ANNALS OF ECONOMICS AND FINANCE **3**, 117–134 (2002)

Generalized Semiparametric Binary Prediction

Jeff Racine

Department of Economics
University of South Florida
Tampa, FL 33620, USA
E-mail: jracine@coba.usf.edu

This paper proposes a semiparametric approach to the estimation of 'generalized' binary choice models. A 'generalized' binary choice model is one with separate indices for each conditioning variable which constitutes a generalization of the standard single-index approach typically employed in applied work. The choice probability distribution is therefore a joint distribution across these indices as opposed to the typical univariate distribution on a scalar index commonly found in applied work. Interest lies in estimating choice probabilities and the gradient of choice probabilities with respect to the conditioning information, and these are estimated nonparametrically using the method of kernels. A data-driven cross-validatory method for bandwidth selection and index-parameter estimation is proposed for maximization of the nonparametric likelihood function. The functional form of the indices enters this nonparametric likelihood function thereby permitting data-driven determination of the index functions in addition to the shape of the joint cumulative distribution function itself. Applications are considered.

Key Words: Semiparametric; Nonparametric methods.
JEL Classification Number: C14

1. INTRODUCTION

Binary choice models are characterized by a dependent variable which can take on only two discrete values. Such models are widely used in a number of disciplines, and occur frequently in economics because the decision of an economic unit frequently involves binary choice, for instance, whether or not a person joins the labor force or makes an automobile purchase, or whether or not a firm issues a bond. The goal of modeling such choices is first to predict whether or not a choice will be made (choice probability) and second to assess the response of the probability of the choice being made to changes in variables believed to influence choice (choice gradient).

1529-7373/2002

Existing approaches to modeling binary choice fall into two camps, parametric and semiparametric. Both camps typically assume the existence of a single threshold given by an index function beyond which it is more likely than not that a choice will be made, hence they are often called 'single-index' models. Such models can be quite powerful and they enjoy widespread use, however, the use of a single-index can impose some strong restrictions on the underlying process. For example, one such restriction is that choice gradients are forced to have identical shape for each variable influencing choice and differing only by the value of a scalar parameter.

This paper considers a generalization of such approaches in which there exist separate thresholds and indices for each variable believed to influence choice. This is quite different from existing 'multiple-index' models which are typically used to model multinomial discrete choice. In this paper the choice remains binary in nature, but multiple thresholds are admitted. The resulting method is more flexible than the standard approach and constitutes an alternative to single-index models when modeling binary choice.

Existing approaches are now briefly summarized in order to compare and contrast them from the proposed approach.

1.1. Parametric Estimation of Single-Index Binary Choice Models

Logit and probit models remain the most widely used parametric methods for the estimation of binary choice models. Such approaches depend on two assumptions; a known index which is assumed to influence choice, and a known parametric form for a distribution function (CDF) which is assumed to yield choice probabilities. The index is assumed to yield a positive choice when it exceeds a threshold (with error), and a negative choice otherwise.

Parametric models are typically chosen due to their tractability and ease of interpretation. For excellent overviews of estimation and inference based upon parametric binary choice models see Amemiya (1981), Mcfadden (1984), Blundell (1987), and Davidson & Mackinnon (1993). Typically, parametric models of binary choice are based on the concept of an unobserved or 'latent' variable Y_i^* (see Davidson & Mackinnon (1993, page 514).

The traditional parametric approach to modeling binary choice is as follows:

1. Assume that the true data generating processes (DGP) is given by

$$Y_i^* = E[Y^*|X_i] + U_i = X_i'\beta + U_i,$$

where $X_i \in \mathbb{R}^k$, but that we only observe $Y_i = 1$ if $Y_i^* > 0$ otherwise we observe $Y_i = 0$.

2. $Y_i = 1$ therefore occurs when $X_i'\beta + U_i > 0$, that is, when $U_i > -X_i'\beta$.

3. $Pr[Y_i = 1]$ is therefore equal to $Pr[U_i > -X_i'\beta] = 1 - Pr[U_i < -X_i'\beta]$.

4. Assuming symmetry of the distribution of U, $1 - Pr[U_i < -X_i'\beta] = Pr[U_i < X_i'\beta] = F(X_i'\beta)$.

5. Assuming that the distribution of U is of known parametric form, then we can estimate β by the maximum likelihood method[1].

6. Empirical choice probabilities are given by $F(x'\tilde{\beta})$, while the gradient of choice probability with respect to the conditioning information is given by $\partial F(x'\tilde{\beta})/\partial x$, where $\tilde{\beta}$ is the maximum likelihood estimator of β.

Clearly this approach has its origins in regression modeling and the use of a scalar index function $X_i'\beta$ plays a key role. The choice probability function $F(\cdot)$ (the CDF of U) is independent of the distribution of those variables believed to influence choice by assumption, and symmetry of $F(\cdot)$ is typically assumed. This assumed symmetry therefore imposes symmetry on the choice probability gradient, each element having identical shape and differing only by $\tilde{\beta}_j$, $j = 2, \dots, k$. In this framework there is no way for us to estimate the shape of the error distribution, either parametrically or nonparametrically since we do not observe U and cannot estimate U since we do not observe Y_i^*. This stands in stark contrast to a standard regression model in which the vector of residuals $\hat{U} = Y - X\hat{\beta}$ can be used to nonparametrically estimate the density of U since we actually observe the underlying dependent variable in this case. If we adopt a parametric approach we must simply presume a parametric distribution for U whose location and scale is determined by the observed choices $Y \in \{0, 1\}$ and by $X_i'\beta$.

1.2. Semiparametric Estimation of Single-Index Binary Choice Models

There exists a rich and very impressive variety of approaches towards semiparametric estimation of discrete choice models including that of Coslett (1983), Ichimura (1986), Manski (1985), Rudd (1986), Ichimura & Lee (1991), Coslett (1991), Klein & Spady (1993), Lee (1995), Chen & Randall (1997), Picone & Butler (1998), and Ichimura & Thompson (1998) to name but a few. For a recent survey of semiparametric approaches to the estimation of discrete choice models see Pagan & Ullah (1999 Chapter 7).

In this framework note that $Pr[Y_i = 1] = E[Y_i | X_i'\beta]$, hence one cannot proceed by estimating $E[Y_i | X_i]$ using standard nonparametric regression

[1] For example, assuming a Gaussian distribution for U leads to the widely-used ‘Probit’ model, while assuming a Logistic distribution yields the widely-used ‘Logit’ model.

techniques as this fails to capture the single-index nature of the model. A number of leading approaches therefore use nonparametric techniques to estimate $E[Y_i|X_i'\beta]$ (see, for example, Ichimura (1986) and Klein & Spady (1993)), and the resulting estimators are essentially ratios of nonparametric density estimates which may require trimming both to deal with behavior of the numerator at boundary points and to obtain asymptotic results. These approaches are very similar in spirit to univariate Nadaraya-Watson regression (Nadaraya (1965), Watson (1964)) with the index function itself serving as conditioning information.

Existing semiparametric approaches share a number of features; they typically assume an underlying latent variable specification; they typically employ a scalar-index which restricts the choice threshold to be a hyperplane; they typically model the (univariate) distribution of the scalar index function in order to generate empirical choice probabilities; data-driven methods for bandwidth selection or its counterpart when using flexible functional forms are not employed, and they do not permit data-driven index choice.

1.3. Semiparametric Estimation of Generalized Binary Choice Models

In this paper, a semiparametric approach to the estimation of 'generalized' binary choice models is proposed. A 'generalized' binary choice model is one with separate indices for each conditioning variable which constitutes a generalization of the standard single-index approach typically employed in applied work. The choice probability distribution is therefore a joint distribution across these indices as opposed to the typical univariate distribution on a scalar index commonly found in applied work.

In this paper the approach taken is to focus on modeling $Pr[Y_i = 1]$ via kernel estimation of a (joint) CDF rather than by modeling $E[Y_i|X_i'\beta]$ via a ratio of density estimates. It was noted that existing semiparametric approaches are similar in spirit to Nadaraya-Watson kernel regression (Nadaraya (1965), Watson (1964)), and the proposed approach is similar in spirit to the Priestly-Chow regression estimator (Priestley & Chao (1972)). One immediate benefit will be the absence of trimming in this approach. In addition, the proposed approach adopts the use of multiple thresholds the modeling of which will be straightforward and natural in this setting.

This paper builds upon existing work and develops a number of ideas in the context of binary choice modeling. First, the joint distribution of transformations of the conditioning variables is estimated nonparametrically using the method of kernels. This extends existing approaches which employ a univariate distribution defined over a scalar index, and the distribution of variables influencing choice influences both choice probabilities and associated gradients. Second, the indices are assumed to be of known form but are generalized to allow for multiple thresholds. Third, both bandwidth and parameter selection is data-driven using a new likelihood-based cross-validatory method. The value added by the proposed approach lies in its ability to model situations wherein a single index may be appropriate, but in addition is capable of handling a richer range of phenomena than those which single-index models are capable of handling. Finally, the functional form of the index can be data-driven since it enters the nonparametric likelihood function thereby generalizing existing approaches in which the index is assumed to be known.

The modeling of binary choice via multiple thresholds is now outlined, and then a semiparametric implementation is proposed.

2. BACKGROUND AND ASSUMPTIONS

Consider a situation for which

$$Y_i = \begin{cases} 1 & \text{if a choice is made} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \dots, n \tag{2}$$

$Y_i \in \mathbb{R}^1$ is assumed to be a random variable that depends on a random vector of characteristics, $X_i \in \mathbb{R}^k$. Interest lies in predicting the probability that $Y_i = 1$ given a realization of the vector of characteristics, x_i, and in assessing how this probability changes with changes in these characteristics.

As noted, this paper considers the case in which choices are governed in general by multiple thresholds defined over the choice variables or transformations thereof, one for each variable influencing choice. In this setting, $Pr[Y_i = 1]$ can be expressed as a joint distribution function given by

$$\begin{aligned} Pr[Y_i = 1] &= Pr[g_1(x_{i1}, \theta^1) \geq g^1 \text{ and } g_2(x_{i2}, \theta^2) \geq g^2 \dots \text{ and } g_k(x_{ik}, \theta^k) \geq g^k] \\ &= F(g^1 - g_1(x_{i1}, \theta^1), \dots, g^k - g_k(x_{ik}, \theta^k)) \end{aligned} \tag{3}$$

where the functions $g_j(x_{ij}, \theta^j)$, $j = 1, \dots, k$, are (unknown) functions which influence choices, θ^j a vector of parameters, g^is thresholds above which choices tend to be made, and $F(\cdot)$ a joint distribution function. It is noted that

$$\begin{aligned} Pr[Y_i = 0] = 1 - Pr[Y_i = 1] = Pr[g_1(x_{i1}, \theta^1) < g^1 \\ \text{and } g_2(x_{i2}, \theta^2) < g^2 \dots \text{ and } g_k(x_{ik}, \theta^k) < g^k] \\ = 1 - F(g^1 - g_1(x_{i1}, \theta^1), \dots, g^k - g_k(x_{ik}, \theta^k)) \end{aligned} \tag{4}$$

As usual, the choice probability gradient is defined as

$$\nabla_x Pr[Y_i = 1] = \frac{\partial Pr[Y_i = 1]}{\partial x} \in \mathbb{R}^k \tag{5}$$

which tells us the response of the choice probability due to changes in the factors influencing choice.

We can write the PDF of the choice Y_i as

$$\begin{aligned} f(y_i) = {} & \left[F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right]^{y_i} \\ & \times \left[1 - F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right]^{(1-y_i)} \end{aligned} \tag{6}$$

where $y_i \in \{0,1\}$, and for independent realizations of Y_i the joint density function is given by

$$\begin{aligned} f(y_1,\ldots,y_n) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \left[F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right]^{y_i} \\ &\quad \times \left[1 - F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right]^{(1-y_i)} \end{aligned} \tag{7}$$

Estimation of binary choice models is typically based upon the method of maximum likelihood where the log-likelihood function is given by

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^{n} & y_i \ln\left[F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right] \\ & + (1-y_i)\ln\left[1 - F(g^1 - g_1(x_{i1},\theta^1),\ldots,g^k - g_k(x_{ik},\theta^k))\right] \end{aligned} \tag{8}$$

and the resulting estimators will be referred to collectively as $\tilde{\theta}$.

Hypothesis testing can often be based on the asymptotic variance-covariance matrix. For example, the variance-covariance matrices for linear indices is given by

$$V[\tilde{\theta}] = \sum_{i=1}^{n} \left[\frac{f^2(x_i,\theta)}{F(x_i,\theta)[1-F(x_i,\theta)]} x_i x_i'\right]^{-1} \tag{9}$$

where θ denotes a vector of all parameters in the model and where x_i a vector of realizations of all variables influencing choice. An estimate is obtained by evaluating Equation 9 at $\tilde{\theta}$. Alternatively, one could use the method of obtaining Equation 9 via computation of the Hessian of the likelihood function which often must be obtained numerically due to nonlinearities in the likelihood function itself.

3. SEMIPARAMETRIC GENERALIZED BINARY CHOICE WITH KNOWN INDICES

The approach taken in this paper involves the nonparametric estimation of a joint distribution function which has as its arguments generalized known indices of each conditioning variable. Estimation proceeds via the method of kernels based on a nonparametric likelihood function which is maximized using cross-validatory techniques. Cross-validation is employed for both bandwidth and parameter selection and, in addition, the parametric form of the index itself can also be selected via cross-validation by letting the nature of the index be an argument of the cross-validatory likelihood function.

Kernel estimation of distribution functions can be based upon an estimator such as the Nadaraya-Watson (Nadaraya (1965), Watson (1964)) estimator of a joint density function. The kernel estimator of a joint density function is given by

$$\hat{f}(x_1,\ldots,x_i) = \frac{1}{n\prod_{j=1}^{k} h_j}\sum_{i=1}^{n} K\left(\frac{x_1 - x_{i1}}{h_1},\ldots,\frac{x_k - x_{ik}}{h_k}\right) \tag{10}$$

The kernel estimator of the joint distribution function (Prakasa & Rao (1983 page 397)) is therefore given by

$$\begin{aligned}\hat{F}(x_1,\ldots,x_k) &= \int_{-\infty}^{x_1,\ldots,x_k} \hat{f}(t_1,\ldots,t_i)\,dt_1,\ldots,dt_k\\ &= \frac{1}{n\prod_{j=1}^{k} h_j}\sum_{i=1}^{n}\int_{-\infty}^{x_1,\ldots,x_k} K\left(\frac{t_1 - x_{i1}}{h_1},\ldots,\frac{t_k - x_{ik}}{h_k}\right)dt_1,\ldots,dt_k\\ &= \frac{1}{n}\sum_{i=1}^{n} K_{int}\left(\frac{x_1 - x_{i1}}{h_1},\ldots,\frac{x_k - x_{ik}}{h_k}\right)\end{aligned} \tag{11}$$

where $K_{int}(\cdot) = \int_{-\infty}^{x_1,\ldots,x_k} K(\cdot)\,dt_1,\ldots,dt_k$ is a kernel distribution function.

Once we assume a specific parametric function for the indices ($g_j(x_{ij},\theta^j)$, $j = 1,2,\ldots,k$) we can then obtain the kernel estimator of the joint distribution function of these indices via simple transformation of variables. For the commonly used linear index this involves nothing more than simply replacing variable j with the index itself.

It should be noted that we have used the standard Rosenblatt-Parzen estimator (Silverman (1986 page 40, 76)), but this approach remains unchanged when using other nonparametric density estimators such as generalized nearest-neighbor (Silverman (1986 page 21)) and adaptive (Silverman (1986 page 21)) approaches.

By way of example, consider a situation in which the indices are linear in nature, hence an index for variable j would be given by $\theta_0^j + \theta_1^j x_{ij}$.

The kernel estimator of the joint distribution function ($Pr[Y_i = 1]$) would be given by

$$\begin{aligned}&\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik}) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{int}\left[\left(\frac{\theta_0^1 + \theta_1^1(x_1 - x_{i1})}{h_1}\right), \ldots, \left(\frac{\theta_0^k + \theta_1^k(x_k - x_{ik})}{h_k}\right)\right]\end{aligned} \tag{12}$$

while the gradient vector $\nabla\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})$ would have typical element

$$\begin{aligned}&\frac{\partial\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})}{\partial x_j} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial x_j} K_{int}\left[\left(\frac{\theta_0^1 + \theta_1^1(x_1 - x_{i1})}{h_1}\right), \ldots, \left(\frac{\theta_0^k + \theta_1^k(x_k - x_{ik})}{h_k}\right)\right]\end{aligned} \tag{13}$$

This estimator is very similar to the Priestly-Chow regression estimator (Priestley & Chao (1972)) outside of the obvious difference in the nature of the kernel function.

4. ESTIMATION OF THE MODEL.

In applied settings we clearly require a method of determining the smoothing parameters h, the parameters of the index function θ, and possibly the functional forms of the indices themselves.

Suppose for the moment that the form of the index is of known linear form. We could therefore express the log-likelihood function as

$$\begin{aligned}\mathcal{L}(\theta, h) = \sum_{i=1}^{n} & y_i \ln\left[\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})\right] \\ & + (1 - y_i)\ln\left[1 - \hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})\right]\end{aligned} \tag{14}$$

We would like to proceed to select h and θ in a manner similar to that used in a parametric framework.

Unfortunately, as is common in nonparametric settings, this objective function is unbounded. We therefore considering maximizing the cross-validatory log-likelihood function using a leave-one-out kernel estimator in which observation i is omitted when computing $\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})$. This now allow us to conduct data driven bandwidth and parameter estimation. Letting $\mathcal{L}_{-i}(\theta, h)$ denote the cross-validatory likelhood function in which observation i is omitted when computing $\hat{F}(\theta_0^1 + \theta_1^1 x_{i1}, \ldots, \theta_0^k + \theta_1^k x_{ik})$, we define

$$(\tilde{\theta}, \tilde{h}) = \underset{\theta,h}{\operatorname{argmax}}\ \mathcal{L}_{-i}(\theta, h) \tag{15}$$

and consistency of this method would be expected to follow given the results of Stone (1974) though a rigorous proof of this statement is not attempted here.

At this point it should be evident that we could admit the functional form of the index as yet another argument entering this likelihood function so that, in addition to the bandwidths h and index parameters θ, choice of the index functions themselves could be data-driven thereby removing the need to assume that the functional form of the index is known prior to estimation.

Much of the focus of a number of excellent recent papers has been on estimation of the index parameters themselves. However, Amemiya (1981 page 1488) notes that "when one wants to compare models with different probability functions, it is generally better to compare probabilities directly rather than comparing the estimates of the coefficients even after an appropriate conversion." He also suggests that "an alternative way of comparing different models is to look at the derivatives of the probabilities with respect to a particular independent variable." In light of this, in this paper interest will center on estimation of the choice probabilities $F(\cdot)$ and their gradient $\partial F(\cdot)/\partial x$ in this generalized setting when the distribution function is not assumed to be of known parametric form.

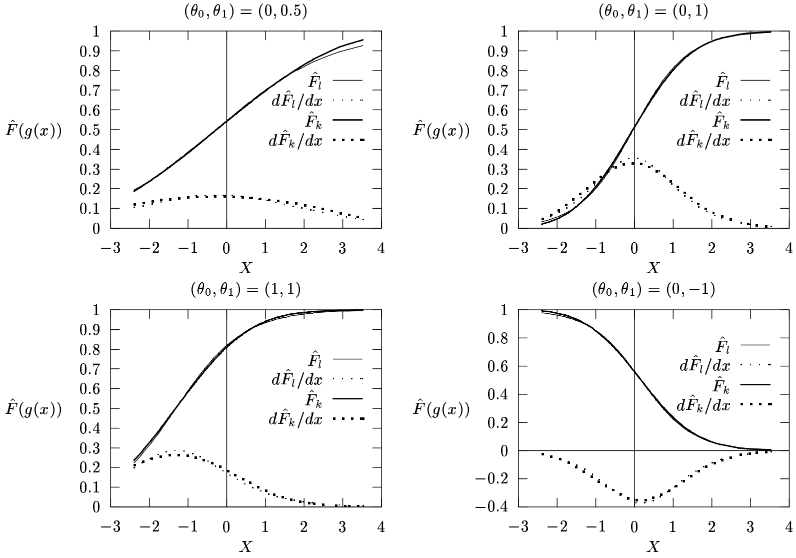
We now turn to the behavior of the proposed method relative to existing single-index methods. For comparison purposes, we consider a traditional parametric approach along with the proposed approach.

5. SOME SIMPLE EXAMPLES

One key difference between many existing approaches and the proposed approach lies in the modeling of the choice distribution itself. For instance, the traditional parametric approach requires the presumption of a known functional form for the distribution of an error term U, and the shape of this distribution is independent of the distribution of the variables which influence choice. However, in the proposed approach the distribution depends on the variables influencing choice, which is also assumed by Klein & Spady (1993) for instance though in a single-index setting.

We first consider the case of one conditioning variable in order to highlight this difference. We first consider the case in which a Logit model would be appropriate in that the distribution of X is symmetric, and then

FIG. 1. Choice probability and gradient vector for simulated normal data $x \sim N(0,1)$, $n = 250$. $\hat{F}_l$ represents that based on the logistic parametric specification, while $\hat{F}_k$ represents that based on a nonparametric kernel estimator. All bandwidths were selected via maximization of the cross-validatory maximum likelihood function.



consider the case where X is bimodal in order to contrast the proposed approach from existing approaches.

5.1. Univariate: Symmetric Distribution of X

We begin by considering the behavior of the proposed estimator relative to that based on the widely-used logistic specification when the distribution of the variables influencing choice is symmetric.

For the experiments, the choice probabilities were determined from the Logit model

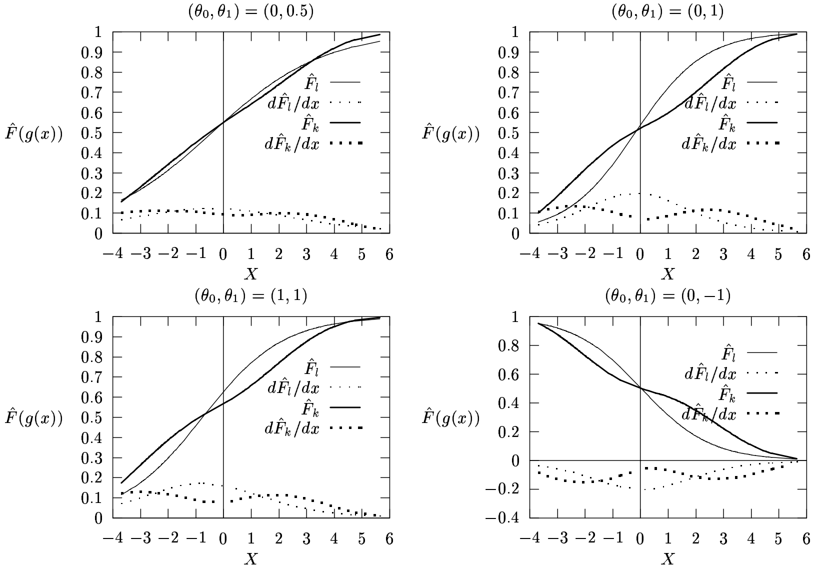
$$y_i = \begin{cases} 1 & \text{if } \left[1/(1+e^{-(\theta_0+\theta_1 x_i)}) + u_i\right] \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n \tag{16}$$

where $x \sim N(0,1)$ and $u \sim N(0, 0.25^2)$. The values of (θ_0, θ_1) were varied, and results for four simulated data sets are plotted in Figure 1.

As can be seen, even for a sample which is small by nonparametric standards, the proposed approach appears to give reasonable estimates of the choice probabilities and gradient which are consistent with the underlying DGP.

However, suppose that the distribution of X was bimodal, for example. The proposed approach would use this information while standard models

FIG. 2. Choice probability and gradient vector for simulated bimodal normal data $x \sim N(-2.5 : 2.5, 0.5^2 : 1.0)$, $n = 250$. $\hat{F}_l$ represents that based on the logistic parametric specification, while $\hat{F}_k$ represents that based on a nonparametric kernel estimator. All bandwidths were selected via maximization of the cross-validatory maximum likelihood function.



$(\theta_0, \theta_1) = (0, 0.5)$

$(\theta_0, \theta_1) = (0, 1)$

$(\theta_0, \theta_1) = (1, 1)$

$(\theta_0, \theta_1) = (0, -1)$

cannot as they assume that choice probabilities are independent of the distribution of the variables influencing choice.

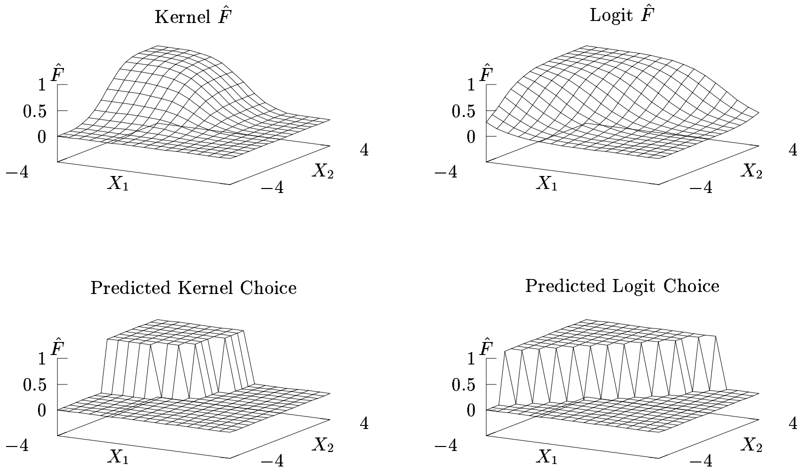
5.2. Univariate: Bimodal Distribution of *X*

We now consider the behavior of the proposed estimator and the logistic estimator when the data on X is bimodal, $x \sim N(-2.5 : 2.5, 0.5^2 : 1.0)$.

Which approach is appropriate? Clearly, if the latent-variable model is driving choices then the Logit model is appropriate. However, if choices are governed by variables influencing choices and not by an independent error process, then the proposed approach will be able to capture such behavior. In either case, predicted choice probabilities will be similar, but choice gradients will differ.

5.3. Multivariate: Symmetric distribution of X_1, X_2

FIG. 3. Choice probability and predicted choices for simulated data with $X_1, X_2 \sim N()$, $n = 500$. The first column contains that for the kernel estimator, the second for the Logit.



For the experiments, the choice probabilities were determined from the Logit model

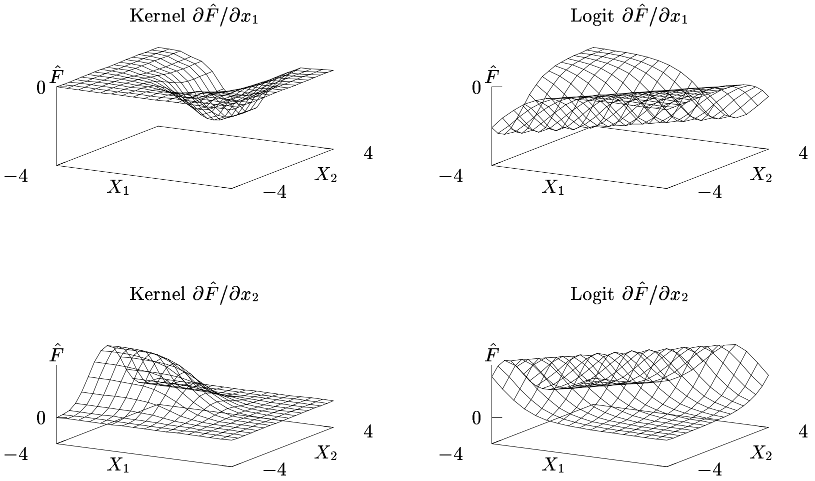
$$y_i = \begin{cases} 1 & \text{if } \left[\frac{1}{(1+e^{-(\theta_{01}+\theta_{11}x_{i1})})}\frac{1}{(1+e^{-(\theta_{02}+\theta_{12}x_{i2})})} + u_i\right] \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n \tag{17}$$

where $x_1, x_2 \sim N(0, 1)$ and $u \sim N(0, 0.5^2)$. The values of $(\theta_{01}, \theta_{11}, \theta_{02}, \theta_{12})$ were varied, and results for four simulated data sets are plotted in Figure 3. Note that the true choices should be $Y_i = 1$ if X_1, X_2 is in the northwest quadrant of the input space. However, there is noise thrown on top of the inputs hence the somewhat gradual response of the probability due to changes in the inputs.

An examination of Figure 3 reveals that models such as the Logit cannot capture this simple thresholding which would be expected in simple economic models. The Logit model can only fit one hyperplane through the input space which falls down the diagonal axis. However, the kernel estimator permits two (K) such hyperplanes thereby permitting consistent estimation of the choice probabilities conditional upon the index specification.

An examination of the choice gradients graphed in Figure 4 reveals both the appeal of a multiple threshold approach and the limitations of single-index models. The gradient with respect to X_1 is everywhere negative and everywhere positive with respect to X_2. By way of example, note that,

FIG. 4. Choice gradients for simulated data with $X_1, X_2 \sim N()$, $n = 250$. The first column contains that for the kernel estimator, the second for the Logit.


Kernel $\partial\hat{F}/\partial x_1$

Logit $\partial\hat{F}/\partial x_1$

Kernel $\partial\hat{F}/\partial x_2$

Logit $\partial\hat{F}/\partial x_2$

when $x_2 = -3.5$, the gradient with respect to X_1 is zero almost everywhere. The kernel estimator picks this up, but the Logit specification imposes non-zero gradient where in fact there is none. The same phenomena can be observed when $X_1 = 3.5$ whereby the true gradient with respect to X_2 is zero almost everywhere, but again the logistic specification imposes a discernibly non-zero gradient.

This effect occurs when the data is in fact jointly symmetrically distributed. When the data is skewed, asymmetric, and/or bimodal, the gradients given by widely-used parametric models should be viewed with extreme caution.

6. SIMULATIONS - LATENT VARIABLE SPECIFICATION

We consider a simple simulation in order to gauge how the proposed method performs relative to widely-used parametric methods. We consider the case for which a Probit model is the correctly-specified single-threshold parametric model. The second-order Epanechnikov kernel was employed for the following set of simulations, and both the bandwidth and index parameters were selected via the proposed method of likelihood cross-validation. For both the parametric and semiparametric approaches, we employ linear indices throughout.

TABLE 1.

DGP is a bivariate latent variable model.

	% Correct		Avg. Grad		$\log \mathcal{L}$	
n	Par	Sem	Par	Sem	Par	Sem
$\sigma_u = 0.5$						
50	85%	85%	0.36	0.10	-15.4	-18.5
100	86%	86%	0.36	0.15	-30.6	-37.0
$\sigma_u = 2.0$						
50	76%	76%	0.28	0.16	-23.8	-24.2
100	75%	75%	0.28	0.20	-49.1	-49.7
$\sigma_u = 2.0$						
50	67%	67%	0.19	0.17	-30.0	-30.1
100	66%	66%	0.18	0.18	-61.2	-61.5

We consider the case where the true DGP is in fact a latent variable model given by

$$Y_i^* = \beta_1 + \beta_2 X_i + U_i \tag{18}$$

where $Y_i = 1$ if $Y_i^* > 0$ and where $(\beta_1, \beta_2) = (0, 1)$, $X_i \sim N(0, 1)$ and $U_i \sim N(0, \sigma_u^2)$.

Since $U \sim N(0, \sigma^2)$ then the Probit model is appropriate. In order to facilitate comparison with the Probit model, we follow the convention of expressing the choice probability gradient as the sample average of the gradient. We therefore will compare percentage of correct predictions and the sample average of the choice probability gradient for the parametric Probit and the semiparametric specification proposed in this paper. 1,000 Monte Carlo replications were drawn from this DGP, and median percentage of correct predictions as well as median choice probability gradient are reported. For each replication, the proposed method of likelihood cross-validation was used to select the bandwidth h and parameter vector $\theta = (\theta_1, \theta_2)$. Sample sizes of $n = 50$ and 100 are considered, and results appear in Table 1 below.

These results suggest that the predictive ability of the semiparametric model does not differ from that of a correctly specified Probit model, while the proposed approach appears to have a slight downward bias in the average choice probability derivative which disappears as the sample size increases. Finally, this bias also disappears as the variance of U increases.

7. APPLICATION - PREDICTING VOTING BEHAVIOR

We consider an example in which we model voting behavior given information on various economic characteristics on individuals. For this example we consider the choice of voting 'yes' or 'no' for a local school tax referendum. Two economic variables used to predict choice outcome in these settings are income and education. This is typically modeled using a Logit or Probit model in which the covariates are expressed in logarithm() form.

Interest lies in the predictive power of the proposed approach versus models such as the widely-used Logit or Probit models. This aim of this modest example is simply to gauge the performance of the proposed method in a real-world setting. Data was taken from Pindyck & Rubinfeld (1998 page 332-333), and there were a total of $n = 95$ observations available. Table 2 summarizes the results from this modest exercise.

TABLE 2.

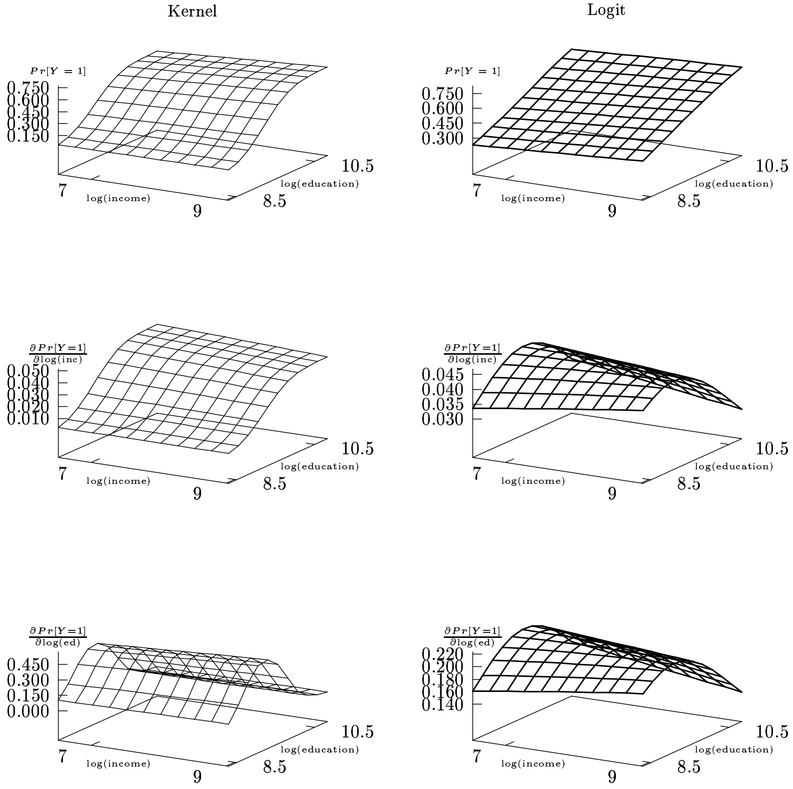
Comparison of models of voting behavior.

		Average Derivative	
Method	% Correct	log(*income*)	log(*education*)
Logit	65.3%	0.04	0.20
Probit	65.3%	0.05	0.20
Kernel	67.4%	0.04	0.19

It appears that, for this example, the proposed method yields improved prediction probabilities. As well, the average derivatives are extremely close for both methods. However, an examination of Figure 5 reveals that the derivatives are quite different which is missed when quoting the average derivative as is typically done. As well, note the well-known fact that Logit models impart the same shape on all derivatives when using the standard linear index, but as can be seen this restriction is not imposed by the proposed method.

This modest example is in no way intended to be a serious investigation into the prediction of voting outcomes. For this simple application, the predictive ability of the model increases when admitting multiple thresholds, and the average choice probability gradient is in line with that from standard models of binary choice. These results suggest that allowing for multiple thresholds may improve predictive ability when modeling binary choice and may constitute a tool which could benefit applied researchers in their quest to model binary choice outcomes.

FIG. 5. Choice probabilities and gradient vectors for voting data where X_1 is log(income) and X_2 is log(education). The light lined surfaces (left hand side graphs) are those for the proposed method while the dark liked surfaces (right hand side graphs) are those for the Logit model. For this example both approaches employ linear indices.



8. CONCLUSION

It is widely known that misspecification of parametric models can yield biased and inconsistent estimates and inference based on such models would be invalid. This paper proposes a method for estimating generalized binary choice models admitting multiple thresholds for which the joint distribution is estimated nonparametrically and the functional form of the index is data-dependent. It is demonstrated how standard parametric models such as the Logit model can distort both the choice probabilities and choice probability gradient in simple situations. The proposed approach uses nonparametric methods for estimation of the choice probabilities and associated gradients.

Future work includes orthogonality tests for a subset of variables and extension of the index to include a fully nonparametric index in addition to the choice probability distribution.

REFERENCES

Amemiya, T., 1981, Qualitative response models: A survey. *Journal of Economic Literature* **19**, 1483-1536.

Blundell, R., ed., 1987, *Journal of Econometrics* **34**, North Holland.

Chen, H. Z. and A. Randall, 1997, Semi-nonparametric estimation of binary response models with an application to natural resource valuation. *Journal of Econometrics* **76**, 323-340.

Coslett, S. R., 1983, Distribution-free maximum likelihood estimation of the binary choice model. *Econometrica* **51**, 765-782.

Coslett, S. R., 1991, Semiparametric Estimation of a Regressioin Model with Sampling Selectivity, Cambridge, pp. 175-198.

Davidson, R. and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics.* Oxford University Press.

Ichimura, H., 1986, Estimation of single index models. Unpublished Manuscript.

Ichimura, H. and L. F. Lee, 1991, Semiparametric Estimation of Multiple Index Models, Cambridge, pp. 3-50.

Ichimura, H. and T. S. Thompson, 1998, Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics* **86(2)**, 269-295.

Klein, R. W. and R. H. Spady, 1993, An efficient semiparametric estimator fo binary response models. *Econometrica* **61**, 387-421.

Lee, L. F., 1995, Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* **65**, 381-428.

Manski, C. F., 1985, Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313-333.

McFadden, D., 1984, Econometric analysis of qualitative response models, in Z. Griliches and M. Intriligator, eds, *Handbook of Econometrics*, North Holland, pp. 1385-1457.

Nadaraya, E., 1965, On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* **1O**, 186-190.

Pagan, A. and A. Ullah, 1999, Nonparametric Econometrics, Cambridge University Press.

Picone, G. and J. S. Butler, 1998, Semiparametric estimation of multiple equation index models. Unpublished Manuscript.

Pindyck, R. S. and D. L. Rubinfeld, 1998, Econometric Models and Economic Forecasts, Irwin McGraw-Hill.

Prakasa Rao, B., 1983, Nonparametric Functional Estimation, Academic Press.

Priestley, M. B. and M. T. Chao, 1972, Nonparametric function fitting. *Journal of the Royal Statistical Society* **34**, 385-392.

Rudd, P., 1986, Consistent estimation of limited dependent variable models despite mis-speification of distribution. *Journal of Econometrics* **32**, 157-187.

Silverman, B. W., 1986, Density Estimation for Statistics and Data Analysis, Chapman and Hall.

Stone, C. J., 1974, Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society* **36**, 111-147.

Watson, G., 1964, Smooth regression analysis. *Sanikhya* **26:15**, 175-184.