Nonparametric Estimation of Conditional Distributions in the Presence of Continuous and Categorical Data*

Jeff Racine
Department of Economics
University of South Florida
Tampa, FL 33620
jracine@coba.usf.edu

Qi Li
Department of Economics
Texas A&M University
College Station, TX 77843
qi@econ.tamu.edu

January 25, 2000

Abstract

A method is proposed for the consistent nonparametric estimation of conditional probability and probability density functions along with associated gradients when both the conditioned and conditioning variables are categorical, continuous, or a mixture of both types. The method builds on the work of Aitchison & Aitken (1976) who proposed a novel method for kernel density estimation when using multinomial categorical data types. Simulations show that the proposed method performs quite well for a number of conditional simulated processes that mix both categorical and continuous variables. Applications of the proposed method to (i) the widely-cited Iris dataset of Fisher (1936), (ii) the female labor supply dataset from the Panel Study on Income Dynamics examined in Mroz (1987), and (iii) the Swiss labor force data studied by Gerfin (1996) all demonstrate that the proposed method performs better than conventional parametric models for predicting multinomial discrete choice. The method extends the realm of nonparametric modeling through the seamless blending of both categorical and continuous variables, and is capable of detecting structure in the data which frequently remains undetected by conventional parametric approaches.

*Li's research is supported by the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanity Research Council of Canada, and by the Bush Program in the Economics of Public Policy. Racine would like to thank the USF Division of Sponsored Programs for their continuing support.

1 Introduction

Conditional probability density functions (CPDF) play a key role in applied statistical analysis. Often the CPDF is itself of direct interest while at other times it is embedded in objects of interest such as a conditional expectation or higher order conditional moment. Unfortunately, a parametric framework is not well-suited for the modeling of a CPDF. In a parametric framework we would require a functional specification for the joint CPDF combining potentially different marginal distributions for both continuous and categorical variables prior to estimation. Compound this with modeling unknown dependence among variables and, in the absence of knowledge regarding the underlying CPDF, parametric approaches quickly become intractable. Aitchison & Aitken (1976, page 419) refer to the simplest of such problems as "parametrically awkward". For this reason the direct modeling of CPDFs has received little attention even though such models could prove to be extremely valuable in a variety of situations. The modeling of labor force participation conditional upon a vector of personal characteristics (Mroz (1987)), the modeling of consumer choice and the response of choices to changes in variables influencing choice (Amemiya (1981), McFadden (1984)), and the modeling of nonlinear discriminant rules (Mardia, Kent & Bibby (1979, Chapter 11)) are all examples of situations which could be modeled via a CPDF.

The intractable nature of modeling a CPDF in a parametric framework arises simply because we are ignorant of nature's data generating process (DGP). Parametric approaches force us to make 'functional guesses' prior to estimation which are unlikely to be correct in this setting, and using an incorrectly specified parametric CPDF will result in biased and inconsistent estimates, while hypothesis tests based upon such estimates will have asymptotically incorrect size and power.

Nonparametric methods, on the other hand, permit us to model a CPDF without requiring that the researcher correctly specify the unknown distribution, and are consistent under

less restrictive assumptions than those required for the consistency of parametric methods though at the cost of rates of convergence which depend on the number of variables involved (often called the 'curse of dimensionality'). This tradeoff is commonly encountered when choosing modeling procedures - if you impose less structure then you need more data in order to achieve the same degree of precision. When faced with discrete data, however, the conventional nonparametric approach uses a 'frequency estimator' to handle the discrete variables by splitting the sample into a number of subsets or 'cells'. When using conventional frequency-based nonparametric approaches one suffers a loss of efficiency arising from a reduction in the sample size due to splitting the sample into a number of cells and also faces the issue of how to assess interaction effects.

This paper considers a kernel-based solution to the problem of modeling a CPDF and related objects when faced with both continuous and categorical data types. We consider hybrid multivariate product kernels in which 'categorical kernels' and 'continuous kernels' can be seamlessly mixed building upon the work of Aitchison & Aitken (1976) who proposed a novel method of nonparametric density estimation for multivariate binary data. Silverman's (1986) book also contains a brief but informative discussion of these issues. The proposed method does not suffer from finite-sample efficiency losses arising from sample splitting, and naturally handles interaction among the discrete and continuous variables. The strength of the proposed method lies in its ability to model situations involving complex dependence among categorical and continuous data in a fully-nonparametric regression framework.

Related work includes that of Hall (1981) who considered bandwidth selection issues which arise when using the method of Aitchison & Aitken (1976) when there exist empty cells for categorical data and who proposed a robust solution to this problem, Hall & Wand (1988) who considered nonparametric discrimination in which the bandwidths for the density for each population are chosen jointly (they model both categorical and continuous variables), and Chaudhuri & Dewanji (1995) who considered theoretical underpinnings of likelihood

cross-validation for both parametric and nonparametric approaches to the estimation of conditional probabilities for continuous data types.

This paper proceeds as follows: in Section 2 we consider the underlying DGP, while in Section 2.2 we outline the hybrid multivariate product kernel which is central to the current work; Section 2.3 outlines the proposed nonparametric estimator of the conditional density function and its gradient in the presence of categorical and continuous data types; Section 4 considers a number of empirical applications of the proposed technique; Section 5 reports simulation results that examine the finite sample performance of the proposed estimator, while Section 6 concludes.

2 Estimating A Conditional Density With Mixed Categorical and Continuous Data

2.1 Background and Notation

Let $(Y, X) = (Y_1, \ldots, Y_k, X_1, \ldots, X_p)$ denote a $(k+p)$-dimensional vector of random variables of interest, and define $y = (y_1, \ldots, y_k)$ and $x = (x_1, \ldots, x_p)$ to be k and p-dimensional realizations of $Y = (Y_1, \ldots, Y_k)$ and $X = (X_1, \ldots, X_p)$ respectively. The joint density of the random vector Y conditional on the vector X is defined as

$$g(Y_1, \ldots, Y_k | X_1, \ldots, X_p) = \frac{f(Y_1, \ldots, Y_k, X_1, \ldots, X_p)}{f_x(X_1, \ldots, X_p)} \tag{1}$$

which we shall write simply as

$$g(Y|X) = \frac{f(Y, X)}{f_x(X)} \tag{2}$$

where $g(\cdot)$ denotes the density of Y conditional upon X, $f(\cdot)$ the joint density of (Y, X), and $f_x(\cdot)$ the marginal density of X.

Often the response of the CPDF with respect to the conditioning data is of interest. We define the gradients of $g(Y|X)$ with respect to X as

$$\nabla_x g(Y|X) = \frac{\partial g(Y|X)}{\partial X} \in \mathbb{R}^p \tag{3}$$

Now we arbitrarily let $Y = (Y_1^d, \ldots, Y_{k_d}^d, Y_{k_d+1}, \ldots Y_k)$ denote the vector $Y = (Y_1, \ldots, Y_k)$ with the first k_d variables being the categorical ones and the remaining $k - k_d$ being continuous. As well, let $X = (X_1^d, \ldots, X_{p_d}^d, X_{p_d+1}, \ldots X_p)$ denote the vector $X = (X_1, \ldots, X_p)$ with the first p_d variables being the categorical ones and the remaining $p - p_d$ being continuous. We use Z to denote (Y, X). Without loss of generality we assume that each categorical variable Z_t^d can assume the c_t discrete values $0, \ldots, c_t - 1$ for $t = 1, \ldots, k_d + p_d$ and $c_t \geq 2$.

We now turn our attention to the consistent nonparametric estimation of $g(Y|X)$ and $\nabla_x g(Y|X)$ in the presence of mixed continuous and categorical variables.

2.2 Hybrid Kernels Admitting Mixed Categorical and Continuous Data

We briefly review the approach of Aitchison & Aitken (1976) towards kernel estimation of probability functions for categorical data and demonstrate how this can be applied to the estimation of conditional probability functions. Extensions to handle more complicated cases involving categorical data such as ordered categories follow naturally and the interested reader is referred to Aitchison & Aitken (1976) and Habbema, Hermans & Remme (1978) for examples of such categorical kernels.

Recall that we defined $Z = (Y, X)$, and we can also partition Z into Z^d and Z^c, where Z^d contains the discrete variables in Z, and Z^c is the remaining continuous variables. Let Z_t^d be the tth component of Z^d and assume Z_t^d is a c_t-category discrete variable. A univariate

kernel for c_t-category data (Bowman (1980)) is given by

$$l(Z^d_{t,i}, Z^d_{t,j}, \lambda_t) = \begin{cases} \frac{1-\lambda_t}{c_t - 1} & \text{if } Z^d_{t,i} = Z^d_{t,j} \\ \lambda_t & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n, \tag{4}$$

where λ_t is a smoothing parameter. One of the interesting features of this kernel is that when $\lambda_t = 0$ we obtain the maximum likelihood estimator (probabilities for each category given by sample relative frequencies) while when $\lambda_t = 1/c_t$ we obtain a uniform distribution across the categorical variable (equal probabilities for each category). The product kernel for Z^d is defined as

$$L(Z^d_i, Z^d_j, \lambda) = \prod_{t=1}^{k_d+p_d} l(Z^d_{t,i}, Z^d_{t,j}, \lambda_t). \tag{5}$$

Let Z^c_t be the tth component of Z^c, let $w(\cdot)$ be a univariate kernel function for a univariate continuous variable, and let $W(\cdot)$ be the product kernel function for Z^c. We define

$$W(Z^c_i, Z^c_j, h) \stackrel{def}{=} \prod_{t=1}^{k+p-k_d-p_d} h_t^{-1} w\left(\frac{Z^c_{t,i} - Z^c_{t,j}}{h_t}\right), \tag{6}$$

where h_t is the smoothing parameter associated with continuous variable Z^c_t.

The product kernel for $Z = (Z^d, Z^c)$ is therefore given by

$$K(Z_i, Z_j, \lambda, h) = L(Z^d_i, Z^d_j, \lambda) W(Z^c_i, Z^c_j, h). \tag{7}$$

The product kernel for X is similarly defined as $K_x(X_i, X_j, \lambda_x, h_x) = L(X^d_i, X^d_j, \lambda_x) W(X^c_i, X^c_j, h_x)$.

2.3 Kernel Estimation of $g(Y|X)$

Define $Y_i = (Y^d_{i1}, \ldots, Y^d_{ik_d}, Y_{ik_d+1}, \ldots, Y_{ik})$ and $X_i = (X^d_{i1}, \ldots, X^d_{ip_d}, X_{ip_d+1}, \ldots, X_{ip})$ to be realizations of Y and X for $i = 1, \ldots, n$ where n denotes the sample size. Define λ_y and λ_x

to be vectors of smoothing parameters for the categorical variables in Y and X, and let h_y and h_x be vectors of smoothing parameters for the continuous variables in Y and X respectively. Letting $K(Y_i, Y_j, X_i, X_j, \lambda_x, h_x, \lambda_y, h_y)$ and $K_x(X_i, X_j, \lambda_x, h_x)$ denote multivariate product kernels using the categorical kernel if the variable is categorical and a continuous kernel for continuous data, the proposed kernel estimator of a conditional density evaluated at the point (Y_i, X_i) is given by

$$\hat{g}(Y_i|X_i) = \frac{\sum_{j=1}^{n} K\left(Y_i, Y_j, X_i, X_j, \lambda_x, h_x, \lambda_y, h_y\right)}{\sum_{j=1}^{n} K_x\left(X_i, X_j, \lambda_x, h_x\right)}, \quad i = 1, \ldots, n. \tag{8}$$

This is essentially the ratio of two Parzen (1962) estimators, the first being the estimator of the joint density of (Y, X) given by

$$\hat{f}(Y_i, X_i) = \frac{1}{n} \sum_{j=1}^{n} K\left(Y_i, Y_j, X_i, X_j, \lambda_x, h_x, \lambda_y, h_y\right), \quad i = 1, \ldots, n. \tag{9}$$

and the second being the estimator of the marginal density of X given by

$$\hat{f}_x(X_i) = \frac{1}{n} \sum_{j=1}^{n} K_x\left(X_i, X_j, \lambda_x, h_x\right), \quad i = 1, \ldots, n. \tag{10}$$

in which the smoothing parameter vector (λ_x, h_x) for the conditioning data is identical for both the joint and marginal density estimators. This mirrors the framework used for the Nadaraya-Watson (Nadaraya (1965), Watson (1964)) estimator of a conditional expectation in which the kernel function in the numerator and denominator employ the same bandwidths.

Bandwidth selection can proceed via likelihood cross-validation. Theoretical results for consistency of likelihood cross-validatory bandwidth selection for nonparametric estimators of a CPDF using continuous data can be found in Chaudhuri & Dewanji (1995). It is known, however, that maximum likelihood cross-validation can break down when the data is drawn

from fat-tailed distributions which is of concern when we mix continuous variables in with the discrete variables (see Hall (1987*a*), Hall (1987*b*)). An alternative to likelihood cross-validation is to choose (λ, h) to minimize the weighted integrated squared difference between $\hat{g}(y|x)$ and $g(y|x)$ which is known not to suffer from the aforementioned problems. Using the notation $\int dy\,dx = \sum_{y^d}\sum_{x^d}\int dy^c dx^c$, then a weighted integrated squared difference between $\hat{g}(\cdot)$ and $g(\cdot)$ is

$$\begin{aligned} I_n &= \int [\hat{g}(y|x) - g(y|x)]^2 f_x(x)\,dy\,dx \\ &= \int [\hat{g}(y|x)]^2 f_x(x)\,dy\,dx - 2\int \hat{g}(y|x)g(y|x)f_x(x)\,dy\,dx + \int [g(y|x)]^2 f_x(x)\,dy\,dx \\ &\equiv I_{1n} - 2I_{2n} + \int [g(y|x)]^2 f_x(x)\,dy\,dx, \end{aligned} \tag{11}$$

where $I_{1n} = \int [\hat{g}(y|x)]^2 f_x(x)\,dy\,dx$, $I_{2n} = \int \hat{g}(y|x)g(y|x)f_x(x)\,dy\,dx$, and the last term on the right-hand-side of Equation (11) does not depend on either λ or h. Define $\hat{G}(x) = \int [\hat{g}(y|x)]^2\,dy$. Then we have $I_{1n} = \int \hat{G}(x)f_x(x)dx = E[\hat{G}(X)]$. Therefore, we estimate I_{1n} by

$$\hat{I}_{1n} = \frac{1}{n}\sum_i \hat{G}(X_i). \tag{12}$$

Note that $I_{2n} = \int \hat{g}(y|x)g(y|x)f_x(x)\,dy\,dx = \int \hat{g}(y|x)f(y,x)\,dy\,dx = E[\hat{g}(Y|X)]$. Hence, we estimate $I_{2n} = E[\hat{g}(Y|X)]$ by

$$\hat{I}_{2n} = \frac{1}{n}\sum_i \hat{g}(Y_i|X_i). \tag{13}$$

Define $K_{x,ij} = K(X_i, X_j, \lambda_x, h_x)$, $K_{y,ij} = K(Y_i, Y_j, \lambda_y, h_y)$, and $K_{y,j} = K(y, Y_j, \lambda_y, h_y)$. Then using Equation (7) we have

$$\hat{G}(X_i) = \int [\hat{g}(y|X_i)]^2\,dy$$

$$
\begin{aligned}
&= \frac{\int [\hat{f}(y, X_i)]^2 \, dy}{[\hat{f}_x(X_i)]^2} \\
&= \frac{n^{-2} \sum_j \sum_l K_{x,ij} K_{x,il} \int K_{y,j} K_{y,l} \, dy}{[\hat{f}_x(x)]^2} \\
&= \frac{n^{-2} \sum_j \sum_l K_{x,ij} K_{x,il} K^{(2)}_{y,jl}}{[\hat{f}_x(X_i)]^2} \qquad (14)
\end{aligned}
$$

where $K^{(2)}_{y,j,l} = L^{(2)}(Y^d_j, Y^d_l, \lambda_y) W^{(2)}(Y^c_i, Y^c_j, h_y)$ with $L^{(2)}(Y^d_j, Y^d_l, \lambda_y) = \sum_{y^d} L(Y^d_j, y^d, \lambda_y) L(Y^d_l, y^d, \lambda_y)$, $W^{(2)}(Y^c_i, Y^c_j, h_y) = \prod_{t=1}^{k_d} h^{-1}_{y,t} w^{(2)}(\frac{Y^c_j - Y^c_l}{h_{y,t}})$ and $w^{(2)}(v) = \int w(u+v) w(u) \, du$ is the second order convolution kernel derived from $w(\cdot)$.

Therefore, we choose (λ, h) to minimize

$$
\begin{aligned}
CV(\lambda, h) &\equiv \hat{I}_{1n} - 2\hat{I}_{2n} \\
&= n^{-1} \sum_{i=1}^{n} \frac{n^{-2} \sum_{j=1}^{n} \sum_{l=1}^{n} K_{x,ij} K_{x,il} K^{(2)}_{y,ij}}{[\hat{f}_x(X_i)]^2} - 2 \frac{\hat{f}(Y_i, X_i)}{\hat{f}_x(X_i)} \qquad (15)
\end{aligned}
$$

We use $(\hat{\lambda}, \hat{h})$ to denote the above cross-validatory choice of (λ, h). The following assumptions are used to derive the rates of convergence of $(\hat{\lambda}, \hat{h})$ and $\hat{f}(z)$.

Assumption (A1) (i) $\{Z_i\}_{i=1}^n = \{X_i, Y_i\}_{i=1}^n$ is i.i.d. as $Z = (X, Y)$, $\mathcal{D}$, the support of Z^d, is finite, and $min_{\{z^d \in \mathcal{D}\}} p(z^d) \geq \delta$ for some $\delta > 0$. (ii) Let $f(y|x)$ denote the conditional density function of Y given $X = x$, assume that $f(\cdot|x)$ is four times differentiable, and assume that $f(y|x)$ and its derivatives are bounded on the support of Z^c for all $z^d \in \mathcal{D}$.

Assumption (A2) (i) The kernel function $w(\cdot)$ is non-negative, bounded and symmetric around zero, also $\int w(v) \, dv = 1$, $\int w(v) v^4 \, dv < \infty$. (ii) $\hat{h}$ lies in a shrinking set $H_n = \{h : h \in \mathcal{R}_+, h = o(1), (nh^p)^{-1} = o(1)\}$ (e.g., Härdle & Marron (1987)).

Theorem 2.1 *Under assumptions (A1) and (A2), we have*

$\hat{g}(y|x) - g(y|x) = o_p(1)$, *provided* $g(y|x) \geq \delta$ *for some* $\delta > 0$.

Proof: The proof for the general case is quite tedious. Here we only provide a proof for

the simple case where Z^d is a multivariate binary variable, $Z^d \in \{0,1\}^{k_d+p_d}$. Also, we will assume $\hat{\lambda}_t = \hat{\lambda}$ for all $t = 1, ..., k_d + p_d$.

First note that when $\lambda = 0$, $\hat{g}(y|x, \lambda = 0, \hat{h})$ becomes the usual frequency kernel estimator of $g(y|x)$, and we know that $g(y|x, \lambda = 0, \hat{h}) = o_p(1)$ because $\hat{h} \in H_n$. Then from $0 \leq I_n(\hat{\lambda}, \hat{h}) \leq I_n(0, \hat{h}) = o_p(1)$, we get

(i) $I_n(\hat{\lambda}, \hat{h}) = o_p(1)$.

Next, one can show that when $\lambda \neq o_p(1)$,

(ii) $I_n(\lambda, \hat{h}) = E[I_n(\lambda, \hat{h})] + o_p(1) = \sum_{s=1}^{2(k_d+p_d)} C_s \lambda^s + o_p(1) = O_p(1) \neq o_p(1)$,

where C_s's are constants and some of them are non-zero. That is, $I_n(\lambda, \hat{h})$ can be expanded as as a polynomial function of λ.

(i) and (ii) imply that $\hat{\lambda} = o_p(1)$. The fact that $\hat{\lambda} = o_p(1)$ and $\hat{h} \in H_n$ imply that $\hat{g}(y|x) - g(y|x) = o_p(1)$. This completes the proof of Theorem 2.1.

3 Kernel Estimation of $\nabla_x g(Y|X)$

Often interest lies in the response of the CPDF to changes in the conditioning variables. We shall call this vector of responses $\nabla_x g(Y|X) = \partial g(Y|X)/\partial X$. For continuous conditioning variables X^c we propose a kernel estimator of $\nabla_x g(Y|X)$ given by

$$\begin{aligned} \widehat{\nabla}_{X_i} \hat{g}(Y_i|X_i) &= \frac{\partial}{\partial X_i^c} \left\{ \frac{\sum_{j=1}^n K\left(Y_i, Y_j, X_i, X_j, \lambda_k^d, \lambda_k, \lambda_p^d, \lambda_p\right)}{\sum_{j=1}^n K_x\left(X_i, X_j, \lambda_p^d, \lambda_p\right)} \right\} \\ &= \frac{\partial}{\partial X_i^c} \left\{ \frac{\sum K(\cdot)}{\sum K_x(\cdot)} \right\} \\ &= \frac{\sum K_x(\cdot) \sum K'(\cdot) - \sum K(\cdot) \sum K_x'(\cdot)}{\left[\sum K_x(\cdot)\right]^2} \in \mathbb{R}^{p-p_d} \end{aligned} \tag{16}$$

where $K'(\cdot) = \partial K\left(Y_i, Y_j, X_i, X_j, \lambda, h\right)/\partial X_i^c$ and $K_x'(\cdot) = \partial K\left(X_i, X_j, \lambda_x, h_x\right)/\partial X_i^c$ are the $(p - p_d)$-dimensional analytical derivatives of the respective product kernel functions.

For categorical variables, arbitrarily focusing upon an assumed categorical variable X_1^d, we define the difference between the CPDF when $X_1^d = 0$ and $X_1^d = X_1^d$ $(0 \leq X_1^d \leq c_1 - 1)$ to be

$$g(Y_i|X_1^d = 0, \ldots, X_p = X_p) - g(Y_i \ldots, Y_k|X_1^d = X_1^d, \ldots, X_p = X_p). \tag{17}$$

which is naturally estimated using

$$\widehat{\nabla}_{X_1}\hat{g}(Y_i|X_i) = \hat{g}(Y_i|X_1^d = 0, \ldots, X_p = X_p) - \hat{g}(Y_i \ldots, Y_k|X_1^d = X_1^d, \ldots, X_p = X_p). \tag{18}$$

That is, for categorical variables we compute response simply by the difference in the CPDF when a categorical variable takes on the value 0 versus its sample realization while all other variables assume their sample realizations.

4 Applications

4.1 Modeling Fisher's Iris Data

We begin with perhaps the most well-worn polychotomous data set in existence, the Iris data set reported in Fisher (1936). The data report four characteristics (sepal width, sepal length, petal width and petal length) of three species of Iris flower, and there are $n = 150$ observations. The goal, given the four measurements, is to predict which one of the three species of Iris flower the measurements are likely to have come from. This is a widely used and publicly available benchmark for various classification and discrimination techniques, and we adopt it for this purpose. Data can be downloaded from the Statlib archives located at

`http://lib.stat.cmu.edu/DASL/Datafiles/Fisher'sIris.html`

The range commonly found by various discriminant methods is 96-99%. Using the proposed method with bandwidths selected via the proposed method of cross-validation, the proposed technique correctly predicts 96.7% of all observations. However, Fisher's measurements were discrete in nature as they were rounded and recorded with no decimal places, and so there were only 22 unique values for sepal width, 43 for sepal length, 23 for petal width and 35 for petal length. It is of interest therefore to model these measurements as categorical rather than continuous. Application of the proposed method treating the measurements as categorical rather than continuous correctly predicts 100% of all observations. This simple application suggests that the proposed method can perform as well or better than conventional parametric models.

4.2 Modeling Swiss Labor Market Participation

For our next application we use the data of Gerfin (1996) who models the labor market participation of married Swiss women using a cross-section data set of size $n = 872$ having six explanatory variables. He uses a Probit model along with three semiparametric specifications, and finds that the Probit specification cannot be rejected and that the models yield similar results. He concludes that "more work is necessary on specification tests of semiparametric models and on simulations using these models". We simply use this dataset to see whether predictions given by the Probit and semiparametric specifications can be substantially improved upon (we do not include Gerfin's (1996) semiparametric results here as they all yielded similar results.)

Data for this study can be found at

`http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/`

The variables used by the Gerfin (1996) study are

1. LFP: Labor force participation dummy.
2. LNNLINC: Log of non-labor income.
3. AGE: Age in years.
4. EDUC: Years of formal education.
5. NYC: Number of young children (younger than 7).
6. NOC: Number of older children.
7. FOREIGN: Dummy, = 1 if obs is not Swiss.

We compare the results of our estimator with those from Gerfin (1996), and the confusion matrices and classification rates for both the proposed and Probit approaches are summarized in Table 1. A confusion matrix is one whose diagonal elements are correctly predicted outcomes and whose off-diagonal elements are incorrectly predicted outcomes. We also report the overall correct classification rate and correct classification rates for each values assumed by the categorical variable[1]. As can be seen from Table 1, the proposed method correctly predicts 74.1% of all observations while a Probit model correctly predicts 66.5% which represents a marked improvement in model performance. To address potential concerns that these results are an artifact of within-sample 'overfitting', we randomized the data and split it into independent estimation and evaluation samples[2]. The predictive ability of the model as measured by performance on the independent data mirrors the within-sample results reported in Table 1 for a large number of different splits indicating that this is indeed a general improvement in predictive ability and not simply an artifact of overfitting.

4.3 Modeling U.S. Female Labor Force Participation

Our final application uses the Mroz (1987) data file which is taken from the 1976 Panel Study of Income Dynamics, and is based on data for 1975. There are 753 observations in

[1]For example, CCR(0) is the number of predicted zeros $\div$ number of zeros in the sample $\times$ 100.

[2]For example, we considered estimation samples of size $n_1 = 700$ and prediction samples of size $n_2 = 172$, $n_1 = 750$ and $n_2 = 122$ and so on.

Kernel			Logit		
A/P	0	1	A/P	0	1
0	360	111	0	358	113
1	115	286	1	179	222
%Correct		74.1%	%Correct		66.5%
%CCR(0)		76.4%	%CCR(0)		76.0%
%CCR(1)		71.3%	%CCR(1)		55.3%

Table 1: Confusion matrix and classification rates for the kernel and Logit models.

this dataset, the first 428 for women with positive hours worked and the remaining 325 observations for women who did not work for pay. For a complete discussion of the data see Mroz (1987, Appendix 1). This is a widely cited study on 'second-generation models of labor supply' and is featured in Berndt (1991, Chapter 11), a popular textbook used to train undergraduate and graduate students of economics. Data and TSP code to replicate the study can be found at

`http://www.stanford.edu/~clint/berndt/`

and we consider an application of the proposed method to modeling the female labor force participation decision following Berndt (1991, Chapter 11, page 654–657).

We replicate the results of Berndt (1991, Chapter 11, page 654–657), and the following variables from Mroz's (1987) data file were therefore used:

1. LFP: A dummy variable equal to 1 if the woman worked in 1975, 0 otherwise.
2. KL6: The number of children less than 6 years old in the household.
3. K618: The number of children between ages 6 and 18 in the household.
4. WA: The wife's age.
5. WE: The wife's educational attainment, in years.
6. CIT: A dummy variable equal to 1 if the woman lives in a large city (SMSA), 0 otherwise.

7. AX: The actual years of the wife's previous labor market experience.
8. UN: The unemployment rate in county of residence, in percentage points. This is taken from bracketed ranges.
9. LWW1: The log of the wage (wife's average hourly earnings, in 1975 dollars) for working women, the log of predicted wage for non-workers.
10. PRIN: The wife's property income computed as total family income minus the labor income earned by the wife.

The Logit and Probit approaches model Item (1) as the dependent variable and items (2) through (10) as explanatory variables in addition to a constant term, while the proposed approach does not require the use of the constant term. For the proposed method we use the fact that items (1) through (7) are categorical while items (8) through (10) are continuous. Again, bandwidths were determined via the proposed method of cross-validation. The confusion matrices and classification rates for both the proposed and Logit approaches are summarized in Table 2.

Kernel			**Logit**		
A/P	0	1	A/P	0	1
0	314	11	0	166	159
1	74	354	1	80	348
%Correct		88.7%	%Correct		68.2%
%CCR(0)		96.6%	%CCR(0)		51.0%
%CCR(1)		82.7%	%CCR(1)		81.3%

Table 2: Confusion matrix and classification rates for the kernel and Logit models.

The estimated Logit and Probit models of labor force participation correctly predict 514 (68.2%) and 512 (68.0%) of the labor force participation decisions respectively. As can be seen from Table 2, the proposed method correctly predicts 668 (88.72%) labor force participation choices which translates into an additional 154 choices being correctly predicted, which is a fairly dramatic improvement in terms of prediction accuracy. On the basis of the simulations presented in Section 5.2 this suggests that the Logit specification is inappropriate

for this DGP though obviously no formal test of this hypothesis is conducted at this point. To address possible concerns that these results are an artifact of within-sample 'overfitting', we again randomized the data and split it into independent estimation and evaluation samples, and the predictive ability on the independent sample reflected the within-sample results reported in Table 2. This is also suggestive that the gradients based upon the parametric models are poorly specified and therefore any inference based up these gradients would be suspect. For the continuous variables we compute the average derivatives of the probability that $Y = 1$ which are typically reported along with Logit and Probit estimates, and this is reported in Table 3.

Variable	Logit dP/dX	Kernel $\widehat{\nabla}_{X_i} \hat{g}(Y_i \mid X_i)$
PRIN	-7.30199×10^{-06}	-3.09158×10^{-07}
LWW1	0.081026	0.157381
UN	-0.0036825	-0.00296792

Table 3: Logit and Kernel Estimates of the average change in the probability of participation ($Y = 1$) with respect to the continuous variables influencing this decision.

Though quantitative results from each approach differ, qualitative results are similar in that the higher the property income or rate of unemployment or the lower the wage the lower is the probability of an individual participating in the labor force. Often interest focuses on how the participation decision is affected by changes in market wages. The proposed method suggests that the average response in the probability of participation is roughly double that based upon the Logit and Probit models (0.16 versus 0.08) suggesting that participation elasticities with respect to wages may be substantially larger than those implied by the Logit or Probit specifications, though again no formal test of this hypothesis is attempted here.

5 Simulations

By way of example we assume that interest lies in predicting $g(Y|X)$ and $\nabla_x g(Y|X)$ for a categorical variable conditional upon a vector of realizations of continuous data. This is a common situation in economics that is frequently modeled used a linear-index Logit model which is adopted here as a parametric benchmark. We assume that interest lies in predicting $Pr[Y = y|X_{i1}, \dots]$ where $Y \in \{0, 1, \dots\}$ and in estimating how this probability responds to changes in the conditioning variables. Note that the proposed method is applicable when the conditioned data set is multivariate, and therefore is more general than the following examples would suggest - the examples are chosen simply due to the popularity of prediction of categorical data and due to potential problems which can arise when using standard parametric approaches.

5.1 Predicting Nonlinear Binary Choice - Univariate Conditioning Set

We begin with a simple example in which X_1 is distributed $U[-4, 4]$ and $Y \in \{0, 1\}$ is a binary variate that is conditionally determined by

$$Y = \begin{cases} 1 & \text{if } X_1 + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where ϵ is a white noise $N(0, \sigma_\epsilon^2)$ error term with $\sigma_\epsilon = 1$. This example is often modeled using either a linear single-index Probit or Logit model. As both give virtually identical results for this DGP and the Logit is more widely used, we consider as a benchmark the

Logit specification given by

$$\begin{aligned} Pr[Y = 1|x] &= \frac{1}{1 + \exp(-(\beta_1 - \beta_2 x))} \\ &= \frac{1}{1 + \exp(-I(x))} \end{aligned} \tag{20}$$

where $I(x)$ denotes an 'index function'. We note that the Logit model assumes that the distribution of choices is symmetric, unimodal, correctly specified by the Logistic distribution with the underlying index being correctly specified by $I(x) = \beta_1 + \beta_2 x$ in this instance. Also, though not modeled here, we note that when categorical variables appear as conditioning variables, the researcher must specify how each value taken on by the categorical variable affects each and every parameter in the model.

We draw random samples from the DGP in Equation (19) and compute the proposed kernel estimator and gradient along with the Logit model and its gradient evaluated at 100 equally spaced points over their support. We repeat this $5,000$ times, compute the median value for all objects at each of the evaluation points, and vary the sample size in order to examine the finite-sample performance of the proposed estimator relative to a correctly specified parametric model. Bandwidth selection is achieved via the proposed method of cross-validation for each Monte Carlo replication for all experiments which follow, and the Gaussian kernel is used throughout (Silverman (1986, page 43)).

The median predicted conditional probabilities that $Y = 1$ and associated median gradients from this Monte Carlo experiment are plotted in Figure 1 for sample sizes of $n = 100$ and $n = 1,000$. As can be seen, the proposed method is capable of consistently modeling this DGP, and it does so without requiring the researcher to specify either the functional form of the unknown index or the functional form of the distribution function as would be required by the Logit method. As with all kernel estimators, there is some finite-sample bias which increases with the curvature of the object being estimated, but this bias vanishes

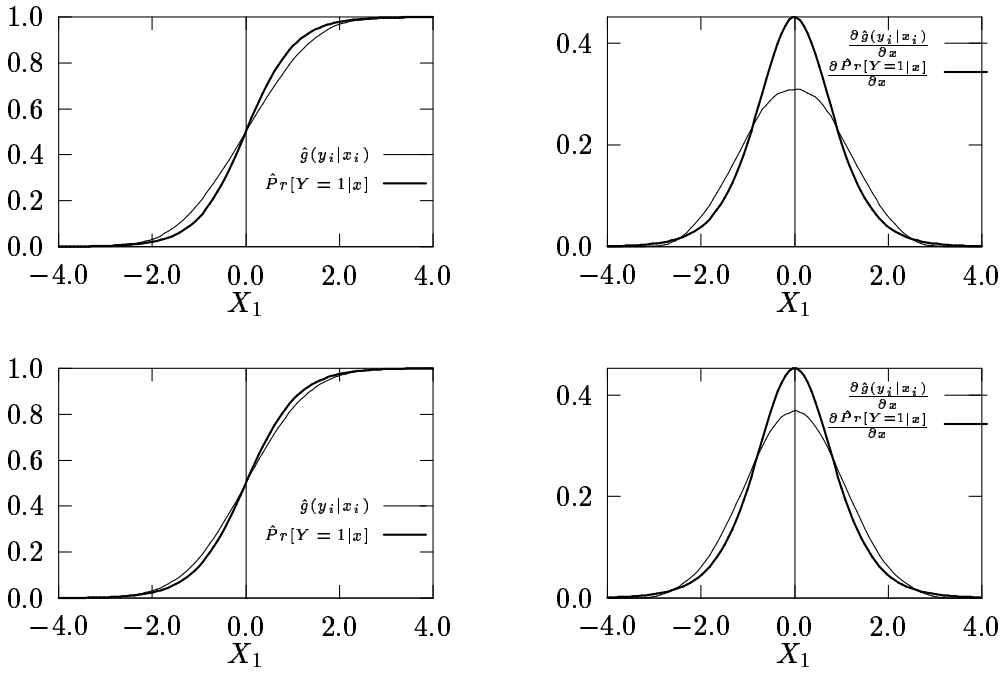

Figure 1: Median kernel and Logit estimates of the conditional probability that $Y = 1$ and the response of the probability with respect to the conditioning variable X_1 for $5,000$ Monte Carlo replications. The figures on the left are the estimated conditional probabilities while those on the right are the gradient of this probability with respect to X_1. The top figures are for a sample size of $n = 100$ while the lower figures are for $n = 1,000$.

asymptotically, and it is evident that going from a sample size of $n = 100$ to $n = 1,000$ results in smaller bias. Of course, the Logit will be inappropriate if the distribution of choices is asymmetric or multimodal or if the underlying index is other than $I(x) = \beta_1 + \beta_2 x$, so one would trade off bias for consistency when moving from correctly specified parametric models to a nonparametric framework.

In addition to examining the performance of the estimator of $g(Y|X)$ and $\nabla_x g(Y|X)$, we consider the predictive performance of the proposed estimator relative to this correctly specified Logit model. For each Monte Carlo replication, $\hat{g}(Y|X)$ and the Logit estimator $\widehat{Pr}[Y = 1|x]$ and their gradients were computed and then predictions were made for an *independent* sample drawn from the same DGP at each of the evaluation points. We then

compute the average confusion matrix for each approach which simply tabulates average predicted versus average actual outcomes for each independent data set and report this in Table 4. The results in Table 4 suggest that the proposed method is capable of mimicking the performance of a correctly specified parametric model even for samples that would be judged to be small in a nonparametric framework (for example, $n = 100$). There is a slight loss in efficiency relative to the correctly specified parametric model as well as finite-sample bias evident in the estimated gradient, but both diminish asymptotically as can be seen from Table 4 and Figure 1.

Kernel			**Logit**		
A/P	0	1	A/P	0	1
0	44.6	5.4	0	44.9	5.1
1	5.3	44.7	1	5.1	44.8
%Correct		89.3%	%Correct		89.8%
%CCR(0)		89.2%	%CCR(0)		89.8%
%CCR(1)		89.4%	%CCR(1)		89.8%

Kernel			**Logit**		
A/P	0	1	A/P	0	1
0	45.1	4.9	0	45.1	4.9
1	4.9	45.0	1	4.9	45.1
%Correct		90.1%	%Correct		90.2%
%CCR(0)		90.1%	%CCR(0)		90.1%
%CCR(1)		90.1%	%CCR(1)		90.2%

Table 4: Confusion matrix and classification rates for the proposed method and that from a Logit model. The top is that for $n = 100$ and the bottom for $n = 1,000$.

We next consider a DGP in which X_1 is again distributed $U[-4, 4]$ and Y is a binary

variate $\in \{0, 1\}$ that is conditionally determined by

$$Y = \begin{cases} 1 & \text{if } -2 < X_1 + \epsilon < 2 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where ϵ is a white noise error term drawn from the skewed $\Gamma(1, 1)$ distribution.

Situations similar to this are often observed in economic settings, an example being consumer goods that are normal for some range of income and inferior for another. As income rises we often observe an increased likelihood of a choice being made but as income continues to rise beyond some range we begin to observe a decreased likelihood of choices being made. Of course, the applied researcher may have no insight into underlying consumer preferences and may prefer to employ estimation techniques that are consistent and do not place rigid parametric restrictions on such behavior.

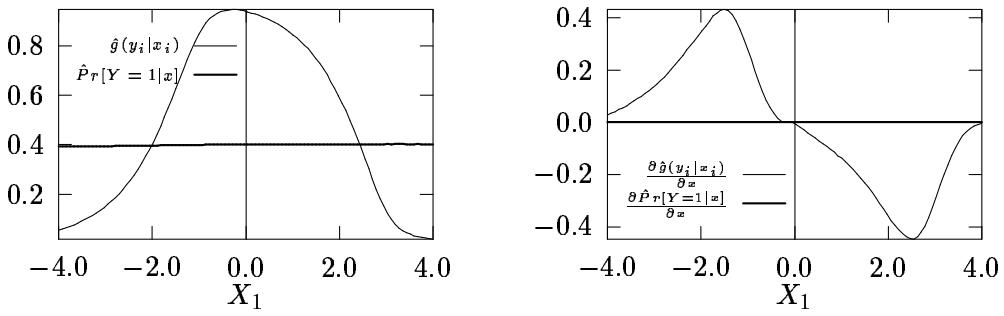

Figure 2: Median kernel and Logit estimates of the conditional probability that $Y = 1$ and the response of the probability with respect to the conditioning variable X_1 for 5,000 Monte Carlo replications.

The median kernel and Logit estimates are plotted in Figure 2, and the Logit model is seen to fail completely in this situation. The median Logit estimate is constant across X_1 and therefore does not use any of the conditioning information - it returns an unconditional prediction, and none of the estimated parameters (except the constant) in the Logit model are significant. As well, the median gradient is everywhere zero as can be seen from Figure

2.

Table 5 presents the confusion matrices for this case which are quite revealing. The Logit model gets quite 'confused' and effectively predicts that every case will be a '0' and yields predictions no better than flipping a coin in this case, while the proposed method is quite successful in detecting both the choice, its asymmetric nature, and the underlying choice gradient as can be seen in Figure 2 and Table 5.

Kernel			**Logit**		
A/P	0	1	A/P	0	1
0	42.0	9.1	0	47.3	3.9
1	8.7	40.2	1	47.0	1.9
%Correct		82.2%	%Correct		49.2%
%CCR(0)		82.2%	%CCR(0)		92.4%
%CCR(1)		82.2%	%CCR(1)		3.8%

Table 5: Confusion matrix and classification rates for the proposed method and that from a Logit model.

We now proceed to the more interesting instance of multivariate conditioning sets and binary prediction and then consider multivariate conditioning sets and multivariate categorical prediction.

5.2 Predicting Nonlinear Binary Choice - Multivariate Conditioning Set

We begin with a simple example in which X_1 and X_2 are both $U[-4, 4]$. Y is a binary variate $\in \{0, 1\}$ and is conditionally determined by

$$Y = \begin{cases} 1 & \text{if } X_1 + X_2 + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where ϵ is a white noise $N(0, \sigma_\epsilon^2)$ error term with $\sigma_\epsilon = 1$.

The median predicted conditional probability and that for the correctly specified Logit model for a sample size of $n = 100$ are plotted in Figure 3, while Table 6 computes the average confusion matrices and classification rates for two sample sizes, $n = 100$ and $n = 1,000$ allowing us to assess the cost of not knowing the parametric form of the underlying DGP.

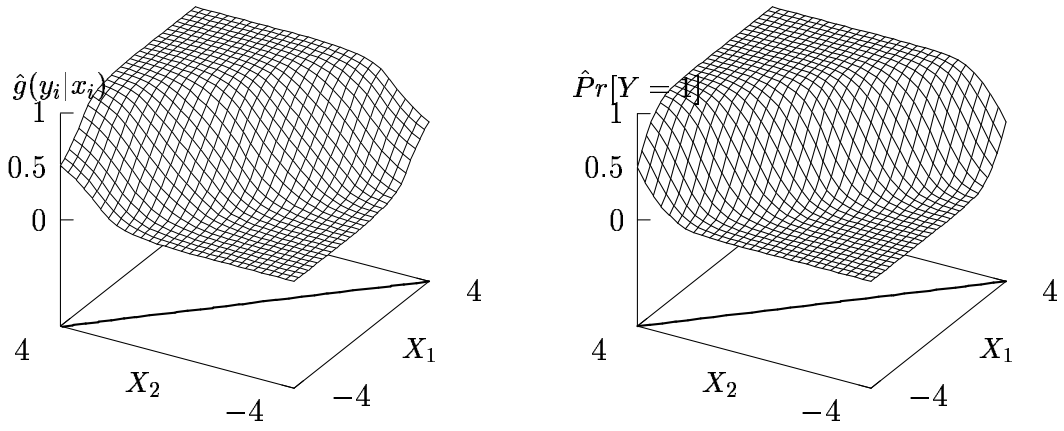

Figure 3: Median kernel and Logit estimates of the conditional probability that $Y = 1$. The Logit estimate is the figure on the right. The contour line on the horizontal plane represents the boundary between the estimated conditional probability that $Y = 0$ and $Y = 1$ for a sample size of $n = 100$ based upon $5,000$ Monte Carlo replications.

This situation is often modeled with a Logit specification. As in Section 5.1, we are interested in how well the proposed method performs relative to a correctly specified parametric model. In Section 5.1 we can see that the proposed method compares favorably to the Logit method when the Logit is correctly specified and there is only one conditioning variable. We are aware of the 'curse of dimensionality' present in the nonparametric approach, and wish to assess its impact in this setting. As can be seen from Table 6, there is more of a loss in terms of predictive accuracy for a given sample size relative to that detailed in Table 4,

Kernel			Logit		
A/P	0	1	A/P	0	1
0	481.0	63.5	0	492.9	51.5
1	63.4	481.2	1	51.7	492.8
%Correct		88.4%	%Correct		90.5%
%CCR(0)		88.3%	%CCR(0)		90.5%
%CCR(1)		88.4%	%CCR(1)		90.5%

Kernel			Logit		
A/P	0	1	A/P	0	1
0	493.5	51.1	0	495.4	49.1
1	51.2	493.2	1	49.1	495.4
%Correct		90.6%	%Correct		91.0%
%CCR(0)		90.6%	%CCR(0)		91.0%
%CCR(1)		90.6%	%CCR(1)		91.0%

Table 6: Confusion matrix and classification rates for the proposed method and that from a Logit model. The upper table is that for $n = 100$ while the lower is for $n = 1,000$.

as is expected, but this sample size of $n = 100$ involving three variables, Y, X_1, and X_2 is extremely small by nonparametric standards. Table 6 considers how this loss behaves as the sample size increases from $n = 100$ to $n = 1,000$, and again we witness the consistent nature of the proposed approach being revealed as the sample size increases.

Next we consider a situation in which X_1 and X_2 are both $U[-4, 4]$. Y is a binary variate $\in \{0, 1\}$ and is conditionally determined by

$$Y = \begin{cases} 1 & \text{if } X_1 + 3\sin(X_2) + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where ϵ is a white noise $N(0, \sigma_\epsilon^2)$ error term with $\sigma_\epsilon^2 = 0.1$. The median predicted conditional probability and that for the Logit model are plotted in Figure 4 while the confusion matrices are presented in Table 7.

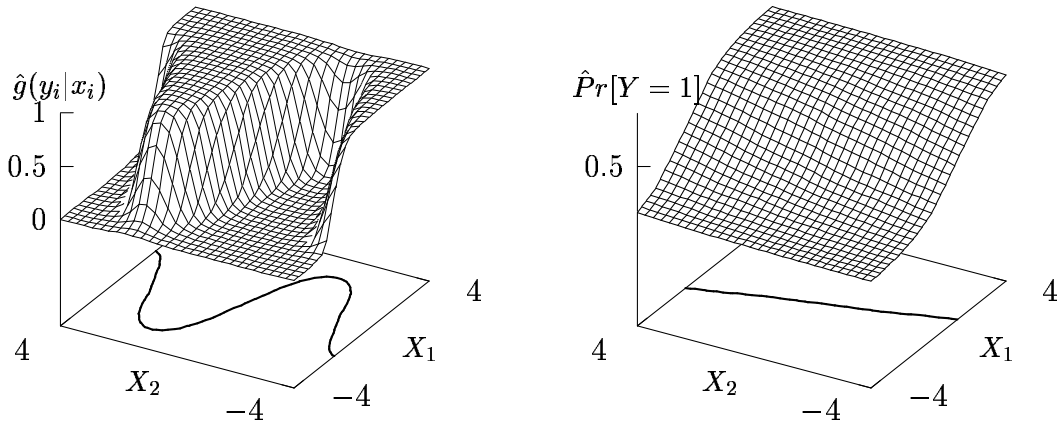

Figure 4: Median kernel and Logit estimates of the conditional probability that $Y = 1$. The Logit estimate is the rightmost figure. The contour line on the horizontal plane represents the boundary between the estimated conditional probability that $Y = 0$ and $Y = 1$ for a sample size of $n = 1,000$ based upon $5,000$ Monte Carlo replications.

Kernel			**Logit**		
A/P	0	1	A/P	0	1
0	504.5	40.0	0	427.0	117.5
1	39.8	504.7	1	116.5	428.0
%Correct		92.7%	%Correct		78.5%
%CCR(0)		92.6%	%CCR(0)		78.4%
%CCR(1)		92.7%	%CCR(1)		78.6%

Table 7: Confusion matrix and classification rates for the proposed method and that from a Logit model.

Modeling this situation with a Logit model would fail except in the situation where the researcher correctly guessed that the index was given by $\beta_0 + \beta_1 X_1 + \beta_2 \sin(X_2)$. It is interesting to consider the estimated gradients for the proposed approach compared with the Logit approach, plots of which can be found in Appendix A. Misspecification of the index function completely distorts the estimated gradient. Often concern lies with how changes

in the conditioning variables affects probabilities, and this would therefore be of concern to applied researchers using parametric approaches.

We now consider a situation in which X_1 and X_2 are both $U[-4, 4]$. Y is a binary variate $\in \{0, 1\}$ and is conditionally determined by

$$Y = \begin{cases} 1 & \text{if } -2 < X_1 + \epsilon_1 < 2 \text{ and } -2 < X_2 + \epsilon_2 < 2 \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

where ϵ_1 and ϵ_2 are white noise $N(0, \sigma_\epsilon^2)$ error terms with $\sigma_\epsilon = 0.1$. The median predicted conditional probability along with the gradient with respect to X_1 are plotted in Figure 4.

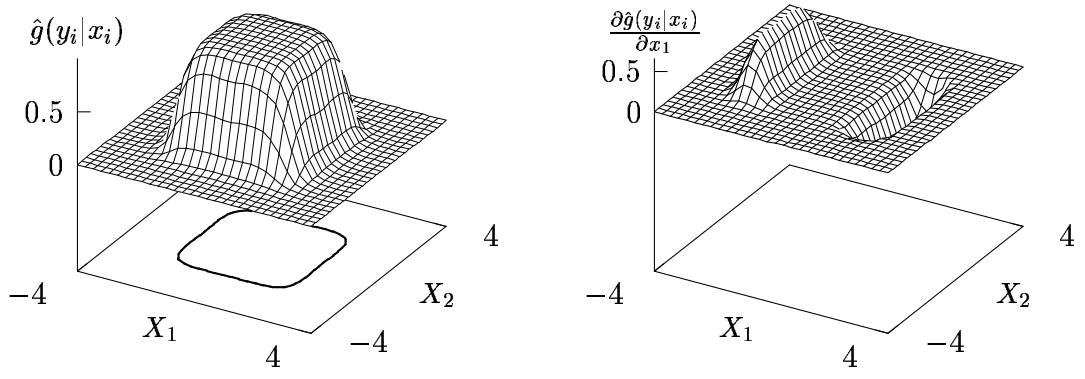

Figure 5: Median kernel estimate of the conditional probability that $Y = 1$ and the gradient with respect to X_1. The contour line on the horizontal plane represents the boundary between the estimated conditional probability that $Y = 0$ and $Y = 1$ for a sample size of $n = 1,000$ based upon $5,000$ Monte Carlo replications.

This is a case in which the Logit model completely breaks down, as can be seen from an examination of Table 8, and is analogous to results found in Figure 2 and Table 5. Again, the Logit specification uses none of the conditioning information contained in X_1 and X_2 and

Kernel			Logit		
A/P	0	1	A/P	0	1
0	799.2	33.8	0	830.5	2.5
1	36.9	219.1	1	256.0	0.0
%Correct		93.5%	%Correct		76.3%
%CCR(0)		95.9%	%CCR(0)		99.7%
%CCR(1)		85.6%	%CCR(1)		0.0%

Table 8: Confusion matrix and classification rates for the proposed method and that from a Logit model.

simply predicts all zeros. The gradients from the Logit model are therefore zero everywhere and again none of the estimated parameters in the Logit model are significant save for the constant.

More interesting cases arise when considering conditional prediction of multinomial categorical data. These situations are frequently encountered in practice. Using a multinomial Logit approach, for example, raises a number of issues such as normalization, identification, and specification of multiple indices. The proposed method does not suffer from any of these issues. Consider the case in which the DGP is given by

$$Y = \begin{cases} 1 & \text{if } X_1 + \epsilon_1 > 0 \text{ and } X_2 + \epsilon_2 > 0 \\ 2 & \text{if } X_1 + \epsilon_1 < 0 \text{ and } X_2 + \epsilon_2 < 0 \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

where again ϵ_1 and ϵ_2 represent white noise $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.1$. error terms.

Both the median kernel and Logit estimators of $Pr[Y = 0|X_1, X_2]$ are plotted in Figure 6 below, while the confusion matrices and classification rates appear in Table 9. As can be seen, the multinomial Logit model cannot consistently model this situation and the gradients in particular from the Logit approach will be totally misleading.

The point to be made is that the proposed estimator can readily model nonlinear con-

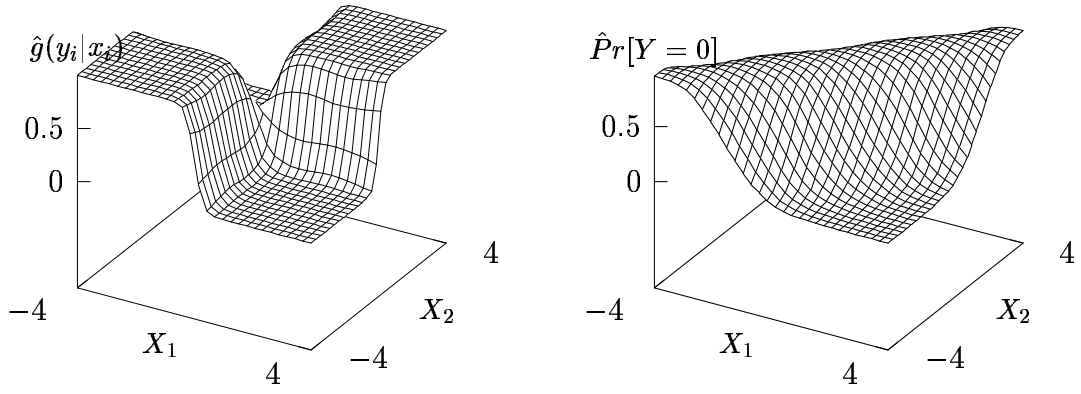

Figure 6: Median kernel and Logit estimates of the conditional probability that $Y = 0$ for a sample size of $n = 100$ based upon $5,000$ Monte Carlo replications. The Logit results are presented in the rightmost figure.

Kernel				**Logit**			
A/P	0	1	2	A/P	0	1	2
0	252.6	19.4	0.3	0	223.5	48.8	0.0
1	19.0	506.6	18.9	1	49.6	446.4	48.6
2	0.3	19.7	252.3	2	1.2	48.5	222.6
%Correct			92.9%	%Correct			82.0%
%CCR(0)			92.8%	%CCR(0)			82.1%
%CCR(1)			93.0%	%CCR(1)			82.0%
%CCR(2)			92.7%	%CCR(2)			81.8%

Table 9: Confusion matrix and classification rates for the proposed method and that from a Logit model.

ditional prediction of binary and multinomial categorical data when the conditioning data are continuous and categorical without requiring the researcher to specify functional forms for indices and distribution functions that, when incorrectly specified, lead to biased inconsistent estimation and hypothesis tests having asymptotically invalid size and power. The

method compares favorably to correctly specified parametric models with an expected loss in efficiency in this instance while it can handily outperform standard parametric models in a wide variety of situations.

6 Conclusion

In reference to parametric models, G. Box wrote that "all models are wrong, but some are useful". However, there are situations in which it is extremely difficult to specify a parametric model, and the modeling of joint conditional distributions in the presence of both categorical and continuous data would be one such instance. Sometimes parametric models of these conditional distributions simply ignore the conditioning information and return unconditional predictions, and therefore are not useful at all.

This paper presents a nonparametric approach to the estimation of a multivariate conditional probability density function and its gradient when faced with mixed categorical and continuous data. The approach can be useful in a wide variety of situations, and does not place the burden of correct specification on the researcher.

The technique can be applied to a number of interesting but parametrically awkward and sometimes parametrically intractable problems. We consider numerous simulations and applications, and offer some general guidelines as to when the technique may outperform parametric approaches when conducting multinomial conditional prediction. The simulations presented in this paper highlight both the consistency and the flexibility of the proposed approach for a number of situations and also examine the statistical finite-sample loss of using the proposed method relative to correctly specified parametric models. One of the benefits of using this approach is best appreciated in comparison to parametric methods such as Logit models for categorical prediction: approaches such as the Logit place rigid restrictions on both the choice probabilities and on the gradient of the predicted conditional

probabilities, while the proposed approach is capable of detecting a wide variety of situations with no functional guessing required by the researcher. For a number of publicly available datasets used for categorical prediction, the proposed method stages a strong performance suggesting that this method may be of value in applied settings.

The main benefit of the proposed approach lies simply in the ability to proceed with estimation of conditional distributions and their gradients without placing the unrealistic burden of correct parametric specification of the underlying DGP upon the researcher.

References

Aitchison, J. & Aitken, C. G. G. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), 413–420.

Amemiya, T. (1981), 'Qualitative response models: A survey', *Journal of Economic Literature* **19**, 1483–1536.

Berndt, E. R. (1991), *The Practice of Econometrics: Classic and Contemporary*, Addison Wesley.

Bowman, A. (1980), 'A note on consistency of the kernel method for the analysis of categorical data', *Biometrika* **67**(3), 682–684.

Chaudhuri, P. & Dewanji, A. (1995), 'On a likelihood-based approach in nonparametric smoothing and cross-validation', *Statistics & Probability Letters* **22**, 7–15.

Fisher, R. A. (1936), 'The use of multiple measurements in axonomic problems', *Annals of Eugenics* **7**, 179–188.

Gerfin, M. (1996), 'Parametric and semiparametric estimation of the binary response model of labour market participation', *Journal of Applied Econometrics* **11**(3), 321–340.

Habbema, J. D. F., Hermans, J. & Remme, J. (1978), 'Variable kernel density estimation in discriminant analysis', *Compstat* pp. 178–185.

Hall, P. (1981), 'On nonparametric multivariate binary discrimination', *Biometrika* **68**(1), 287–294.

Hall, P. (1987*a*), 'On kullback-leibler loss and density estimation', *The Annals of Statistics* **15**, 1491–1519.

Hall, P. (1987*b*), 'On the use of compactly supported densities in problems of discrimination', *Journal of Multivariate Analysis* **23**, 131–158.

Hall, P. & Wand, M. P. (1988), 'On nonparametric discrimination using density differences', *Biometrika* **75**(3), 541–547.

Härdle, W. & Marron, S. (1987), 'Optimal bandwidth selection in nonparametric regression function estimation', *Annuls of Statistics* **13**, 1465–1481.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

McFadden, D. (1984), Econometric analysis of qualitative response models, *in* Z. Griliches & M. Intriligator, eds, 'Handbook of Econometrics', North Holland, pp. 1385–1457.

Mroz, T. A. (1987), 'The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions', *Econometrica* **55**(4), 765–799.

Nadaraya, E. (1965), 'On nonparametric estimates of density functions and regression curves', *Theory of Applied Probability* **10**, 186–190.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *Annals of Mathematical Statistics* **33**, 105–131.

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

Watson, G. (1964), 'Smooth regression analysis', *Sanikhya* **26:15**, 175–184.

A Gradient Estimation

One of the important features of the proposed method is the consistent estimation of the gradient. Often incorrectly specified parametric models are adequate for prediction, but completely misleading when the gradient is of interest. Figure 7 plots the median probability gradient with respect to X_1 for the proposed method and the Logit method.

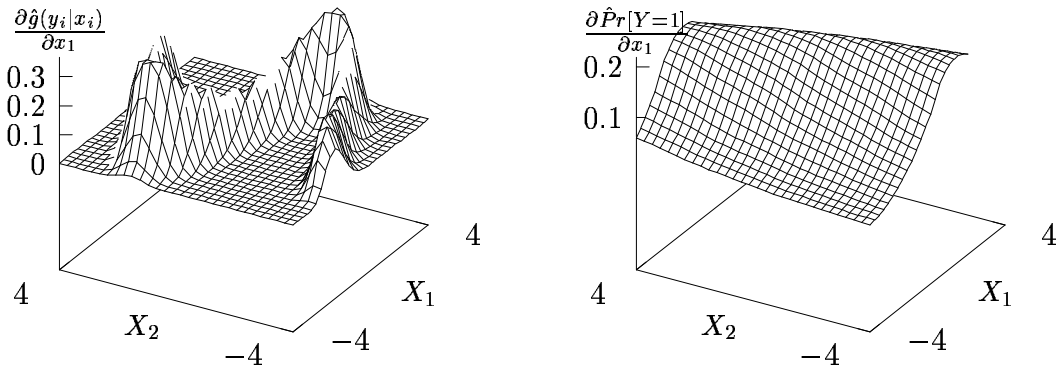

Figure 7: Median kernel and Logit estimates of the gradient vector of the conditional probability that $Y = 1$ with respect to X_1 for a sample size of $n = 100$. The figure on the right is that from a Logit model.

Note that the gradient of the Logit model with respect to X_j is simply $Pr[Y = 1|X_1, X_2] \times (1 - Pr[Y = 1|X_1, X_2])\hat{\beta}_j$. Any misspecification of either the error distribution or the index function will result in biased inconsistent estimates of this response.

Figure 8 plots the median probability gradient with respect to X_1 for the proposed method and the Logit method. The limitations of the Logit approach quickly become apparent since all responses are required to have the same 'shape' differing only by the magnitude/sign of the associated parameter in the index function. The proposed method places no such limitations on the underlying gradient.

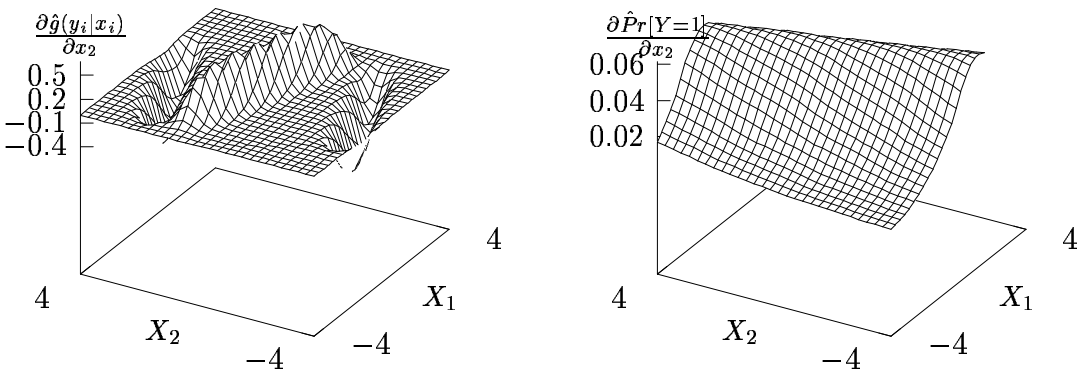

Figure 8: Median kernel and Logit estimates of the gradient vector of the conditional probability that $Y = 1$ with respect to X_2 for a sample size of $n = 100$. The figure on the right is that from a Logit model.