Nonparametric Estimation of Distributions with Categorical and Continuous Data *

Qi Li
Department of Economics, Texas A&M University
College Station, TX 77843 USA

Jeff Racine
Department of Economics, University of South Florida
Tampa, FL 33620 USA

*Running head: Estimation with Discrete and Continuous Data. The corresponding author: Qi Li, email: qiecon.tamu.edu, Tel: 979-845-7349, Fax: 979-847-8757.
The authors would like to thank Peter Hall for directing us to a number of useful and relevant papers. Li thanks the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanity Research Council of Canada, and the Bush Program in the Economics of Public Policy for Research Support. Racine would like to thank the Division of Sponsored Programs at the University of South Florida for their ongoing support.

ABSTRACT

We consider the problem of estimating an unknown joint distribution which is defined over mixed discrete and continuous variables. We use a nonparametric kernel approach with smoothing parameters obtained from the cross-validated minimization of the integrated squared error of the kernel estimator. We prove that our approach is consistent and derive its rate of convergence. Monte Carlo simulations demonstrate that the proposed approach does not suffer from known limitations of likelihood cross-validation method which breaks down with commonly used kernels when the continuous variables are drawn from fat-tailed distributions. The simulations also show that the new estimator performs much better than the conventional nonparametric frequency estimator. An empirical application demonstrates that the proposed method can yield superior out-of-sample predictions relative to commonly used parametric approaches.

Key words: Discrete and continuous variables, density estimation, nonparametric smoothing, cross-validation.

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 well known that one can use a frequency estimator to obtain consistent nonparametric estimates of the 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.

Aitchison & Aitken (1976) 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 (1980). 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 Hall (1987a,1987b)). For related work on issues surrounding the kernel estimation of distributions defined over discrete data the reader is referred to Hall (1981) and Hall and Wand (1988). In related papers, Grund (1993) and Grund and Hall (1993) 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 (1991), while more in-depth treatments of the subject can be found in Hart (1997), Fahrmeir and Tutz (1994), Scott (1992), and Simonoff (1996).

While there exist a number of theoretical papers on the properties of cross-validation methods with only discrete variables (e.g., Hall (1981), Grund (1993) and Grund and Hall (1993)), or with only continuous variables (Marron and Härdle (1985)), little attention has been paid to the more general and interesting case of mixed discrete and continuous variables. The exceptions are the papers by Ahmad and Cerrito (1994) and Tutz (1991) who have considered cross-validation for estimating regression functions and conditional density functions (with mixed variables), respectively. However, both Ahmad and Cerrito (1994) and Tutz (1991) only demonstrate the consistency of their cross-validation methods, while no *rate of convergence* is

established in either paper. Establishing of the rate of convergence is much more demanding than establishing only consistency. In this paper we aim to close the gap and we propose a consistent kernel method for estimating the joint PDF of mixed discrete and continuous data based upon *least-squares* cross-validation which minimizes the integrated squared error of the estimate. We provide the theoretical foundations for this method, obtain rates of convergence, and consider both simulations and applications of the proposed approach. To our knowledge, our work is the first to establish the *rate of convergence* with mixed discrete and continuous variables using cross-validation methods.

In Section 2 we consider the multivariate discrete variables case and propose estimating a joint PDF using least-squares cross-validation, and we establish the consistency and rate of convergence of the proposed estimator. Section 3 deals with the general mixed discrete and continuous variables case. Section 4 reports on simulations designed to assess the finite-sample performance of the estimator. Section 5 considers an empirical application which demonstrates that the proposed approach can yield superior (out-of-sample) predictions relative to commonly used parametric models of binary choice. Finally, Section 6 concludes and discusses possible extensions of the proposed approach.

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 denote $\{0,1\}^k$ by $\mathcal{D}$. 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 $d_{ix} = (X_i - x)'(X_i - x)$ equals the number of disagreement components between X_i and x. Note that d_{ix} takes values in $\{0, 1, 2, \ldots, k\}$.

It is straightforward to generalize the above to the case of a k-dimensional vector of λ. For simplicity of presentation, only scalar λ is treated here.

Some Notation: We will use the summation indices i, j, l to denote observations,

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

We estimate $p(x)$ by

$$\hat{p}(x) = \frac{1}{n} \sum_i L_{ix}, \tag{2.2}$$

where $L_{ix} = L(X_i, x, \lambda)$ is defined in Equation (2.1).

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

$$\begin{aligned} I_n &= \sum_{x\in\mathcal{D}} [\hat{p}(x) - p(x)]^2 = \sum_x [\hat{p}(x)]^2 - 2\sum_x \hat{p}(x)p(x) + \sum_x [p(x)]^2 \\ &\equiv I_{1n} - 2I_{2n} + \sum_x [p(x)]^2, \end{aligned} \tag{2.3}$$

where $I_{1n} = \sum_x [\hat{p}(x)]^2$ and $I_{2n} = \sum_x \hat{p}(x)p(x)$. Note that the last term on the right-hand-side of Equation (2.3) is unrelated to the choice of λ. Note that $I_{2n} = \sum_x \hat{p}(x)p(x) = E[\hat{p}(X)]$. Therefore, we estimate $I_{2n} = E[\hat{p}(X)]$ by

$$\hat{I}_{2n} = n^{-1} \sum_i \hat{p}(X_i) = n^{-2} \sum_i \sum_{j\neq i} L_{ij}, \tag{2.4}$$

where $L_{ij} = L(X_i, X_j, \lambda)$ and $\hat{p}(X_i) = n^{-1} \sum_{j=1, j\neq i}^n L_{ij}$ is the leave-one-out kernel estimator of $p(X_i)$. Using Equation (2.2), we have

$$\begin{aligned} I_{1n} &= \sum_x [\hat{p}(x)]^2 = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_x L_{ix} L_{jx} \\ &\equiv n^{-2} \sum_{i=1}^n \sum_{j=1}^n L_{ij}^{(2)}, \end{aligned} \tag{2.5}$$

where

$$L_{ij}^{(2)} = \sum_x L_{ix} L_{jx}. \tag{2.6}$$

Therefore, we choose λ to minimize the cross-validated integrated squared error of the

kernel estimator given by

$$\begin{aligned} CV(\lambda) &\stackrel{def}{=} I_{1n} - 2\hat{I}_{2n} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^{(2)} - 2n^{-2}\sum_{i}\sum_{j\neq i} L_{ij} \\ &= n^{-2}\sum_{i} L_{ii}^{(2)} + n^{-2}\sum_{i}\sum_{j\neq i}[L_{ij}^{(2)} - 2L_{ij}] \\ &\equiv J_{1n} + J_{2n}, \end{aligned} \tag{2.7}$$

where $J_{1n} = n^{-2}\sum_i L_{ii}^{(2)}$ and $J_{2n} = n^{-2}\sum_i \sum_{j\neq i}[L_{ij}^{(2)} - 2L_{ij}]$.

We use $\tilde{\lambda}$ to denote the cross-validated choice of λ. The following assumption is used to derive the rate at which $\tilde{\lambda}$ converges to zero as well as the rate of convergence of $\hat{p}(x)$ to $p(x)$.

Assumption (A): (i) X_i is independent and identically distributed (i.i.d.) as X, and $min_{\{x\in\mathcal{D}\}}p(x) \geq \delta$ for some $\delta > 0$.

Theorem 2.1. *Under assumption (A), we have*
(i) $\tilde{\lambda} = O_p(n^{-1})$,
(ii) $\hat{p}(x) - p(x) = O_p(n^{-1/2})$.

The proof of Theorem 2.1 is given in Appendix A.

Theorem 2.1 shows that our cross validation $\tilde{\lambda}$ converges to zero at the rate of n^{-1}, which is the same convergence rate as the maximum likelihood cross-validation choice of λ (see Hall (1981)). In the next section, we will show that the cross validation choice of λ has a much slower rate for the mixed discrete and continuous variable case.

3 Estimating A Joint Density with Mixed Categorical and Continuous Data

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

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

We also 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 where $L_{ix} = L(X_i, x, \lambda)$ is defined in Equation (2.1).

Using the notation $\int dz = \sum_x \int dy$, then the integrated squared difference between $\hat{f}(\cdot)$ and $f(\cdot)$ is

$$\begin{aligned} I_n &= \int [\hat{f}(z) - f(z)]^2 dz = \int [\hat{f}(z)]^2 dz - 2\int \hat{f}(z)f(z)dz + \int [f(z)]^2 dz \\ &\stackrel{def}{=} I_{1n} - 2I_{2n} + \int [f(z)]^2 dz, \end{aligned} \tag{3.3}$$

where $I_{1n} = \int [\hat{f}(z)]dz$ and $I_{2n} = \int \hat{f}(z)f(z)dz$, and we observe that the last term on the right-hand-side of Equation (3.3) is not related to (λ, h). Note that

$$I_{2n} = \int \hat{f}(z)f(z)dz = E[\hat{f}(Z)].$$

Therefore, we estimate $I_{2n} = E[\hat{f}(Z)]$ by

$$\hat{I}_{2n} = \frac{1}{n}\sum_i \hat{f}(Z_i) = \frac{1}{n^2}\sum_i \sum_{j\neq i} K_{h,ij}, \tag{3.4}$$

where $K_{h,ij} = L_{ij}W_{h,ij}$ with $L_{ij} = L(X_i, X_j, \lambda)$ and $W_{h,ij} = h^{-p}W\left(\frac{Y_i - Y_j}{h}\right)$. Using Equation (3.2), we have

$$I_{1n} = \int [\hat{f}(z)]^2 dz = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K_{h,iz}K_{h,jz}dz \stackrel{def}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(2)}_{h,ij}, \tag{3.5}$$

where $K^{(2)}_{h,ij} = \int K_{h,iz}K_{h,jz}dz = \sum_x L_{ix}L_{jx}\int W_{h,iy}W_{h,jy}\,dy \equiv L^{(2)}_{ij}W^{(2)}_{h,ij}$ with $L^{(2)}_{ij} = \sum_x L_{ix}L_{jx}$ and $W^{(2)}_{h,ij} = \int W_{h,iy}W_{h,jy}\,dy$. It is easy to show that

$$W^{(2)}_{h,ij} \equiv W^{(2)}_h(Y_i, Y_j) = h^{-p}W^{(2)}\left(\frac{Y_i - Y_j}{h}\right), \tag{3.6}$$

where $W^{(2)}(v) = \int W(u)W(u+v)\,du$ is the second order convolution kernel derived from $W(\cdot)$.

Therefore, we choose (λ, h) to minimize

$$\begin{aligned} CV(\lambda, h) &\equiv I_{1n} - 2\hat{I}_{2n} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{h,ij}^{(2)} - 2\frac{1}{n^2}\sum_{i}\sum_{j\neq i} K_{h,ij} \\ &= \frac{1}{n^2}\sum_{i} K_{h,ii}^{(2)} + \frac{1}{n^2}\sum_{i}\sum_{j\neq i}[K_{h,ij}^{(2)} - 2K_{h,ij}] \equiv J_{1n} + J_{2n}, \end{aligned} \tag{3.7}$$

where $J_{1n} = \frac{1}{n^2}\sum_i K_{h,ii}^{(2)}$ and $J_{2n} = \frac{1}{n^2}\sum_i \sum_{j\neq i}[K_{h,ij}^{(2)} - 2K_{h,ij}]$.

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

Assumption (B1) (i) $\{Z_i\}_{i=1}^n = \{X_i, Y_i\}_{i=1}^n$ is i.i.d. as $Z = (X, Y)$, X_i satisfies the condition in Assumption (A). (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 $x \in \mathcal{D}$. $f(y|x)$ and its derivatives are all bounded and continuous on the support of Y for all $x \in \mathcal{D}$.

Assumption (B2) (i) The kernel function $w(\cdot)$ is non-negative, bounded and symmetric around zero, also $\int w(v)\,dv = 1$, $\int w(v)v^4\,dv < \infty$. (ii) $\hat{h}$ lies in a shrinking set $H_n = [\underline{h}, \bar{h}]$, where $\underline{h} \geq C^{-1}n^{\delta-1/p}$, $\bar{h} \leq Cn^{-\delta}$ for some C, $\delta > 0$

The conditions in (B2) (ii) are similar to those used in Härdle and Mammen (1985), and they are equivalent to $n^{1-\delta p}\underline{h}^p \geq C^{-1}$ and $n^{\delta}\bar{h} \leq C$. Thus, by choosing a very small value of δ, these conditions are virtually equivalent to the usual conditions of $h \to 0$ and $nh^p \to \infty$ as $n \to \infty$.

Theorem 3.1. *Under assumptions (B1) and (B2), we have*
(i) $\hat{h} = O_p(n^{-1/(4+p)})$ *and* $\hat{\lambda} = O_p(n^{-2/(4+p)})$,
(ii) $\hat{f}(z) - f(z) = O_p(n^{-2/(4+p)})$, *provided* $f(z) \geq \delta$ *for some* $\delta > 0$.

The proof of Theorem 3.1 is given in Appendix B.

Comparing Theorem 3.1 and Theorem 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 case. The slower rate is not a disadvantage for finite sample applications. The conventional frequency estimator corresponds to $\lambda = 0$. Therefore, in order for the cross-validation method to significantly out-perform the case of $\lambda = 0$, it is desirable that $\hat{\lambda}$ does not converge to zero too fast in finite sample applications.

4 Monte Carlo Simulation Results

We begin with a simple Monte Carlo experiment designed to demonstrate that the likelihood cross-validation method of bandwidth selection will break down with commonly used kernels when one or more of the continuous data types are drawn from fat-tailed distributions. This situation is sometimes encountered when dealing with economic and financial data, among others. The experiment reveals simply that the proposed method, unlike existing likelihood cross-validation methods, is robust to the underlying distribution.

We consider two simulations (two data generating processes (DGPs)) involving the estimation of a joint distribution, and for each DGP 1,000 replications are drawn. For each case, the binary variable is generated with $Pr[X = 0] = Pr[X = 1] = 0.5$, while the continuous variable is generated independently being either Gaussian or Cauchy, the latter having fat-tails leading to the breakdown of likelihood cross-validation as will be seen. For each of the 1,000 replications, smoothing parameters are obtained in two ways, first using likelihood cross-validation and then using the proposed least-squares cross-validation method. We use the Gaussian kernel for the continuous variable, while the kernel for the discrete variable is that defined in Equation (2.1). The cross-validated choices of (λ, h) are based on minimizing the cross-validation function with respect to λ and h using a conjugate gradient search algorithm. For each replication we compute the MSE defined by $n^{-1}\sum_i(\hat{f}(X_i, Y_i) - f(X_i, Y_i))^2$. Median values of the smoothing parameters and of MSE calculated over the 1,000 replications are listed in Table 1. The cross-validated choices of $\hat{\lambda}$ take values around 0.5 for almost all cases and we omit them from the table for space considerations. This arises because we use equal (uniform) probabilities for $X = 0$ and $X = 1$, and the choice of $\hat{\lambda} = 1/2$ leads to a constant (uniform) discrete kernel function $L(X_i, X_j) = 1/2$ for all X_i and X_j. Thus there is no sample splitting for this case when using cross-validation methods (for both LS-CV and ML-CV).

Table 1: Median Values of hs and MSEs Generated from 1,000 Replications.

n	X_2 Density	h(ML)	h(LS)	MSE(ML)	MSE(LS)	MSE(ML$_{\lambda=0}$)	MSE(LS$_{\lambda=0}$)
50	Gaussian/Bernoulli	0.497	0.538	7.342e-04	8.612e-04	1.222e-03	1.306e-03
100	Gaussian/Bernoulli	0.432	0.463	4.249e-04	5.324e-04	7.629e-04	8.253e-04
250	Gaussian/Bernoulli	0.367	0.379	2.056e-04	2.460e-04	3.868e-04	4.173e-04
50	Cauchy/Bernoulli	4.431	0.621	5.322e-03	5.871e-04	5.319e-03	9.035e-04
100	Cauchy/Bernoulli	5.272	0.505	6.080e-03	3.644e-04	6.066e-03	5.578e-04
250	Cauchy/Bernoulli	5.696	0.405	7.018e-03	1.701e-04	7.004e-03	2.582e-04

Based on the results reported in Table 1 we make the following observations. First, our

proposed LS-CV method performs much better than the conventional frequency estimator (with $\lambda = 0$) in terms of the MSE criterion. This arises because the conventional frequency estimator splits the sample into two parts, one for which $X = 0$ and one for which $X = 1$. Thus we can assess the extent to which frequency estimators suffer from finite-sample efficiency losses arising from sample splitting. Secondly, when the continuous variable is drawn from the fat-tailed Cauchy distribution, the ML-CV method breaks down having a much larger MSE relative to the proposed LS-CV method. The ML-CV choice of h is about 7 to 14 times as large as that given by the LS-CV method, and this extreme over-smoothing shows that the ML-CV method indeed break down for fat-tailed distributions when using popular kernels. Moreover, the 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. Third, when the continuous variable is Gaussian, the LS-CV estimator is slightly less efficient than the ML-CV estimator. This suggests that, for well-behaved thin-tailed distributions, the likelihood cross-validation method has a slight MSE advantage in small samples, but this quickly disappears and is felt to be completely offset by the ability of the proposed approach to model situations involving fat-tailed distributions. The results summarized in Table 1 clearly show that the proposed LS-CV estimator has a distinct edge over the ML-CV estimator in the sense that the former is robust to fat-tailed distributions, while it also performs much better than the conventional frequency-based estimator.

We also evaluate the estimated density on a grid with support $[-3.5, 3.5]$. Figures 1 through Figure 6 plot the median values of the joint PDF evaluated on this grid. Figure 1 and Figure 2 plot, respectively, the LS-CV and the ML-CV curves when the underlying distribution is Cauchy for a sample size of $n = 50$. Figure 1 shows that our LS-CV approach is well-behaved when the underlying distribution is fat-tailed. In contrast, Figure 2 shows that the ML-CV is hopelessly oversmoothed and totally fails to detect the nature of the underlying distribution. To determine whether the behavior of likelihood cross-validation improves if we increase the sample size, we also conducted the experiment with a sample size of $n = 250$ and again evaluate the PDF on a grid and present these results in Figure 3 and Figure 4. We observe that the discrepancy between the LS-CV estimated curve and the true PDF curve quickly disappear as n gets large, while the ML-CV approach appears to increase the amount of oversmoothing as n increases. This clearly shows that the likelihood-based method of bandwidth selection breaks down for fat-tailed distributions such as the Cauchy distribution.

Figure 5 and Figure 6 plot the cases for a Gaussian continuous variable with $n = 50$. Figure 5 reveals that the ML-CV method is consistent for thin-tailed distributions such as the standard normal (with a Gaussian kernel). The median plots are virtually identical for the

LS-CV and the ML-CV curves as can be seen from the Figure 5 and Figure 6.

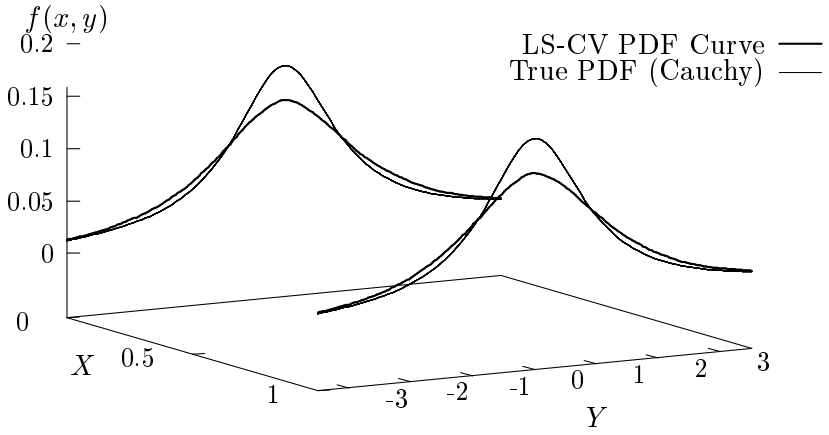

Figure 1: Estimated Joint PDF - Joint Binary/Cauchy, $n = 50$

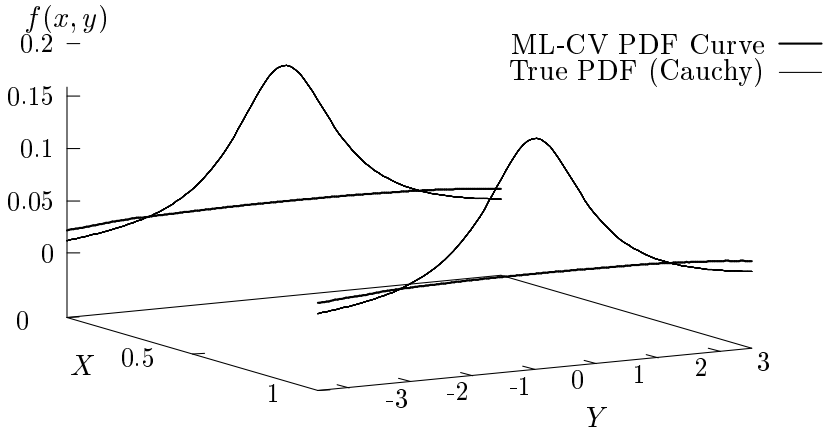

Figure 2: Estimated Joint PDF - Joint Binary/Cauchy, $n = 50$

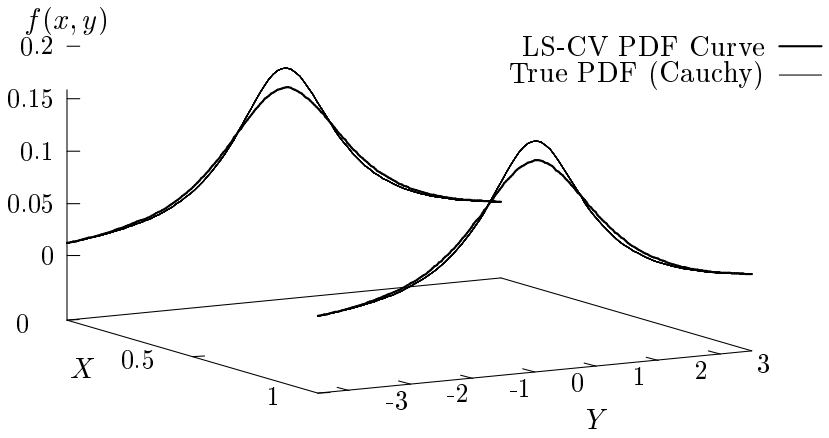

Figure 3: Estimated Joint PDF - Joint Binary/Cauchy, $n = 250$

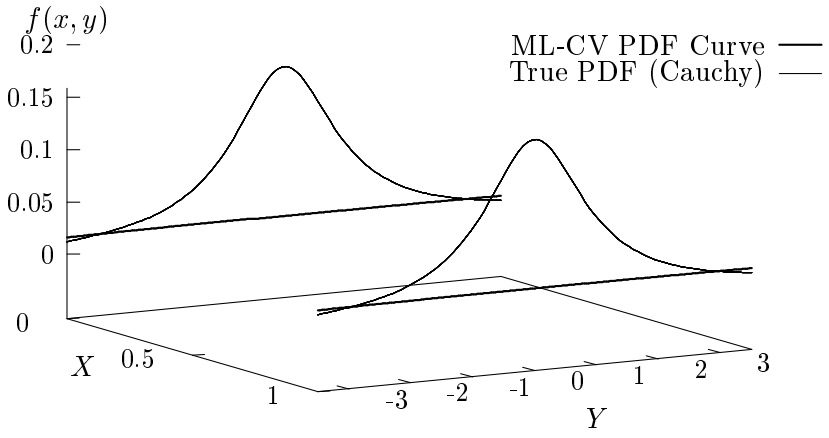

Figure 4: Estimated Joint PDF - Joint Binary/Cauchy, $n = 250$

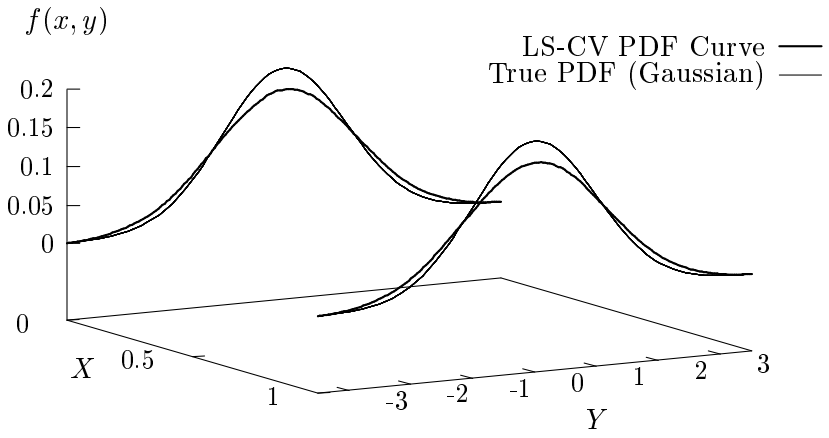

Figure 5: Estimated Joint PDF - Joint Binary/Gaussian, $n = 50$

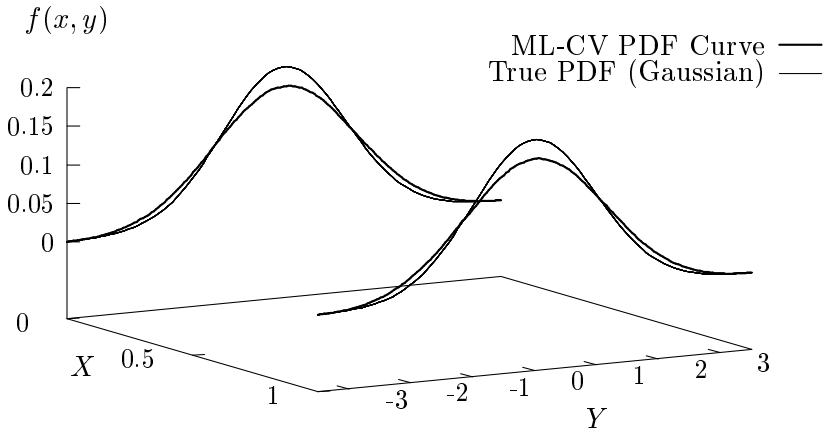

Figure 6: Estimated Joint PDF - Joint Binary/Gaussian, $n = 50$

5 An Empirical Application

We now consider an application of the proposed approach to modeling discrete choice. This example shows how the proposed estimator can be used to obtain superior (out-of-sample) predictive performance relative to commonly used parametric models of discrete choice.

We use the data of Gerfin (1996) who models the labor market participation of married Swiss women using a cross-section data set of size $n = 872$ having six explanatory variables. He uses a Probit model along with three semiparametric specifications, and finds that the Probit specification cannot be rejected and that the models yield similar results. He concludes that "more work is necessary on specification tests of semiparametric models and on simulations using these models". We simply use this dataset to see whether predictions given by the Probit and semiparametric specifications can be substantially improved upon (we do not include Gerfin's (1996) semiparametric results here as they all yielded similar results.) Data for this study can be found at `http://qed.econ.queensu.ca/jae/1996-v11.3/gerfin/`.

The variables used by the Gerfin (1996) study are

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

We compute the conditional distribution as the ratio of the joint distribution of variables 1 through 7 and the marginal distribution of variables 2 through 7,

$$\hat{f}(\text{LFP}|\text{LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN}) = \frac{\hat{f}(\text{LFP, LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN})}{\hat{f}_1(\text{LNNLINC, AGE, EDUC, NYC, NOC, FOREIGN})} \quad (3.8)$$

and bandwidths are chosen via cross-validation. Finally, we predict LFP=1 if $\hat{f}(\text{LFP} = 1|\cdot) > \hat{f}(\text{LFP} = 0|\cdot)$ where $|\cdot)$ denotes the conditioning variables, otherwise we predict LFP=0.

We compare the results of our estimator with those from Gerfin (1996), and the confusion matrices and classification rates for both the proposed and Probit approaches are summarized

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

Kernel			**Probit**		
A/P	0	1	A/P	0	1
0	360	111	0	358	113
1	115	286	1	179	222
%Correct		74.1%	%Correct		66.5%
%CCR(0)		76.4%	%CCR(0)		76.0%
%CCR(1)		71.3%	%CCR(1)		55.4%

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

This application simply demonstrates 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 of the present paper can be extended. We briefly mention a few of them below.

1. Semiparametric estimation of a density function with mixed discrete and continuous data.

[1] For example, CCR(0) is the number of predicted zeros that are in fact zeros ÷ number of zeros in the sample × 100.

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

2. Estimation of joint density function with mixed discrete and continuous variables when the discrete variables contain ordered categorical data.

3. Consistent model specification tests with mixed discrete and continuous regressors, including testing for correct parametric density or a semiparametric density functional form.

The rates of convergence established in this paper will be very helpful when extending the results to estimating a semiparametric density function, especially when deriving the asymptotic distribution of the estimator of the parametric component. With ordered categorical data, it is known that boundary kernels (Dong and Simonoff (1994)), local polynomials (Aerts, Augustyns and Janssen (1997a,b)), penalized likelihood (Simonoff (1983)), and local likelihood methods have better properties than standard kernel estimators (they avoid boundary bias). It will be very useful to extend the result of this paper to cover the case of ordered categorical data. Specification tests (with mixed data types) based on a data-driven choice of smoothing parameters would be significantly more powerful than tests based on frequency estimators as the former do not use sample splitting in finite-sample applications.

Recently, Racine and Li (2000) have considered the problem of nonparametric estimation of regression functions with mixed discrete and continuous regressors and have established the convergence rate of their proposed estimator. Yet another direction is to consider semiparametric regression models with mixed regressors, including partially linear models and additive models, and specification tests for parametric/semiparametric regression functional forms. The authors are currently working on a number of these exciting extensions.

Appendix A

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

Proof: Let $I_n(\lambda) = I_{1n}(\lambda) - 2I_{2n}(\lambda) + E[p(X)]$ be defined as in Equation (2.3), and define $\hat{I}_n(\lambda) = I_{1n}(\lambda) - 2\hat{I}_{2n}(\lambda) + E[p(X)]$, where $\hat{I}_{2n}(\lambda)$ is defined in Equation (2.4). Obviously $\hat{I}_{2n}(\lambda) - I_{2n}(\lambda) = o_p(1)$, which implies that (a): $\hat{I}_n(\lambda) = I_n(\lambda) + o_p(1)$.

Next, $0 \leq \hat{I}_n(\tilde{\lambda}) \leq \hat{I}_n(0) = o_p(1)$ 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): $\hat{I}_n(\tilde{\lambda}) = o_p(1)$.

(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 $0 \leq l \leq 2k$.

Hence, we have (d): $I_n(\lambda) = O_p(1) \neq o_p(1)$ for $\lambda \neq o(1)$.

(c) and (d) imply that $\tilde{\lambda} = o_p(1)$.

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

In lemmas A.1 through A.4 below, we use the property that $\lambda = o(1)$ and obtain expansions of $J_{1n}(\lambda)$ and $J_{2n}(\lambda)$. 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.

Lemma A.1. *$J_{1n}(\lambda) \equiv n^{-2} \sum_{i=1}^{n} L_{ii}^{(2)} = A_1 n^{-1} - A_2 \lambda n^{-1} + O_p(n^{-3/2}) + O_p(n^{-3/2}\lambda + n^{-1}\lambda^2)$, where $A_1 = \sum_x p(x)$ and $A_2 = 2k \sum_x p(x)$ are positive constants.*

Proof: By Equations (2.6) and (2.7), $J_{1n} \equiv n^{-2} \sum_{i=1}^{n} L_{ii}^{(2)} = n^{-2} \sum_{i=1}^{n} \sum_x L_{ix}^2$. Hence,

$E[J_{1n}] = n^{-1} \sum_x E[L_{ix}^2]$

$= n^{-1} \sum_x \{E[L_{ix}^2 | d_{ix} = 0] P(d_{ix} = 0) + E[L_{ix}^2 | d_{ix} \geq 1] P(d_{ix} \geq 1)\}$

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

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

Also,

$J_{1n} - E[J_{1n}] = n^{-1}\{n^{-1} \sum_i [L_{ii}^{(2)} - E(L_{ii}^{(2)})]\}$

$= n^{-1}\{n^{-1} \sum_x \sum_i [L_{ix}^2 - E(L_{ix}^2)]\}$

$= n^{-1}\{\sum_x n^{-1} \sum_i L_{ix}^2 \mathbf{1}(X_i = x) - \sum_x E[L_{ix}^2 | d_{ix} = 0] p(d_{X_i,x} = 0)\}$

$\quad + n^{-1}\{\sum_x n^{-1} \sum_i L_{ix}^2 \mathbf{1}(d_{X_i,x} \geq 1) - \sum_x E[L_{ix}^2 | d_{X_i,x} \geq 1] p(d_{X_i,x} \geq 1)]\}$

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

$= \{(1 - 2k\lambda) n^{-3/2} \{n^{1/2} \sum_x [\bar{p}(x) - p(x)]\} + O_p(n^{-1}\lambda^2)$

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

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

Hence, $J_{1n} = E(J_{1n}) + [J_{1n} - E(J_{1n})] = n^{-1}(1 - 2k\lambda) \sum_x p(x) + O_p(n^{-3/2}) + O_p(n^{-3/2}\lambda + n^{-1}\lambda^2)$.

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

$E[H(X_i, X_j)] = A_3 + A_4\lambda^2 + O(\lambda^3)$,

where A_3 and A_4 are some constants with $A_4 > 0$.

Proof: $E[H(X_i, X_j)] = E[L_{ij}^{(2)}] - 2E[L_{ij}]$. We compute $E[L_{ij}^{(2)}]$ and $E[L_{ij}]$ separately below. Define $p_s(x) = Prob[d(X,x) = s] = \sum_{\{x', d_{x',x}=s\}} p(x')$, $s = 1, 2$.

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

Next,
$E[L_{ij}] = \sum_{x_1} \sum_{x_2} p(x_1)p(x_2) L_{x_1,x_2}$
$= (1-\lambda)^k \sum_{x_1} p(x_1) \sum_{\{x_2, d_{x_1,x_2}=0\}} p(x_2) + \lambda(1-\lambda)^{k-1} \sum_{x_1} p(x_1) \sum_{\{x_2, d_{x_1,x_2}=1\}} p(x_2)$
$+\lambda^2(1-\lambda)^{k-2} \sum_{x_1} p(x_1) \sum_{\{x_2, d_{x_1,x_2}=2\}} p(x_2) + O(\lambda^3)$
$= (1 - k\lambda + \lambda^2 k(k-1)/2) E[p(X)] + \lambda(1 - (k-1)\lambda) E[p_1(X)] + \lambda^2 E[p_2(X)] + O(\lambda^3)$
$= E[p(X)] + \lambda\{E[p_1(X)] - kE[p(X)]\} + \lambda^2\{E[p_2(X)] - (k-1)E[p_1(X)] + [k(k-1)/2]E[p(X)]\}$.

Summarizing the above results, we get
$E[H(X_1, X_2)] = E[L_{ij}^{(2)}] - 2E[L_{ij}]$
$= -E[p(X)] + \lambda^2\{k^2 E[p(X)] - 2kE[p_1(X)] + E[(p_1(X))^2/p(X)]\} + O(\lambda^3)$
$\equiv A_3 + A_4\lambda^2 + O(\lambda^3)$.

Lemma A.3. *$E[H(X_i, X_j)|X_i] = -p(X_i) + O(\lambda^2)$, where $H(X_i, X_j) = L_{ij}^{(2)} - 2L_{ij}$.*

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

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

Next,
$E[L_{ij}^{(2)}|X_i] = \sum_x E[L_{X_i,x} L_{X_j,x}|X_i] = \sum_x \sum_{x_1} p(x_1) L_{X_i,x} L_{x_1,x}$

$= (1-\lambda)^{2k} \sum_{\{x, d_{X_i,x}=0\}} \sum_{x, d_{X_i,x_1}=0} p(x_1)$
$+\lambda(1-\lambda)^{2k-1}\{\sum_{\{x,d_{X_i,x}=0\}} \sum_{\{x_1,d_{X_i,x_1}=1\}} p(x_1) + \sum_{\{x,d_{X_i,x}=1\}} \sum_{\{x_1,d_{X_i,x_1}=0\}} p(x_1)\} + O(\lambda^2)$
$= (1-2k\lambda)p(X_i) + 2\lambda p_1(X_i) + O(\lambda^2)$
$= p(X_i) + 2\lambda[p_1(X_i) - kp(X_i)] + O(\lambda^2).$

Hence, we have

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

Lemma A.4. $J_{2n}(\lambda) = A_3 + A_4\lambda^2 - n^{-1/2}\mathcal{Z}_n + O_p(n^{-3/2}) + o_p(\lambda n^{-1} + \lambda^2)$,
where $\mathcal{Z}_n = n^{-1/2}\sum_{i=1}^n [p(X_i) - E(p((X_i))]$ *is a zero mean* $O_p(1)$ *random variable.*

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

$J_{2n} = n^{-2}\sum_i \sum_{j\neq i} H(X_i, X_j)$
$= E[H(X_i, X_j)] + n^{-1}\sum_{i=1}^n \{E[H(X_i, X_j)|X_i] - E[H(X_i, X_j)]\} + (s.o.)$
$= A_3 + A_4\lambda^2 + O_p(n^{-3/2}) + O_p(n^{-3/2}\lambda + \lambda^3)$
$\quad -n^{-1/2}\{n^{-1/2}\sum_{i=1}^n [p(X_i) - E(p(X_i))\,] + O_p(n^{-1/2}\lambda^2)\} + (s.o.)$
$= A_3 + A_4\lambda^2 - n^{-1/2}\mathcal{Z}_n + O_p(n^{-3/2}) + o_p(\lambda n^{-1} + \lambda^2),$

where $\mathcal{Z}_n = n^{-1/2}\sum_{i=1}^n [p(X_i) - E(p(X_i))]$ is a zero mean $O_p(1)$ random variable.

Proof of Theorem 2.1

First for (i), by lemmas A.2 - A.4, we have

$$\begin{aligned} CV(\lambda) &= J_{1n} + J_{2n} = A_1 n^{-1} - A_2\lambda n^{-1} + A_3 + A_4\lambda^2 - n^{-1/2}\mathcal{Z}_n + o_p(n^{-1}) + o_p(\lambda^2 + n^{-1}\lambda) \\ &= A_4[\lambda - A_2 n^{-1}/(2A_4)]^2 - A_2^2 n^{-2}/(4A_4^2) + A_1 n^{-1} + A_3 + (s.o.) \qquad \text{(A.1)} \end{aligned}$$

Minimization of Equation (A.1) over λ leads to $\tilde{\lambda} = A_2 n^{-1} + o_p(n^{-1}) = O_p(n^{-1})$.

Next, for (ii), similar to (i), one can show that $[\hat{p}(x) - p(x)]^2 = O_p(n^{-1}) + O_p(n^{-1}\hat{\lambda}) + O_p(\hat{\lambda}^2) = O_p(n^{-1})$. Hence, $\hat{p}(x) - p(x) = O_p(n^{-1/2})$.

Appendix B

Note that $\hat{h} = o(1)$ by assumption (B2) (ii). 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 Lemma B.1 to Lemma A.4 below we use $h = o(1)$ and $\lambda = o(1)$ to obtain expansions of $J_{1n}(\lambda, h)$ and $J_{2n}(\lambda, h)$.

Lemma B.1. $J_{1n}(\lambda,h) \equiv n^{-2}\sum_{i=1}^n K_{h,ii}^{(2)} = (nh^p)^{-1}[B_1 - B_2\lambda + O(\lambda^2)]$,
where B_1 and B_2 are some positive constants.

Proof: First from Equation (3.6) we have $W_{h,ii}^{(2)} = h^{-p}W^{(2)}(0) = h^{-p}\int W^2(v)\,dv$. Hence,

$J_{1n} = n^{-2}\sum_{i=1}^n K_{h,ii}^{(2)} = n^{-2}\sum_{i=1}^n L_{ii}^{(2)}W_{h,ii}^{(2)}$
$= [\int W^2(v)\,dv]h^{-p}[n^{-2}\sum_{i=1}^n L_{ii}^{(2)}] = [\int W^2(v)\,dv](nh^p)^{-1}[A_1 - A_2\lambda + O(\lambda^2)]$
$= (nh^p)^{-1}[B_1 - B_2\lambda + O(\lambda^2)]$ by Lemma A.1,

where $B_1 = [\int W^2(v)\,dv]A_1 > 0$ and $B_2 = [\int W^2(v)\,dv]A_2 > 0$.

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

Then $E[H(Z_i,Z_j)] = B_3 + B_4\lambda^2 - B_5\lambda h^2 + B_6h^4 + o_p(\lambda^4 + \lambda^2h^2 + h^4)$,
where B_j $(j = 3,..,6)$ are some constants with $B_4 > 0$ and $B_6 > 0$.

Proof: $E[H(Z_i,Z_j)] = E[K_{h,ij}^{(2)}] - 2E[K_{h,ij}]$. 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 $W_h(y_1,y_2) = h^{-p}W\left(\frac{y_1-y_2}{h}\right)$. We have

$E[K_{h,ij}] = \sum_{x_1}\sum_{x_2} p(x_2)p(x_2)L(x_1,x_2)\int W_h(y_1,y_2)f(y_1|x_1)f(y_2|x_2)\,dy_1\,dy_2$
$\equiv \sum_{x_1}\sum_{x_2} p(x_1)p(x_2)L_{x_1,x_2}G_h(x_1,x_2) \equiv \sum_x\sum_{x_1} p(x)p(x_1)L_{x,x_1}G_h(x,x_1)$
$= (1-\lambda)^k\sum_x[p(x)]^2G_h(x,x) + \lambda(1-\lambda)^{k-1}\sum_x p(x)\sum_{\{x_1,d_{x,x_1}=1\}} p(x_1)G_h(x,x_1)$
$\quad +\lambda^2(1-\lambda)^{k-2}\{\sum_x p(x)\sum_{\{x_1,d_{x,x_1}=2\}} p(x_1)G_h(x,x_1) + O(\lambda^3)$
$= (1 - k\lambda + \lambda^2k(k-1)/2)\sum_x[p(x)]^2G_h(x,x)$
$\quad +\lambda(1-\lambda(k-1))\sum_x p(x)\sum_{\{x_1,d_{x,x_1}=1\}} p(x_1)G_h(x,x_1)$
$\quad +\lambda^2\sum_x p(x)\sum_{\{x_1,d_{x,x_1}=2\}} p(x_1)G_h(x,x_1) + O(\lambda^3)$
$\equiv (1 - k\lambda + \lambda^2k(k-1)/2)T_0 + \lambda(1-\lambda(k-1))T_1 + \lambda^2T_2 + O(\lambda^3)$
$= T_0 + \lambda(T_1 - kT_0) + \lambda^2\{T_2 - (k-1)T_1 + [k(k-1)/2]T_2\} + O(\lambda^3)$,

$$\begin{aligned} T_0 &= \sum_x [p(x)]^2 G_h(x,x), \\ T_1 &= \sum_x p(x) \sum_{\{x_1,d_{x,x_1}=1\}} p(x_1)G_h(x,x_1), \\ T_2 &= \sum_x p(x) \sum_{\{x_1,d_{x,x_1}=2\}} p(x_1)G_h(x,x_1). \end{aligned} \tag{B.1}$$

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

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

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:

$$G_h(x, x_1) = G_0(x, x_1)+h^2G_2(x, x_1)+h^4G_4(x, x_1)+o_p(h^4), \tag{B.3}$$

where $G_0(x, x_1) = \int f(y|x)f(y|x_1)W(v)\,dv\,dy = \int f(y|x)f(y|x_1)\,dy$,
$G_2(x, x_1) = (1/2)\int f(y|x)v'\nabla_y^2 f(y|x_1)vW(v)\,dv\,dy$, 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 $\int W(v)v_l^2v_{l'}^2\,dv$, where v_l is the lth component of $v \in R^p$ $(l = 1, \ldots, p)$.

Next, we consider $E(K_{h,ij}^{(2)})$.

Defining $G_{h,1,2}^{(2)} \equiv G_h^{(2)}(x_1, x_2) \stackrel{def}{=} \int f(y_1|x_1)f(y_2|x_2)W_h^{(2)}(y_1, y_2)\,dy_1\,dy_2$, we have

$$\begin{aligned}
&E[L_{ij}^{(2)}] = E[L_{ij}^{(2)}W_{h,ij}^{(2)}] = \textstyle\sum_x E[L_{ix}L_{jx}W_{h,ij}^{(2)}] \\
&= \textstyle\sum_x \sum_{x_1} \sum_{x_2} p(x_1)p(x_2)L_{x_1,x}L_{x_2,x}\int f(y_1|x_1)f(y_2|x_2)W_h^{(2)}(y_1, y_2)\,dy_1\,dy_2 \\
&\equiv \textstyle\sum_x \sum_{x_1} \sum_{x_2} p(x_1)p(x_2)L_{x_1,x}L_{x_2,x}G_h^{(2)}(x_1, x_2) \\
&= (1-\lambda)^{2k}\textstyle\sum_x \sum_{\{x_1,d_{x_1,x}=0\}} p(x_1) \sum_{\{x_2,d_{x_2,x}=0\}} p(x_2)G_{h,1,2}^{(2)} \\
&\quad +\lambda(1-\lambda)^{2k-1}\textstyle\sum_x \{ \sum_{x_1,d_{x_1,x}=0} p(x_1) \sum_{x_1,d_{x_2,x}=1} p(x_2)G_{h,1,2}^{(2)} \\
&\quad + \textstyle\sum_{x_1,d_{x_1,x}=1} p(x_1) \sum_{x_1,d_{x_2,x}=0} p(x_2)G_{h,1,2}^{(2)} \} \\
&\quad +\lambda^2(1-\lambda)^{2(k-1)}\textstyle\sum_x \{ \sum_{x_1,d_{x_1,x}=2} p(x_1) \sum_{x_1,d_{x_2,x}=0} p(x_2)G_{h,1,2}^{(2)} \\
&\quad + \textstyle\sum_{x_1,d_{x_1,x}=0} p(x_1) \sum_{x_1,d_{x_2,x}=2} p(x_2)G_{h,1,2}^{(2)}+\sum_{x_1,d_{x_1,x}=1} p(x_1) \sum_{x_1,d_{x_2,x}=1} p(x_2)G_{h,1,2}^{(2)} \}+O(\lambda^3) \\
&= (1-2k\lambda+k(2k-1)\lambda^2)\textstyle\sum_x [p(x)]^2G_h^{(2)}(x, x) \\
&\quad +2\lambda(1-(2k-1)\lambda)\textstyle\sum_x p(x) \sum_{x_1,d_{x_1,x}=1} p(x_1)G_h^{(2)}(x, x_1) \\
&\quad +\lambda^2\textstyle\sum_x \{ p(x) \sum_{\{x_1,d_{x_1,x}=2\}} p(x_1)G_h^{(2)}(x_1, x) + p(x) \sum_{\{x_1,d_{x_2,x}=2\}} p(x_2)G_h^{(2)}(x, x_2) \\
&\quad + \textstyle\sum_{\{x_1,d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2,d_{x_2,x}=1\}} p(x_2)G_h^{(2)}(x_1, x_2) \} + O(\lambda^3) \\
&\equiv (1-2k\lambda+k(2k-1)\lambda^2)T_0^{(2)} + \lambda(1-(2k-1)\lambda)T_1^{(2)} + \lambda^2T_2^{(2)} + O(\lambda^3) \\
&= T_0^{(2)} + \lambda(T_1^{(2)} - 2kT_0^{(2)}) + \lambda^2\{T_2^{(2)} - (2k-1)T_1^{(2)} + k(2k-1)T_0^{(2)}\} + O(\lambda^3),
\end{aligned}$$

where

$$\begin{aligned}
T_0^{(2)} &= \sum_x [p(x)]^2 G_h^{(2)}(x,x), \\
T_1^{(2)} &= 2\sum_x p(x) \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) G_h^{(2)}(x,x_1) \\
T_2^{(2)} &= 2\sum_x p(x) \sum_{\{x_1, d_{x_1,x}=2\}} p(x_1) G_h^{(2)}(x,x_1) \\
&\quad + \sum_x \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=1\}} p(x_2) G_h^{(2)}(x_1,x_2).
\end{aligned} \tag{B.4}$$

We summarize the above result in the following equation

$$E[K_{ij}^{(2)}] = T_0^{(2)} + \lambda(T_1^{(2)} - 2kT_0^{(2)}) + \lambda^2\{T_2^{(2)} - (2k-1)T_1^{(2)} + k(2k-1)T_0^{(2)}\} + O(\lambda^3). \tag{B.5}$$

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:

$$G_h^{(2)}(x_1,x_2) = G_0^{(2)}(x_1,x_2) + h^2 G_2^{(2)}(x_1,x_2) + h^4 G_4^{(2)}(x_1,x_2) + o_p(h^4), \tag{B.6}$$

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 between $W(\cdot)$ and $W^{(2)}(\cdot)$ hold.

$$\begin{aligned}
&\int W^{(2)}(v)\, dv = \int W(v)\, dv = 1, \\
&\int W^{(2)}(v) v v'\, dv = 2\int W(v) v v'\, dv, \\
&\int W^{(2)}(v) v_l^4\, dv > 2\int W(v) v_l^4\, dv \qquad\qquad l = 1, \ldots, p.
\end{aligned} \tag{B.7}$$

From Equation (B.3), Equation (B.6) and Equation (B.7), we immediately get

$$G_0^{(2)}(x_1,x_2) = G_0(x_1,x_2), \quad G_2^{(2)}(x_1,x_2) = 2G_2(x_1,x_2), \quad G_4^{(2)}(x_1,x_2) > 2G_4(x_1,x_2). \tag{B.8}$$

Below we write $E[H(Z_i, Z_j)] = EH_0 + \lambda EH_1 + \lambda^2 EH_2 + O(\lambda^3)$ and we will first obtain EH_0, the component of $E[H(Z_i, Z_j)] = E[K_{ij}^{(2)}] - 2E[K_{ij}]$ that is independent of λ. From Equation (B.2), Equation (B.5) and Equation (B.8), we have

$$\begin{aligned} EH_0 &= T_0^{(2)} - 2T_0 = \sum_x [p(x)]^2 [G_h^{(2)}(x,x) - 2G_h(x,x)] \\ &= \sum_x [p(x)]^2 \{-G_0(x,x) + (0)h^2 + h^4 [G_4^{(2)}(x,x) - 2G_4(x,x)]\} \\ &\equiv B_3 + B_4 h^4, \end{aligned} \tag{B.9}$$

where $B_3 = -\sum_x [p(x)]^2 G_0(x,x)$ and $B_4 = \sum_x [p(x)]^2 [G_4^{(2)}(x,x) - 2G_4(x,x)] > 0$ by Equation (B.8).

Next, we compute EH_1, the component of $E[H(Z_i, Z_j)|Z_i]$ that is linear in λ. Using Equation (B.2), Equation (B.5) and Equation (B.8), we have

$$\begin{aligned} EH_1 &= T_1^{(2)} - 2kT_0^{(2)} - 2[T_1 - kT_0] = [T_1^{(2)} - 2T_1] + 2k[T_0 - T_0^{(2)}] \\ &= 2\sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1)[G_h^{(2)}(x,x_1) - G_h(x,x_1)] + 2k \sum_x [p(x)]^2 [G_h(x,x) - G_h^{(2)}(x,x)] \\ &= 2\sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1)[0 + h^2 G_2(x,x_1) + O(h^4)] \\ &\quad + 2k \sum_x [p(x)]^2 [0 - h^2 G_2(x,x) + O(h^4)] \\ &= h^2(-2)\{k \sum_x [p(x)]^2 G_2(x,x) - \sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1) G_2(x,x_1)\} + O(h^4) \\ &\equiv -B_5 h^2 + O_p(h^4), \end{aligned} \tag{B.10}$$

where $B_5 = 2\{k\sum_x [p(x)]^2 G_2(x,x) - \sum_x p(x) \sum_{x_1, d_{x_1,x}=1} p(x_1) G_2(x,x_1)\}$.

Finally, we compute EH_2, the component of $E[H(Z_i, Z_j)|Z_i]$ that is linear in λ^2. Using Equation (B.2), Equation (B.5) and Equation (B.8), we have

$$\begin{aligned}
EH_2 &= [T_2^{(2)} - (2k-1)T_1^{(2)} - k(2k-1)T_0^{(2)}] - 2\{T_2 - (k-1)T_1 + [k(k-1)/2]T_0\} \\
&= [T_2^{(2)} - 2T_2] + [2(k-1)T_1 - (2k-1)T_1^{(2)}] + [k(2k-1)T_0^{(2)} - k(k-1)T_0] \\
&= [\sum_x \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) \sum_{\{x_2, d_{x_2,x}=1\}} p(x_2) G_0^{(2)}(x_1, x_2)] \\
&\quad + [-2k \sum_x p(x) \sum_{\{x_1, d_{x_1,x}=1\}} p(x_1) G_0^{(2)}(x, x_1)] \\
&\quad + [2k^2 \sum_x p(x) \sum_{\{x_1, d_{x_1,x}=2\}} p(x_1) G_0^{(2)}(x, x_2)] + O_p(h^2) \\
&\equiv B_6 + O_p(h^2) \qquad \text{(B.11)}
\end{aligned}$$

where the definition of B_6 should be apparent. Note that B_6 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 Equation (B.8) is used in computing B_6.

By equations (B.9)–(B.11), we immediately obtain

$$\begin{aligned}
E[H(Z_i, Z_j)] &= E[K_{h,ij}^{(2)}] - E[K_{h,ij}] \\
&= B_3 + B_4 h^4 - B_5 \lambda h^2 + B_6 \lambda^2 + o_p(h^4 + \lambda h^2 + \lambda^2). \qquad \text{(B.12)}
\end{aligned}$$

Lemma B.3. $E[H(Z_i, Z_j)|Z_i] = p(X_i) + O_p(h^4 + \lambda h^2 + \lambda^2)$,
where $Z_i = (X_i, Y_i)$.

Proof: $E[H(Z_i, Z_j)|Z_i] = E[K_{h,ij}^{(2)}|Z_i] - 2E[K_{h,ij}|Z_i]$. We consider $E[K_{h,ij}|Z_i]$ and $E[K_{h,ij}^{(2)}|Z_i]$ separately below. Defining $M_h(x, Y_i) = \int W_h(Y_i, y) f(y|x)\, dy$, we have

$E[K_{h,ij}|Z_i] = E[L_{ij} W_{h,ij}|Z_i]$
$= \sum_x p(x) L(X_i, x) \int W_h(Y_i, y) f(y|x)\, dy$
$\equiv \sum_x p(x) L(X_i, x) M_h(x, Y_i)$
$= (1-\lambda)^k p(X_i) M_h(Z_i) + \lambda(1-\lambda)^{k-1} \sum_{\{x, d_{X_i,x}=1\}} p(x) M_h(x, Y_i) + O(\lambda^2)$
$= (1 - k\lambda) p(X_i) M_h(Z_i) + \lambda \sum_{\{x, d_{X_i,x}=1\}} p(x) M_h(x_1, Y_i) + O(\lambda^2)$
$= p(X_i) M(Z_i) + \lambda[\sum_{\{x, d_{X_i,x}=1\}} p(x) M_h(x, Y_i) - k\, p(X_i) M_h(Z_i)] + O(\lambda^2)$
$\equiv C_0(Z_i) + \lambda C_1(Z_i) + O(\lambda^2)$,

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

Summarizing the above result, we have

$$E[K_{ij}|Z_i] = C_0(Z_i) + \lambda C_1(Z_i) + O(\lambda^2). \tag{B.13}$$

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

$$M_h(Z_i) = M_0(Z_i) + h^2 M_2(Z_i) + O_p(h^4) = 1 + h^2 M_2(Z_i) + O_p(h^4), \tag{B.14}$$

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

Next, we consider $E[K_{h,ij}^{(2)}|Z_i]$. Defining $M_h^{(2)}(x, Y_i) = \int W_h^{(2)}(Y_i, y)f(y|x)\,dy$, we have

$$\begin{aligned}
&E[L_{h,ij}^{(2)}|Z_i] = \textstyle\sum_x E[L_{ix}L_{jx}W_h^{(2)}(Y_i, Y_j)|Z_i] \\
&= \textstyle\sum_x \sum_{x_1} p(x_1)L_{X_i,x}L_{x_1,x}\int W_h^{(2)}(Y_i, y_1)f(y_1|x_1)\,dy_1 \\
&\equiv \textstyle\sum_x \sum_{x_1} p(x_1)L_{X_i,x}L_{x_1,x}M_h^{(2)}(x_1, Y_i) \\
&= (1-\lambda)^{2k}p(X_i)M_h^{(2)}(X_i, Y_i) + \lambda(1-\lambda)^{2k-1}\{\textstyle\sum_{\{x,d_{X_i,x}=1\}}\sum_{\{x_1,d_{x_1,x}=0\}} p(x_1)M_h^{(2)}(x_1, Y_i) \\
&\qquad + \textstyle\sum_{\{x,d_{X_i,x}=0\}}\sum_{\{x_1,d_{x_1,x}=1\}} p(x_1)M_h^{(2)}(x_1, Y_i)\} + O(\lambda^2) \\
&= (1-2k\lambda)p(X_i)M_h^{(2)}(Z_i) + 2\lambda\textstyle\sum_{\{x,d_{X_i,x}=1\}} p(x)M_h^{(2)}(x, Y_i) + O(\lambda^2) \\
&= p(X_i)M_h^{(2)}(Z_i) + 2\lambda\{\textstyle\sum_{\{x,d_{X_i,x}=1\}} p(x)M_h^{(2)}(x, Y_i) - k\,p(X_i)M_h^{(2)}(Z_i)\} + O(\lambda^2) \\
&\equiv C_0^{(2)}(Z_i) + \lambda C_1^{(2)}(Z_i) + O(\lambda^2),
\end{aligned}$$

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

Summarizing the above result, we have

$$E[K_{ij}^{(2)}|Z_i] = C_0^{(2)}(Z_i) + \lambda C_1^{(2)}(Z_i) + O(\lambda^2). \tag{B.15}$$

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

$$M_h^{(2)}(Z_i) = M_0^{(2)}(Z_i) + h^2 M^{(2)}(Z_i) + O_p(h^4) = 1 + h^2 M^{(2)}(Z_i) + O_p(h^4) \tag{B.16}$$

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

By Equation (B.7), we know that

$$M_2^{(2)}(Z_i) = 2M_2(Z_i). \tag{B.17}$$

Using Equation (B.13), Equation (B.14), Equation (B.15), Equation (B.16) and Equation (B.17), we obtain

$$\begin{aligned}
& E[H(Z_i,Z_j)|Z_i] = E[K^{(2)}_{h,ij}|Z_i] - 2E[K_{h,ij}|X_i] \\
= \; & [C_0^{(2)}(Z_i) + \lambda C_1^{(2)}(Z_i)] - 2[C_0(Z_i) + \lambda C_1(Z_i)] + O(\lambda^2) \\
= \; & [C_0^{(2)}(Z_i) - 2C_0(Z_i)] + \lambda[C_1^{(2)}(Z_i) - 2C_1(Z_i)] + O(\lambda^2) \\
= \; & -p(X_i)\{1-2] + h^2[M_2^{(2)}(Z_i) - 2M_2(Z_i)] + O(h^4)\} \\
& +2\lambda[\sum_{\{x,d_{X_i,x}=1\}} p(x)\{[M_2^{(2)}(x) - M_2(x)] + h^2 k\, p(X_i)[M_2^{(2)}(Z_i)] - M_2(Z_i)]\} + O_p(h^4)] + O(\lambda^2) \\
= \; & [-p(X_i) + h^2(0) + O(h^4)] + 2\lambda[0 + h^2 k\, p(X_i) M_2(Z_i) + O_p(h^4)] + O(\lambda^2) \\
= \; & p(X_i) + O_p(h^4 + \lambda h^2 + \lambda^2). \qquad \text{(B.18)}
\end{aligned}$$

Lemma B.4. *$J_{2n}(\lambda, h) = B_3 + B_4 h^4 - B_5 \lambda h^2 + B_6 \lambda^2 - n^{-1/2}\mathcal{Z}_n + (s.o.)$, where B_j $(j = 3, \ldots, 6)$ are constants defined in Lemma B.2 and $\mathcal{Z}_n = n^{-1/2}\sum_i [p(X_i) - E(p(X_i))]$.*

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

$$\begin{aligned}
& J_{2n} = n^{-2}\sum_i \sum_{j\neq i} H(X_i, X_j) \\
& = E[H(Z_i, Z_j)] + n^{-1}\sum_i \sum_i \{E[H(Z_i, Z_j)|Z_i] - E[H(Z_i, Z_j)]\} + (s.o.) \\
& = B_3 + B_4 h^4 - B_5 \lambda h^2 + B_6 \lambda^2 - n^{-1/2}\{n^{-1/2}\sum_i [p(X_i) - E(p(X_i))]\} + (s.o.) \\
& = B_3 + B_4 h^4 - B_5 \lambda h^2 + B_6 \lambda^2 - n^{-1/2}\mathcal{Z}_n + (s.o.).
\end{aligned}$$

Proof of Theorem 3.1

First for (i), by Lemma B.1 and Lemma B.4, we have

$$\begin{aligned}
CV_L(\lambda, h) &= J_{1n} + J_{2n} \\
&= B_1(nh^p)^{-1} - B_2\lambda(nh^p)^{-1} + B_3 + B_4 h^4 - B_5\lambda h^2 + B_6\lambda^2 - n^{-1/2}\mathcal{Z}_n + (s.o.) \\
&= B_1(nh^p)^{-1} + B_4 h^4 - B_5\lambda h^2 + B_6\lambda^2 + (\text{terms independent of } (\lambda, h)) + (s.o.) \\
&= B_6[\lambda - B_5 h^2/(2B_6)]^2 + [B_4 - B_5^2/(4B_6)]h^4 + B_1(nh^p)^{-1} \\
&\quad + (\text{terms independent of } (\lambda, h)) + (s.o.). \qquad \text{(B.19)}
\end{aligned}$$

Minimizing Equation (B.19) over (λ, h) leads to $h^4 = O_p((nh^p)^{-1})$ and $\lambda = O_p(h^2)$, which lead to $\hat{h} = O_p(n^{-1/(4+p)})$ and $\hat{\lambda} = O_p(n^{-2/(4+p)})$.

Next for (ii), similar to (i) above, one can easily show that $[\hat{f}(z) - f(z)]^2 = O_p(\hat{\lambda}^2 + \hat{h}^4 + \hat{\lambda}\hat{h}^2 + (n\hat{h}^p)^{-1}) = O_p(\hat{h}^4) = O_p(n^{-4/(4+p)})$ by (i). Hence, $\hat{f}(z) - f(z) = O_p(n^{-2/(4+p)})$.

References

Aerts, M., Augustyns, I., and Janssen, P. (1997a) Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*, 8, 127-147.

Aerts, M., Augustyns, I., and Janssen, P. (1997b) Local polynomial estimation of contingency table cell probabilities. *Statistics*, 30, 127-148.

Ahmad, I.A. and P.B. Cerrito (1994) Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference* 41, 349-364.

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

Bowman, A.W. (1985). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika* 67, 682-684.

Dong, J. and J.S. Simonoff (1994). The construction and properties of boundary kernels for sparse multinomials. *Journal of Computational and Graphical Statistics* 3, 57-66.

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

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

Grund, B. (1993) Kernel estimators for cell probabilities. *Journal of Multivariate Analysis* 46, 283-308.

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. (1987a). "On Kullback-Leibler loss and density estimation," *Ann. Statist.* 15, 1491-1519.

Hall, P. (1987b). "On the use of compactly supported densities in problems of discrimination," *J. Multivar. Anal.* 23, 131-158.

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

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

Hart, J.D. (1997) Nonparametric Smoothing and Lack-of-fit Tests. New York: Springer-Verlag.

Izenman, A. (1991) "Recent developments in nonparametric density estimation," *Journal of the American Statistical Association* 413, 205–224.

Scott, D. (1992) Multivariate density estimation: theory, practice, and visualization. *John Wiley and Sons.*

Simonoff, J.S. (1983) A penalty function approach to smoothing to smoothing large sparse contingency tables. *Annals of Statistics* 11, 208-218.

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

Tutz, G. (1991) "Consistency of cross-validatory choice of smoothing parameters for direct kernel estimates," *Computational Statistics Quarterly* 4, 295-314.