Nonparametric estimation of distributions with categorical and continuous data

Qi Li[a],* and Jeff Racine[b]

[a] *Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA*
[b] *Department of Economics, Syracuse University, Syracuse, NY 13244-1020, USA*

Abstract

In this paper we consider the problem of estimating an unknown joint distribution which is defined over mixed discrete and continuous variables. A nonparametric kernel approach is proposed with smoothing parameters obtained from the cross-validated minimization of the estimator's integrated squared error. We derive the rate of convergence of the cross-validated smoothing parameters to their 'benchmark' optimal values, and we also establish the asymptotic normality of the resulting nonparametric kernel density estimator. Monte Carlo simulations illustrate that the proposed estimator performs substantially better than the conventional nonparametric frequency estimator in a range of settings. The simulations also demonstrate that the proposed approach does not suffer from known limitations of the likelihood cross-validation method which breaks down with commonly used kernels when the continuous variables are drawn from fat-tailed distributions. An empirical application demonstrates that the proposed method can yield superior predictions relative to commonly used parametric models.
© 2003 Elsevier Science (USA). All rights reserved.

*Corresponding author.
E-mail address:* qi@econmail.tamu.edu (Q. Li).

1. Introduction and background

Nonparametric kernel methods are frequently used to estimate joint distributions, however, conventional approaches do not handle mixed discrete and continuous data in a satisfactory manner. Although it is widely appreciated that one can use a frequency estimator to obtain consistent nonparametric estimates of a joint probability density function (PDF) in the presence of discrete variables, this frequency-based approach splits the sample into many parts ('cells') and the number of observations lying in each cell may be insufficient to ensure the accurate nonparametric estimation of the PDF of the remaining continuous variables. Furthermore, it is not uncommon to encounter situations in which the number of cells exceeds the number of observations hence the conventional frequency estimator cannot even be applied.

Aitchison and Aitken [4] proposed a novel nonparametric kernel method for estimating a joint distribution defined over binary data in a multivariate binary discrimination context. They also proposed a data-dependent *likelihood-based* method of bandwidth selection which has been shown to be consistent by Bowman [5]. One advantage that their method has over the conventional frequency estimator is that it does not split the sample into cells in finite-sample applications. A weakness of their method becomes apparent, however, in mixed discrete and continuous variable settings. This weakness results in part from the use of likelihood cross-validatory bandwidth selection which is known to break down when modeling 'fat-tailed' 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 [13,14]). For related work on issues surrounding the kernel estimation of distributions defined over discrete data the reader is referred to Hall [11] and Hall and Wand [15]. In related papers, Grund [9] and Grund and Hall [10] investigated the kernel estimation of a PDF defined over k-dimensional multivariate binary data using *least-squares* cross-validation. In particular, they looked at both the situation with fixed k and the case where $k \to \infty$ as the sample size $n \to \infty$. For an excellent survey on kernel density estimation methods see Izenman [18], while more in-depth treatments of the subject can be found in [7,17,20,22].

While there exist a number of theoretical papers on the properties of cross-validation methods with only discrete variables (e.g., [9–11]), or with only continuous variables [16], little attention has been paid to the more general and interesting case of mixed discrete and continuous variables. The exceptions are the papers by Tutz [23] and Ahmad and Cerrito [3] who have considered cross-validation for estimating conditional density functions and regression functions (with mixed variables), respectively. However, both Tutz [23] and Ahmad and Cerrito [3] only demonstrate that their estimators are consistent—they have not established the asymptotic distributions of their estimators. It is appreciated that establishing the asymptotic distribution of an estimator is typically a more formidable task than that of establishing consistency alone.

In this paper we aim to close this gap by providing the theoretical foundations for a consistent kernel estimator of a joint PDF defined over mixed discrete and

continuous data employing *least-squares* cross-validation selection of the smoothing parameters. In particular, we obtain rates of convergence of the smoothing parameters to some benchmark optimal values, and we establish the asymptotic normality of the estimator. We also provide simulations and applications of the proposed approach designed to examine its finite-sample performance. To the best of our knowledge, our work is the first to establish *asymptotic normality* results for kernel density estimators with mixed discrete and continuous variables using *cross-validation* methods.

The rest of this paper proceeds as follows. In Section 2 we restrict attention to the multivariate discrete variables case and consider estimating a joint PDF using least-squares cross-validation. We establish the convergence rate of the cross-validated smoothing parameters and the asymptotic normality of the resulting kernel probability estimator. Section 3 builds on these results for the general mixed discrete and continuous variables case. We again obtain convergence rates for the cross-validated smoothing parameters and establish the asymptotic normality of the resulting estimator. Section 4 reports on simulations designed to illuminate the finite-sample performance of the estimator. Section 5 considers an empirical application which demonstrates how the proposed approach can be used to yield superior predictions relative to commonly used parametric models of binary choice. Finally, Section 6 concludes and discusses a number of possible extensions.

2. Estimating a joint density with categorical data

In this section we consider the estimation of a joint PDF defined over discrete data. Let X denote a $k \times 1$ vector of discrete variables. For expositional simplicity we consider the case where X is a k-dimensional binary variable, $X \in \{0,1\}^k$ (we discuss the more general case at the end of Section 3). We denote $\{0,1\}^k$ by $\mathscr{D}$ and let $p(\cdot)$ denote the probability function of X. We use $X_{i,t}$ and x_t to denote the tth component of X_i and x ($i = 1, \ldots, n$), respectively. For x_t, $X_{i,t} \in \{0,1\}$, define a univariate kernel function $l(X_{i,t}, x_t) = 1 - \lambda$ if $X_{i,t} = x_t$, and $l(X_{i,t}, x_t) = \lambda$ if $X_{i,t} \neq x_t$, where λ is a smoothing parameter.

For multivariate data we use a standard product kernel given by

$$L(X_i, x, \lambda) = \prod_{t=1}^{k} l(X_{t,i}, x_t) = (1 - \lambda)^{k - d_{ix}} \lambda^{d_{ix}}, \tag{2.1}$$

where $d_{ix} = k - \mathbf{1}(X_{i,t} - x_t)$ equals the number of 'disagreement components' between X_i and x, $\mathbf{1}(A)$ is the usual indicator function, which equals one if A holds, and zero otherwise. Note that d_{ix} takes values in $\{0, 1, 2, \ldots, k\}$.

We would like to emphasize that we use a scalar λ for expositional simplicity. In practice, one would use a different smoothing parameter λ for each different component of x, i.e., λ should be a k-dimensional vector, and any multidimensional search algorithm will do so. Dealing with a k-dimensional vector λ will make the

notation and proofs much more cumbersome. Therefore, only the scalar λ case is treated in this paper.

We estimate $p(x)$ by

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, x, \lambda). \tag{2.2}$$

The sum of squared differences between $\hat{p}(\cdot)$ and $p(\cdot)$ is given by

$$I_n = \sum_{x \in \mathscr{D}} [\hat{p}(x) - p(x)]^2 = \sum_{x \in \mathscr{D}} [\hat{p}(x)]^2 - 2 \sum_{x \in \mathscr{D}} \hat{p}(x)p(x) + \sum_{x \in \mathscr{D}} [p(x)]^2. \tag{2.3}$$

Using Eq. (2.2) we have $\sum_{x \in \mathscr{D}} [\hat{p}(x)]^2 = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{ij}^{(2)}$, where $L_{ij}^{(2)} = \sum_{x \in \mathscr{D}} L_{ix}L_{jx}$. We estimate $\sum_x \hat{p}(x)p(x) = E[\hat{p}(X)]$ by $n^{-1} \sum_{i=1}^{n} \hat{p}_{-i}(X_i) = [n(n-1)]^{-1} \sum_{i=1}^{n} \sum_{j=1,j \neq i}^{n} L_{ij}$, where $\hat{p}_{-i}(X_i) = (n-1)^{-1} \sum_{j=1,j \neq i}^{n} L_{ij}$ is the leave-one-out kernel estimator of $p(X_i)$, $L_{ij} = L(X_i, X_j, \lambda)$. The last term on the right-hand side of Eq. (2.3) is unrelated to λ, therefore, we choose λ to minimize the cross-validated integrated squared error given by

$$CV(\lambda) \stackrel{\text{def}}{=} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{ij}^{(2)} - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j \neq i}^{n} L_{ij}. \tag{2.4}$$

We let $\tilde{\lambda}$ denote the cross-validated choice of λ. The following assumption is used to derive the rate at which $\tilde{\lambda}$ converges to zero along with the asymptotic normality of $\sqrt{n}(\hat{p}(x) - p(x))$.

Assumption (A). (i) X_i is independent and identically distributed (i.i.d.) as X, (ii) $p(x)$ is not a constant function in $x \in \mathscr{D}$, (iii) $\min_{\{x \in \mathscr{D}\}} p(x) \geqslant \delta$ for some $\delta > 0$.

Theorem 2.1. *Under Assumption* (A), *we have*

(i) $\tilde{\lambda} = O_p(n^{-1})$,
(ii) *For any* $x \in \mathscr{D}$, $\sqrt{n}(\hat{p}(x) - p(x)) \to N(0, p(x)(1 - p(x)))$ *in distribution.*

The proof of Theorem 2.1 is given in Appendix A. This theorem demonstrates that our cross-validation choice of $\tilde{\lambda}$ converges to zero at the rate of n^{-1}, the same rate as the maximum likelihood cross-validation choice of λ (see [11]). Next, we turn our attention to the mixed discrete and continuous variables case.

3. Estimating a joint density with mixed data

We now consider the case involving mixed discrete and continuous data. As in Section 2, $X \in \mathscr{D}$ represents the discrete variables, and we use $Y \in \mathscr{R}^p$ to denote the

continuous variables. Let $Y_{i,t}$ denote the tth component of Y_i, let $w(\cdot)$ be a univariate kernel function, and let $W(\cdot)$ be the product kernel function for the continuous variables. We define

$$W_{h,ij} \equiv W_h(Y_i, Y_j) \stackrel{\text{def}}{=} h^{-p} W\left(\frac{Y_i - Y_j}{h}\right) = h^{-p} \prod_{t=1}^{p} w\left(\frac{Y_{i,t} - Y_{j,t}}{h}\right), \tag{3.1}$$

where h is the smoothing parameter. We only consider a scalar h case for expositional simplicity. In applications, h should be a $p \times 1$ vector. We further define $Z = (X, Y)$, and we use $f(z) = f(x, y)$ to denote the joint PDF of (X, Y). We estimate $f(z)$ by

$$\hat{f}(z) = \frac{1}{n} \sum_{i=1}^{n} K_{h,iz}, \tag{3.2}$$

where $K_{h,iz} = L_{ix} W_{h,iy}$, $W_{h,iy} = h^{-p} W\left(\frac{Y_i - y}{h}\right)$ and $L_{ix} = L(X_i, x, \lambda)$ is that defined in Eq. (2.1). Using the notation $\int dz = \sum_{x \in \mathscr{D}} \int dy$, the integrated squared difference between $\hat{f}(\cdot)$ and $f(\cdot)$ is

$$J_n = \int [\hat{f}(z) - f(z)]^2 \, dz = \int [\hat{f}(z)]^2 \, dz - 2 \int \hat{f}(z) f(z) \, dz + \int [f(z)]^2 \, dz. \tag{3.3}$$

Using Eq. (3.2) we have $\int [\hat{f}(z)]^2 \, dz = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h,ij}^{(2)}$, where $K_{h,ij}^{(2)} = L_{ij}^{(2)} W_{h,ij}^{(2)}$ with $L_{ij}^{(2)} = \sum_{x \in \mathscr{D}} L_{ix} L_{jx}$ and $W_{h,ij}^{(2)} = \int W_{h,iy} W_{h,jy} \, dy$. We estimate $\int \hat{f}(z) f(z) \, dz \equiv E[\hat{f}(Z)]$ by $n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(Z_i) \equiv \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} K_{h,ij}$, where $\hat{f}_{-i}(Z_i) \equiv \frac{1}{(n-1)} \sum_{j=1, j \neq i}^{n} K_{h,ij}$, $K_{h,ij} = L_{ij} W_{h,ij}$, $L_{ij} = L(X_i, X_j, \lambda)$ and $W_{h,ij} = h^{-p} W\left(\frac{Y_i - Y_j}{h}\right)$. Given that the last term on the right-hand side of Eq. (3.3) is unrelated to (λ, h), we therefore choose (λ, h) to minimize

$$CV(h, \lambda) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h,ij}^{(2)} - 2 \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} K_{h,ij}. \tag{3.4}$$

Let $(\hat{\lambda}, \hat{h})$ denote the above cross-validated choices of (λ, h). The following assumptions are used to derive the rates of convergence of $(\hat{\lambda}, \hat{h})$ to (λ_o, h_o), and $\hat{f}(z)$ to $f(z)$, respectively.

Assumption (B1). (i) $\{Z_i\}_{i=1}^{n} = \{X_i, Y_i\}_{i=1}^{n}$ is i.i.d. where $Z = (X, Y)$, (ii) let $f(y|x)$ denote the conditional density function of Y given $X = x$. $f(\cdot|x)$ is four times continuously differentiable on the support of Y for all $x \in \mathscr{D}$. $f(y|x)$ and its derivatives are all bounded and continuous on the support of Y for all $x \in \mathscr{D}$, (iv) there exists $x, x' \in \mathscr{D}$ such that $f(x, y) \neq f(x', y)$ for y on a set with positive Lebesgue measure.

Assumption (B2). (i) The kernel function $w(\cdot)$ is nonnegative, 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 $H_n = [\underline{h}, \bar{h}]$, where $\underline{h} \geqslant C^{-1} n^{\delta - 1/p}$, $\bar{h} \leqslant C n^{-\delta}$ for some $C, \delta > 0$.

The conditions given in Assumption (B2)(ii) are similar to those used in [16], and are equivalent to $n^{1-\delta_p}\underline{h}^p \geqslant C^{-1}$ and $n^\delta \bar{h} \leqslant C$. Thus, by choosing a very small value of δ, these conditions are virtually equivalent to the standard assumptions that $h \to 0$ and $nh^p \to \infty$ as $n \to \infty$.

In Appendix B we show that the leading term of $CV(h, \lambda)$ is $CV_L(h, \lambda)$ given by

$$CV_L(\lambda, h) = B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + B_4 (nh^p)^{-1}, \tag{3.5}$$

where the B_j's are constants ($j = 1, \ldots, 4$). Let (h_o, λ_o) denote the values of (h, λ) that minimize $CV_L(h, \lambda)$. Then some simple calculus shows that

$$h_o = c_1 n^{-1/(p+4)} \quad \text{and} \quad \lambda_o = c_2 n^{-2/(p+4)}, \tag{3.6}$$

where c_1 and c_2 are constants defined in Appendix B.

The next theorem establishes the rate of convergence of $(\hat{\lambda}, \hat{h})$ to (λ_o, h_o) along with the asymptotic normal distribution of $\hat{f}(z)$.

Theorem 3.1. *Under assumptions* (B1) *and* (B2), *and if* $f(z) \geqslant \delta > 0$, *we have*

(i) $(\hat{h} - h_o)/h_o = O_p(n^{-\alpha/(4+p)})$ *and* $\hat{\lambda} - \lambda_o = O_p(n^{-\beta/(4+p)})$, *where* $\alpha = \min\{2, p/2\}$ *and* $\beta = \min\{1/2, 4/(4+p)\}$.

(ii) $\sqrt{nh^p}(\hat{f}(z) - f(z) - \hat{h}^2 \mathscr{B}_1(z) - \hat{\lambda}\mathscr{B}_2(z)) \to N(0, V(z))$ *in distribution, where* $\mathscr{B}_1(z) = (1/2)tr\{\nabla^2 f(z)\}[\int w(v)v^2 \, dv]$, $\mathscr{B}_2(z) = \sum_{x' \in \mathscr{D}, d_{x,x'}=1}[f(x', y) - f(x, y)]$, *and* $V(z) = f(z)[\int W^2(v) \, dv]$.

The proof of Theorem 3.1 is given in Appendix B. Comparing Theorems 3.1 and 2.1, we see that, for the mixed variable case, the convergence rate of $\hat{\lambda}$ is much slower than that of $\tilde{\lambda}$ for the discrete variable only case.

Let $\bar{f}(z)$ denote the density estimator with $\lambda = 0$ and $h = cn^{-1/(4+p)}$ ($c > 0$ is a constant). Then $\bar{f}(z)$ is the conventional frequency kernel estimator for $f(z)$. It is well established that $\sqrt{nh^p}(\bar{f}(z) - f(z) - h^2 \mathscr{B}_1(z)) \to N(0, V(z))$ in distribution. We see that our cross-validation-based estimator has the same asymptotic variance as that of the conventional estimator. However, as we show in Section 4, our cross-validation based estimator can substantially outperform the conventional frequency-based estimator in finite-sample settings.

The general multivariate discrete variable case: We have only considered the case whereby the discrete variable X is a multivariate binary variable. We now discuss the general multivariate discrete variable case. Let x_t be the tth component of x and suppose that x_t can assume $c_t \geqslant 2$ different values ($t = 1, \ldots, k$). Following Aitchison and Aitken [4], we define the kernel weight function $l(X_{i,t}, x_t, \lambda) = 1 - \lambda$ if $X_{i,t} = x_t$ and $l(X_{i,t}, x_t, \lambda) = \lambda/(c_t - 1)$ if $X_{i,t} \neq x_t$. In this case the product kernel becomes

$$L(X_i, x, \lambda) = \prod_{t=1}^k l(X_{i,t}, x_t, \lambda) = c_0(1 - \lambda)^{k - d_{ix}} \lambda^{d_{ix}}, \tag{3.7}$$

where $c_0 = \prod_{t=1}^{k} [1/(c_t - 1)]$ is a constant, and d_{ix} is the same as that defined in Eq. (2.1). Comparing Eq. (3.7) with (2.1) we see that, for the general multivariate discrete variable case, the only difference is that the kernel function has an extra multiplicative constant c_0. By inspection of the proofs of Theorems 2.1 and 3.1, we know that this extra multiplicative constant does not affect any of the results in the appendices. Therefore, the conclusions of Theorems 2.1 and 3.1 remain unchanged when one has a general multivariate discrete variable, provided one uses the kernel function defined in (3.7) in such instances.

4. Monte Carlo simulation results

For the simulations that follow, we draw 1000 replications from each DGP. For each of the 1000 Monte Carlo replications, smoothing parameters are selected via cross-validation, and then we estimate the joint distribution. We use the second-order Gaussian kernel for the continuous variable, while the kernel for the discrete variable is that defined in Eq. (2.1). The cross-validated choices of (λ, h) are based upon minimizing the cross-validation function with respect to λ and h using a conjugate gradient search algorithm. We also compute the conventional frequency estimator for comparison purposes whereby univariate cross-validation is conducted for the continuous variable using only those observations lying in each cell. For each replication we compute the MSE defined by $n^{-1} \sum_{i=1}^{n} (\hat{f}(X_i, Y_i) - f(X_i, Y_i))^2$ where $f(X_i, Y_i)$ is the true DGP and $\hat{f}(X_i, Y_i)$ is its kernel estimate. Median values and the 5th and 95th percentiles of the MSE generated from the 1000 replications are summarized in tabular form.

4.1. Finite-sample performance: independent identical distributions

We first assess the potential finite-sample efficiency gains exhibited by our method relative to the conventional frequency estimator. For the frequency method, $\lambda = 0$, and the smoothing parameter h is selected via the method of least-squares cross-validation method (using the data in each discrete cells). We begin with a case for which the density for the continuous variable is the same regardless of the realization taken on by the binary variable, hence $Y \sim N(\mu, \sigma^2)$ independent of X. We consider two cases, one for which $Pr[X = 1] = 0.7$, and one for which $Pr[X = 1] = 0.9$. Results are summarized in Table 1, and columns with headings 'LS_{freq}' contain results for the conventional frequency estimator, while the 'LS' denotes the proposed least-squares cross-validation method.

From Table 1 we see that, as expected, our cross-validation method performs much better than the conventional frequency estimator. The median MSE of the proposed method is only 1/2 to 1/3 of the median MSE of the conventional frequency-based method.

Table 1
Median MSE values

n	$Pr[X=1]=0.7$		$Pr[X=1]=0.9$	
	$MSE(LS)$	$MSE(LS_{freq})$	$MSE(LS)$	$MSE(LS_{freq})$
50	8.23e-04	1.67e-03	7.39e-04	2.20e-03
	(5.23e-04,1.33e-03)	(9.52e-04,2.97e-03)	(4.12e-04,1.20e-03)	(1.13e-03,3.74e-03)
100	4.61e-04	9.79e-04	4.73e-04	1.40e-03
	(3.05e-04,7.13e-04)	(6.05e-04,1.63e-03)	(3.02e-04,7.73e-04)	(8.20e-04,2.44e-03)
200	2.57e-04	5.40e-04	2.61e-04	7.64e-04
	(1.71e-04,3.82e-04)	(3.35e-04,9.06e-04)	(1.66e-04,4.24e-04)	(4.52e-04,1.34e-03)

The 5th and 95th percentiles appear in parentheses. $Y \sim$ Gaussian (μ, σ^2).

Table 2
Median values of h and MSE

n	$Pr[X=1]=0.7$		$Pr[X=1]=0.9$	
	$MSE(LS)$	$MSE(LS_{freq})$	$MSE(LS)$	$MSE(LS_{freq})$
50	7.50e-04	1.53e-03	7.77e-04	2.33e-03
	(4.60e-04,1.15e-03)	(7.73e-04,2.57e-03)	(4.51e-04,1.28e-03)	(1.23e-03,3.94e-03)
100	4.27e-04	8.73e-04	4.48e-04	1.31e-03
	(2.70e-04,6.49e-04)	(5.15e-04,1.47e-03)	(2.48e-04,7.31e-04)	(6.75e-04,2.26e-03)
200	2.47e-04	4.89e-04	2.72e-04	8.14e-04
	(1.60e-04,3.69e-04)	(2.92e-04,8.29e-04)	(1.71e-04,4.23e-04)	(4.66e-04,1.33e-03)

The 5th and 95th percentiles appear in parentheses. $Y \sim$ Gaussian $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

4.2. Finite-sample performance: shifted conditional densities

Next, we consider the case where the density for the continuous variable is shifted both in mean and variance conditional on the values assumed by the binary variable. $Y \sim N(\mu_1, \sigma_1^2)$ when $X = 0$ and $Y \sim N(\mu_2, \sigma_2^2)$ when $X = 1$ with $(\mu_1, \mu_2) = (-1, 1)$ and $(\sigma_1, \sigma_2) = (1, 2)$. We consider two cases, one for which $Pr[X = 1] = 0.7$ and the other for which $Pr[X = 1] = 0.9$. Results are summarized in Table 2.

Examining Table 2 we again observe that the finite-sample efficiency gains associated with the proposed method relative to the conventional frequency estimator are substantial. The median MSE of the proposed method is around 1/2–1/3 of the median MSE of the conventional frequency-based method.

4.3. Finite-sample performance: least-squares versus likelihood cross-validation with fat-tailed distributions

It is known that likelihood cross-validation can breakdown with commonly used kernels when one or more of the continuous data types are drawn from fat-tailed distributions, a situation frequently encountered when dealing with economic and

financial data. In order to verify that the proposed method does not suffer from this defect, we consider a continuous variable Y drawn from the Cauchy distribution and a discrete variable X that is independent of Y having $Pr[X = 1] = 0.7$. Results are summarized in Table 3. Columns labeled 'ML' correspond to likelihood cross-validation and those labeled 'LS' again are those for the proposed least-squares cross-validation method.

Based on Table 3 we observe that, when the continuous variable is drawn from the Cauchy distribution, the likelihood cross-validation (ML-CV) method breaks down as expected while the proposed method does not. The ML-CV choice of h for the Cauchy example is an order of magnitude larger than that given by the proposed least-squares cross-validation (LS-CV) method, while the median MSE of the ML-CV estimator does not decrease as n increases which illustrates the inconsistency of the ML-CV estimator for fat-tailed distributions. To further demonstrate the extent of the over-smoothing exhibited by ML-CV when the underlying DGP is Cauchy, we evaluate the estimated density on a grid with support $[-3.5, 3.5]$ and plot the median values from the Monte Carlo simulation in Fig. 1. From Fig. 1 we see that the ML-CV method completely breaks down, giving a flat estimated density curve, while the proposed method is well behaved.

4.4. Discussion

The three simulation exercises described above illustrate how the proposed method can be of value in common situations where interest lies in estimating a joint

Table 3
Median values of h and MSE

Likelihood cross-validation ($Y \sim$ Cauchy, $X \sim$ Binomial)			
n	$h(ML)$	$MSE(ML)$	$MSE(ML_{freq})$
50	5.58	3.50e-03	6.98e-03
	(2.73,13.80)	(2.18e-03,4.92e-03)	(4.02e-03,9.98e-03)
100	8.79	4.22e-03	8.68e-03
	(3.90,20.90)	(2.81e-03,5.38e-03)	(5.64e-03,1.12e-02)
200	10.20	4.46e-03	9.75e-03
	(5.01,27.20)	(3.24e-03,5.61e-03)	(6.94e-03,1.18e-02)

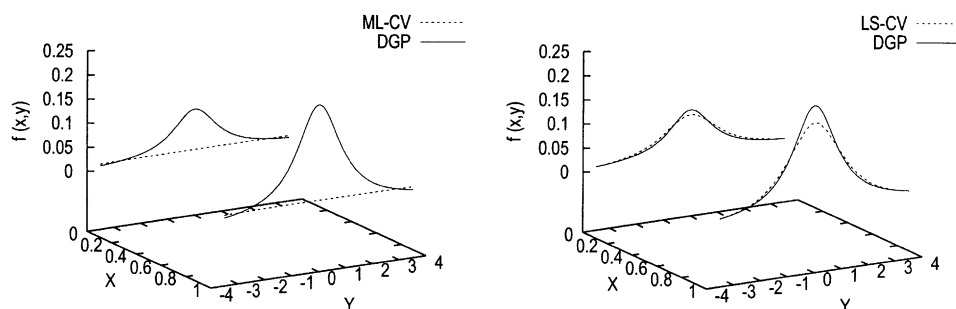| Least squares cross-validation ($Y \sim$ Cauchy, $X \sim$ Binomial)			
n	$h(LS)$	$MSE(LS)$	$MSE(LS_{freq})$
50	0.67	6.91e-04	1.21e-03
	(0.50,0.84)	(4.62e-04,1.02e-03)	(7.54e-04,1.91e-03)
100	0.57	4.08e-04	7.23e-04
	(0.44,0.67)	(2.71e-04,6.00e-04)	(4.67e-04,1.11e-03)
200	0.48	2.28e-04	3.92e-04
	(0.40,0.55)	(1.59e-04,3.34e-04)	(2.56e-04,6.09e-04)

The 5th and 95th percentiles appear in parentheses.

Fig. 1. Estimated joint PDF with a Cauchy continuous variable, $n = 200$. The figure on the left is that for likelihood cross-validation while the figure on the right is that for the proposed method. The solid curve represents the true Cauchy density function

distribution defined over a mix of continuous and binary data. The conventional frequency estimator is clearly less efficient in finite-sample applications. Also, we note that the proposed method does not suffer from the known limitations of likelihood cross-validation in the presence of 'fat-tailed' distributions which can be encountered when analyzing economic and financial data for instance. Note that we have only considered the simple case with one binary discrete variable and one continuous variable. With multivariate discrete data, the relative efficiency gains exhibited by the proposed method can be even more substantial.

5. An empirical application—modeling labor market participation

We now apply the proposed approach to modeling discrete choice, and we use Gerfin's [8] cross-section data set containing $n = 872$ records and seven variables used to model the labor market participation of married Swiss women. Gerfin [8] uses a Probit model along with three semiparametric specifications, and finds that the Probit specification cannot be rejected and that all 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 data set to see whether predictions given by the Probit and semiparametric specifications can be substantially improved upon (we do not include Gerfin's [8] semiparametric results here as they all yielded similar results). Data for this study can be found at qed.econ.queensu.ca/jae/1996-v11.3/gerfin/.

The variables used by the Gerfin [8] study are

1. LFP: Labor force participation dummy (0/1).
2. FOREIGN: Dummy if observation is not Swiss (0/1).
3. NYC: Number of young children (younger than 7) $(0, 1, 2, 3)$.
4. NOC: Number of older children $(0, 1, \ldots, 6)$.
5. EDUC: Years of formal education $(1, 2, \ldots, 21)$.
6. AGE: Age in years $(20, 21, \ldots, 62)$.
7. LNNLINC: Log of non-labor income $(7.1869 - 12.3757$ with 840 unique values).

Table 4
Confusion matrix and classification rates for the kernel and Probit models

	Kernel			Probit	
Act/Pred	0	1	Act/Pred	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)		55.4%

Act = actual sample realization, Pred = predicted outcome.

Let U denote variables 2–7. We compute the conditional probability of *LFP* given U defined as

$$\hat{f}(LFP|U) = \frac{\hat{f}(LFP, U)}{\hat{f}_1(U)}, \tag{3.8}$$

where $f_1(\cdot)$ is the marginal density function of U.

We treat the variables AGE and LNINC as continuous and the rest as categorical, and bandwidths are chosen via cross-validation using a conjugate gradient search algorithm.[1] Note that the use of a multivariate search algorithm naturally yields different smoothing parameters for each variable as discussed in Section 2. Using the cross-validated bandwidths, we then predict $LFP = 1$ if $\hat{f}(LFP = 1|U) > \hat{f}(LFP = 0|U)$, otherwise we predict $LFP = 0$.

We compare the predictions based upon our estimator with those from the Probit model used in [8], and the confusion matrices and classification rates for both approaches are summarized in Table 4 (a confusion matrix is one whose diagonal elements are correctly predicted outcomes and whose off-diagonal elements are incorrectly predicted outcomes). As can be seen from Table 4, the proposed method correctly predicts 74.1% of all observations while the Probit model correctly predicts 66.5%. We also report the correct classification rates for each value assumed by the categorical dependent variable. For example, $CCR(0) = 76.4\%$ means that, considering the subset of observations for which $LFP = 0$, we correctly predict 76.4% of them. To address potential concerns that these results might be an artifact of within-sample 'over-fitting', 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

---

[1] Some discrete variables take more than two different values, thus we use the kernel function defined in Eq. (3.7).

[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.

in Table 4 for a large number of different splits indicating that this is indeed a general improvement in predictive ability and not simply an artifact of over-fitting.

This application is simply intended to illustrate how the proposed method can be used to obtain superior predictions of categorical variables relative to predictions based upon commonly used parametric specifications such as the Probit model.

6. Possible extensions

There are numerous ways in which the results developed in this paper can be extended including (i) semiparametric estimation of a density function with mixed data, (ii) consistent model specification tests with mixed discrete and continuous regressors, including testing for a parametric or a semiparametric density functional form, and (iii) estimation of a joint density function with mixed discrete and continuous variables when the discrete variables contain ordered categorical data.

With ordered categorical data, it is known that boundary kernels [6], local polynomials [1,2], penalized likelihood [21], and local likelihood methods have better properties than standard kernel estimators as they are designed explicitly to counteract boundary bias associated with standard kernel estimators. It will be fruitful to extend the current results to the case of ordered categorical data. Specification tests (with mixed data types) based on a data-driven choice of smoothing parameters are expected to be significantly more powerful than existing tests based on frequency estimators as the former do not use sample splitting in finite-sample applications.

Recently, Racine and Li [19] have considered the problem of nonparametric estimation of regression functions with mixed discrete and continuous regressors and have established the asymptotic distribution of their proposed estimator. Yet another extension is to consider semiparametric regression models with mixed regressors, including partially linear models and additive models, along with specification tests for parametric/semiparametric regression functional forms. The authors are currently working on a number of related extensions having widespread potential application.

Appendix A

This appendix[3] contains the proof of Theorem 2.1. In Lemma A.0 we first show that $\tilde{\lambda} = o_p(1)$. Then Lemmas A.1–A.5 use the property that $\lambda = o(1)$ to obtain a λ power series expansion of $CV(\lambda)$, which is then used to prove Theorem 2.1.

Some Notation. We will use the summation indices i, j, and l to denote observations, whereby $\sum_i = \sum_{i=1}^n$, $\sum_i \sum_{i \neq j} = \sum_{i=1}^n \sum_{j=1, j \neq 1}^n$, $\sum \sum \sum_{i \neq j \neq l} = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \sum_{l=1, l \neq i, l \neq j}^n$. We use the summation indices x, x_1, and x_2 to denote the sum over the support of $x, x_1, x_2 \in \mathscr{D}$, i.e., $\sum_x = \sum_{x \in \mathscr{D}}$.

From Eq. (2.4) we get

$$CV(\lambda) = \frac{1}{n^2} \sum_i L_{ii}^{(2)} + \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} [L_{ij}^{(2)} - 2L_{ij}] - \frac{1}{n^2(n-1)} \sum_i \sum_{j \neq i} L_{ij}^{(2)}$$

$$\equiv I_{1n} + I_{2n} - I_{3n}, \tag{A.1}$$

where the definitions of I_{jn} $(j = 1, 2, 3)$ should be apparent.

Proof of Theorem 2.1. (i) Using Eq. (A.1) and by Lemmas A.2, A.4 and A.5, we have

$$CV(\lambda) = I_{1n} + I_{2n} - I_{3n}$$

$$= A_1 \lambda^2 - A_2 \lambda n^{-1} + o_p(\lambda^2 + n^{-1}\lambda) + (\text{terms unrelated to } \lambda), \tag{A.2}$$

where $A_2 = 2k = \tilde{A}_2$, while A_1 and $\tilde{A}_2$ are two constants defined in Lemmas A.2 and A.5, respectively. Minimizing Eq. (A.2) over λ leads to $\tilde{\lambda} = [A_2/(2A_1)]n^{-1} + o_p(n^{-1}) = O_p(n^{-1})$.

(ii) Define $\bar{p}(x) = n^{-1} \sum_i \mathbf{1}(d_{ix} = 1) \equiv n^{-1} \sum_i \mathbf{1}(X_i = x)$, which is a frequency estimator of $p(x)$ (corresponding to $\lambda = 0$). It is well established that $\sqrt{n}(\bar{p}(x) - p(x)) \to N(0, (1 - p(x))p(x))$ in distribution. Now, using Eq. (A.3) (see below) and the fact that $\tilde{\lambda} = O_p(n^{-1})$, we have $\hat{p}(x) - \bar{p}(x) = O_p(\tilde{\lambda}) = O_p(n^{-1})$. Hence, we have

$$\sqrt{n}(\hat{p}(x) - p(x)) = \sqrt{n}(\bar{p}(x) - p(x)) + O_p(n^{-1/2}) \to N(0, (1 - p(x))p(x)) \text{ is distribution.} \quad \square$$

Below we prove some lemmas that are used to prove Theorem 2.1. We will write $B_n = D_n + (s.o.)$ to indicate that D_n is the leading term of B_n (D_n and B_n have the same order), and (s.o.) denotes terms having order strictly smaller than D_n.

Defining $\mathbf{1}_{d_{ix}=l} \equiv \mathbf{1}(d_{ix} = l)$, the discrete variable kernel $L(X_i, x, \lambda)$ defined in Eq. (2.1) can be written as a power series expansion in λ^l $(0 \leq l \leq k)$.

$$L(X_i, x, \lambda) = \mathbf{1}_{d_{ix}=0}(1 - \lambda)^k + \mathbf{1}_{d_{ix}=1}(1 - \lambda)^{k-1}\lambda + \mathbf{1}_{d_{ix}=2}(1 - \lambda)^{k-2}\lambda^2 + O_p(\lambda^3).$$

$$\tag{A.3}$$

---

[3] A longer version of Appendices A and B containing more detailed proofs of the main results are available from the authors upon request.

In Lemmas A.1–A.5 we evaluate the orders of I_{ln} $(l = 1, 2, 3)$ defined in Eq. (A.1). I_{ln} contains terms with two and three summations. We will use the U-statistic H-decomposition together with the expansion found in Eq. (A.3) to obtain the leading order terms of I_{ln}.

Lemma A.0. $\tilde{\lambda} = o_p(1)$.

Proof. Note that $\sum_x p(x)\hat{p}(x) = E[\hat{p}(x)]$. We observe from Eq. (2.3) that $I_n(\lambda) = \sum_x [\hat{p}(x)]^2 - 2E[\hat{p}(x)] + E[p(X)]$. Now define $\hat{I}_n(\lambda) = \sum_x [\hat{p}(x)]^2 - 2n^{-1} \sum_i \hat{p}_{-i}(X_i) + E[p(X)]$. Obviously $n^{-1} \sum_i \hat{p}_{-i}(X_i) - E[\hat{p}(X)] = o_p(1)$, which implies that (a): $\hat{I}_n(\lambda) = I_n(\lambda) + o_p(1)$.

Next, $0 \leqslant \hat{I}_n(\tilde{\lambda}) \leqslant \hat{I}_n(0) = o_p(1)$ because $\lambda = 0$ corresponds to the usual frequency estimator and it is well established that $\hat{I}_n(0) = o_p(1)$. Thus, we have (b): $\hat{I}_n(\tilde{\lambda}) = o_p(1)$.

Conditions (a) and (b) leads to (c): $I_n(\tilde{\lambda}) = o_p(1)$.

Finally, for $\lambda \neq o(1)$, using the H-decomposition of U-statistic theory, it is easy to show that $I_n(\lambda) = E(I_n(\lambda)) + o_p(1) = \sum_{l=1}^{2k} C_l \lambda^l + o_p(1) \neq o_p(1)$ because $C_l \neq 0$ for some $0 \leqslant l \leqslant 2k$.

Hence, we have (d): $I_n(\lambda) = O_p(1) \neq o_p(1)$ for $\lambda \neq o(1)$. Conditions (c) and (d) imply that $\tilde{\lambda} = o_p(1)$.

Note that Lemma A.0 implies the consistency of $\hat{p}(x)$, i.e., $\hat{p}(x) - p(x) = o_p(1)$.

Lemma A.1. $I_{1n}(\lambda) = -2k\lambda n^{-1} + O_p(n^{-3/2}\lambda + n^{-1}\lambda^2) + (terms\ unrelated\ to\ \lambda)$.

Proof. By Eq. (A.1) we know that $I_{1n} \overset{\text{def}}{=} n^{-2} \sum_i L_{ii}^{(2)} \equiv n^{-2} \sum_i \sum_x L_{ix}^2$.

Using the expansion given in Eq. (A.3) we obtain a λ power expansion of $E[I_{1n}]$.

$$E[I_{1n}] = n^{-1} \sum_x E[L_{1x}^2] = n^{-1} \sum_x \{E[L_{1x}^2 \mathbf{1}(d_{1x} = 0)] + E[L_{1x}^2 \mathbf{1}(d_{1x} \geqslant 1)]\}$$

$$= n^{-1}(1 - \lambda)^{2k} \sum_x p(x) + O(n^{-1}\lambda^2) = n^{-1}(1 - 2k\lambda) + O(n^{-1}\lambda^2).$$

Similarly, we have (again using the expansion found in {Eq. (A.3)),

$$I_{1n} - E[I_{1n}] = n^{-1} \left\{ \sum_x n^{-1} \sum_i L_{ix}^2 \mathbf{1}(d_{ix} = 0) - \sum_x E[L_{ix}^2 \mathbf{1}(d_{ix} = 0)] \right\} + O_p(n^{-1}\lambda^2)$$

$$= (1 - \lambda)^{2k} n^{-1} \sum_x [\bar{p}(x) - p(x)] + O_p(n^{-1}\lambda^2)$$

$$= (1 - 2k\lambda)n^{-3/2}\{\mathscr{V}_n + O_p(n^{-1}\lambda^2),$$

where $\bar{p}(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_i = x)$ (because $\mathbf{1}(d_{ix} = 0) = \mathbf{1}(X_i = x)$) is a frequency estimator of $p(x)$ and $\mathcal{V}_n = n^{1/2} \sum_{x \in \mathcal{D}} [\bar{p}(x) - p(x)]$ is a zero mean $O_p(1)$ random variable.

Summarizing the above we have shown that

$$I_{1n} = E(I_{1n}) + [I_{1n} - E(I_{1n})] = n^{-1}(1 - 2k\lambda) + O_p(n^{-3/2}\lambda + n^{-1}\lambda^2)$$

$$+ \text{ terms unrelated to } \lambda.$$

Lemma A.2. *Define* $H_n(X_i, X_j) = L_{ij}^{(2)} - 2L_{ij}$.

Then $E[H_n(X_i, X_j)] = A_1\lambda^2 + O(\lambda^3) + (\text{terms unrelated to } \lambda)$, *where* $A_1 = k^2 E[p(X)] - 2k E[p_1(X)] + E[(p_1(X))^2/p(X)], p_1(x) = \sum_{\{x' \in \mathcal{D}, d_{x',x}=1\}} p(x')$.

Proof. $E[H_n(X_i, X_j)] = E[L_{ij}^{(2)}] - 2E[L_{ij}]$. We compute $E[L_{ij}^{(2)}]$ and $E[L_{ij}]$ separately.

In the proof below we will use Eq. (A.3) frequently. Since the proof is relatively tedious and lengthy, we will often incorporate the indicator function restriction in the summation index, for example we will write $\sum_{x_1 \in \mathcal{D}} \mathbf{1}(d_{x_1,x} = 1)p_1(x_1) = \sum_{x_1,d_{x_1,x}=1} p(x_1)$ to save space.

We consider $E[L_{ij}]$ first. Define $p_s(x) = \sum_{\{x',d_{x',x}=s\}} p(x')(s = 1, 2)$. Using an expansion of Eq. (A.3), we get (note that $L_{ij} = L(X_i, X_j, \lambda)$, and $L_{x_1,x_2} = L(x_1, x_2, \lambda)$)

$$E[L_{ij}] = E[L(X_i, X_j, \lambda)] = \sum_{x_1} \sum_{x_2} p(x_1)p(x_2)L_{x_1,x_2}$$

$$= (1 - \lambda)^k \sum_{x_1} p(x_1) \sum_{\{x_2,d_{x_1,x_2}=0\}} p(x_2) + \lambda(1 - \lambda)^{k-1} \sum_{x_1} p(x_1)$$

$$\times \sum_{\{x_2,d_{x_1,x_2}=1\}} p(x_2) + O(\lambda^3)$$

$$+ \lambda^2(1 - \lambda)^{k-2} \sum_{x_1} p(x_1) \sum_{\{x_2,d_{x_1,x_2}=2\}} p(x_2) + O(\lambda^3)$$

$$= (1 - \lambda)^k \sum_{x_1} [p(x_1)]^2 + \lambda(1 - \lambda)^{k-1} \sum_{x_1} p(x_1)p_1(x_1) + \lambda^2(1 - \lambda)^{k-2}$$

$$\times \sum_{x_1} p(x_1)p_2(x_1) + O(\lambda^3)$$

$$= E[p(X)] + \lambda\{E[p_1(X)] - kE[p(X)]\} + \lambda^2\{E[p_2(X)] - (k - 1)E[p_1(X)]$$

$$+ [k(k - 1)/2]E[p(X)]\}.$$

Next, using the expansion contained in Eq. (A.3) we have

$$
\begin{aligned}
E[L_{ij}^{(2)}] &= \sum_x E[L_{X_i,x} L_{X_j,x}] = \sum_x \sum_{x_1} \sum_{x_2} p(x_1) p(x_2) L_{x_1,x} L_{x_2,x} \\
&= (1-\lambda)^{2k} \sum_x \sum_{\{x_1, d_{x_1,x}=0\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=0\}} p(x_2) \\
&\quad + \lambda(1-\lambda)^{2k-1} \sum_x \Bigg\{ \sum_{\{x_1, d_{x_1,x}=0\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=1\}} p(x_2) \\
&\quad + \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=0\}} p(x_2) \Bigg\} \\
&\quad + \lambda^2(1-\lambda)^{2(k-1)} \sum_x \Bigg\{ \sum_{\{x_1, d_{x_1,x}=2\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=0\}} p(x_2) \\
&\quad + \sum_{\{x_1, d_{x_1,x}=0\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=2\}} p(x_2) \\
&\quad + \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=1\}} p(x_2) \Bigg\} + O(\lambda^3) \\
&= E[p(X)] + 2\lambda\{E[p_1(X)] - kE[p(X)]\} \\
&\quad + \lambda^2\{2E[p_2(X)] + E[(p_1(X))^2/p(X)] - 2(2k-1)E[p_1(X)] \\
&\quad + k(2k-1)E[p(X)]\} + O(\lambda^3).
\end{aligned}
$$

Summarizing the above results, we get

$$
\begin{aligned}
E[H_n(X_1, X_2)] &= E[L_{ij}^{(2)}] - 2E[L_{ij}] \\
&= -E[p(X)] + \lambda^2\{k^2 E[p(X)] - 2kE[p_1(X)] + E[(p_1(X))^2/p(X)]\} + O(\lambda^3) \\
&\equiv A_1 \lambda^2 + O(\lambda^3) + (\text{terms unrelated to } \lambda). \quad \square
\end{aligned}
$$

Lemma A.3. *Define* $H_n(X_i, X_j) = L_{ij}^{(2)} - 2L_{ij}$. *Then*

$$
E[H_n(X_i, X_j)|X_i] = -p(X_i) + O(\lambda^2).
$$

Proof. $E[H_n(X_i, X_j)|X_i] = E[L_{ij}^{(2)}|X_i] - 2E[L_{ij}|X_i]$. We compute $E[L_{ij}|X_i]$ and $E[L_{ij}^{(2)}|X_i]$ separately below. Using the expansion given in Eq. (A.3) we have

$$
\begin{aligned}
E[L_{ij}|X_i] &= \sum_x p(x)L(X_i, x) \\
&= (1 - \lambda)^k \sum_{\{x, d_{X_i, x} = 0\}} p(x) + \lambda(1 - \lambda)^{k-1} \sum_{\{x_2, d_{X_i, x} = 1\}} p(x) + O(\lambda^2) \\
&= p(X_i) + \lambda[p_1(X_i) - kp(X_i)] + O(\lambda^2).
\end{aligned}
$$

Next,

$$
\begin{aligned}
E[L_{ij}^{(2)}|X_i] &= \sum_x E[L_{X_i, x} L_{X_j, x}|X_i] = \sum_x \sum_{x_1} p(x_1) L_{X_i, x} L_{x_1, x} \\
&= (1 - \lambda)^{2k} \sum_{\{x, d_{X_i, x} = 0\}} \sum_{\{x, d_{X_i, x_1} = 0\}} p(x_1) \\
&\quad + \lambda(1 - \lambda)^{2k-1} \left\{ \sum_{\{x, d_{X_i, x} = 0\}} \sum_{\{x_1, d_{X_i, x_1} = 1\}} p(x_1) \right. \\
&\quad \left. + \sum_{\{x, d_{X_i, x} = 1\}} \sum_{\{x_1, d_{X_i, x_1} = 0\}} p(x_1) \right\} + O(\lambda^2) \\
&= p(X_i) + 2\lambda[p_1(X_i) - kp(X_i)] + O(\lambda^2).
\end{aligned}
$$

Hence, we have

$$
E[H_n(X_i, X_j)|X_i] = E[L_{ij}^{(2)}|X_i] - 2E[L_{ij}|X_i] = -p(X_i) + O(\lambda^2).
$$

Note that the terms which are linear in λ cancel out in $E[L_{ij}^{(2)}|X_i]$ and $2E[L_{ij}|X_i]$. $\square$

Lemma A.4. $I_{2n}(\lambda) = A_1\lambda^2 + o_p(\lambda n^{-1} + \lambda^2) + (\textit{terms unrelated to } \lambda)$.

Proof. By Lemmas A.2, A.3 and the H-decomposition, we have

$$
\begin{aligned}
I_{2n} &= n^{-2} \sum_i \sum_{j \neq i} H_n(X_i, X_j) \\
&= E[H_n(X_i, X_j)] + 2n^{-1} \sum_i \{E[H_n(X_i, X_j)|X_i] - E[H_n(X_i, X_j)]\} + (\textit{s.o.}) \\
&= A_1\lambda^2 + o_p(\lambda n^{-1} + \lambda^2) + (\text{terms unrelated to } \lambda). \quad \square
\end{aligned}
$$

Lemma A.5. $I_{3n} = \tilde{A}_2\lambda n^{-1} + O_p(\lambda^2 n^{-1}) + (\textit{terms unrelated to } \lambda)$, where $\tilde{A}_2 = 2\{E[p_1(X)] - kE[p(X)]\}$.

Proof. Define $\mathscr{A}_n = (n(n-1))^{-1} \sum_i \sum_{j \neq i} L_{ij}^{(2)}$. Then $\mathscr{A}_n$ is a second order U-statistic. $E(\mathscr{A}_n) = E[L_{ij}^{(2)}] = E[p(X)] + \tilde{A}_2\lambda + O(\lambda^2)$ is proved in the proof of Lemma A.2. By the U-statistic H-decomposition, $\mathscr{A}_n = \tilde{A}_2\lambda + O_p(\lambda^2) + $ (terms unrelated to λ). Thus, $I_{3n} = n^{-1}\mathscr{A}_n = \tilde{A}_2\lambda n^{-1} + O_p(\lambda^2 n^{-1}) + $ (terms unrelated to λ). $\quad\square$

Appendix B

Note that $\hat{h} = o(1)$ by Assumption (B2)(ii). Along lines similar to the proof of Lemma A.0, one can show that $\hat{\lambda} = o_p(1)$. $\hat{h} \in [\underline{h}, \bar{h}]$ and $\hat{\lambda} = o_p(1)$ imply the consistency of $\hat{f}(x,y)$. In Lemmas B.1–B.4 below we use $h = o(1)$ and $\lambda = o(1)$ to obtain expansions of $CV(h,\lambda)$.

From Eq. (3.4) we get

$$CV(h,\lambda) = \frac{1}{n^2}\sum_i K_{h,ii}^{(2)} + \frac{1}{n(n-1)}\sum_i \sum_{j \neq i}[K_{h,ij}^{(2)} - 2K_{h,ij}] - \frac{1}{n^2(n-1)}\sum_i \sum_{j \neq i} K_{h,ij}^{(2)}$$

$$\equiv J_{1n} + J_{2n} - J_{3n}, \tag{B.1}$$

where the definitions of J_{ln} ($l = 1, 2, 3$) should be apparent.

Proof of Theorem 3.1. (i) By Lemmas B.1 and B.5, we have (ignoring the terms unrelated to (h,λ))

$$CV(h,\lambda) = J_{1n} + J_{2n} - J_{3n}$$

$$= B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + B_4(nh^p)^{-1} + O_p((h^2 + \lambda)^3 + \lambda(nh^p)^{-1} + (nh^{p/2})^{-1})$$

$$\equiv CV_L(h,\lambda) + O_p((h^2 + \lambda)^3 + \lambda(nh^p)^{-1} + (nh^{p/2})^{-1}), \tag{B.2}$$

where $CV_L(h,\lambda) = B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + B_4(nh^p)^{-1}$ is the leading term of $CV(.,.)$. Letting (h_o, λ_o) denote the values of (h,λ) that minimize $CV_L(h,\lambda)$, simple calculus shows that

$$h_o = c_1 n^{-1/(4+p)} \quad \text{and} \quad \lambda_o = c_2 n^{-2/(4+p)}, \tag{B.3}$$

where $c_1 = \{pB_4/(4[B_1 - B_2^2/(4B_3)])\}^{1/(4+p)}$ and $c_2 = c_1^2 B_2/(2B_3)$. Obviously $(\hat{h}, \hat{\lambda})$ will converge to (h_o, λ_o). To obtain the rates of $(\hat{h} - h_o)/h_o$ and $\hat{\lambda} - \lambda_o$, we need to consider the higher order terms in the expansion of $CV(h,\lambda)$. By inspection of the proofs of Lemmas B.1–B.5, we know that

$$CV(h,\lambda) = CV_L(h,\lambda) + C_1 h^6 + C_2 h^4 \lambda + C_3 h^2 \lambda^2 + C_4 \lambda^3 + C_5 \lambda(nh^p)^{-1}$$

$$+ \mathscr{V}_n(nh^{p/2})^{-1} + (s.o.), \tag{B.4}$$

where C_j's are some constants ($j = 1, \ldots, 5$) and $\mathscr{V}_n$ is a zero mean $O_p(1)$ random variable ($\mathscr{V}_n$ is a degenerate U-statistic—see Lemma B.4's proof for further explanation). We need to consider two cases: (i) $p \leqslant 3$ and (ii) $p \geqslant 4$.

Case (i) $p \leqslant 3$, $(nh^{p/2})^{-1}$ has an order larger than h^6 (because $(nh^p)^{-1} = O(h^4)$). Following exactly the same arguments as in [19],[4] using Eq. (B.4) one can show that

$$(\hat{h} - h_o)/h_o = O_p(h_o^{p/2}) \quad \text{and} \quad \lambda - \lambda_o = O_p(n^{-1/2}). \tag{B.5}$$

For case (ii) of $p \geqslant 4$, h^6 has an order at least as large as $O\left((nh^{p/2})^{-1}\right)$, and again by following the same arguments as in [19], one can show that

$$(\hat{h} - h_o)/h_o = O_p(h_o^2) \quad \text{and} \quad \lambda - \lambda_o = O_p(h_o^4). \tag{B.6}$$

Summarizing Eqs. (B.5) and (B.6), and noting that $h_o = O(n^{-1/(4+p)})$, we have

$$(\hat{h} - h_o)/h_o = O_p(n^{-\alpha/(p+4)}) \quad \text{and} \quad \lambda - \lambda_o = O_p(n^{-\beta}), \tag{B.7}$$

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

(ii) Define $\tilde{\tilde{f}}(z)$ in the same manner as $\tilde{f}(z)$ but with $(\hat{h}, \hat{\lambda})$ being replaced by the nonstochastic smoothing parameters (h_o, λ_o). Then it is straightforward to show that

$$\sqrt{nh_o^p}(\tilde{\tilde{f}}(z) - h_o^2 \mathscr{B}_1(z) - \lambda_o \mathscr{B}_2(z)) \to N(0, V(z)) \quad \text{in distribution} \tag{B.8}$$

by Liapunov's central limit theorem. Using Eq. (B.7) and a standard Taylor expansion argument, it is easy to show that

$$\hat{f}(z) - \tilde{f}(z) = o_p((nh_o)^{-1/2}). \tag{B.9}$$

Eqs. (B.7), (B.8) and (B.9) imply that

$$\sqrt{nh^p}(\hat{f}(z) - \hat{h}^2 \mathscr{B}_1(z) - \hat{\lambda} \mathscr{B}_2(z)) \to N(0, V(z)) \quad \text{in distribution.} \quad \square \tag{B.10}$$

Below we present some lemmas that are used for proving Theorem 3.1. The idea of the proof is similar to that contained in Appendix A, but now our cross-validation function $CV(h, \lambda)$ is more involved as it depends on both λ and h. In the proofs below, we first use Eq. (A.3) to obtain an expansion of $CV(h, \lambda)$ in a power series of λ up to the order of λ^2 plus some $o_p(\lambda^2)$ terms. Then we apply the standard change-of-variable argument to the continuous variable to obtain an expansion of $CV(h, \lambda)$ in a power series of h^2, up to the order of h^4, plus some $o_p(h^4)$ terms.

Lemma B.1. $J_{1n}(\lambda, h) \equiv n^{-2} \sum_i K_{h,ii}^{(2)} = (nh^p)^{-1}[B_4 - B_5\lambda + O(\lambda^2)]$, where B_4 and B_5 are two positive constants.

---

[4] A PDF file of this paper is available from econweb.tamu.edu/li/papers/htm

Proof. First note that $W_{h,ii}^{(2)} \stackrel{\text{def}}{=} \int W_{h,iy}^2 \, dy = h^{-p} \int W^2(v) \, dv$. Hence,

$$
\begin{aligned}
J_{1n} &= n^{-2} \sum_i K_{h,ii}^{(2)} = n^{-2} \sum_i L_{ii}^{(2)} W_{h,ii}^{(2)} = \left[\int W^2(v) \, dv \right] h^{-p} \left[n^{-2} \sum_i L_{ii}^{(2)} \right] \\
&= \left[\int W^2(v) \, dv \right] (nh^p)^{-1} [1 - 2k\lambda + O(\lambda^2)] = (nh^p)^{-1} [B_4 - B_5 \lambda + O(\lambda^2)] \quad \text{by Lemma A.1,}
\end{aligned}
$$

where $B_4 = [\int W^2(v) \, dv] > 0$ and $B_5 = 2k[\int W^2(v) \, dv] > 0$. □

Lemma B.2. *Define* $\mathscr{H}_n(Z_i, Z_j) = K_{h,ij}^{(2)} - 2K_{h,ij}$.

Then $E[\mathscr{H}_n(Z_i, Z_j)] = B_0 + B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + o_p(\lambda^2 + \lambda h^2 + h^4)$,

where B_j $(j = 0, \ldots, 3)$ *are some constants with* $B_1 > 0$ *and* $B_3 > 0$.

Proof. $E[\mathscr{H}_n(Z_i, Z_j)] = E[K_{h,ij}^{(2)}] - 2E[K_{h,ij}]$. We compute $E[K_{h,ij}]$ and $E[K_{h,ij}^{(2)}]$ separately below. We will use $f(y|x)$ to denote the conditional probability density function of Y given $X = x$. Define $G_h(x_1, x_2) = \int W_h(y_1, y_2) f(y_1|x_1) f(y_2|x_2) \, dy_1 \, dy_2$, where $W_h(y_1, y_2) = h^{-p} W(\frac{y_1 - y_2}{h})$. We will first use Equation (A.3) to expand $E[K_{h,ij}]$ in terms of λ^l $(l = 0, 1, \ldots, k)$

$$
\begin{aligned}
E[K_{h,ij}] &= \sum_x \sum_{x_1} p(x) p(x_1) L_{x,x_1} \int W_h(y, y_1) f(y|x) f(y_1|x_1) \, dy \, dy_1 \\
&= \sum_x \sum_{x_1} p(x) p(x_1) L_{x,x_1} G_h(x, x_1) \\
&= (1 - \lambda)^k \sum_x [p(x)]^2 G_h(x, x) + \lambda(1 - \lambda)^{k-1} \sum_x p(x) \sum_{\{x_1, d_{x,x_1} = 1\}} p(x_1) G_h(x, x_1) \\
&\quad + \lambda^2 (1 - \lambda)^{k-2} \left\{ \sum_x p(x) \sum_{\{x_1, d_{x,x_1} = 2\}} p(x_1) G_h(x, x_1) \right\} + O(\lambda^3) \\
&= T_0 + \lambda(T_1 - kT_0) + \lambda^2 \{ T_2 - (k - 1)T_1 + [k(k - 1)/2] T_2 \} + O(\lambda^3),
\end{aligned}
$$

where

$$
T_0 = \sum_x [p(x)]^2 G_h(x, x), \quad T_1 = \sum_x p(x) \sum_{\{x_1, d_{x,x_1} = 1\}} p(x_1) G_h(x, x_1),
$$

$$
T_2 = \sum_x p(x) \sum_{\{x_1, d_{x,x_1} = 2\}} p(x_1) G_h(x, x_1). \tag{B.11}
$$

For ease of reference we summarize the above result in the following equation:

$$
E[K_{ij}] = T_0 + \lambda(T_1 - kT_0) + \lambda^2 \{ T_2 - (k - 1)T_1 + [k(k - 1)/2] T_2 \} + O(\lambda^3). \tag{B.12}
$$

Eq. (B.12) gives an expansion of $E[K_{ij}]$ in λ^s $(s = 0, 1, 2, 3)$. T_0, T_1 and T_2 in Eq. (B.12) can be expanded as power series in h because $G_h(x_1, x_2)$ depends on h.

From the definition of $G_h(x, x_1)$ and the fact that $W(\cdot)$ is a symmetric function, it is easy to see that it admits the following expansion (in terms of powers of h):

$$G_h(x, x_1) = G_0(x, x_1) + h^2 G_2(x, x_1) + h^4 G_4(x, x_1) + o_p(h^4), \tag{B.13}$$

where $G_0(x, x_1) = \int f(y|x) f(y|x_1) W(v) \, dv \, dy = \int f(y|x) f(y|x_1) \, dy$, $G_2(x, x_1) = (1/2) \int f(y|x) v' \nabla^2_{y'} f(y|x_1) v W(v) \, dv \, dy$, and $G_4(x, x_1)$ involves the fourth order derivatives of $f(y|x)$ with respect to y, and factors like $\int W(v) v_l^4 \, dv$ or $\int W(v) v_l^2 v_{l'}^2 \, dv$, where v_l is the lth component of $v \in R^p$ $(l = 1, \dots, p)$.

Eq. (B.13) gives an expansion of $G_h(x, x_1)$ in h^l $(l = 0, 2, 4)$. If one substitutes Eq. (B.13) into Eq. (B.11), and then substitutes Eq. (B.11) into Eq. (B.12), one can get a power series expansion in $(h^2)^l \lambda^s$ $(l, s = 0, 1, 2, \dots)$. Below we will conduct some similar calculations for $E(K^{(2)}_{h,ij})$.

Next, we consider $E(K^{(2)}_{h,ij})$. Define $G^{(2)}_h(x_1, x_2) = \int f(y_1|x_1) f(y_2|x_2) W^{(2)}_h(y_1, y_2) \, dy_1 \, dy_2$. We use Eq. (A.3) to obtain an expansion of $E[K^{(2)}_{h,ij}]$ in terms of λ^l $(l = 0, 1, \dots, k)$

$$
\begin{aligned}
E[K^{(2)}_{h,ij}] &= E[L^{(2)}_{ij} W^{(2)}_{h,ij}] = \sum_x E[L_{ix} L_{jx} W^{(2)}_{h,ij}] \\
&= \sum_x \sum_{x_1} \sum_{x_2} p(x_1) p(x_2) L_{x_1,x} L_{x_2,x} \int f(y_1|x_1) f(y_2|x_2) W^{(2)}_h(y_1, y_2) \, dy_1 \, dy_2 \\
&= \sum_x \sum_{x_1} \sum_{x_2} p(x_1) p(x_2) L_{x_1,x} L_{x_2,x} G^{(2)}_h(x_1, x_2) \\
&= T^{(2)}_0 + \lambda(T^{(2)}_1 - 2k T^{(2)}_0) + \lambda^2 \{T^{(2)}_2 - (2k-1) T^{(2)}_1 + k(2k-1) T^{(2)}_0\} + O(\lambda^3),
\end{aligned}
$$

where

$$T^{(2)}_0 = \sum_x [p(x)]^2 G^{(2)}_h(x, x), \quad T^{(2)}_1 = 2 \sum_x p(x) \sum_{\{x_1, d_{x_1, x} = 1\}} p(x_1) G^{(2)}_h(x, x_1),$$

$$
\begin{aligned}
T^{(2)}_2 = {} &2 \sum_x p(x) \sum_{\{x_1, d_{x_1, x} = 2\}} p(x_1) G^{(2)}_h(x, x_1) \\
&+ \sum_x \sum_{\{x_1, d_{x_1, x} = 1\}} p(x_1) \sum_{\{x_2, d_{x_2, x} = 1\}} p(x_2) G^{(2)}_h(x_1, x_2). \tag{B.14}
\end{aligned}
$$

We summarize the above result in the following equation:

$$
\begin{aligned}
E[K^{(2)}_{ij}] = {} &T^{(2)}_0 + \lambda(T^{(2)}_1 - 2k T^{(2)}_0) \\
&+ \lambda^2 \{T^{(2)}_2 - (2k-1) T^{(2)}_1 + k(2k-1) T^{(2)}_0\} + O(\lambda^3). \tag{B.15}
\end{aligned}
$$

From the definition of $G_h^{(2)}(x_1, x_2)$ and the fact that $W^{(2)}(\cdot)$ is a symmetric function, it is easy to see that it admits the following expansion (in terms of powers of h):

$$G_h^{(2)}(x_1, x_2) = G_0^{(2)}(x_1, x_2) + h^2 G_2^{(2)}(x_1, x_2) + h^4 G_4^{(2)}(x_1, x_2) + o_p(h^4), \quad \text{(B.16)}$$

where $G_0^{(2)}(x_1, x_2) = \int f(y|x_1) f(y|x_2) W^{(2)}(v) \, dv \, dy = \int f(y|x_1) f(y|x_2) \, dy$, $G_2^{(2)}(x_1, x_2)$ $= (1/2) \int f(y_1|x_1) v' \nabla_{y_1}^2 f(y_1|x_2) v W^{(2)}(v) \, dv \, dy_1$ and $G_4^{(2)}(x_1, x_2)$ involve the fourth order derivatives of $f(y|x)$ with respect to y, and factors like $\int W^{(2)}(v) v_l^4 \, dv$ or $\int W^{(2)}(v) v_l^2 v_{l'}^2 \, dv$, where v_l is the lth component of $v \in R^p$ $(l = 1, \ldots, p)$.

From the definition of $W^{(2)}(\cdot)$, it is easy to check that the following relationships hold:

$$\int W^{(2)}(v) \, dv = \int W(v) \, dv = 1, \quad \int W^{(2)}(v) vv' \, dv = 2 \int W(v) vv' \, dv,$$

$$\int W^{(2)}(v) v_l^4 \, dv > 2 \int W(v) v_l^4 \, dv \qquad l = 1, \ldots, p. \quad \text{(B.17)}$$

From Eqs. (B.13), (B.16) and (B.17), we immediately get

$$G_0^{(2)}(x_1, x_2) = G_0(x_1, x_2), G_2^{(2)}(x_1, x_2) = 2 G_2(x_1, x_2), G_4^{(2)}(x_1, x_2) > 2 G_4(x_1, x_2). \quad \text{(B.18)}$$

Below we will obtain an expansion of $E[\mathscr{H}_n(Z_i, Z_j)] = E[K_{ij}^{(2)}] - 2E[K_{ij}]$ in the powers of h and λ. We write $E[\mathscr{H}_n(Z_i, Z_j)] = H_0(h) + \lambda H_1(h) + \lambda^2 H_2(h) + O(\lambda^3)$, where $H_l(h)$ can be written as a power expansion of h and where the subscript l means the power expansion of λ^l $(l = 0, 1, 2)$. We will first obtain an expansion for $H_0(h)$, the component of $E[\mathscr{H}_n(Z_i, Z_j)] = E[K_{ij}^{(2)}] - 2E[K_{ij}]$ that is independent of λ. From Eqs. (B.11)–(B.18), we know that

$$H_0(h) = T_0^{(2)} - 2T_0 = \sum_x [p(x)]^2 [G_h^{(2)}(x, x) - 2G_h(x, x)]$$

$$= \sum_x [p(x)]^2 \{-G_0(x, x) + (0)h^2 + h^4 [G_4^{(2)}(x, x) - 2G_4(x, x)]\}$$

$$\equiv B_0 + B_1 h^4, \quad \text{(B.19)}$$

where $B_0 = -\sum_x [p(x)]^2 G_0(x, x)$, $B_1 = \sum_x [p(x)]^2 [G_4^{(2)}(x, x) - 2G_4(x, x)]$. We see that, due to cancellations between $T_0^{(2)}$ and $2T^0$, there is no h^2 term in the above expansion.

Next, we compute $H_1(h)$, the component of $E[\mathcal{H}_n(Z_i, Z_j)]$ that is linear in λ. From Eqs. (B.11)–(B.18) we have

$$
\begin{aligned}
H_1(h) &= T_1^{(2)} - 2kT_0^{(2)} - 2[T_1 - kT_0] = [T_1^{(2)} - 2T_1] + 2k[T_0 - T_0^{(2)}] \\
&= 2\sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1)[G_h^{(2)}(x, x_1) - G_h(x, x_1)] \\
&\quad + 2k \sum_x [p(x)]^2 [G_h(x, x) - G_h^{(2)}(x, x)] \\
&= -B_2 h^2 + O_p(h^4),
\end{aligned}
\tag{B.20}
$$

where $B_2 = 2\{k \sum_x [p(x)]^2 G_2(x, x) - \sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1)G_2(x, x_1)\}$. Due to some cancellations, there is no constant term in the above expansion.

Finally, we compute $H_2(h)$, the component of $E[\mathcal{H}_n(Z_i, Z_j)]$ that is linear in λ^2. Again from Eqs. (B.11)–(B.18), we obtain

$$
\begin{aligned}
H_2(h) &= [T_2^{(2)} - (2k-1)T_1^{(2)} - k(2k-1)T_0^{(2)}] - 2\{T_2 - (k-1)T_1 + [k(k-1)/2]T_0\} \\
&= \sum_x \left\{ \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=1\}} p(x_2)G_0^{(2)}(x_1, x_2) \right. \\
&\quad \left. - 2kp(x) \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1)G_0^{(2)}(x, x_1) \right\} \\
&\quad + \left[2k^2 \sum_x p(x) \sum_{\{x_1, d_{x_1,x}=2\}} p(x_1)G_0^{(2)}(x, x_2) \right] + O_p(h^2) \\
&\equiv B_3 + O_p(h^2),
\end{aligned}
\tag{B.21}
$$

where the definition of B_3 should be apparent. Note that B_3 is obtained by replacing $G_h(x, x_1)$ and $G_h^{(2)}(x, x_1)$ by $G_0(x, x_1)$ and $G_0^{(2)}(x, x_1)$ in T_j and $T_j^{(2)}$ ($j = 1, 2, 3$), respectively. Also, $G_0(x, x_1) = G_0^{(2)}(x, x_1)$ by Eq. (B.18) is used in computing B_3.

By Eqs. (B.19)–(B.21), we immediately obtain

$$
\begin{aligned}
E[\mathcal{H}_n(Z_i, Z_j)] &= E[K_{h,ij}^{(2)}] - 2E[K_{h,ij}] \\
&= B_0 + B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + o_p(h^4 + \lambda h^2 + \lambda^2). \quad \square
\end{aligned}
\tag{B.22}
$$

Lemma B.3. $E[\mathcal{H}_n(Z_i, Z_j)|Z_i] = p(X_i) + O_p(h^4 + \lambda h^2 + \lambda^2)$, where $\mathcal{H}_n(Z_i, Z_j) = K_{h,ij}^{(2)} - 2K_{h,ij}$ with $Z_i = (X_i, Y_i)$.

Proof. $E[\mathcal{H}_n(Z_i, Z_j)|Z_i] = E[K_{h,ij}^{(2)}|Z_i] - 2E[K_{h,ij}|Z_i]$. We consider $E[K_{h,ij}|Z_i]$ and $E[K_{h,ij}^{(2)}|Z_i]$ separately below. Define $M_h(x, Y_i) = \int W_h(Y_i, y)f(y|x)dy$. We will first use Eq. (A.3) to expand $E[K_{h,ij}|Z_i]$ in terms of λ^l ($l = 0, 1, ..., k$).

$$E[K_{h,ij}|Z_i] = \sum_x p(x)L(X_i, x)\int W_h(Y_i, y)f(y|x)dy \equiv \sum_x p(x)L(X_i, x)M_h(x, Y_i)$$

$$= (1 - \lambda)^k p(X_i)M_h(Z_i) + \lambda(1 - \lambda)^{k-1} \sum_{\{x, d_{X_i, x}=1\}} p(x)M_h(x, y_i) + O(\lambda^2)$$

$$\equiv C_0(Z_i) + \lambda C_1(Z_i) + O(\lambda^2),$$

where $C_0(Z_i) = p(X_i)M_h(Z_i)$ and $C_1(Z_i) = \sum_{\{x_1, d_{X_i, x}=1\}} p(x)M_h(x, Y_i) - kp(X_i)M_h(Z_i)$.

Summarizing the above result, we have

$$E[K_{ij}|Z_i] = C_0(Z_i) + \lambda C_1(Z_i) + O(\lambda^2). \tag{B.23}$$

Using the usual change-of-variable method, it is easy to see that $M_h(Z_i)$ admits the following expansion (an expansion in powers of h):

$$M_h(Z_i) = M_0(Z_i) + h^2 M_2(Z_i) + O_p(h^4) = 1 + h^2 M_2(Z_i) + O_p(h^4), \tag{B.24}$$

where $M_0(Z_i) = \int f(y|X_i)W(v)\,dv\,dy = 1$ and $M_2(Z_i) = (1/2)\int v'\nabla_y^2 f(y|X_i)vW(v)\,dv\,dy$.

Next, we consider $E[K_{h,ij}^{(2)}|Z_i]$. Define $M_h^{(2)}(x, Y_i) = \int W_h^{(2)}(Y_i, y)f(y|x)\,dy$. We use Eq. (A.3) to expand $E[K_{h,ij}^{(2)}|Z_i]$ in powers of λ^l ($l = 0, 1, ..., k$).

$$E[K_{h,ij}^{(2)}|Z_i] = \sum_x E[L_{ix}L_{jx}W_h^{(2)}(Y_i, Y_j)|Z_i]$$

$$= \sum_x \sum_{x_1} p(x_1)L_{X_i, x}L_{x_1, x}M_h^{(2)}(x_1, Y_i)$$

$$= (1 - \lambda)^{2k}p(X_i)M_h^{(2)}(X_i, Y_i) + \lambda(1 - \lambda)^{2k-1}$$

$$\times \left\{ \sum_{\{x, d_{X_i, x}=1\}} \sum_{\{x_1, d_{x_1, x}=0\}} p(x_1)M_h^{(2)}(x_1, Y_i) \right.$$

$$\left. + \sum_{\{x, d_{X_i, x}=0\}} \sum_{\{x_1, d_{x_1, x}=1\}} p(x_1)M_h^{(2)}(x_1, Y_i) \right\} + O(\lambda^2)$$

$$= C_0^{(2)}(Z_i) + \lambda C_1^{(2)}(Z_i) + O(\lambda^2),$$

where $C_0^{(2)}(Z_i) = p(X_i)M_h^{(2)}(Z_i)$ and $C_1^{(2)}(Z_i) = 2\{\sum_{\{x_1, d_{X_i, x}=1\}} p(x)M_h^{(2)}(x, Y_i) - kp(X_i)M_h^{(2)}(Z_i)\}$.

Summarizing the above result, we have

$$E[K_{ij}^{(2)}|Z_i] = C_0^{(2)}(Z_i) + \lambda C_1^{(2)}(Z_i) + O(\lambda^2). \tag{B.25}$$

It is easy to show that $M_h^{(2)}(Z_i)$ has the following expansion (a power expansion in h)

$$M_h^{(2)}(Z_i) = M_0^{(2)}(Z_i) + h^2 M^{(2)}(Z_i) + O_p(h^4) = 1 + h^2 M^{(2)}(Z_i) + O_p(h^4) \tag{B.26}$$

where $\quad M_0^{(2)}(Z_i) = \int f(y|X_i) W^{(2)}(v) \, dv \, dy = 1 \quad$ and $\quad M_2^{(2)}(Z_i) = (1/2) \int v' \nabla_y^2 \times f(y|X_i) v W^{(2)}(v) \, dv \, dy$.

By Eq. (B.17), we know that

$$M_2^{(2)}(Z_i) = 2M_2(Z_i). \tag{B.27}$$

Using Eqs. (B.23)–(B.27) we obtain

$$
\begin{aligned}
E[\mathscr{H}_n(Z_i, Z_j)|Z_i] &= E[K_{h,ij}^{(2)}|Z_i] - 2E[K_{h,ij}|X_i] \\
&= [C_0^{(2)}(Z_i) - 2C_0(Z_i)] + \lambda[C_1^{(2)}(Z_i) - 2C_1(Z_i)] + O(\lambda^2) \\
&= [p(X_i) + h^2(0) + O(h^4)] + 2\lambda[0 + h^2 kp(X_i) M_2(Z_i) + O_p(h^4)] + O(\lambda^2) \\
&= p(X_i) + O_p(h^4 + \lambda h^2 + \lambda^2). \quad \square
\end{aligned}
\tag{B.28}
$$

We observe that, due to some cancellations between $E[K_{h,ij}^{(2)}|Z_i]$ and $2E[K_{h,ij}|X_i]$, there are no h^2 and λ terms in the expansion given in Eq. (B.28).

Lemma B.4. $J_{2n}(\lambda, h) = B_0 + B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + O_p((h^2 + \lambda)^2 + (nh^{p/2})^{-1}) +$ *terms unrelated to* (h, λ), *where* B_j ($j = 0, \dots, 3$) *are constants defined in Lemma* B.2.

Proof. $J_{2n} = n^{-2} \sum_i \sum_{j \neq i} H_n(Z_i, Z_j)$, where $H_n(Z_i, Z_j) = K_{h,ij}^{(2)} - 2K_{h,ij}$. By H-decomposition and the results of Lemmas B.2 and B.3, we have

$$
\begin{aligned}
J_{2n} =\ & n^{-2} \sum_i \sum_{j \neq i} \mathscr{H}_n(X_i, X_j) \\
=\ & E[\mathscr{H}_n(Z_i, Z_j)] + (2/n) \sum_i \{E[\mathscr{H}_n(Z_i, Z_j)|Z_i] - [\mathscr{H}_n(Z_i, Z_j)]\} \\
& + (2/n^2) \sum_i \sum_{j > i} \{[\mathscr{H}_n(Z_i, Z_j)] - [\mathscr{H}_n(Z_i, Z_j)|Z_i] - [\mathscr{H}_n(Z_i, Z_j)|Z_j] \\
& + E[\mathscr{H}_n(Z_i, Z_j)]\}
\end{aligned}
$$

$$= (B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2)$$

$$+ \left(-(2/n) \left\{ \sum_i [p(X_i) - E(p(X_i))] + O_p((h^2 + \lambda)^2) \right\} \right)$$

$$+ ((nh^{p/2})^{-1} \mathscr{V}_n)$$

$$= B_1 h^4 - B_2 \lambda h^2 + B_3 \lambda^2 + O_p((h^2 + \lambda)^3 + (nh^{p/2})^{-1})$$

$$+ \text{ terms unrelated to } (h, \lambda),$$

where $\mathscr{V}_n$ is a zero mean $O_p(1)$ random variable obtained from the last term in the H-decomposition, being a degenerate U-statistic. It is straightforward to show that the second moment of this degenerate U-statistic is of the order $(1/n^2) E[H^2(Z_i, Z_j)] = O((n^2 h^p)^{-1})$ (e.g., Theorem 1 of [12]). Therefore, this last term in the H-decomposition has an order of $O_p((nh^{p/2})^{-1})$. So we write it as $(nh^{p/2})^{-1} \mathscr{V}_n$, where $\mathscr{V}_n$ is a zero mean $O_p(1)$ random variable. $\square$

Lemma B.5. $J_{3n}(\lambda, h) = O_p(n^{-1}(h^2 + \lambda)) +$ *terms unrelated of* (h, λ).

Proof. First define $\mathscr{W}_n = (n(n-1))^{-1} \sum_i \sum_{j \neq i} K_{ij}^{(2)}$, which is a second order U-statistic. The proof of Lemma B.2 implies that $E[K_{ij}^{(2)}] = \tilde{B}_0 + \tilde{B}_1 h^2 + \tilde{B}_2 \lambda + (s.o.)$ for some constants $\tilde{B}_j$ $(j = 0, 1, 2)$. Hence, by the U-statistic H-decomposition we have $J_{3n} = n^{-1} \mathscr{W}_n = n^{-1}[E(\mathscr{W}_n) + (s.o.)] = n^{-1}[\tilde{B}_0 + \tilde{B}_1 h^2 + \tilde{B}_2 \lambda + (s.o.)] = O_p(n^{-1}(h^2 + \lambda)) +$ terms unrelated to (h, λ). $\square$

References

[1] M. Aerts, I. Augustyns, P. Janssen, Smoothing sparse multinomial data using local polynomial fitting, J. Nonparametric Statist. 8 (1997) 127–147.

[2] M. Aerts, I. Augustyns, P. Janssen, Local polynomial estimation of contingency table cell probabilities, Statistics 30 (1997) 127–148.

[3] I.A. Ahmad, P.B. Cerrito, Nonparametric estimation of joint discrete-continuous probability densities with applications, J. Statist. Plann. Inference 41 (1994) 349–364.

[4] J. Aitchison, C.G.G. Aitken, Multivariate binary discrimination by the kernel method, Biometrika 63 (1976) 413–420.

[5] A.W. Bowman, A note on consistency of the kernel method for the analysis of categorical data, Biometrika 67 (1980) 682–684.

[6] J. Dong, J.S. Simonoff, The construction and properties of boundary kernels for sparse multinomials, J. Comput. Graph. Statist. 3 (1994) 57–66.

[7] L. Fahrmeir, G. Tutz, Multivariate Statistical Modeling Based on Generalized Models, Springer, New York, 1994.

[8] M. Gerfin, Parametric and semiparametric estimation of the binary response model of labour market participation, J. Appl. Econometrics 11 (3) (1996) 321–340.

[9] B. Grund, Kernel estimators for cell probabilities, J. Multivariate Anal. 46 (1993) 283–308.

[10] B. Grund, P. Hall, On the performance of kernel estimators for high-dimensional sparse binary data, J. Multivariate Anal. 44 (1993) 321–344.

[11] P. Hall, On nonparametric multivariate binary discrimination, Biometrika 68 (1981) 287–294.

[12] P. Hall, Central limit theorem for integrated square error of multivariate nonparametric density estimators, J. Multivariate Anal. 14 (1984) 1–16.

[13] P. Hall, On Kullback–Leibler loss and density estimation, Ann. Statist. 15 (1987) 1491–1519.

[14] P. Hall, On the use of compactly supported densities in problems of discrimination, J. Multivariate Anal. 23 (1987) 131–158.

[15] P. Hall, M. Wand, On nonparametric discrimination using density differences, Biometrika 75 (1988) 541–547.

[16] W. Härdle, J.S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, Ann. Statist. 13 (1985) 1465–1481.

[17] J.D. Hart, Nonparametric Smoothing and Lack-of-fit Tests, Springer, New York, 1997.

[18] A. Izenman, Recent developments in nonparametric density estimation, J. Amer. Statist. Assoc. 413 (1991) 205–224.

[19] J. Racine, Q. Li, Nonparametric estimation of regression functions with both categorical and continuous data, J. Econometrics (2001) submitted.

[20] D. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, New York, 1992.

[21] J.S. Simonoff, A penalty function approach to smoothing to smoothing large sparse contingency tables, Ann. Statist. 11 (1983) 208–218.

[22] J.S. Simonoff, Smoothing Methods in Statistics, Springer, New York, 1996.

[23] G. Tutz, Consistency of cross-validatory choice of smoothing parameters for direct kernel estimates, Comput. Statist. Quart. 4 (1991) 295–314.