ANNALS OF ECONOMICS AND FINANCE **5**, 211–235 (2004)

Kernel Estimation of Multivariate Conditional Distributions

Jeff Racine

Department of Economics & Center for Policy Research
Syracuse University, Syracuse, NY 13244
E-mail: jracine@maxwell.syr.edu

Qi Li

Department of Economics, Texas A&M University
College Station, TX 77843
E-mail: qi@econmail.tamu.edu

and

Xi Zhu

Department of Economics, Tsinghua University Beijing, 100084 PRC
E-mail: s972250@em.tsinghua.edu.cn

We consider the problem of estimating conditional probability distributions that are multivariate in both the conditioned and conditioning variable sets. This is an extension of Hall, Racine, and Li (forthcoming), who considered the case of a univariate conditioned variable but who also considered the more general case of both irrelevant and relevant conditioning variables. Following Hall et al. (forthcoming), we use the kernel method with the smoothing parameters selected from the cross-validated minimization of a weighted integrated squared error of the kernel estimator. We derive the rate of convergence of the smoothing parameters to some non-stochastic optimal smoothing parameter values, and establish the asymptotic normal distribution of the resulting nonparametric conditional probability (density) estimator. Simulations show that the proposed method performs quite well with a mixture of categorical and continuous variables.

Key Words: Estimation; Multivariate conditional distributions.
JEL Classification Numbers: C51, C30.

1529-7373/2004

1. INTRODUCTION

In this paper we consider the problem of estimating conditional probability (density) functions that are multivariate in both the conditioned and conditioning variable sets. Likelihood cross-validation is known to break down when modeling 'fat-tail' continuous data with commonly used compact support kernels such as the Epanechnikov kernel or thin-tailed kernels such as the widely used Gaussian kernel (see Hall (1987a,1987b)), and so we select the smoothing parameters by cross-validated minimization of a weighted integrated squared error of the kernel estimator. We derive the rate of convergence of the smoothing parameters to some benchmark nonstochastic optimal smoothing parameters, and establish the asymptotic normal distribution of the resulting nonparametric conditional probability (density) estimator. This paper extends results found in Hall, Racine, and Li (forthcoming), who consider the case of univariate conditioned variables and do not derive the *rate* of convergence of the cross validation selected smoothing parameters to some benchmark optimal values. However, Hall et al. (forthcoming) consider both irrelevant and relevant conditioning variables that we do not address here.

Related work includes that of Hall (1981), who considered bandwidth selection issues that arise when using the method of Aitchison and Aitken (1976) when there exist empty cells for categorical data, and who proposed a robust solution to this problem, Titterington (1980), Wang and Ryzin, (1981), Hall and Wand (1988), Scott (1992), Simonoff (1996), Li and Racine (2003), and Racine and Li (2004), to mention only a few. We note that Tutz (1991) has considered cross-validation for estimating conditional density functions with mixed variables, though he only shows the consistency of his proposed estimator and does not establish rates of convergence or asymptotic distributions.

This paper proceeds as follows. In Section 2 we consider the proposed nonparametric estimator of the conditional density function in the presence of categorical and continuous data types; Section 3 reports simulation results that examine the finite-sample performance of the proposed estimator. Proofs of the main results are given in Appendices A and B.

2. ESTIMATION OF CONDITIONAL DISTRIBUTIONS

Let $Z = (X, Y)$ denote a vector of random variables. We assume that Z consists of k discrete variables and q continuous variables, and we use Z^d to denote a $k \times 1$ vector of discrete variables. In this section, for expositional simplicity, we will first consider the case where $Z^d \in \{0, 1\}^k$. We use $Z^c \in \mathcal{R}^q$ to denote the continuous components of Z. We also write $X = (X^c, X^d)$, where $X^c \in \mathcal{R}^p$ is the continuous components of

X, and $X^d \in \{0,1\}^r$ is the discrete components of X. Similarly we have $Y = (Y^c, Y^d)$, $Y^c \in \mathcal{R}^{q-p}$ and $Y^d \in \{0,1\}^{k-r}$.

Let $f(z) = f(x,y)$ denote the joint density function of $Z = (X,Y)$, let $m(x)$ denote the marginal density function of X, and let $g(y|x) = f(x,y)/m(x)$ denote the conditional density of Y given $X = x$.

We use $Z^d_{t,i}$ to denote the tth component of Z^d_i. For $Z^d_{t,i}$, $Z^d_{t,j} \in \{0,1\}$, define a univariate kernel function $l(Z^d_{t,i}, Z^d_{t,j}) = 1-\lambda$ if $Z^d_{t,i} = Z^d_{t,j}$, and $l(Z^d_{t,i}, Z^d_{t,j}) = \lambda$ if $Z^d_{t,i} \neq Z^d_{t,j}$, where λ is a smoothing parameter.

Define $d_{z_i,z_j} = (Z^d_i - Z^d_j)'(Z^d_i - Z^d_j)$. d_{z_i,z_j}; takes values in $\in \{0,1,2,\ldots,k\}$, and it equals the number of disagreement components between Z^d_i and Z^d_j. The product kernel is given by

$$L(Z^d_i, Z^d_j, \lambda) = \prod_{t=1}^{k} l(Z^d_{t,i}, Z^d_{t,j}) = (1-\lambda)^{k-d_{z_i,z_j}} \lambda^{d_{z_i,z_j}}. \tag{1}$$

It is straightforward to generalize the above to the case of a k-dimensional vector of smoothing parameters λ. For simplicity of presentation and without loss of generalization, only scalar λ is treated here. In practice, we employ multidimensional numerical search routines that indeed allow λ to differ across variables.

Letting $Z^c_{i,t}$ denote the tth component of Z^c_i, letting $w(\cdot)$ be a univariate kernel function for a univariate continuous variable, and letting $W(\cdot)$ be the product kernel function for the continuous variables, we have

$$W_h(Z^c_i, Z^c_j) \equiv h^{-q} W\left(\frac{Z^c_i - Z^c_j}{h}\right) \stackrel{def}{=} h^{-q} \prod_{t=1}^{q} w\left(\frac{Z^c_{i,t} - Z^c_{j,t}}{h}\right). \tag{2}$$

To avoid introducing too much notation, we shall use the same notation $L(\cdot)$ and $W(\cdot)$ to denote the product kernel for X^d and X^c, i.e.,

$$L(X^d_i, X^d_j, \lambda) = \prod_{t=1}^{r} l(X^d_{t,i}, X^d_{t,j}) = (1-\lambda)^{r-d_{x_i,x_j}} \lambda^{d_{x_i,x_j}}, \tag{3}$$

where $d_{x_i,x_j} = (X^d_i - X^d_j)'(X^d_i - X^d_j)$ equals the number of disagreement components between X^d_i and X^d_j, and

$$W_h(X^c_i, X^c_j) \stackrel{def}{=} h^{-p} W\left(\frac{X^c_i - X^c_j}{h}\right) = h^{-p} \prod_{t=1}^{p} w\left(\frac{X^c_{i,t} - X^c_{j,t}}{h}\right). \tag{4}$$

Similarly we define

$$L(Y^d_i, Y^d_j, \lambda) = \prod_{t=1}^{k-r} l(Y^d_{t,i}, Y^d_{t,j}) = (1-\lambda)^{r-d_{y_i,y_j}} \lambda^{d_{y_i,y_j}}, \tag{5}$$

and

$$W_h(Y_i^c, Y_j^c) = h^{-(q-p)} W\left(\frac{Y_i^c - Y_j^c}{h}\right) = h^{-(q-p)} \prod_{t=1}^{q-p} w\left(\frac{Y_{i,t}^c - Y_{j,t}^c}{h}\right). \quad (6)$$

We estimate $f(z)$ by

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} K_{Z_i,z}, \quad (7)$$

where $K_{Z_i,z} = L_{Z_i^d,z^d} W_{Z_i^c,z^c}$, $L_{z_i^d,z^d} = L(Z_i^d, z^d, \lambda)$, and $W_{Z_i^c,z^c} = W_h(Z_i^c, z^c)$ are defined in (1) and (2), respectively.

Similarly, we estimate the marginal density $m(x)$ by

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{X_i,x}, \quad (8)$$

where $K_{X_i,x} = L_{X_i^d,x^d} W_{X_i^c,x^c}$, $L_{X_i^d,x^d} = L(X_i^d, x^d, \lambda)$ and $W_{X_i^c,x^c} = W_h(X_i^c, x^c)$ are defined in (3) and (4), respectively.

Therefore, we estimate $g(y|x) = f(x,y)/m(x)$ by

$$\hat{g}(y|x) = \frac{\hat{f}(x,y)}{\hat{m}(x)}. \quad (9)$$

It is well established that maximum-likelihood cross-validation methods do not lead to consistent estimation for fat-tail distributions with the kernel functions typically used in practice (Hall (1987a,b)). Therefore, we will choose the smoothing parameters by cross-validation methods that involve the minimization of a weighted integrated square error. We first introduce some notation. We will use subscripts i, j, and l to denote observations (i.e., $\sum_i = \sum_{i=1}^n$, $\sum_i \sum_{j \neq i} = \sum_{i=1}^n \sum_{j=1, j\neq i}^n$, etc.). When z^d, x^d, and y^d appear as the summation index, it runs over the support of z^d: $\mathcal{D}_z = \{0,1\}^k$, the support of x^d: $\mathcal{D}_x = \{0,1\}^r$, and the support of y^d: $\mathcal{D}_y = \{0,1\}^{k-r}$, i.e., $\sum_{z^d} = \sum_{z^d \in \mathcal{D}_z}$, $\sum_{x^d} = \sum_{x^d \in \mathcal{D}_x}$, and $\sum_{y^d} = \sum_{y^d \in \mathcal{D}_y}$.

Using the notation $\int dz = \sum_{z^d} \int dz^c$, a weighted integrated square difference between $\hat{g}(\cdot)$ and $g(\cdot)$ is given by

$$\begin{aligned} I_n &= \int [\hat{g}(y|x) - g(y|x)]^2 m(x)\, dz \\ &= \int [\hat{g}(y|x)]^2 m(x)\, dz - 2\int \hat{g}(y|x) g(y|x) m(x)\, dz + \int [g(y|x)]^2 m(x)\, dz \\ &\equiv I_{1n} - 2I_{2n} + I_{3n}, \end{aligned} \quad (10)$$

where $I_{1n} = \int [\hat{g}(y|x)]^2 m(x)\,dz$, $I_{2n} = \int \hat{g}(y|x) g(y|x) m(x)\,dz$, and $I_{3n} = \int [g(y|x)]^2 m(x)\,dz$. The reason for choosing $m(x)$ as the weight function in (9) will become apparent later. Note that I_{3n} is independent of (h, λ). Therefore, minimizing I_n over (h, λ) is equivalent to minimizing $I_{1n} - 2I_{2n}$.

Define

$$\begin{aligned} \hat{G}(x) &= \int [\hat{f}(x,y)]^2 dy = n^{-2} \sum_i \sum_j K_{X_i,x} K_{X_j,x} \int K_{Y_i,y} K_{Y_j,y} dy \\ &= n^{-2} \sum_i \sum_j K_{X_i,x} K_{X_j,x} K^{(2)}_{Y_i,Y_j}, \end{aligned} \tag{11}$$

where $K^{(2)}_{Y_i,Y_j} = \int K_{Y_i,y} K_{Y_j,y} dy \equiv \sum_{y^d} \int K_{Y_i,y} K_{Y_j,y} dy^c$ is the second order convolution kernel, $K_{Y_i,y} = W_{Y_i,y} L_{Y_i,y}$, $L_{Y_i,y} = L(Y_i, y, \lambda)$, and $W_{Y_i,y} = h^{-(q-p)} W\left(\frac{Y_i - y}{h}\right)$ are defined by (5) and (6), respectively.

Using (10), we have

$$\begin{aligned} I_{1n} &= \int \int [\hat{g}(y|x)]^2 m(x)\,dz = \int \frac{\int [\hat{f}(x,y)]^2 dy}{[\hat{m}(x)]^2} m(x)\,dx \\ &= \int \frac{\hat{G}(x)}{[\hat{m}(x)]^2} m(x)\,dx = E_X \left[\frac{\hat{G}(X)}{[\hat{m}(X)]^2} \right], \end{aligned} \tag{12}$$

where $E_X(\cdot)$ denotes the expectation with respect to X only (not with respect to the random observations $\{Z_i\}_{i=1}^n$).

Also,

$$\begin{aligned} I_{2n} &= \int \hat{g}(y|x) g(y|x) m(x)\,dz = \int \hat{g}(y|x) f(x,y)\,dxdy \\ &= \int \left[\frac{\hat{f}(x,y)}{\hat{m}(x)} \right] f(x,y)\,dx = E_Z \left[\frac{\hat{f}(Z)}{\hat{m}(X)} \right], \end{aligned} \tag{13}$$

where E_Z denotes the expectation with respect to Z only (not with respect to the random observations $\{Z_i\}_{i=1}^n$).

From (11) and (12) we see that by choosing $m(x)$ as the weighting function, we can write I_{1n} and I_{2n} in simple forms, enabling us to construct simple estimators for them.

Therefore, minimizing I_n is equivalent to minimizing $I_{1n} - 2I_{2n}$ given by

$$I_{1n} - 2I_{2n} = E_X \left\{ \frac{\hat{G}(X)}{[\hat{m}(X)]^2} \right\} - 2E_Z \left[\frac{\hat{f}(X,Y)}{\hat{m}(X)} \right]. \tag{14}$$

Equation (14) suggests that in practice, one can replace the expectations E_X and E_Z by their sample analogues. However, some caution is needed.

Let us consider I_{2n} first. When replacing $E_Z[\hat{f}(X,Y)/\hat{m}(X)]$ by its sample analogue $n^{-1}\sum_{l=1}^{n}\hat{f}(X_l,Y_l)/\hat{m}(X_l)$, one needs to use the leave-one-out estimators for $\hat{f}(X_l,Y_l)$ and $\hat{m}(X_l)$ given by

$$\hat{f}_{-l}(X_l,Y_l)=n^{-1}\sum_{i=1,i\neq l}^{n}K_{Z_i,Z_l}, \tag{15}$$

and

$$\hat{m}_{-l}(X_l)=n^{-1}\sum_{i=1,i\neq l}^{n}K_{X_i,X_l}. \tag{16}$$

This is because, in the definition of $E_Z[\hat{f}(X,Y)/\hat{m}(X)]$, the Z variable must be treated as independent of the observations that are used to estimate $\hat{f}(Z)$ and $\hat{m}(X)$. The leave-one-out estimator insures that Z_i and Z_l are independent of each other (since $i\neq l$).

Similarly, one should also use a leave-one-out estimator for $G(X_l)$ given by

$$\hat{G}_{-l}(X_l)=n^{-2}\sum_{i\neq l}\sum_{j\neq l}K_{X_i,X_l}K_{X_j,X_l}K^{(2)}_{Y_i,Y_j}. \tag{17}$$

Therefore, replacing $E_X(\cdot)$ and $E_Z(\cdot)$ by their sample analogues in (14), we obtain

$$CV(h,\lambda)\stackrel{def}{=}\frac{1}{n}\sum_{l=1}^{n}\frac{\hat{G}_{-l}(X_l)}{[\hat{m}_{-l}(X_l)]^2}-\frac{2}{n}\sum_{l=1}^{n}\frac{\hat{f}_{-l}(X_l,Y_l)}{\hat{m}_{-l}(X_l)}, \tag{18}$$

where $\hat{f}_{-l}(X_l,Y_l)$, $\hat{m}_{-l}(X_l)$, and $\hat{G}_{-l}(X_l)$ are the leave-one-out estimators given in (15), (16), and (17), respectively.

We will choose (λ,h) to minimize $CV(h,\lambda)$, defined in (18), and we will use $(\hat{h},\hat{\lambda})$ to denote this cross-validation choice of (h,λ).

The following assumptions will be used.

(A1) (i) $\{Z_i\}_{i=1}^n=\{X_i,Y_i\}_{i=1}^n$ is i.i.d. as $Z=(X,Y)$. (ii) Let $f(z)$ be the joint density of Z, and $m(x)$ be the marginal density of X, $f(z^c,z^d)$ (or $m(x^c,x^d)$) is four times continuously differentiable with respect to its continuous arguments for all $z^d\in\mathcal{D}_z$ ($x^d\in\mathcal{D}_x$). (iii) $inf_{x\in\mathcal{S}_x}m(x)\geq\delta>0$ for some positive δ.

(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) $\tilde{h}$ lies in a shrinking set $H_n=[\underline{h},\bar{h}]$, where $\underline{h}\geq Cn^{\delta-q}$, $\bar{h}\leq Cn^{-\delta}$ for some $C>0$ and $\delta>0$.

(A3) Define $m_\lambda(x^c,x^d)=\sum_{s=0}^{p}\sum_{x_1^d,d_{x,x_1}=s}(1-\lambda)^{1-s}\lambda^s m(x^c,x_1^d)$, $f_\lambda(z^c,z^d)=\sum_{s=0}^{q}\sum_{z_1^d,d_{z,z_1}=s}(1-\lambda)^{1-s}\lambda^s f(z^c,z_1^d)$, and

$g_\lambda(y|x) = f_\lambda(x,y)/m_\lambda(x)$. Then $\int [g_\lambda(y|x) - g(y|x)]^2 m(x)\,dxdy > 0$ for $\lambda \neq 0$.

(A1) (iii) rules out the case where X has an unbounded support. This assumption is not crucial and can be relaxed. When X has an unbounded support, one needs to introduce a trimming parameter to trim out observations near the boundary. The proof will be more tedious. Roughly speaking (A2) (ii) requires h satisfy the usual conditions of $h = o(1)$ and $(nh^q)^{-1} = o(1)$ (e.g., Härdle and Marron (1985)). (A3) is only used to prove that $\hat{\lambda} = o_p(1)$. It can be removed by assuming that $\hat{\lambda}$ takes values in a shrinking set, say, $\Lambda_n = [0, C_0/\log(n)]$ for some $C_0 > 0$.

Letting $CV_0(h,\lambda)$ denote the leading term of $CV(h,\lambda)$, in Appendix A we show that

$$CV_0(h,\lambda) = D_1 h^4 - D_2 h^2 \lambda + D_3 \lambda^2 + D_4 (nh^q)^{-1}, \tag{19}$$

where D_j's are some constants defined in Appendix A. Letting (h_o, λ_o) denote the values of (h,λ) that minimize $CV_0(h,\lambda)$, simple calculus shows that

$$h_o = c_1 n^{-1/(4+q)} \text{ and } \lambda_o = c_2 n^{-2/(4+q)}, \tag{20}$$

where $c_1 = \{pD_4/(4[D_1 - D_2^2/(4D_3)])\}^{1/(4+p)}$ and $c_2 = D_2 c_1^2/(2D_3)$. We interpret h_o and λ_o as *non-stochastic* optimal smoothing parameters.

Theorem 1 below establishes the rate of convergence of $(\hat{h}, \hat{\lambda})$ to (h_o, λ_o).

THEOREM 1. *Under assumptions (A1) to (A3), we have*

$$(\hat{h} - h_o)/h_o = O_p(n^{-\alpha/(4+q)}) \text{ and } \hat{\lambda} - \lambda_o = O_p(n^{-\beta}),$$

where $\alpha = \min\{2, q/2\}$ *and* $\beta = \min\{1/2, 4/(4+q)\}$.

The proof of Theorem 1 is given in Appendix A. By the result of Theorem 1, it is easy to show that

THEOREM 2.
Under assumptions (A1) to (A3), we have

$$\sqrt{n\hat{h}^p}(\hat{g}(y|x) - g(y|x) - \hat{h}^2 \mathcal{B}_1(z) - \hat{\lambda}\mathcal{B}_2(z)) \to N(0, \Omega(z)) \text{ in distribution,}$$

where

$$\begin{aligned} \mathcal{B}_1(z) &= (1/2)(1/m(z)) tr[\nabla^2 f(z)][\int w(v) v^2 dv], \\ \mathcal{B}_2(z) &= (1/m(z)) \sum_{\tilde{z}^d, d_{z,\tilde{z}}=1} [f(z^c, \tilde{z}^d) - f(z^c, z^d)], \end{aligned}$$

and $\Omega(z) = [f(z)/m^2(x)][\int W^2(v)dv]$ *(*∇^2 *is with respect to* z^c*).*

Up to now we have assumed that the discrete variable z^d is a multivariate binary variable. It is straightforward to generalize our results to the more general case to which we now turn.

The General Categorical Data Case

Assume that $Z^d_{t,i}$ takes $c_t \geq 2$ different values, i.e., $Z^d_{t,i} \in \{0, 1, ..., c_t - 1\}$, $t = 1, ..., k$. We use $\mathcal{D}_z = \prod_{t=1}^{k}\{0, 1, ..., c_t - 1\}$ to denote the range assumed by Z^d_i. For Z^d_i, $Z^d_j \in \mathcal{D}_z$. Following Aitchison and Aitken (1976) we use a univariate kernel function: $l(Z^d_{t,i}, Z^d_{t,j}, \lambda) = 1 - \lambda$ if $Z^d_{t,i} = Z^d_{t,j}$, and $l(Z^d_{t,i}, Z^d_{t,j}, \lambda) = \lambda/(c_t - 1)$ if $Z^d_{t,i} \neq Z^d_{t,j}$. Define an indicator function $\mathbf{1}(Z^d_{t,i} \neq Z^d_{t,j})$, which takes value 1 if $Z^d_{t,i} \neq Z^d_{t,j}$, and 0 otherwise. Also, define $d_{z_i,z_j} = \sum_{t=1}^{k} \mathbf{1}(Z^d_{t,i} \neq Z^d_{t,j})$, which equals the number of disagreement components between Z^d_i and Z^d_j. Then the product kernel for the discrete variables is defined by

$$L(Z^d_i, Z^d_j, \lambda) = \prod_{t=1}^{k} l(Z^d_{t,i}, Z^d_{t,j}, \lambda) = c_0(1 - \lambda)^{k - d_{z_i,z_j}} \lambda^{d_{z_i,z_j}}, \tag{21}$$

where $c_0 = \prod_{l=1}^{k} \mathbf{1}(Z^d_{t,i} \neq Z^d_{t,j})/(c_t - 1)$. The product kernels $L(X^d_i, X^d_j, \lambda)$ and $L(Y^d_i, Y^d_j, \lambda)$ are similarly defined. One can show that the results of Theorem 1 and Theorem 2 remain unchanged with the above product kernels, and the above definition of d_{z_i,z_j}.

In the above we have assumed that the discrete variables do not have a natural ordering, examples of which would include different regions, ethnicity, and so on. In practice, discrete variables may have some natural orderings, examples of which would include preference orderings (like, indifference, dislike), health (excellent, good, poor), and so forth. In this case Aitchison and Aitken (1976, p.29) suggest using the kernel weight function: $l(Z^d_{t,i}, Z^d_{t,j}, \lambda) = c(c_t, s)\lambda^s(1 - \lambda)^{c_t - s}$ when $|Z^d_{t,i} - Z^d_{t,j}| = s$ ($0 \leq s \leq c_t$), where $(c(c_t, s) = c_t!/[s!(c_t - s)!]$. The results of Theorem 1 and Theorem 2 can also be easily extended to cover the case for which some of the discrete variables have natural orderings while others do not.

3. SIMULATIONS

We now consider the finite-sample performance of the proposed method under a variety of scenarios. Though the theory we present is an extension of Hall at al. (forthcoming) to multivariate conditioned sets, we restrict attention in the following simulations to a univariate conditioned set for ease of interpretation. While Hall et al. (forthcoming) consider simulations

involving continuous Y, here we consider those involving discrete Y, a popular setting in economic applications. We assume that interest lies in predicting $Pr[Y = y|X_{i1}, \ldots]$, and in estimating how this probability responds to changes in the conditioning variables. The kernel estimator $\hat{g}(Y|x)$ is given in (9) and the gradient estimator is given by

$$\nabla_x \hat{g}(y|x) = \frac{\hat{m}(x)\nabla_x \hat{f}(x,y) - \hat{f}(x,y)\nabla_x \hat{m}(x)}{[\hat{m}(x)]^2}. \tag{22}$$

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

$$\text{DGP1}: \quad Y = \begin{cases} 1 & \text{if } X_1 + X_2 + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

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

The median predicted conditional probability and that for the Probit model for a sample size of $n = 100$ are plotted in Figure 1, while Table 1 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.

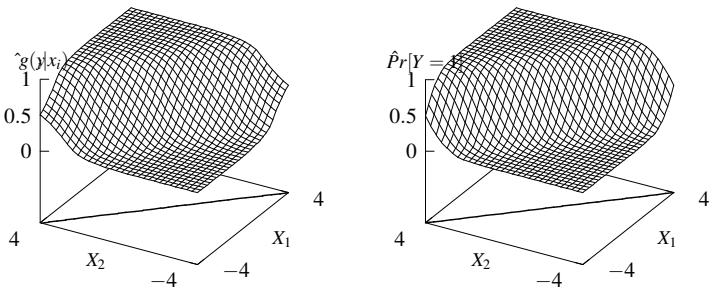

FIG. 1. Median kernel and Probit estimates of the conditional probability that $Y = 1$. The Probit estimate is the figure on the right. The contour line on the horizontal plane represents the boundary between the estimated conditional probability that $Y = 0$ and $Y = 1$ for a sample size of $n = 100$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

This situation is often modeled with a Probit specification. We are interested in how well the proposed method performs relative to a parametric model. As expected from Table 1, we observe that the parametric methods perform better than the nonparametric approach. Table 1 considers how this efficiency loss behaves as the sample size increases from $n = 100$

TABLE 1.

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

Kernel			**Probit**		
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			**Probit**		
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%

to $n = 1,000$, and we witness the consistent nature of the nonparametric 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

$$\text{DGP2}: \quad Y = \begin{cases} 1 & \text{if } -2 < X_1 + \epsilon_1 < 2 \text{ and } -2 < X_2 + \epsilon_2 < 2 \\ 0 & \text{otherwise} \end{cases}, \tag{24}$$

where ϵ_1 and ϵ_2 are white noise $N(0, \sigma_\epsilon^2)$ error terms with $\sigma_\epsilon = 0.1$. Note that the Probit model is misspecified for DGP2 because it uses a misspecified index function $\beta_1 X_1 + \beta_2 X_2$. The median predicted conditional probability along with the gradient with respect to X_1 are plotted in Figure 2.

This is a case in which the Probit model completely breaks down, as can be seen from an examination of Table 2. The Probit specification uses none of the conditioning information contained in X_1 and X_2 and simply predicts all zeros. The gradients from the Probit model are therefore zero everywhere and again none of the estimated parameters in the Probit model is significant except 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 Probit approach, for example, raises a

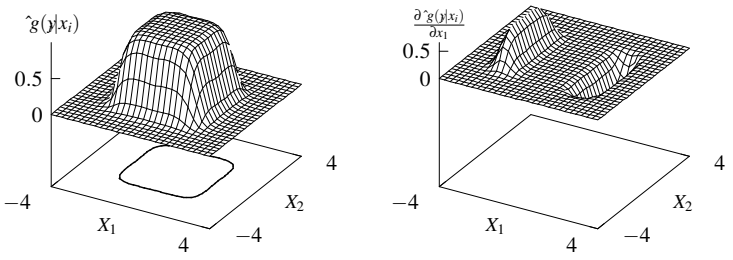

FIG. 2. Median kernel estimate of the conditional probability that $Y = 1$ and the gradient with respect to X_1. The contour line on the horizontal plane represents the boundary between the estimated conditional probability that $Y = 0$ and $Y = 1$ for a sample size of $n = 1,000$ based on $5,000$ Monte Carlo replications.

TABLE 2.

Confusion matrix and classification rates for the proposed method and that from a Probit model.

Kernel			**Probit**		
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%

number of issues such as normalization, identification, and specification of multiple indices. The proposed method does not suffer from any of these issues. Below we consider a multinomial categorical data case.

$$\text{DGP3}: \quad Y = \begin{cases} 1 & \text{if } X_1 + \epsilon_1 > 0 \text{ and } X_2 + \epsilon_2 > 0 \\ 2 & \text{if } X_1 + \epsilon_1 < 0 \text{ and } X_2 + \epsilon_2 < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

where ϵ_1 and ϵ_2 represent white noise $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.1$. For DGP3 a standard multinomial Probit model is misspecified because (25) does not have the conventional index functional form.

Both the median kernel and Probit estimators of $Pr[Y = 0|X_1, X_2]$ are plotted in Figure 3 below, while the confusion matrices and classification rates appear in Table 3. As can be seen, the multinomial Probit model cannot consistently model this situation and the gradients in particular from the Probit approach will be totally misleading.

The proposed estimator can readily model nonlinear conditional prediction of binary and multinomial categorical data without requiring the

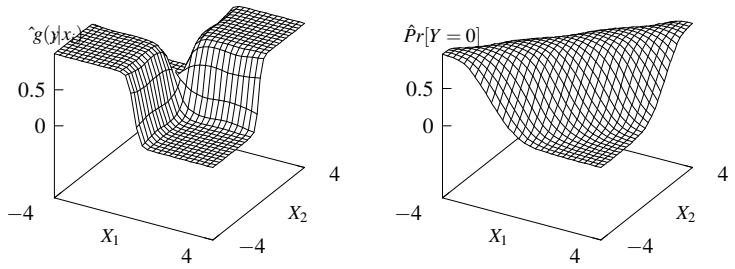

FIG. 3. Median kernel and Probit estimates of the conditional probability that $Y = 0$ for a sample size of $n = 100$ based on $5,000$ Monte Carlo replications. The Probit results are presented in the rightmost figure.

TABLE 3.

Confusion matrix and classification rates for the proposed method and that from a Probit model.

Kernel				**Probit**			
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%

researcher to specify functional forms for indices and distributions of the errors. The method only has a slight finite-sample efficiency loss compared to parametric estimators based on correctly specified models, while it completely dominates parametric estimators when the parametric model is misspecified.

4. CONCLUSION

This paper presents a nonparametric approach to the estimation of a multivariate conditional probability density function when faced with mixed categorical and continuous data and multivariate conditioned and conditioning variable sets. The approach can be useful in a wide variety of situations, and does not place the burden of correct specification on the researcher. The simulations presented in this paper highlight both the consistency and the flexibility of the proposed approach for a variety of situations.

APPENDIX A

Proof of Theorem 1.

In Appendix A we will use $(s.o.)$ to denote terms of smaller orders, or terms independent of (h, λ). For example, for $A_n = A_n(h, \lambda)$ and $B_n = B_n(h, \lambda)$, if we write $A_n = B_n + (s.o.)$, then $(s.o.)$ contains terms of smaller orders than B_n and the terms that are independent of (h, λ).

In order to save space, we will not distinguish between n^{-1} and $(n-1)^{-1}$, etc., since these will not change the conclusions in the proofs below. Also, we will write $\hat{m}(X_l)$ to denote $\hat{m}_{-l}(X_l)$, etc.

The random denominator $\hat{m}$ in $CV(h, \lambda)$ is difficult to handle from a theoretical point of view. This is dealt with by using the following identity:

$$\frac{1}{\hat{m}(X_l)} = \frac{1}{m(X_l)} + \frac{\hat{m}(X_l) - m(X_l)}{m(X_l)\hat{m}(X_l)}. \tag{A.1}$$

By the uniform consistency of $\hat{m}$ to m and given that m is bounded below in its support (see Lemma A.1), the second term is negligible compared to the first. Using $CV_1(h, \lambda)$ to denote $CV(h, \lambda)$ when $\hat{m}$ is replaced by m, from (18) we have

$$CV_1(h, \lambda) = n^{-1} \sum_l \frac{\hat{G}(X_l)}{[m(X_l)]^2} - 2n^{-1} \sum_l \frac{\hat{f}(X_l, Y_l)}{m(X_l)}. \tag{A.2}$$

Using (17), we have

$$\begin{aligned} E\left\{ \frac{\hat{G}(X_l)}{[m(X_l)]^2} \right\} &= E\left[\frac{n^{-2} \sum_{i \neq l} \sum_{j \neq l} K^{(2)}_{Y_i, Y_j} K_{X_i, X_l} K_{X_j, X_l}}{m^2(X_l)} \right] \\ &= n^{-1} E\left[\frac{K^{(2)}_{Y_i, Y_i} (K_{X_i, X_l})^2}{m^2(X_l)} \right] \\ &+ E\left[\frac{K^{(2)}_{Y_i, Y_j} K_{X_i, X_l} K_{X_j, X_l}}{m^2(X_l)} \right], \end{aligned} \tag{A.3}$$

where the first term corresponds to $i = j$ and the second term corresponds to $i \neq j$. In the above we ignore the difference between n, and $(n-1)$ since they will not change the order of the quantities we analyze.

Defining $J_n = E[CV_1(h,\lambda)]$, then, by (A.2) and (A.2), we have

$$\begin{aligned} J_n &\stackrel{def}{=} E(CV_1) \\ &= n^{-1}E\left[\frac{K^{(2)}_{Y_i,Y_i}(K_{X_i,X_l})^2}{m^2(X_l)}\right] + E\left[\frac{K^{(2)}_{Y_i,Y_j}K_{X_i,X}K_{X_j,X_l}}{m^2(X_l)}\right] \\ &\quad -2E\left[\frac{K_{Z_i,Z_l}}{m(X_l)}\right] \\ &= J_{n,1} + J_{n,2} - 2J_{n,3}, \end{aligned} \tag{A.4}$$

where the definition of $J_{n,j}$ $(j = 1, 2, 3)$ should be apparent.

From Lemma 2 and Lemma 3, we know that

$$J_n = J_{n,1}+J_{n,2}-2J_{n,3} = D_1h^4-D_2h^2\lambda+D_3\lambda^2+D_4(nh^q)^{-1}+(s.o.), \tag{A.5}$$

where $(s.o.)$ denote terms of smaller orders, or terms independent of (h,λ).

Lemma 4 shows that

$$\begin{aligned} CV_1 &\equiv \hat{J}_{n,1} + \hat{J}_{n,2} - 2\hat{J}_{n,3} \\ &= J_{n,1} + J_{n,2} - 2J_{n,3} \\ &+ O_p\left((h^2+\lambda)^3 + n^{-1/2}(h^2+\lambda) + (nh^{q/2})^{-1}\right). \end{aligned} \tag{A.6}$$

Define $CV_2 = CV - CV_1$. Using (A.1) and Lemma 1, one can easily show that

$$CV_2 = O_p(h^2+\lambda)O_p(CV_1) = O_p\left((h^2+\lambda)^3\right). \tag{A.7}$$

(A.5) and (A.7) give us

$$\begin{aligned} CV(h,\lambda) &= CV_1 + CV_2 \\ &= CV_0 + O_p((h^2+\lambda)^3 + n^{-1/2}\left(h^2+\lambda+(nh^q)^{-1/2}\right), \end{aligned} \tag{A.8}$$

where $CV_0 = J_{n,1} + J_{n,2} + J_{n,3}$.

From (A.7) one can show that $(\hat{h}-h_o)/h_o = O_p(n^{-\alpha/(4+q)})$ and $\hat{\lambda}-\lambda_o = O_p(n^{-\beta})$, where α and β are defined as in Theorem 1. We briefly discuss how this is done.

From (A.7) we know that $\hat{h} - h_o = o_p(h_o)$ and $\hat{\lambda} - \lambda_o = o_p(\lambda_o)$. Note that when $q \leq 3$, $(\hat{h}^2+\hat{\lambda})^3 = o_p\left(n^{-1/2}(h^2+\lambda+(nh^{q/2})^{-1}\right)$. Therefore, we have

$$CV(h,\lambda) = CV_0 + O_p\left(n^{-1/2}(h^2+\lambda+(nh^q)^{-1/2})\right) + (s.o.). \tag{A.9}$$

Define $h_1 = \hat{h} - h_o$ and $\lambda_1 = \hat{\lambda} - \lambda_o$, and note that h_1 (λ_1) has an order smaller than h_o (λ_o). Since $(\hat{h}, \hat{\lambda})$ minimizes (A.9), we must have $(\hat{h})^4 - h_o^4 = (h_o + h_1)^4 - h_o^4 = 4h_o^3 h_1 + (s.o.) = O\left(n^{-1/2}\hat{h}^2\right) = O(n^{-1/2}h_o^2)$, which gives $h_1 h_o = O_p(n^{-1/2})$, or $h_1/h_o \equiv (\hat{h} - h_o)/h_o = O_p(n^{-1/[2(4+q)]})$. Similarly, we have $\hat{\lambda}^2 - \lambda_o^2 = 2\hat{\lambda}\lambda_o + (s.o.) = O_p\left(n^{-1/2}\hat{\lambda}\right) = O_p\left(n^{-1/2}\lambda_o\right)$, which gives $\lambda_1 \equiv \hat{\lambda} - \lambda_o = O_p(n^{-1/2})$. Summarizing the above we have, for $q \leq 3$,

$$(\hat{h} - h_o)/h_o = O_p\left(n^{-1/[2(4+q)]}\right) \quad \text{and} \quad \hat{\lambda} - \lambda_o = O_p\left(n^{-1/2}\right). \tag{A.10}$$

When $q \geq 4$, we have

$$CV(h, \lambda) = CV_0 + O_p\left((h^2 + \lambda)^3\right) + (s.o.). \tag{A.11}$$

From (A.11) it is easy to see that $(\hat{h})^4 - h_o^4 = 4h_o^3 h_1 + (s.o.) = O(\hat{h}^6) = O(h_o^6)$, which leads to $h_1 = O_p(h_o^3)$, or $h_1/h_o = O_p(h^3)$. Also, $\hat{\lambda}^2 - \lambda_o^2 = 2\hat{\lambda}\lambda_o + (s.o.) = O_p(\hat{\lambda}^3) = O_p(\lambda_o^3)$, which gives $\lambda_1 \equiv \hat{\lambda} - \lambda_o = O_p(\lambda_o^2) = O_p(h_o^4)$ (because $\lambda_o = O(h_o^2)$). Thus we have for $q \geq 4$,

$$(\hat{h} - h_o)/h_o = O_p\left(n^{-2/(4+q)}\right) \quad \text{and} \quad \hat{\lambda} - \lambda_o = O_p\left(n^{-4/(4+q)}\right). \tag{A.12}$$

(A.10) and (A.12) prove Theorem 1.

Proof of Theorem 2

Define $\tilde{f}(z)$ and $\tilde{m}(x)$ the same way as $\hat{f}(z)$ and $\hat{m}(z)$ but with $(\hat{h}, \hat{\lambda})$ being replaced by (h_o, λ_o). Then it is easy to show that

$$E[\tilde{f}(z)] - f(z) = h_o^2 \mathcal{B}_1(z) + \lambda_o \mathcal{B}_2(z) + O((h_o^2 + \lambda)^2), \tag{A.13}$$

$$Var(\tilde{f}(z)) = (nh^q)^{-1}[\Omega(z) + O(h^2 + \lambda_o)], \tag{A.14}$$

and

$$\tilde{m}(x) - m(x) = O_p(h_o^2 + \lambda_o). \tag{A.15}$$

(A.13), (A.14), and (A.15) imply that (using Lyapunov's CLT)

$$\begin{aligned} &\sqrt{nh^q}[\tilde{g}(z) - g(z) - (h_o^2 \mathcal{B}_1(z) + \lambda_o \mathcal{B}_2(z))m(z)] \\ &\to N(0, \Omega(z)) \text{ in distribution,} \end{aligned} \tag{A.16}$$

where $\tilde{g}(y|x) = \tilde{f}(z)/\tilde{m}(x)$, and where $\mathcal{B}_1(z)$ and $\mathcal{B}_2(z)$ are defined as in Theorem 2.

Using Theorem 1, (A.15), and a Taylor expansion argument, one can easily show that

$$\begin{aligned}&\sqrt{n\hat{h}^q}(\hat{g}(z)-g(z)-\hat{h}^2\mathcal{B}_1(z)-\hat{\lambda}\mathcal{B}_2(z))\\ &\to\ N(0,\Omega(z)) \text{ in distribution.}\end{aligned} \tag{A.17}$$

This completes the proof of Theorem 2.

APPENDIX B

LEMMA 1. *(i)* $\sup_{x\in\mathcal{D}_x}|\hat{m}(x)-m(x)|=O(h)$ *a.s.*
(ii) $\sup_{z\in\mathcal{D}_z}|\hat{g}(y|x)-g(y|x)|=O(h)$ *a.s.*

Proof: First note that $\hat{h}=o(1)$ by Assumption (A2), and using Assumption (A3), one can show that $\hat{\lambda}=o_p(1)$. The remaining steps are similar to the proof of Lemma 1 of Härdle and Marron (1985), and are therefore omitted here.

LEMMA 2. $J_{n,1}=D_4(nh^q)^{-1}+O((nh^q)^{-1}(h^2+\lambda))$,
where D_4 *is constant defined in the proof below.*

Proof: Define

$$G_h(z^d,z_1^d)=h^{-2q}\int W^2\left(\frac{z_1^c-z^c}{h}\right)f(z_1^c,z_1^d)m^{-1}(x^c,x^d)\,dz^c\,dz_1^c. \tag{B.1}$$

From $J_{n,1}=n^{-1}E\left[K^{(2)}_{Y_i,Y_i}(K_{X_i,X_l})^2/m^2(X_l)\right]$, and $K^{(2)}_{Y_i,Y_i}=\int[K_{Y_i,y}]^2dy$, we have

$$\begin{aligned}&nJ_{n,1}\\ =&E\left[K^{(2)}_{Y_i,Y_i}(K_{X_i,X_l})^2/m^2(X_l)\right]=\int E\left\{[K_{y,Y_1}]^2[K_{X_1,X_2}]^2/m^2(X_2)\right\}dy\\ =&\int\left[K^2_{y,y_1}K^2_{x_1,x}/m(x)\right]f(z_1)\,dz_1dy\,dx=\int\left[K^2_{z,z_1}/m(x)\right]f(z_1)\,dz_1\,dz\\ =&\sum_{z^d}\sum_{z_1^d}L^2_{z^d,z_1^d}h^{-2q}\int W^2((z_1^c-z^c)/h)f(z^c,z^d)m^{-1}(xz_1^c,x_1^d)\,dz^c\,dz_1^c\\ =&\sum_{z^d}\sum_{z_1^d}L^2_{z^d,z_1^d}G_h(z^d,z_1^d)\\ =&(1-\lambda)^{2q}\sum_{z^d}G_h(z^d,z^d)+\lambda(1-\lambda)^{2q-1}\sum_{z^d}\sum_{z_1^d,d_{z_1,z}=1}G_h(z^d,z_1^d)+O(\lambda^2)\end{aligned}$$

$$
\begin{aligned}
&=(1-2q\lambda)\sum_{z^d} G_h(z^d,z^d)+\lambda\sum_{z^d}\sum_{z_1^d,d_{z_1,z}=1} G_h(z^d,z_1^d)+O(\lambda^2)\\
&=(1-2q\lambda)T_{0,h}+\lambda T_{1,h}+O(\lambda^2)\\
&=T_{0,h}+\lambda(T_{1,h}-2qT_{0,h})+O(\lambda^2), \qquad \text{(B.2)}
\end{aligned}
$$

where

$$
\begin{aligned}
T_{0,h} &= \sum_{z^d} G_h(z^d,z^d)\\
T_{1,h} &= \sum_{z^d}\sum_{z_1^d,d_{z_1,z}=1} G_h(z^d,z_1^d). \qquad \text{(B.3)}
\end{aligned}
$$

Applying change-of-variables to (B.1), we have

$$
\begin{aligned}
G_h(z^d,z_1^d) &= h^{-2q}\int W^2\left(\frac{z_1^c-z^c}{h}\right) f(z_1^c,z_1^d)m^{-1}(x^c,x^d)\,dz^c\,dz_1^c\\
&= h^{-q}\int W^2(v) f(z^c+hv,z_1^d)m^{-1}(x^c,x^d)\,dz^c\,dv\\
&= h^{-q}\left[G_0(z^d,z_1^d)+O(h^2)\right], \qquad \text{(B.4)}
\end{aligned}
$$

where

$$
G_0(z^d,z_1^d)=\left[\int f(z^c,z_1^d)m^{-1}(x^c,x^d)\,dz^c\right]\left[\int W^2(v)\,dv\right]. \qquad \text{(B.5)}
$$

Substituting (B.3) into (B.2), we get

$$
\begin{aligned}
T_{0,h} &= h^{-q}\sum_{z^d} G_0(z^d,z^d)+O(h^{2-q}) \equiv h^{-q}T_{0,0}+O(h^{2-q})\\
T_{1,h} &= h^{-q}\sum_{z^d}\sum_{z_1^d,d_{z_1,z}=1} G_0(z^d,z_1^d)+O(h^{2-q})\\
&\equiv h^{-q}T_{1,0}+O(h^{2-q}), \qquad \text{(B.6)}
\end{aligned}
$$

where $T_{0,0} = \sum_{z^d} G_0(z^d,z^d)$, $T_{1,0} = \sum_{z^d}\sum_{z_1^d,d_{z_1,z}=1} G_0(z^d,z_1^d)$ with $G_0(z^d,z_1^d)$ given in (B.5).

Substituting (B.5) into (B.2), we have

$$
\begin{aligned}
J_{n,1} &= n^{-1}\left[T_{0,h}+\lambda(T_{1,h}-2qT_{0,h})+O(\lambda^2)\right]\\
&= (nh^q)^{-1}\left[T_{0,0}+\lambda(T_{1,0}-2qT_{0,0})+O(h^2)+O(\lambda^2)\right]\\
&= D_4(nh^q)^{-1}+O((nh^q)^{-1}(\lambda+h^2)), \qquad \text{(B.7)}
\end{aligned}
$$

where $D_4 = T_{0,0}$ ($D_4 > 0$).

LEMMA 3. *$J_{n,2} - 2J_{n,3} = D_0 + D_1 h^4 - D_2 \lambda h^2 + D_3 \lambda^2 + O((h^2+\lambda)^2)$, where D_j's ($j = 0, 1, ..., 4$) are some constants defined in the proof below.*

Proof: We first consider $J_{n,3}$. Define

$$M_h(z^d, z_1^d) = \int \left[W(z^c, z_1^c)/m(x^c, x^d)\right] f(z_1^c, z_1^d) f(z^c, z^d)\, dz^c\, dz_1^c. \quad \text{(B.8)}$$

We have

$$\begin{aligned}
J_{n,3} &= E\left[K_{Z_i,Z_l}/m(X_l)\right] = E\left[L_{Z_i^d,Z_l^d} W_{Z_i^c,Z_l^c}/m(X_l)\right] \\
&= \sum_{z_1^d}\sum_{z^d} L_{z_1^d,z^d} \int \left[W(z^c, z_1^c)/m(x^c, x^d)\right] f(z_1^c, z_1^d) f(z^c, z^d)\, dz^c\, dz_1^c \\
&\equiv \sum_{z_1^d}\sum_{z^d} L_{z_1^d,z^d} M_h(z^d, z_1^d) \\
&= (1-\lambda)^q \sum_{z^d} M_h(z^d, z^d) + \lambda(1-\lambda)^{q-1} \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} M_h(z^d, z_1^d) \\
&\quad + \lambda^2 (1-\lambda)^{q-2} \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=2} M_h(z^d, z_1^d) + O(\lambda^3) \\
&= (1 - q\lambda + q(q-1)\lambda^2/2) \sum_{z^d} M_h(z^d, z^d) \\
&\quad + \lambda(1-(q-1)\lambda) \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} M_h(z^d, z_1^d) \\
&\quad + \lambda^2 \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=2} M_h(z^d, z_1^d) + (s.o.) \\
&= (1 - q\lambda + q(q-1)\lambda^2/2) A_{0,h} + \lambda(1-(q-1)\lambda) A_{1,h} + \lambda^2 A_{2,h} + (s.o.) \\
&= A_{0,h} + \lambda(A_{1,h} - qA_{0,h}) \\
&+ \lambda^2 \left\{A_{2,h} - (q-1)A_{1,h} + [q(q-1)/2]\, A_{0,h}\right\} + (s.o.), \quad \text{(B.9)}
\end{aligned}$$

where

$$\begin{aligned}
A_{0,h} &= \sum_{z^d} M_h(z^d, z^d), \\
A_{1,h} &= \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} M_h(z^d, z_1^d) \\
A_{2,h} &= \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=2} M_h(z^d, z_1^d). \quad \text{(B.10)}
\end{aligned}$$

Applying change-of-variables to (B.8), we get

$$\begin{aligned} M_h(z^d, z_1^d) &= h^{-q} \int \left[W\left(\frac{z^c - z_1^c}{h} \right) / m(x^c, x^d) \right] f(z_1^c, z_1^d) f(z^c, z^d)\, dz^c\, dz_1^c \\ &= \int W(v) \left[m(x^c, x^d) \right]^{-1} f(z^c + hv, z_1^d) f(z^c, z^d)\, dz^c\, dv \\ &= M_0(z^d, z_1^d) + h^2 M_2(z^d, z_1^d) + h^4 M_4(z^d, z_1^d) + o(h^4), \quad \text{(B.11)} \end{aligned}$$

where

$$\begin{aligned} M_0(z^d, z_1^d) &= \int [m(x^c, x^d)]^{-1} f(z^c, z_1^d) f(z^c, z^d)\, dz^c, \\ M_2(z^d, z_1^d) &= (1/2) \int [m(x^c, x^d)]^{-1} W(v) v' \nabla^2 f(z^c, z_1^d) v f(z^c, z^d)\, dz^c\, dv, \\ M_4(z^d, z_1^d) &= \int [m(x^c, x^d)]^{-1} W(v) v^{(4)} \nabla^4 f(z^c, z_1^d) v f(z^c, z^d)\, dz^c\, dv, \quad \text{(B.12)} \end{aligned}$$

where

$$v^{(4)} \nabla^4 f(z^c, z^d) = (1/4!) \sum_{k_1+k_2+k_3+k_4=4} \frac{\prod_{s=1}^k (v_s)^{k_s} \partial^4 f(z^c, z^d)}{\prod_{s=1}^k \partial (z_s^c)^{k_s}}$$

denotes the fourth order Taylor expansion (v_s and z_s^c are the sth components of v and z^c, respectively).

Next we consider $J_{n,2}$. Define

$$Q_h(z^d, z_1^d, z_2^d) = \int W_{z_1^c, z^c} W_{z_2^c, z^c} [m(x^c, x^d)]^{-1} f(z_1^c, z_1^d) f(z_2^c, z_2^d) dz_1^c dz_2^c\, dz^c. \quad \text{(B.13)}$$

We have

$$\begin{aligned} & J_{n,2} \\ =& E\left[K^{(2)}_{Y_i, Y_j} K_{X_i, X_l} K_{X_j, X_l} / m^2(X_l) \right] \\ =& \int E\left[K_{Y_i, y} K_{Y_j, y} K_{X_i, X_l} K_{X_j, X_l} / m^2(X_l) \right] dy \\ =& \int \left[K_{y_1, y} K_{y_2, y} K_{x_1, x} K_{x_2, x} / m(x) \right] f(z_1) f(z_2)\, dz_1\, dz_2\, dx dy \\ =& \int [K_{z_1, z} K_{z_2, z} / m(x)] f(z_1) f(z_2)\, dz_1\, dz_2\, dz \\ =& \sum_{z^d} \sum_{z_1^d} \sum_{z_2^d} L_{z^d, z_1^d} L_{z^d, z_2^d} \int [W(z_1^c, z_1^c) W(z^c, z_2^c) / m(x)] f(z_1) f(z_2)\, dz_1^c\, dz_2^c\, dz^c \end{aligned}$$

$$
\begin{aligned}
&= \sum_{z^d} \sum_{z_1^d} \sum_{z_2^d} L_{z^d, z_1^d} L_{z^d, z_2^d} Q_h(z^d, z_1^d, z_2^d) \\
&= (1-\lambda)^{2q} \sum_{z^d} Q_h(z^d, z^d, z^d) \\
&+ \lambda(1-\lambda)^{2q-1} \left\{ \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} Q_h(z^d, z_1^d, z^d) + \sum_{z^d} \sum_{z_2^d, d_{z,z_2}=1} Q_h(z^d, z^d, z_2^d) \right\} \\
&+ \lambda^2(1-\lambda)^{2q-2} \left\{ \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=2} Q_h(z^d, z_1^d, z^d) + \sum_{z^d} \sum_{z_2^d, d_{z,z_2}=2} Q_h(z^d, z^d, z_2^d) \right. \\
&\left. + \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} \sum_{z_2^d, d_{z_2,z}=1} Q_h(z^d, z_1^d, z_2^d) \right\} + O(\lambda^3) \\
&= (1 - 2q\lambda + q(2q-1)\lambda^2) \sum_{z^d} Q_h(z^d, z^d, z^d) \\
&+ \lambda(1 - (2q-1)\lambda) \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} 2Q_h(z^d, z_1^d, z^d) \\
&+ \lambda^2 \left\{ \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=2} 2Q_h(z^d, z_1^d, z^d) + \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} \sum_{z_2^d, d_{z,z_2}=1} Q_h(z^d, z_1^d, z_2^d) \right\} \\
&\quad + O(\lambda^3) \\
&= (1 - 2q\lambda + q(2q-1)\lambda^2) B_{0,h} + \lambda(1 - (2q-1)\lambda) B_{1,h} + \lambda^2 B_{2,h} \\
&\quad + O(\lambda^3) \\
&= B_{0,h} + \lambda[B_{1,h} - 2qB_{0,h}] \\
&\quad + \lambda^2[B_{2,h} - (2q-1)B_{1,h} + q(2q-1)B_{0,h}] + O(\lambda^3), \qquad \text{(B.14)}
\end{aligned}
$$

where

$$
\begin{aligned}
B_{0,h} &= \sum_{z^d} Q_h(z^d, z^d, z^d), \\
B_{1,h} &= 2 \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} Q_h(z^d, z_1^d, z^d) \\
B_{2,h} &= 2 \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=2} Q_h(z^d, z_1^d, z^d) \\
&\quad + \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} \sum_{z_2^d, d_{z,z_2}=1} Q_h(z^d, z_1^d, z_2^d). \qquad \text{(B.15)}
\end{aligned}
$$

Applying change-of-variables to (B.13), it is easy to see that $Q_h(z^d, z_1^d, z_2^d)$ has the following expansion:

$$\begin{aligned} Q_h(z^d, z_1^d, z_2^d) =& Q_0(z^d, z_1^d, z_2^d) + h^2 Q_2(z^d, z_1^d, z_2^d) \\ &+ h^4 Q_4(z^d, z_1^d, z_2^d) + o(h^4), \end{aligned} \tag{B.16}$$

where

$$\begin{aligned} Q_0(z^d, z_1^d, z_2^d) &= \int [m(x^c, x^d)]^{-1} f(z^c, z_1^d) f(z^c, z_2^d)\, dz^c, \\ Q_2(z^d, z_1^d, z_2^d) &= (1/2) \int [m(x^c, x^d)]^{-1} W(v) [v' \nabla^2 f(z^c, z_1^d) v f(z^c, z_2^d) \\ &+ f(z^c, z_1^d) v' \nabla^2 f(z^c, z_2^d) v]\, dv\, dz^c, \\ Q_4(z^d, z_1^d, z_2^d) &= \int m^{-1}(x^c, x^d) W(v) W(u) [v' \nabla^2 f(x^c, z_1^d) v u' \nabla^2 f(x^c, z_2^d) u \\ &+ f(z^c, z_1^d) u^{(4)} \nabla^4 f(z^c, z_2^d) \\ &+ v^{(4)} \nabla^4 f(z^c, z_1^d) f(z^c, z_2^d)]\, du\, dv\, dz^c, \end{aligned} \tag{B.17}$$

where $v^{(4)} \nabla^4 f(z^c, z_1^d)$ is defined below (B.11), and $u^{(4)} \nabla^4 f(z^c, z_2^d)$ is similarly defined.

From (B.11), (B.16), and (B.17), we immediately obtain the following:

$$\begin{aligned} Q_0(z^d, z_1^d, z^d) &= M_0(z^d, z_1^d), \\ Q_2(z^d, z^d, z^d) &= 2M_2(z^d, z^d), \\ Q_4(z^d, z^d, z^d) &> 2M_4(z^d, z^d), \\ \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} Q_2(z^d, z_1^d, z^d) &= 2 \sum_{z^d} \sum_{z_1^d, d_{z,z_1}=1} M_2(z^d, z_1^d). \end{aligned} \tag{B.18}$$

From (B.8) and (B.14), we get

$$J_{n,2} - 2J_{n,3} = C_{0,h} + \lambda C_{1,h} + \lambda^2 C_{2,h} + O(\lambda^3), \tag{B.19}$$

where $C_{0,h} = B_{0,h} - 2A_{0,h}$, $C_{1,h} = (B_{1,h} - 2qB_{0,h}) - 2(A_{1,h} - qA_{0,h})$, and $C_{2,h} = [B_{2,h} - (2q-1)B_{1,h} + q(2q-1)B_{0,h}] - 2\{A_{2,h} - (q-1)A_{1,h} + q(q-1)/2]A_{0,h}\}$.

Using (B.9), (B.14) and (B.17), we have

$$\begin{aligned} C_{0,h} &= B_{0,h} - 2A_{0,h} = \sum_{z^d}[Q_h(z^d, z^d, z^d) - 2M_h(z^d, z^d)] \\ &= \sum_{z^d}[Q_0(z^d, z^d, z^d) - 2M_0(z^d, z^d)] \\ &\quad + h^2 \sum_{z^d}[Q_2(z^d, z^d, z^d) - 2M_2(z^d, z^d)] \\ &\quad + h^4 \sum_{z^d}[Q_4(z^d, z^d, z^d) - 2M_4(z^d, z^d)] + o(h^4) \\ &= -\sum_{z^d} M_0(z^d, z^d) + h^2(0) + h^4 \sum_{z^d}[Q_4(z^d, z^d) - 2M_4(z^d, z^d)] + o(h^4) \\ &\equiv D_0 + D_1 h^4, \end{aligned} \tag{B.20}$$

where $D_0 = -\sum_{z^d} Q_0(z^d, z^d)$ and $D_1 = \sum_{z^d}[Q_4(z^d, z^d) - 2M_4(z^d, z^d)]$. $D_1 > 0$ by (B.17).

By (B.9), (B.14) and (B.17), we have

$$\begin{aligned} C_{1,h} &= 2q(A_{0,h} - B_{0,h}) + (B_{1,h} - 2A_{1,h}) \\ &= 2q \sum_{z^d}[M_h(z^d, z^d) - Q_h(z^d, z^d, z^d)] \\ &\quad + \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} [Q_h(z^d, z_1^d, z^d) - 2M_h(z^d, z_1^d)] \\ &= 2q \sum_{z^d}\{0 + h^2[M_2(z^d, z^d) - Q_2(z^d, z_1^d, z^d)] + O(h^4)\} \\ &\quad + \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} \{0 + h^2[Q_2(z^d, z_1^d, z^d) - M_2(z^d, z_1^d)] + O(h^4)\} \\ &\quad - h^2(2q) \left\{ \sum_{z^d} M_2(z^d, z^d) - \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} M_2(z^d, z_1^d) \right\} + O(h^4) \\ &= -h^2 D_2 + O(h^4), \end{aligned} \tag{B.21}$$

where $D_2 = 2q\{\sum_{z^d} M_2(z^d, z^d) - \sum_{z^d} \sum_{z_1^d, d_{z_1,z}=1} M_2(z^d, z_1^d)\}$.

Define $A_{j,0}$ the same way as $A_{j,h}$ except that $Q_h(\cdot)$ in $A_{j,h}$ is replaced by $Q_0(\cdot)$ ($Q_0(\cdot)$ defined in (B.16)). Also define $B_{j,0}$ the same way as $B_{j,h}$ except that $M_h(\cdot)$ in $B_{j,h}$ is replaced by $M_0(\cdot)$ ($M_0(\cdot)$ defined in (B.11)) ($j = 1, 2, 3$). Then we have

$$\begin{aligned} A_{j,h} &= A_{j,0} + O(h^2), \\ B_{j,h} &= B_{j,0} + O(h^2). \end{aligned} \tag{B.22}$$

Using (B.9), (B.14) and (B.21), we get

$$\begin{aligned} C_{2,h} &= [B_{2,h} - 2A_{2,h}] + [2(q-1)A_{1,h} - (2q-1)B_{1,h}] \\ &+ q[(2q-1)B_{0,h} - (q-1)A_{0,h}/2 \\ &= [B_{2,0} - 2A_{2,0}] + [2(q-1)A_{1,0} - (2q-1)B_{1,0}] \\ &+ q[(2q-1)B_{0,0} - (q-1)A_{0,0}/2 + O(h^2) \\ &\equiv D_3 + O(h^2), \end{aligned} \tag{B.23}$$

where we define $D_3 = [B_{2,0} - 2A_{2,0}] + [2(q-1)A_{1,0} - (2q-1)B_{1,0}] + q[(2q-1)B_{0,0} - (q-1)A_{0,0}/2$.

Summarizing (B.19) through (B.22) we have shown that

$$\begin{aligned} J_{n,2} - 2J_{n,3} &= C_{0,h} + \lambda C_{1,h} + \lambda^2 C_{2,h} + O(\lambda^3) \\ &= D_0 + D_1 h^4 - D_2 h^2 \lambda + D_3 \lambda^2 + O((h^2+\lambda)^3). \end{aligned} \tag{B.24}$$

This completes the proof of Lemma 3.

LEMMA 4. $CV_1 = J_{n,1} + J_{n,2} - 2J_{n,3} + O_p((h^2+\lambda)^3) + O_p(n^{-1/2}(h^2 + \lambda + (nh^q)^{-1/2}) + (s.o.)$.

Proof: Lemmas 2 and 3 have shown that

$$\begin{aligned} E(CV_L) =& D_0 + D_1 h^4 - D_2 h^2 \lambda + D_3 \lambda^2 \\ &+ D_4 (nh^q)^{-1} + O((h^2+\lambda)^3 + (nh^q)^{-1}(h^2+\lambda)). \end{aligned}$$

It is easy to see that $\hat{h}$ needs to balance terms of order h^4 and $(nh^q)^{-1}$. Therefore, h^2 has an order larger than $n^{-1/2}$, or $n^{-1/2} = o(h^2)$. Below we will show that $CV_1 - E(CV_1) = O_p(n^{-1/2}(\lambda + h^2)) + O_p(nh^{q/2})$.

Substituting (17) and (15) into (19),

$$\begin{aligned} CV_1 &= n^{-3} \sum_l \sum_{i \neq l} \sum_{j \neq l} \left[\frac{K^{(2)}_{Y_i,Y_j} K_{X_i,X_l} K_{X_j,X_l}}{m^2(X_l)} \right] - 2n^{-3} \sum_l \sum_{i \neq l} \left[\frac{K_{Z_i,Z_l}}{m(X_l)} \right] \\ &= n^{-1} \left[n^{-2} \sum_l \sum_{i \neq l} \frac{K^{(2)}_{Y_i,Y_i} K^2_{X_i,X_l}}{m^2(X_l)} \right] \\ &+ n^{-3} \sum_l \sum_{i \neq l} \sum_{j \neq l, j \neq i} \frac{K^{(2)}_{Y_i,Y_j} K_{X_i,X} K_{X_j,X_l}}{m^2(X_l)} - 2n^{-3} \sum_l \sum_{i \neq l} \left[\frac{K_{Z_i,Z_l}}{m(X_l)} \right] \\ &\equiv \hat{J}_{n,1} + \hat{J}_{n,2} - 2\hat{J}_{n,3}, \end{aligned} \tag{B.25}$$

where the definitions of $\hat{J}_{n,s}$ $(s = 1, 2, 3)$ should be apparent. $\hat{J}_{n,1}$ and $\hat{J}_{n,3}$ can be written as second-order U-statistics, and $\hat{J}_{n,2}$ as a third order U-statistic. Below we work on $\hat{J}_{n,3}$ first. Note that we can write $\hat{J}_{n,3}$ as (ignoring the difference between n and $(n-1)$)

$$\hat{J}_{n,3} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} \mathcal{H}_n(Z_i, Z_j), \tag{B.26}$$

where $\mathcal{H}_n(Z_i, Z_j) = (1/2)K_{Z_i,Z_j}[m^{-2}(X_i) + m^{-2}(X_j)]$. Letting $\theta = E[\mathcal{H}_n(Z_i, Z_j)]$, by the H-decomposition of U-statistics, we know that

$$\begin{aligned} \hat{J}_{n,3} &= \theta + \frac{2}{n} \sum_i [\mathcal{H}_{n,1}(Z_i) - \theta] \\ &+ \frac{2}{n(n-1)} \sum_i \sum_{j>i} [\mathcal{H}_n(Z_i, Z_j) - \mathcal{H}_{n,1}(Z_i) - \mathcal{H}_{n,1}(Z_j) + \theta]. \end{aligned} \tag{B.27}$$

By the proof of Lemma A.3, we know that $\theta = E[\mathcal{H}_n(Z_i, Z_j)] = \alpha_1\lambda + \alpha_2 h^2 + (s.o.)$ for some constants α_j's $(j = 1, 2$; recall that $(s.o.)$ also includes terms that are independent of $(h, \lambda))$. By similar arguments, it is easy to see that $\mathcal{H}_{n,1}(Z_i) = \beta_{1i}\lambda + \beta_{2i}h^2 + (s.o.)$ for some functions $\beta_{j,i} = \beta_j(Z_i)$ $(j = 1, 2)$. Therefore,

$$n^{-1} \sum_i [\mathcal{H}_{n,1}(Z_i) - \theta] = n^{-1/2}[O_p(\lambda + h^2)] + (s.o.).$$

Also, the last term in the H-decomposition is a degenerate U-statistic and it is easy to show that it has an order of $O_p((nh^{q/2})^{-1})$. By noting that $J_{n,3} = E[\hat{J}_{n,3}] = \theta$, we have shown that

$$\hat{J}_{n,3} = J_{n,3} + O_p(n^{-1/2}(h^2 + \lambda) + O_p((nh^{q/2})^{-1})) + (s.o.). \tag{B.28}$$

By exactly the same arguments, one can show that

$$\hat{J}_{n,2} = J_{n,2} + O_p(n^{-1/2}(h^2 + \lambda) + O_p((nh^{q/2})^{-1}) + (s.o.). \tag{B.29}$$

For $\hat{J}_{n,1}$, we know from Lemma 2 that $J_{n1} = E(\hat{J}_{n,1}) = O((nh^q)^{-1})$. Hence, by H-decomposition, it is easy to show that

$$\hat{J}_{n,1} = E(\hat{J}_{n,1}) + n^{-1/2}O((nh^q)^{-1})) = J_{n,1} + O_p(n^{-1/2}(nh^q)^{-1}). \tag{B.30}$$

(B.28) through (B.30) therefore give us the result

$$\begin{aligned} CV_1 &\equiv \hat{J}_{n,1} + \hat{J}_{n,2} - 2\hat{J}_{n,3} \\ &= J_{n,1} + J_{n,2} - 2J_{n,3} \\ &+ O_p\left((h^2 + \lambda)^3 + n^{-1/2}(h^2 + \lambda) + (nh^{q/2})^{-1}\right). \end{aligned} \tag{B.31}$$

REFERENCES

Aitchison, J. and C.G.G. Aitken, 1976, Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413-420.

Bowman, A.W., P. Hall, and T. D.M. Titterington, 1984, Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* **71**, 341-351.

Fahrmeir, L. and G. Tutz, 1994, *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag: New York.

Grund, B. and P. Hall, 1993, On the performance of kernel estimators for high-dimensional sparse binary data. *Journal of Multivariate Analysis* **44**, 321-344.

Hall, P., 1981, On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287-294.

Hall, P. and J. S. Racine, and Q. Li (forthcoming), Cross-Validation and the Estimation of Conditional Probability Densities. *Journal of The American Statistical Association*.

Hall, P. and M. Wand, 1988, On nonparametric discrimination using density differences. *Biometrika* **75**, 541-547.

Härdle, W., P. Hall, and J.S. Marron, 1988, How far are automatically chosen regression smoothing parameters from their optimum? *Journal of American Statistical Association* **83**, 86-99.

Härdle, W., P. Hall, and J.S. Marron, 1992, Regression smoothing parameters that are not far from their optimum. *Journal of American Statistical Association* **87**, 227-233.

Härdle, W. and J.S. Marron, 1985, Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics* **13**, 1465-1481.

Kalbfleisch, J.D. and R.L. Prentice, 1980, *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Li, Q. and J. S. Racine, 2003, Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* **86**, 266-292.

Racine, J. S. and Q. Li, 2004, Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**, 99-130.

Scott, D., 1992, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons.

Simonoff, J.S., 1996, *Smoothing Methods in Statistics*. Springer: New York.

Titterington, D.M., 1980, A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22**, 259-268.

Wang, M.C., and J. Ryzin, 1981, A class of smooth estimators for discrete distributions. *Biometrika* **68**, 301-309.