parametric model and avoid at least in part the aforementioned drawbacks of the nonparametric methods.

In this paper we review some semiparametric regression methods that apply to *scoring*, that is to determine how likely an individual will 'not default'. The starting point of the review is the logistic regression. The power of the semiparametric methods is investigated using data on bankruptcy of French companies and publicly available data on credit-scoring from a German bank.

2 Semiparametric models for binary response variables

Let Y be a random variable taking the values 0 ('bad' or 'default' individual) or 1 ('good' or 'not default' individual). The problem on hand to estimate the probability of the event $\{Y = 1\}$ given a vector of explanatory variables $\mathbf{X}$. The logistic regression is a particular case of the so-called *generalized linear model* (see [McCullagh and Nelder, 1989]) where the conditional mean of Y given $\mathbf{X}$ has the form

$$E(Y \mid \mathbf{X}) = G(c + \mathbf{X}^\top \beta) \tag{1}$$

with a known monotone function G ($G(x) = \{1 + \exp(-x)\}^{-1}$ for the logistic regression) and an unknown parameters $(c, \beta^\top)^\top$. This model can be interpreted as follows: there exists a latent variable Y^* that can be related to $\mathbf{X}$ through a linear model $Y^* = c + \mathbf{X}^\top \beta + u$ with u an error term with cumulative distribution function (cdf) G. The observation Y is nothing but $\mathbf{1}_{\{Y^* \geq 0\}}$ where $\mathbf{1}_{\{\cdot\}}$ equals one if the condition inside the curly brackets holds,

and zero otherwise. The model (1) is purely parametric in the sense that one only has to estimate the vector of coefficients $(c, \beta^\top)^\top$.

Several semiparametric extensions of model (1) have been proposed. A natural idea is to relax the hypothesis of a linear regression model for the latent variable Y^*. [Härdle *et al.*, 1998] proposed to separate the explanatory variables into two groups, that is $\mathbf{X} = (\mathbf{Z}^\top, \mathbf{T}^\top)^\top$ with $\mathbf{Z} \in \mathbf{R}^{p_1}$, $\mathbf{T} \in \mathbf{R}^{p_2}$, and to suppose that $Y^* = \mathbf{Z}^\top\beta + m(\mathbf{T}) + u$, where the error term has a logistic law (the constant c appearing in model (1) is absorbed by the function $m(\cdot)$). The function m is unknown and it must be estimated nonparametrically. In this settings one has a *generalized partially linear model*

$$E(Y \mid \mathbf{Z}, T) = G(\mathbf{Z}^\top\beta + m(\mathbf{T})) \tag{2}$$

with $G(x) = \{1 + \exp(-x)\}^{-1}$. This model is semiparametric in the sense that in addition to the finite dimensional vector β, one has to estimate also the function m. If one wants to assume that several explanatory variables have a nonlinear effect on the conditional mean of Y^*, one has to estimate nonparametrically a multivariate function m. In order to avoid the course of dimensionality, [Härdle *et al.*, 2004] considered that $m(\mathbf{T}) = m(T_1, ..., T_{p_2}) = m_1(T_1) + ... + m(T_{p_2})$. Another approach that avoids nonparametric estimation of a multivariate function is to suppose that there exists a vector $(\alpha_1, ..., \alpha_{p_2})^\top$ (identifiable up to a scaling factor) such that

$$m(T_1, ..., T_{p_2}) = m(\alpha_1 T_1 + ... + \alpha_{p_2} T_{p_2}).$$

See [Carroll *et al.*, 1997]. In all these models the coefficients β (β and α for the model of [Carroll *et al.*, 1997]) can be estimated with a precision of order $n^{-1/2}$ where n is the sample size, that is the usual precision of a parametric model.

Another natural extension of the parametric model goes as follows. Assume that $Y^* = c + \mathbf{X}^\top\beta + u$ with u an error term with *unknown* law independent of $\mathbf{X}$ given $\mathbf{X}^\top\beta$. Then,

$$E(Y \mid \mathbf{X}) = r(\mathbf{X}^\top\beta) \tag{3}$$

with $r(\cdot)$ an unknown function that has to be estimated nonparametrically. The constant c is absorbed by $r(\cdot)$. Moreover, the vector β can only be determined up to a scaling factor. The model (3) belongs to a general class of semiparametric models called *single-index models (SIM)*. In such models one only assumes that when computing the conditional expectation of Y given $\mathbf{X}$, all the relevant information carried by $\mathbf{X}$ is contained in a linear combination of the components of $\mathbf{X}$. In the following we shall concentrate on inference methods for model (3). Note that model (3) can be obtained as a particular case of the model of [Carroll *et al.*, 1997] by taking $\beta = 0$ and setting $r = G \circ m$ (and relabelling the explanatory variables).

Several semiparametric approaches for consistent estimation of β in SIM have been proposed including $M-$estimation, average derivative methods

and iterative methods. See [Delecroix *et al.*, 2004] for a review. Here, we focus on $M-$estimation. Typically, if $(Y_1, \mathbf{X}_1^\top)^\top, ..., (Y_n, \mathbf{X}_n^\top)^\top$ denote the observations, a semiparametric M-estimator of β is defined as

$$\widehat{\beta} = \arg\min_\beta \frac{1}{n} \sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \beta; \beta\right)\right) \tau_n(\mathbf{X}_i), \tag{4}$$

where $\widehat{r}\,(t;\beta)$ is a nonparametric estimator, for instance the Nadaraya-Watson estimator, of the regression function $r\,(t;\beta) = E\left(Y \mid \mathbf{X}^\top \beta = t\right)$, ψ is a contrast function and $\tau_n(\cdot)$ is a so-called trimming function introduced to guard against small values of the denominators appearing in the nonparametric estimator. Finally, the conditional mean of Y is estimated by $\widehat{r}(x^\top \widehat{\beta}; \widehat{\beta})$. [Klein and Spady, 1993] considered the case $\psi(y, r) = -\{y \log(r) + (1-y) \log(1-r)\}$ which yields the semiparametric maximum likelihood estimate of β. [Dominitz and Sherman, 2003] considered the case $\psi(y, r) = (y - r)^2$ and proposed a nice iterative method that avoids optimization with respect to both occurrences of β in equation (4). [Delecroix *et al.*, 2004] suggested other choices for $\psi(y, r)$ that improve the performances of the estimator $\widehat{\beta}$ in the presence of outliers.

The large sample properties of the estimates $\widehat{\beta}$ and $\widehat{r}(x^\top \widehat{\beta}; \widehat{\beta})$ obtained from optimization procedures as (4) are now well known in the case of independent, identically distributed observations of $(Y, \mathbf{X}^\top)^\top$. In particular, this allows to obtain significance tests for the coefficients β and confidence intervals for the conditional probability of 'not default' given $\mathbf{X}$. Extensions to the case of dependent data have been also studied. See [Xia *et al.*, 2002] for a description of the techniques of proof that apply for dependent data and for a list of references.

A crucial problem associated with the estimator $\widehat{\beta}$ is the choice of the smoothing parameter for the nonparametric estimator of the regression function $r\,(t;\beta)$. One may consider the smoothing parameter as another parameter of interest which can be estimated at the same time as β, that is one can optimize the objective function in (4) simultaneously with respect to β and the smoothing parameter. In general, to avoid degenerate problems when optimizing simultaneously with respect to β and the smoothing parameter, a leave-one-out version of the nonparametric estimator should replace $\widehat{r}$ in equation (4). See, e.g., [Delecroix *et al.*, 2004] for the theoretical properties of the simultaneous optimization approach.

Despite the fact that the regression function is supposed unknown, a SIM still imposes that all the relevant information carried by $\mathbf{X}$ is contained in one factor that is obtained as a linear combination of the components of $\mathbf{X}$. A natural idea is to investigate whether more than one factor is necessary to capture the information contained in $\mathbf{X}$. For instance, one may consider the model

$$E(Y \mid \mathbf{X}) = r(\mathbf{X}^\top \beta^1, \mathbf{X}^\top \beta^2) \tag{5}$$

with $r(\cdot,\cdot)$ an unknown bivariate function that has to be estimated nonparametrically and β^1, β^2 two vectors of unknown coefficients. (Suitable normalization conditions are necessary to make the vectors β^1, β^2 identifiable.) The unknown parameters can be estimated by an extension of (4), that is

$$(\widehat{\beta^1}, \widehat{\beta^2}) = \arg\min_{(\beta^1,\beta^2)} \frac{1}{n} \sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \beta^1, \mathbf{X}_i^\top \beta^2; (\beta^1, \beta^2)\right)\right) \tau_n(\mathbf{X}_i), \quad (6)$$

where $\widehat{r}\left(t, s; (\beta^1, \beta^2)\right)$ is a nonparametric estimator of the regression function $r(t, s; (\beta^1, \beta^2)) = E\left(Y \mid (\mathbf{X}^\top \beta^1, \mathbf{X}^\top \beta^2) = (t, s)\right)$. The smoothing parameters of the bivariate estimator of r can be selected by simultaneous optimization in (6) with respect to (β^1, β^2) and the smoothing parameters. See [Xia *et al.*, 2002] and [Delecroix *et al.*, 2004].

An alternative procedure for finding β^1, β^2 is to search these directions one by one: first, search $\widehat{\beta^1}$ like in (4); second, search $\widehat{\beta^2}$ orthogonal to $\widehat{\beta^1}$ and solution of the problem

$$\widehat{\beta^2} = \arg\min_{\beta^2} \frac{1}{n} \sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}\left(\mathbf{X}_i^\top \widehat{\beta^1}, \mathbf{X}_i^\top \beta^2; (\widehat{\beta^1}, \beta^2)\right)\right) \tau_n(\mathbf{X}_i).$$

This procedure simplifies the optimization problem. It can be shown that, under mild conditions, the linear subspace generated by directions obtained by sequential search is the same as the linear subspace generated by the directions obtained from (6). One may search for more than two directions β, either by joint maximization as in (6) or by sequential search after finding the first two directions $\widehat{\beta^1}$, $\widehat{\beta^2}$, but the results will become much more difficult to interpret.

The last theoretical issue we shall discuss here is the problem of testing in the semiparametric models mentioned above. There are at least two types of test problems one may consider. First, it is important to be able to test the traditional parametric binary response regression models, typically the logistic regression, using semiparametric models. [Härdle *et al.*, 1998] started from model (2) and tested the logistic regression model by setting the null hypothesis $m(\mathbf{T}) = c + \mathbf{T}^\top \gamma$ for some constant c and some vector γ. [Härdle and Spokoiny, 1997] considered the SIM framework described by equation (3) and proposed a test procedure for checking whether the function r has a given form (typically, whether r is the logistic function or not).

If the parametric model is rejected in favor of a more flexible semiparametric specification, the next step is to test the semiparametric model itself against more general semiparametric or nonparametric alternatives. It is only recently that promising testing procedures for SIM have been proposed. See [Xia *et al.*, 2004] and [Stute and Wang, 1994].

3 The data

The stakes of a reliable, interpretable, easy to implement and easy to update scoring method are important. The semiparametric methods represent an alternative stream of dealing with these aspects. Their power was relatively little investigated in scoring applications. Our aim is to provide additional empirical evidence on the utility of the semiparametric methods in scoring problems. The semiparametric techniques mentioned above are tested and compared with the benchmark parametric models. For this purpose we use two types of data. First, we work with a sample from a database of *Banque de France*. Our dataset contains the accounting balances of French companies from one economic sector during several years. The task is to asses the risk of bankruptcy for a company given the information provided by the company.

For our second application we use data on private loans from a German bank. The data are presented in [Fahrmeir and Tutz, 1994] and are publicly available. In credit business, banks are interested in information whether prospective consumers will pay back their credit or not. The aim of credit scoring is to predict the probability that a consumer with certain characteristics is to be considered as a potential risk. The dataset we consider consists of 1000 consumer credits. For each consumer the binary response variable "creditability" is available, together with a set of covariates that are assumed to influence creditability.

References

[Carroll *et al.*, 1997]R.J. Carroll, J. Fan, I. Gijbels, and M.P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

[Delecroix *et al.*, 2004]M. Delecroix, M. Hristache, and V. Patilea. On semiparametric m-estimation. *Journal of Statistical Planning and Inference*, in press, 2004.

[Dominitz and Sherman, 2003]J. Dominitz and R.P. Sherman. Some convergence theory for iterative estimation procedures. *Working Paper, California Institute of Technology*, 2003.

[Fahrmeir and Tutz, 1994]L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New-York, 2nd edition, 1994.

[Hand and Henley, 1997]D.J. Hand and W.E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, 160(4):523–541, 1997.

[Härdle and Spokoiny, 1997]W. Härdle and S. Spokoiny, V.and Sperlich. Semiparametric single index versus fixed function modelling. *The Annals of Statistics*, 25:212–243, 1997.

[Härdle *et al.*, 1998]W. Härdle, E. Mammen, and M. Müller. Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93(444):1461–1474, 1998.

[Härdle *et al.*, 2004]W. Härdle, S. Huet, E. Mammen, and M. Sperlich. Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20:265–300, 2004.

[Klein and Spady, 1993]R.W. Klein and R.H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421, 1993.

[McCullagh and Nelder, 1989]P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probabilites. Chapman and Hall, London, 2nd edition, 1989.

[Stute and Zhu, 2004]W. Stute and L.X. Zhu. Nonparametric checks for single-index models. *The Annals of Statistics*, in press, 2004.

[Xia *et al.*, 2002]Y. Xia, H. Tong, W.K. Li, and L.X. Zhu. An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B*, 64(3):363–410, 2002.

[Xia *et al.*, 2004]Y. Xia, W.K. Li, H. Tong, and D. Zhang. A goodness-of-fit test for single index models. *Statistica Sinica*, 14:1–39, 2004.