Available online at www.sciencedirect.com
SCIENCE DIRECT®
JOURNAL OF Econometrics
Journal of Econometrics 119 (2004) 99–130
www.elsevier.com/locate/econbase

Nonparametric estimation of regression functions with both categorical and continuous data

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

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

Accepted 19 March 2003

Abstract

In this paper we propose a method for nonparametric regression which admits continuous and categorical data in a natural manner using the method of kernels. A data-driven method of bandwidth selection is proposed, and we establish the asymptotic normality of the estimator. We also establish the rate of convergence of the cross-validated smoothing parameters to their benchmark optimal smoothing parameters. Simulations suggest that the new estimator performs much better than the conventional nonparametric estimator in the presence of mixed data. An empirical application to a widely used and publicly available dynamic panel of patent data demonstrates that the out-of-sample squared prediction error of our proposed estimator is only 14–20% of that obtained by some popular parametric approaches which have been used to model this data set.

JEL classification: C14; C13

Keywords: Discrete variables; Nonparametric smoothing; Cross-validation; Asymptotic normality

1. Introduction and background

One of the most appealing features of nonparametric estimation techniques is that, by allowing the data to model the relationships among variables, they are robust to functional form specification and therefore have the ability to detect structure which sometimes remains undetected by traditional parametric estimation techniques. In light of this feature, it is not surprising that nonparametric techniques have attracted the attention of econometricians as is underscored by the tremendous literature on

* Corresponding author. Tel.: +1-979-845-9954; fax: +1-979-847-8757.
E-mail addresses: jracine@maxwell.syr.edu (J. Racine), qi@econmail.tamu.edu (Q. Li).

0304-4076/$ - see front matter
doi:10.1016/S0304-4076(03)00157-X

nonparametric estimation and inference which has recently appeared in leading economics journals.

Along with the development of nonparametric techniques, it is evident that applications of nonparametric methods are also on the rise as is witnessed by the recent special issue on *Application of Semiparametric Methods for Micro-Data* in the Journal of Applied Econometrics (Vol. 13, 1998), and the monograph of Horowitz (1998) which contains some interesting empirical applications.

When compared with the vast theoretical literature, however, the number of empirical applications of nonparametric techniques appears to be relatively sparse. One frequently cited reason as to why nonparametric techniques have not been more widely used is because economic data frequently contain both continuous and categorical variables such as gender, family size, or choices made by economic agents, and standard nonparametric estimators do not handle categorical variables satisfactorily. The conventional nonparametric approach uses a 'frequency estimator' to handle the categorical variables which involves splitting the sample into a number of subsets or 'cells'. When the number of cells in a data set is large, each cell may not have enough observations to nonparametrically estimate the relationship among the remaining continuous variables. Perhaps for this reason many authors suggest treating categorical variables as parametric components, thereby retreating to a semiparametric framework from a fully nonparametric one. For example, Stock (1989) proposed the estimation of a partially linear model where the discrete variables enter the model in a linear fashion while the continuous variables enter the model nonparametrically, while Fan et al. (1998) considered the estimation of additive partially linear models where the discrete variables again enter the model in a linear fashion.

It is evident, therefore, that the recurring issue of how best to handle mixed categorical and continuous data in a nonparametric framework remains unsettled. In this paper we shall draw upon the work of Aitchison and Aitken (1976) who proposed a novel extension of the kernel method of density estimation to a discrete data setting in a multivariate binary discrimination context. A key feature of their technique is that it allows the data points themselves to determine any dependencies and interactions in the estimated density function. We continue with this line of inquiry and propose a natural extension of Aitchison and Aitken's (1976) work to the problem of mixed categorical and continuous data in a nonparametric regression framework.[1] The proposed method does not split the sample into cells in finite-sample applications and it handles interaction among the categorical and continuous variables in a natural manner. The strength of the proposed method lies in its ability to model situations involving complex dependence among categorical and continuous data in a fully nonparametric regression framework.

[1] There is a rich literature in statistics on smoothing discrete variables (see Hall (1981), Hall and Wand (1988), Grund and Hall (1993), Fahrmeir and Tutz (1994), Scott (1992), and Simonoff (1996), among others). When faced with a mix of discrete and continuous regressors, the only *theoretical* work on smoothing the mixed regressors that we are aware of are the works by Bierens (1983, 1987), and Ahmad and Cerrito (1994). However, neither of these articles study the fundamental issue of data-driven selection of smoothing parameters. Delgado and Mora (1995) consider a semiparametric partially linear specification with discrete regressors, but they did not smooth the discrete regressors.

The paper is organized as follows. Section 2 presents our kernel estimator of a conditional mean function and establishes the asymptotic normality of the proposed estimator. We provide the rate of convergence of the cross-validated smoothing parameters to their optimal values, and in the case of $p \leqslant 3$ (p is the dimension of the continuous regressors), we also obtain asymptotic normality results for these cross-validated smoothing parameters. Section 3 reports some simulation results which examine the finite-sample performance of the proposed estimator. We apply the new estimation method to a publicly available data set in Section 4, whereby we consider the nonparametric estimation of a dynamic panel of patent data to generate out-of-sample predictions. We show that the out-of-sample squared prediction error of our proposed estimator is only 14–20% of that obtained by some popular parametric approaches which have been used to model this data set. Section 5 concludes the paper and also suggests some future research topics. All technical proofs are relegated to two appendices.

2. Consistent kernel regression with discrete and continuous variables

We consider a nonparametric regression model where a subset of regressors is categorical and the remaining are continuous. Let X_i^d denote a $k \times 1$ vector of regressors that assume discrete values and let $X_i^c \in R^p$ denote the remaining continuous regressors. We use $X_{t,i}^d$ to denote the tth component of X_i^d, and we assume that $X_{t,i}^d$ can assume $c_t \geqslant 2$ different values, i.e., $X_{t,i}^d \in \{0, 1, \ldots, c_t - 1\}$ for $t = 1, \ldots, k$. Define $X_i = (X_i^d, X_i^c)$. We consider a nonparametric regression model given by

$$Y_i = g(X_i) + u_i, \tag{2.1}$$

where $g(\cdot)$ has an unknown functional form. We use $f(x) = f(x^c, x^d)$ to denote the joint density function of (X_i^c, X_i^d).

For the discrete variables X_i^d, we will first consider the case for which there is no natural ordering in X_i^d. The extension to the general case whereby some of the discrete regressors have natural orderings will be discussed at the end of this section.

We use $\mathscr{D} = \prod_{t=1}^k \{0, 1, \ldots, c_t - 1\}$ to denote the range assumed by X_i^d. For x^d, $X_i^d \in \mathscr{D}$. Aitchison and Aitken (1976) suggested smoothing the discrete regressors x_t^d by using a univariate kernel function given by $\tilde{l}(X_{t,i}^d, x_t^d, \lambda) = 1 - \lambda$ if $X_{t,i}^d = x_t^d$, and $\tilde{l}(X_{t,i}^d, x_t^d, \lambda) = \lambda/(c_t - 1)$ if $X_{t,i}^d \neq x_t^d$, where λ is a smoothing parameter. The product kernel for the discrete variables is then defined to be $\tilde{L}(X_i^d, x^d, \lambda) = \prod_{t=1}^k \tilde{l}(X_{t,i}^d, x_t^d, \lambda)$. In this paper we use a different kernel function which has a simpler form than the one suggested by Aitchison and Aitken (1976), and the simpler form makes it much easier to generalize our results to cover the ordered categorical variable case. Define[2]

$$l(X_{t,i}^d, x_t^d, \lambda) = \begin{cases} 1 & \text{if } X_{t,i}^d = x_t^d, \\ \lambda & \text{if } X_{t,i}^d \neq x_t^d. \end{cases} \tag{2.2}$$

[2] Note that the kernel weights add up to $1 + (c_t - 1)\lambda \neq 1$ for $\lambda \neq 0$, but this does not affect the nonparametric estimator defined in Eq. (2.6) because the kernel function appears in both the numerator and the denominator of Eq. (2.6), thus the kernel function can be multiplied by any positive constant without changing the definition of $\hat{g}(x)$.

Note that when $\lambda = 0$, the above kernel function $l(X_{t,i}^d, x^d, 0)$ becomes an indicator function which takes value 1 if $X_{t,i}^d = x_t^d$, and 0 otherwise. If $\lambda = 1$, $l(X_{t,i}^d, x^d, 1) \equiv 1$ becomes a constant. The range of λ is $[0, 1]$.

Define an indicator function $\mathbf{1}(X_{t,i}^d \neq x_t^d)$, which takes value 1 if $X_{t,i}^d \neq x_t^d$ and 0 otherwise. Also, define $d_{x_i,x} = \sum_{t=1}^k \mathbf{1}(X_{t,i}^d \neq x_t^d)$, which equals the number of disagreeing components between X_i^d and x^d. Then the product kernel for the discrete variables is defined by

$$L(X_i^d, x^d, \lambda) = \prod_{t=1}^k l(X_{t,i}^d, x_t^d, \lambda) = 1^{k - d_{x_i,x}} \lambda^{d_{x_i,x}} = \lambda^{d_{x_i,x}}. \tag{2.3}$$

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

We use $W(\cdot)$ to denote the kernel function associated with the continuous variables x^c and h to denote the smoothing parameters for the continuous variables. Using the shorthand notation $K_{h,ix} = W_{h,ix} L_{\lambda,ix}$, where $W_{h,ix} = h^{-p} W((X_i^c - x^c)/h)$ and $L_{\lambda,ix} = L(X_i^d, x^d, \lambda)$, the kernel estimator of $f(x)$, the joint density function of (X_i^c, X_i^d), is given by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_{h,ix}. \tag{2.4}$$

It is well known that one needs conditions such as $h \to 0$ and $nh^p \to \infty$ as $n \to \infty$ in order to obtain a consistent estimator when using kernel methods with only continuous variables. For the discrete variable case we need $\lambda \to 0$, while for the mixed continuous and discrete variable case we need both sets of conditions.

Let $v(y, x)$ denote the joint density function of (Y_i, X_i). First, we consider the case where the endogenous variable Y_i is continuous. In this case we estimate the joint density of (Y_i, X_i) by

$$\hat{v}(y, x) = n^{-1} \sum_{i=1}^n h^{-1} w((Y_i - y)/h) K_{h,ix}, \tag{2.5}$$

where $w(\cdot)$ is a univariate kernel function satisfying assumption (A1) (ii) (see below). Then we will estimate $g(x) = \mathrm{E}[Y_i | X_i = x]$ by

$$\hat{g}(x) = \frac{\int y \hat{v}(y, x)\, \mathrm{d}y}{\hat{f}(x)} = \frac{n^{-1} \sum_{i=1}^n Y_i K_{h,ix}}{\hat{f}(x)}, \tag{2.6}$$

where we have used $\int w(v)\, \mathrm{d}v = 1$ and $\int v\, w(v)\, \mathrm{d}v = 0$. When $\lambda = 0$, our estimator reverts back to the conventional approach whereby one uses a frequency estimator to deal with the discrete variables. The conventional frequency estimator possesses a major weakness, however, being that it often cannot be applied when the number of cells is large relative to the sample size since one may not have enough (any) observations in each cell to conduct nonparametric estimation. In contrast, smoothing the discrete variables will be seen to avoid the problem, while the resulting finite-sample efficiency gains can be quite substantial.[3]

Ahmad and Cerrito (1994) consider using more general discrete kernel functions which include the kernel function (2.3) as a special case (see also Wang and Van Ryzin (1981) for various approaches to the smoothing of discrete variables). Therefore, using $\hat{g}(x)$ defined in (2.6) as a kernel estimator for $g(x)$ can be viewed as a special case of Ahmad and Cerrito (1994). However, the more general kernel functions used in Ahmad and Cerrito (1994) also render the asymptotic analysis more difficult. With the simple kernel function defined in (2.3), we are able to derive the rates of convergence of the cross-validated smoothing parameters to some benchmark optimal values (see Theorems 2.2 and 2.4). In the continuous variable case, it is known that the choice of the kernel function is not very important, while the selection of smoothing parameters is of crucial importance for the behavior of the nonparametric estimator. Our experience with the discrete variable case is similar: the choice of the discrete kernel function (such as the more general ones in Ahmad and Cerrito (1994) or the simple one given in this paper) is of much less importance than the selection of the smoothing parameters. In this paper we advocate the use of the least-squares cross-validation method for selecting both h and λ, and we demonstrate that this method works quite well for simulated and real data (see also Li and Racine (2001, 2003)).

Next we consider the case where Y_i is a discrete variable. Let $D_y = \{0, 1, \ldots, c_y - 1\}$ denote the range of Y_i. In this case we estimate $v(y,x)$ by $\bar{v}(y,x) = n^{-1}\sum_{i=1}^n \mathbf{1}(Y_i = y)K_{h,ix}$, where $\mathbf{1}(.)$ denotes an indicator function, $\mathbf{1}(Y_i = y) = 1$ if $Y_i = y$ and 0 otherwise. Therefore, we estimate $g(x)$ by

$$\bar{g}(x) = \frac{\sum_{y\in D_y} y\bar{v}(y,x)}{\hat{f}(x)} = \frac{n^{-1}\sum_{i=1}^n \sum_{y\in D_y} y\mathbf{1}(Y_i = y)K_{h,ix}}{\hat{f}(x)}$$
$$= \frac{n^{-1}\sum_{i=1}^n Y_i K_{h,ix}}{\hat{f}(x)}, \tag{2.7}$$

where we have used $\sum_{y\in D_y} y\mathbf{1}(Y_i = y) = Y_i$. We see that $\bar{g}(x)$ has exactly the same form as $\hat{g}(x)$ defined in (2.6). Hence, we will use $\hat{g}(x)$ to denote the kernel estimator of $g(x)$ regardless of whether or not Y_i is continuous or discrete.

As noted above, it is known that the choice of smoothing parameters is of crucial importance in nonparametric kernel estimation. We choose (λ, h) to minimize

$$CV(\lambda, h) = \sum_{i=1}^n [Y_i - \hat{g}_{-i}(X_i)]^2 M(X_i), \tag{2.8}$$

[3] One can also view this method as the classic trade-off between bias and variance. Although the frequency estimator is unbiased (in the case with only discrete regressors), it can have a huge variance. The new estimator on the other hand introduces some finite-sample bias, but it can reduce the variance significantly resulting in much better finite-sample performance than that of the conventional frequency estimator.

where $M(\cdot)$ is a weight function that trims out boundary observations,

$$\hat{g}_{-i}(X_i) = \frac{n^{-1}\sum_{j\neq i} Y_j K_{h,ij}}{\hat{f}_{-i}(X_i)} \tag{2.9}$$

is the leave-one-out kernel estimator of $g(X_i)$, $K_{h,ij}=L_{\lambda,ij}W_{h,ij}$, $L_{\lambda,ij}=L(X_i^d,X_j^d,\lambda)$, $W_{h,ij} = h^{-p}W((X_i^c - X_j^c)/h)$, and

$$\hat{f}_{-i}(X_i) = \frac{1}{n}\sum_{j\neq i} K_{h,ij} \tag{2.10}$$

is the leave-one-out estimator of $f(X_i)$. Note that usually one uses the factor $n-1$ rather than n in defining the leave-one-out kernel estimators $\hat{g}_{-i}(X_i)$ and $\hat{f}_{-i}(X_i)$. However, this will not change the result as the factor $n-1$ (or n) cancels out in the definition of $\hat{g}_{-i}(X_i)$. Using the factor of n rather than $n-1$ simplifies the notation and saves space as can be seen in Appendix B where we provide proofs of the main results of the paper.

We now list the assumptions that will be used to establish the asymptotic distribution of $\hat{g}(x)$.

Let $\mathscr{G}_\mu^\alpha$ denote the class of smooth functions introduced in Robinson (1988) ($\alpha > 0$, μ is a positive integer). That is, if $m(x^c)\in\mathscr{G}_\mu^\alpha$ (recall that x^c is a continuous variable), then $m(x^c)$ is μ times differentiable, and $m(x^c)$ and its partial derivatives (up to order μ) are all bounded by functions that have finite αth moment (e.g., Robinson, 1988).

(A1) (i) We restrict $(\hat{\lambda},\hat{h})$ to lie in a shrinking set $\Lambda_n\times H_n$, where $\Lambda_n=[0,\min\{1,C_0(\log n)^{-1}\}]$ and $H_n=[\underline{h},\bar{h}]$, $\underline{h}\geqslant C^{-1}n^{\delta-1/p}$, $\bar{h}\leqslant Cn^{-\delta}$ for some $C_0,C,\delta>0$. (ii) The kernel function $W(\cdot)$ is the product kernel defined by $W(v)=\prod_{t=1}^{p} w(v_t)$ (v_t is the tth component of v), while the univariate function $w(\cdot)$ is nonnegative, symmetric and bounded with $\int w(v)v^4\,\mathrm{d}v<\infty$. Moreover, $w(\cdot)$ is m times differentiable. Letting $w^{(s)}(\cdot)$ denote the sth-order derivative of $w(\cdot)$, then $\int|w^{(s)}(v)v^s|\,\mathrm{d}v<\infty$ for all $s=1,\dots,m$, where $m>\max\{2+4/p,1+p/2\}$ is a positive integer. (iii) For all $x^d\in\mathscr{D}$, $M(\cdot,x^d)$ is bounded and supported on a compact set with nonempty interior for all $x^d\in\mathscr{D}$. (iv) $f(x)$ is bounded below on the support of $M(\cdot)$.

(A2) (i) $\{X_i,Y_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) as (X,Y), $u_i = Y_i - g(X_i)$ has finite fourth moment. (ii) Defining $\sigma^2(x)=\mathrm{E}[u_i^2|X_i=x]$, $\sigma^2(\cdot,x^d)$, $g(\cdot,x^d)$ and $f(\cdot,x^d)$ all belong to $\mathscr{G}_2^4$ for all $x^d\in\mathscr{D}$. (iii) Denote by $\nabla g_i=\partial g(v,X_i^d)/\partial v]|_{v=X_i^c}$, and $\nabla f_i=\partial f(v,X_i^d)/\partial v]|_{v=X_i^c}$. Define $B_{1,1}(X_i)=\{\nabla f_i'\nabla g_i+(1/2)f(X_i)\int tr(\nabla^2 g_i)\}[\int w(v)v^2\,\mathrm{d}v]$ and $B_{1,2}(X_i)=\mathrm{E}[g(X_i^c,X_j^d)-g(X_i^c,X_i^d)|X_i,d_{ij}=1]\mathrm{P}(d_{ij}=1|X_i)$ ($d_{ij}=d_{x_i,x_j}$), $B_1=\mathrm{E}[(B_{1,1}(X_i)/f_i)^2]$, $B_2=2\mathrm{E}[B_{1,1}(X_i)B_{1,2}(X_i)/f_i^2]$ and $B_3=\mathrm{E}[(B_{1,2}(X_i)/f_i)^2]$, where $\nabla^2 g_i=\partial g(v,X_i^d)/\partial v\,\partial v']|_{v=X_i^c}$. Then $4B_1B_3-B_2^2>0$.

The requirement that $\hat{\lambda}$ and $\hat{h}$ lie in a shrinking set is not as restrictive as it may appear, since otherwise the kernel estimator will have a nonvanishing bias term resulting in an inconsistent estimator. The two conditions on h in (A1) (i) are also used in Härdle and Marron (1985), and these 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 identical to the usual conditions $h\to 0$ and $nh^p\to\infty$. (A1) (ii) requires that the kernel function is differentiable up to order m, and this condition is used to show that a remainder term in a Taylor expansion of $W((X_i^c - x^c)/\hat{h})$ will have a negligible order, where $\hat{h}$ is the cross-validation choice of h. We note that the widely used standard normal kernel satisfies (A1) (ii). This condition can be replaced with a compactly supported kernel function that is Hölder continuous requiring a different type of proof such as Lemma 2 in Härdle et al. (1988). (A1) (iii) and (iv) allow a uniform convergence rate for $\hat{f}(x)$ and $\hat{g}(x)$. (A2) (i) and (ii) contain some standard moment and smoothness conditions. By Cauchy's inequality we know that $4B_1B_3 - B_2^2 \geqslant 0$. (A2) (iii) rules out the case of a regression function that is in fact independent of x^c.

An anonymous referee has correctly noted that if the regression function is independent of a discrete variable, then the cross-validation method will not select a small value of λ for this variable (it is widely known that this will also occur for h for a continuous variable in such instances). Instead the cross-validation method will tend to select $\lambda = 1$ (this discrete variable is smoothed out) and lead to a more efficient estimation result. However, the distribution of $\hat{\lambda}$ in this case is quite complicated and is beyond the scope of this paper.

In Appendix A we show that the leading term of $CV(h,\lambda)$ is CV_0 which is given by

$$CV_0 = B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + B_4(nh^p)^{-1}, \tag{2.11}$$

where the B_j's are some constants. Letting λ_0 and h_0 denote the values of λ and h that minimize $CV_0(h,\lambda)$, then it is easy to show that $h_0 = c_1n^{-1/(4+p)}$ and $\lambda_0 = c_2n^{-2/(4+p)}$, where c_1 and c_2 are some constants which are defined in Appendix A. The values of h_0 and λ_0 can be interpreted as the nonstochastic optimal smoothing parameters because it can be shown that h_0 and λ_0 also minimize the leading term of the nonstochastic objective function $\mathrm{E}[CV(h,\lambda)]$.

In Appendix A we also show that $(\hat{h} - h_0)/h_0 = o_p(1)$ and $(\hat{\lambda} - \lambda_0)/\lambda_0 = o_p(1)$. Therefore, both $\hat{h}/h_0$ and $\hat{\lambda}/\lambda_0$ converge to one in probability. Let $\tilde{g}(x)$ be defined the same way as $\hat{g}(x)$ except that $(\hat{h},\hat{\lambda})$ are replaced by (h_0,λ_0), i.e.,

$$\tilde{g}(x) = \frac{n^{-1}\sum_i Y_i W_{h_0}((X_i^c - x^c)/h_0)L(X_i^d, x^d, \lambda_0)}{\tilde{f}(x)} \tag{2.12}$$

where $W_{h_0}((X_i^c - x^c)/h_0) = h_0^{-p}W((X_i^c - x^c)/h_0)$ and

$$\tilde{f}(x) = \frac{1}{n}\sum_i W_{h_0}\left(\frac{X_i^c - x^c}{h_0}\right)L(X_i^d, x^d, \lambda_0) \tag{2.13}$$

is the kernel estimator of $f(x)$ using the nonstochastic smoothing parameters (h_0,λ_0).

We first present the asymptotic distribution of $\tilde{g}(x)$, and then we will show that $\hat{g}(x)$ has the same asymptotic distribution as $\tilde{g}(x)$.

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

$$\sqrt{nh_0^p}(\tilde{g}(x) - g(x) - B(h_0,\lambda_0)) \to \mathrm{N}(0,\Omega(x)) \textit{ in distribution},$$

where $B(h_0,\lambda_0) = h_0^2\{\nabla f(x)'\nabla g(x)/f(x) + tr[\nabla^2 g(x)]/2\}[\int w(v)v^2\,\mathrm{d}v] + \lambda_0 \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} [g(x^c,\tilde{x}^d) - g(x)]f(x^c,\tilde{x}^d)/f(x)$, *and* $\Omega(x) = \sigma^2(x)[\int W^2(v)\,\mathrm{d}v]/f(x)$.

In order to establish the asymptotic distribution of $\hat{g}(x)$, we will first derive the rates of convergence of $(\hat{h} - h_0)/h_0$ and $(\hat{\lambda} - \lambda_0)$.

Theorem 2.2. *Under the same conditions as in Theorem* 2.1, *we have*

(i) *If* $p \leqslant 3$, $(\hat{h} - h_0)/h_0 = O_p(n^{-p/[2(4+p)]})$ *and* $\hat{\lambda} - \lambda = O_p(n^{-1/2})$.
(ii) *If* $p \geqslant 4$, $(\hat{h} - h_0)/h_0 = O_p(h_0^2) = O_p(n^{-2/(4+p)})$ *and* $\hat{\lambda} - \lambda_0 = O_p(h_0^4) = O_p(n^{-4/(4+p)})$.

Using the result of Theorem 2.2 and a Taylor expansion argument, we show that $\hat{g}(x) - g(x) - B(h_0,\lambda_0) = \tilde{g}(x) - g(x) - B(h_0,\lambda_0) + (s.o.)$, where $(s.o.)$ means smaller order terms. Hence, $\hat{g}(x)$ has the same asymptotic distribution as that of $\tilde{g}(x)$. We give this result in the next theorem.

Theorem 2.3. *Let* $\hat{\lambda}$ *and* $\hat{h}$ *denote the cross-validation choices of* λ *and* h *that minimize Eq.* (2.8). *Under assumptions* (A1) *and* (A2), *we have*

(i) $\sqrt{n\hat{h}^p}(\hat{g}(x) - g(x) - B(h_0,\lambda_0)) = \sqrt{nh_0^p}(\tilde{g}(x) - g(x) - B(h_0,\lambda_0)) + o_p(1) \to N(0,\Omega(x))$ *in distribution, where* $B(h_0,\lambda_0)$ *and* $\Omega(x)$ *are defined in Theorem* 2.1.
(ii) *Define* $\hat{B}(\hat{h},\hat{\lambda}) = \hat{h}^2\{\nabla\hat{f}(x)'\nabla\hat{g}(x)/\hat{f}(x) + tr[\nabla^2\hat{g}(x)]/2\}[\int w(v)v^2\,\mathrm{d}v] + \hat{\lambda}\sum_{\tilde{x}^d, d_{\tilde{x},x}=1} [\hat{g}(x^c,\tilde{x}^d) - \hat{g}(x^c,x^d)]\hat{f}(x^c,\tilde{x}^d)/\hat{f}(x)$, $\hat{\Omega}(x) = \hat{\sigma}^2(x)[\int W^2(v)\,\mathrm{d}v]/\hat{f}(x)$ *and* $\hat{\sigma}^2(x) = n^{-1}\sum_i [Y_i - \hat{g}(X_i)]^2 W_{\hat{h},ix} L_{\hat{\lambda},ix}/\hat{f}(x)$. *Then*

$$\sqrt{n\hat{h}^p}(\hat{g}(x) - g(x) - \hat{B}(\hat{h},\hat{\lambda}))/\sqrt{\hat{\Omega}(x)} \to N(0,1) \text{ in distribution.}$$

Theorem 2.3 demonstrates that the convergence rate of $\hat{g}(x)$ is the same as the case where there are continuous regressors x^c only. Indeed, when there are no discrete variables $(x = x^c)$, Theorems 2.2 and 2.3 collapse to the well-known case with only continuous regressors. However, when there are no continuous regressors, it can be shown that the cross-validation choice of λ will converge to zero at the rate of $O_p(n^{-1})$. This result cannot be easily obtained as a corollary of Theorem 2.3, and a separate proof is needed to show this. This proof for the discrete-regressor-only case is available from the authors upon request.

When proving Theorem 2.2 for the rates of convergence of $(\hat{h} - h_0)/h_0$ and $\hat{\lambda} - \lambda_0$, we have shown that, for $p \leqslant 3$, the leading terms of both $\sqrt{n}(\hat{\lambda} - \lambda_0)$ and $n^{p/[2(4+p)]}(\hat{h} - h_0)/h_0$ are some mean-zero $O_p(1)$ random variables. In fact, one can further show that these mean-zero $O_p(1)$ random variables have asymptotic normal distributions.

Theorem 2.4. *Under the same conditions as in Theorem* 2.2 *and for* $p \leqslant 3$, *we have*

$$\sqrt{n}(\hat{\lambda} - \lambda_0) \to N(0, V_1) \text{ in distribution,}$$

and

$$n^{p/[2(4+p)]}(\hat{h} - h_0) \rightarrow \mathrm{N}(0, V_2) \text{ in distribution,}$$

where V_1 and V_2 are two finite positive constants.

The exact expressions of V_1 and V_2 are complicated, therefore, we do not give the explicit expressions for them here. But we give sufficient details about them in the proof of Theorem 2.4 in Appendix A, where we show that they are the asymptotic variances of some $\mathrm{O_p}(1)$ U-statistics.

Härdle et al. (1988) derived the asymptotic distribution of $(\hat{h} - h_0)/h_0$ for a model with a univariate nonstochastic regressor (see also Härdle et al. (1992) on the use of 'double smoothing' to improve the rate of convergence of $(\hat{h} - h_0)/h_0$). Here, we generalize the result of Härdle et al. (1988) to the case of $p \leqslant 3$ augmented by a $k \times 1$ vector of discrete regressors. Upon inspection of the proofs of Theorems 2.2 and 2.4, it can be seen that even for $p = 4$, one can still establish the asymptotic normality of $\sqrt{n}(\hat{\lambda} - \lambda_0 - \mu_1)$ and $n^{p/[2(4+p)]}(\hat{h} - h_0 - \mu_2)/h_0$, where μ_1 and μ_2 are some constants. The extra nonzero center terms μ_1 and μ_2 come from the contribution of the A_{2n} term because, when $p = 4$, A_{2n} has the same order as A_{1n} (see Appendix A for the definitions of A_{1n} and A_{2n}). We do not formally establish this result for space considerations.

2.1. The general categorical data case: some regressors have a natural ordering

Up to now we have assumed that the discrete variables do not have a natural ordering, examples of which would include different regions, ethnicity, and so forth. We now examine the extension of the above results to the case where a discrete variable has a natural ordering, examples of which would include preference orderings (like, indifference, dislike), health (excellent, good, poor), or discrete representations of some inherently continuous variables.[4]

Using the same notation as above, let x_t be the tth component of x and suppose that x_t can assume $c_t \geqslant 2$ different values ($t = 1, \ldots, k$). Aitchison and Aitken (1976, p. 29) suggest the kernel weight function given by $l(X_{i,t}, x_t, \lambda) = \binom{c_t}{s} \lambda^s (1-\lambda)^{c_t - s}$ when $|X_{i,t} - x_t| = s$ $(0 \leqslant s \leqslant c_t)$, where $\binom{c_t}{s} = c_t!/[s!(c_t - s)!]$. These weights add up to one because $1 = [(1-\lambda) + \lambda]^{c_t}$. While there is no doubt that one can extend the results of Theorems 2.1–2.4 to cover this case, such an extension would be quite tedious. Therefore, we suggest the use of a simple kernel function defined by $l(X_{i,t}, x_t) = \lambda^s$ when $|X_{i,t} - x_t| = s$ $(0 \leqslant s \leqslant c_t)$, where λ is the smoothing parameter. In this case the

[4] In the linear regression model case, it may be possible to retrieve information related to the original continuous variables such as partial effects (see Hsiao (1983), and Hsiao and Mountain (1985)). However, in the nonparametric regression case, this type of information retrieval does not seem possible because we do not impose any functional form assumptions.

product kernel function is given by

$$L(X_i, x, \lambda) = \prod_{t=1}^{k} \lambda^{|X_{i,t}^d - x_t^d|} = \lambda^{\delta_{x_i,x}}, \tag{2.14}$$

where $\delta_{x_i,x} = \sum_{t=1}^{k} |X_{i,t}^d - x_t^d|$ is the L_1 distance between X_i^d and x^d.

We see that Eq. (2.14) has a form identical to that of Eq. (2.3) except that $d_{x_i,x}$ is replaced by $\delta_{x_i,x}$. In particular, the estimation bias will be of order $O(\lambda)$.

In practice, it is likely that some of the discrete variables have natural orderings while others will not. Let $\tilde{X}_i^d$ denote a $k_1 \times 1$ vector (say, the first k_1 components of X_i^d) of discrete regressors that have a natural ordering ($1 \leqslant k_1 \leqslant k$), and let $\bar{X}_i^d$ denote the remaining discrete regressors that do not have a natural ordering. In this case, the product kernel will be of the form

$$L(X_i^d, x^d, \lambda) = \left[\prod_{t=1}^{k_1} \lambda^{|\tilde{X}_{i,t}^d - \tilde{x}_t^d|}\right] [\lambda^{d_{\bar{x}_i,\bar{x}}}] = \lambda^{\delta_{\tilde{x}_i,\tilde{x}} + d_{\bar{x}_i,\bar{x}}}, \tag{2.15}$$

where $\delta_{\tilde{x}_i,\tilde{x}} = \sum_{t=1}^{k_1} |\tilde{X}_{t,i}^d - \tilde{x}_t^d|$ is the L_1 distance between $\tilde{X}_i^d$ and $\tilde{x}^d$, and $d_{\bar{x}_i,\bar{x}}$ equals the number of disagreeing components between $\bar{X}_i^d$ and $\bar{x}^d$.

The results of Theorem 2.3 can be easily extended to the general case when some (or all) of the discrete regressors have a natural ordering as the following corollary demonstrates.

Corollary 2.1. *Under the same conditions found in Theorem* 2.3 *with the first* k_1 *components of* X_i^d *being ordered discrete variables* ($1 \leqslant k_1 \leqslant k$), *let* $\hat{g}(x)$ *be defined as in Eq.* (2.6) *with the kernel function* $L(\cdot)$ *being defined by Eq.* (2.15).

Then the conclusion of Theorem 2.3 *remains unchanged.*

The proof of Corollary 2.1 is identical to the proof of Theorem 2.3 and is thus omitted.

We now turn our attention to the finite-sample behavior of the proposed estimator.

3. Monte Carlo results—finite-sample performance

For what follows we shall compute the out-of-sample mean-square error using $n_2^{-1} \sum_i^{n_2} (Y_i - \hat{Y}_i)^2$ where Y_i and $\hat{Y}_i$ are the actual and predicted values for an independent evaluation sample.

The first data generating process (DGP) which we consider is given by

$$Y_i = \sum_{t=1}^{4} \beta_t X_{t,i} + \sum_{t=1}^{4} \sum_{s \neq t, s=1}^{4} \beta_{t,s} X_{t,i} X_{s,i} + \sum_{t=1}^{4} X_{t,i} m_1(Z_i) + m_2(Z_i) + u_i, \tag{3.1}$$

where, for $t = 1, \dots, 4$, $X_{t,i} \in \{0, 1\}$ with $P(X_{ti} = l) = 0.5$ for $l = 0, 1$, $m_1(Z_i) = \sin(Z_i \pi)$, and $m_2(Z_i) = Z_i - 0.5Z_i^2 + 0.3Z_i^3$, Z_i is uniformly distributed on the interval $[0, 2]$, and u_i is $N(0, 1)$. We choose $\beta_t = 1$ and $\beta_{t,s} = 0.5$ for all $t, s = 1, \dots, 4$.

Table 1
Finite-sample estimator comparison

n_1	Model	Mean MSE	Median MSE	SD (MSE)	IQR(MSE)
100	NP	2.30	2.24	0.40	0.47
	NP-FREQ	2.59	2.47	0.65	0.43
	PAR	1.22	1.21	0.08	0.10
	PAR-LIN	2.92	2.91	0.14	0.18
200	NP	1.64	1.62	0.16	0.18
	NP-FREQ	1.77	1.75	0.15	0.20
	PAR	1.10	1.10	0.04	0.05
	PAR-LIN	2.84	2.83	0.11	0.14
500	NP	1.27	1.26	0.05	0.06
	NP-FREQ	1.30	1.29	0.05	0.06
	PAR	1.04	1.04	0.01	0.02
	PAR-LIN	2.79	2.78	0.09	0.12

We compute predictions from our nonparametric model (NP), the nonparametric models with $\lambda = 0$ (NP-FREQ) which is the conventional frequency estimator, a correctly specified parametric model (PAR), and a misspecified linear parametric model with no interaction terms (PAR-LIN). We report the mean, median, standard error, and interquartile range of MSE over the 1,000 Monte Carlo replications. The estimation samples are of size n_1 (100, 200, and 500), and the independent evaluation sample is always of size $n_2 = 1,000$.

From Table 1 we observe that our proposed nonparametric estimator dominates both the conventional frequency nonparametric estimator and the estimator based on a misspecified linear model, while it converges quite quickly to the correctly specified benchmark parametric model.

3.1. A comparison of unordered and ordered kernel types

The second DGP which we consider is given by

$$Y_i = Z_{i1} + Z_{i2} + X_{i1} + X_{i2} + u_i, \tag{3.2}$$

where $X_{ij} \sim \mathrm{N}(0,1)$, $Z_{ij} \in \{0,1,\dots,5\}$ with $\mathrm{P}(Z_{ij} = l) = 1/6$ for $l = 0,\dots,5$ and $u_i \sim \mathrm{N}(0,1)$.

We consider three nonparametric estimators differing by their kernel functions—the unordered kernel, ordered kernel, and the frequency approach. We expect that, the smaller the ratio of the sample size to the number of 'cells', the worse the nonparametric frequency approach relative to our proposed estimator. Also, the ordered kernel should dominate the unordered kernel estimator in finite-sample applications since the data indeed have a natural ordering. We again consider the out-of-sample performance given by $MSE = n_2^{-1}\sum_{i=1}^{n_2}(Y_i - \hat{Y}_i)^2$ ($n_2 = 1,000$). The number of Monte Carlo replications is again 1,000, and results are presented in Table 2.

From Table 2 we observe that both the ordered and the unordered kernel estimators dominate the frequency estimator on MSE grounds. Also, it is evident that when we

Table 2
Comparison of out-of-sample MSE for each model

n_1/c	n_1	Ordered NP	Unordered NP	NP-frequency
2.78	100	1.98	2.90	6.12
5.56	200	1.64	2.12	2.80
13.9	500	1.39	1.66	1.78
27.8	1,000	1.27	1.42	1.48

take into account the natural ordering of the data, we achieve further finite-sample efficiency gains relative to the unordered kernel estimator. Finally, the smaller the ratio of the sample size to the number of cells, the better the performance of estimator relative to the frequency estimator.

3.2. Ordered categorical variables with unequal distances

Next we consider a situation in which there exists an ordered categorical variable; however, the true distance between categories is not as coded. We have in mind situations such as coding years of education with 0 denoting a high school diploma, 1 denoting some post-secondary education but not having graduated, 2 denoting a bachelors degree or equivalent, 3 denoting some graduate work but not having graduated, and 4 denoting a graduate degree. Such situations are common in applied settings. The salient feature is simply that there indeed exists an underlying order; however, the true underlying distance between categories is not as coded and is somewhat arbitrary. The aim of this simulation is simply to assess whether the proposed approach performs adequately in such settings, in particular, whether accounting for a natural order in such settings can result in improved finite-sample performance relative to the conventional frequency approach.

The third DGP we consider is given by

$$Y_i = 1 + \sqrt{Z_i} + X_i + u_i, \quad i = 1, \dots, n, \tag{3.3}$$

where $Z_i \in \{0, 1, \dots, 4\}$ having $\mathrm{P}(Z_i = l) = 1/5$ for $l = 0, 1, \dots, 4$, $X_i \sim \mathrm{N}(0, 1)$, and $u_i \sim \mathrm{N}(0, 1)$. Note that for this DGP the true distance between category 0 and 1 is 1 but between 1 and 2 is 0.41, 2 and 3 0.32, and 3 and 4 0.27; however, the data on Z_i are $\in \{0, 1, \dots, 4\}$.

We shall consider a number of approaches: the frequency approach, an unordered approach, an ordered approach, and also a 'dummy variable' approach in which five 0/1 dummy variables are created corresponding to the 0/1/2/3/4 coding of the variable of interest (for this method there will be five different λ smoothing parameters). We compute the out-of-sample $MSE = n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2$ of each approach using estimation samples of size n_1 $(50, 100, 200)$, independent evaluation samples of size $n_2 = 1,000$, and we conduct 1,000 Monte Carlo replications for each estimation sample size. Table 3 presents the mean out-of-sample MSE for each approach evaluated over the 1,000 Monte Carlo replications.

Table 3
Comparison of out-of-sample MSE for each model

n_1	Ordered NP	Unordered NP	Dummy NP	NP-frequency
50	1.33	1.40	1.41	1.50
100	1.19	1.23	1.23	1.26
200	1.11	1.13	1.13	1.14

It can be seen from Table 3 that the proposed method in which one accounts for the natural ordering yields significantly improved finite-sample estimates relative to the frequency estimator, and also performs better than the case whereby one does not account for the natural ordering. In addition, taking a dummy-variable approach yields estimates comparable to the *unordered* case but which are inferior to the proposed method. On the basis of this modest simulation we argue that even when there exists a natural ordering which has been coded arbitrarily but where the coding preserves the order itself, the proposed approach outperforms the conventional frequency estimator in finite-sample settings.

3.3. A semiparametric index model

Often semiparametric index models are used when one deals with a data set for which the 'curse-of-dimensionality' is present. We investigate the performance of our proposed estimator relative to the semiparametric index model in a small sample setting having four explanatory variables.

We consider two more DGPs which are given by

$$\text{DGP4: } Y_i = Z_{i1} + Z_{i2} + X_{i1} + X_{i2} + u_i, \tag{3.4}$$

$$\text{DGP5: } Y_i = Z_{i1} + Z_{i2} + Z_{i1}Z_{i2} + X_{i1} + X_{i2} + X_{i1}X_{i2} + u_i, \tag{3.5}$$

where $X_{ij} \sim \mathrm{N}(0,1)$ and $Z_{ij} \in \{0,1\}$ with $\mathrm{P}(z_{ij} = l) = 0.5$ for $l = 0,1$, and where $u_i \sim \mathrm{N}(0,1)$.

For DGP4 we compare the NP, single-index, and correctly specified parametric model, while for DGP5 we compare the NP, misspecified single-index (ignoring the interaction terms), and correctly specified parametric model. We consider the out-of-sample $MSE = n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2$ ($n_2 = 1,000$). The number of Monte Carlo replications is again 1,000, and results are summarized in Table 4.

As expected, the above simulations show that the single-index model performs better than our nonparametric estimator for DGP4, and our proposed estimator outperforms the misspecified single-index estimator for DGP5.

Having investigated the finite-sample performance of our estimator in a range of simulated settings, we now consider its out-of-sample performance for a widely used and publicly available data set.

Table 4
Comparison of out-of-sample MSE for each model

	DGP4			DGP5		
n_1	NP	SP	PAR	NP	SP	PAR
100	1.43	1.16	1.04	1.83	2.18	1.07
200	1.27	1.08	1.02	1.52	1.95	1.03
500	1.15	1.03	1.01	1.29	1.84	1.01

4. An empirical application: modeling count panel data

We consider the data used by Hausman et al. (1984) in which they model the number of successful patent applications made by firms in scientific and non-scientific sectors across a 7-year period. The variables they use are the following:

PATENTS: number of successful patent applications in the year, CUSIP: firm identifier, YEAR: year of data, SCISECT: dummy for firms in scientific sector, LOGR: log of R&D spending, LOGK: log of R&D stock at beginning of year.

This data set is a balanced panel containing 896 observations on six variables, and the first four variables are categorical while the last two are continuous. For the categorical variables there were 227 unique values for the variable PATENTS, 128 values for CUSIP (128 firms), seven for YEAR, and two for SCISECT. For this data set, the number of discrete cells exceeds the sample size, therefore, the conventional frequency estimator cannot be used which is not uncommon in practice.

We wish to assess the dynamic predictive ability of the proposed method for this time-series count panel, and we use the first 6 years of data for estimation purposes and the remaining seventh year for evaluation purposes leaving $n_1 = 768$ (128 firm $\times$ 6 years) and $n_2 = 128$ (128 firm $\times$ 1 year).

For comparison purposes we consider three parametric models found in the literature: (1) A nonlinear OLS regression of log(PATENTS) on the explanatory variables, where log(PATENTS) is set to zero and a dummy variable used when PATENTS $=0$ (Hausman et al., 1984, p. 912); (2) a pooled Poisson count panel model; and (3) a Poisson count panel model with firm-specific effects. We apply the proposed nonparametric method first ignoring the natural ordering in the discrete regressor YEAR and then accounting for the ordering by using an ordered kernel.

We again assess predictive ability on the independent data using $MSE = n_2^{-1}\sum_{i=1}^{n_2}(\mathrm{PATENTS}_i - \widehat{\mathrm{PATENTS}}_i)^2$ where $\widehat{\mathrm{PATENTS}}_i$ denotes the predicted values generated from each model. As well, we compute the correlation coefficient between the actual and predicted values of PATENTS, $\hat{\rho}_{\hat{y},y}$. Results appear in Table 5.

The results presented in Table 5 show that the new approach completely dominates the parametric specifications and that accounting for the natural order in the discrete explanatory variable YEAR leads to further finite-sample efficiency gains. The squared prediction error of our nonparametric estimator (using the ordered kernel) is only

Table 5
Comparison of out-of-sample performance for each model

Model	Prediction MSE	$\hat{\rho}_{\hat{y},y}$
OLS	2618.3	0.86
Poisson (pooled)	1915.9	0.87
Poisson (firm-effects)	2834.7	0.82
Unordered kernel	403.4	0.97
Ordered kernel	385.2	0.98

14–20% of those obtained by various parametric methods which have been used to model this data set.

Additional empirical applications demonstrating how smoothing discrete variables can lead to superior out-of-sample predictions compared to commonly used parametric methods are available from the website http://econweb.tamu.edu/li/papers.htm under the title “Empirical Applications of Smoothing Discrete Variables”.

5. Concluding remarks

In this paper we propose a nonparametric kernel estimator for the case where the regressors contain a mix of continuous and categorical variables. A data-driven method of bandwidth selection is proposed, and we establish the asymptotic normality of the estimator. Simulations show that the new estimator performs substantially better than the conventional nonparametric estimator which has been used to handle the presence of categorical variables. An empirical application demonstrates the usefulness of the proposed method in practice.

Li and Racine (2003) have considered the nonparametric estimation of joint density functions with mixed discrete and continuous variables, and have shown that the proposed method performs significantly better than the conventional frequency estimator. Ker and Racine (2001), and Li and Racine (2001) show that, via many types of empirical data (e.g., U.S. crop yield data, female labor market participation data, marketing data, medical treatment data, etc.), smoothing the discrete variables often leads to much better out-of-sample predictions than the conventional sample-splitting nonparametric method and some commonly used parametric methods. One should expect that the smoothing method outperforms the frequency method in general, since the former includes the latter as a special case (when $\lambda = 0$). However, when the sample size is very large, the computational cost can be high for the cross-validation-based smoothing method. Therefore, in practice one may want to use the frequency method when the sample size is much larger than the number of discrete cells due to the computational simplicity of the frequency method. But even in such a situation the efficiency gain of the smoothing method over the frequency method can be substantial because the cross-validation method may choose large values of λ for some discrete variables (e.g., Insik et al., 2002).

There are numerous ways in which the results of the present paper can be extended, and we briefly mention a few of them at this point.

1. Using a local polynomial nonparametric approach rather than a local constant approach.
2. Nonparametric estimation of a *conditional* density with mixed discrete and continuous data.
3. Consistent model specification tests with mixed discrete and continuous regressors, including the testing of parametric functional forms, nonparametric significance testing, and so forth.

Acknowledgements

We would like to thank three referees and an editor for helpful comments that greatly improve the paper. Li's research is supported by the Bush Program in the Economics of Public Policy and the Private Enterprises Research Center, Texas A&M University.

Appendix A.

A.1. Proof of Theorem 2.1

Write $\tilde{g}(x)-g(x)=(\tilde{g}(x)-g(x))\tilde{f}(x)/\tilde{f}(x)$. We first consider the numerator $(\tilde{g}(x)-g(x))\tilde{f}(x)$:

$$
\begin{aligned}
(\tilde{g}(x)-g(x))\tilde{f}(x) &= \frac{1}{n}\sum_i [Y_i-g(x)]W_{h_0}\left(\frac{X_i^c-x^c}{h_0}\right)L(X_i^d,x^d,\lambda_0)\\
&= \frac{1}{n}\sum_i [g(X_i)-g(x)]W_{h_0}\left(\frac{X_i^c-x^c}{h_0}\right)L(X_i^d,x^d,\lambda_0)\\
&\quad + \frac{1}{n}\sum_i u_i W_{h_0}\left(\frac{X_i^c-x^c}{h_0}\right)L(X_i^d,x^d,\lambda_0)\\
&\equiv I_{1n}(x)+I_{2n}(x), \qquad\qquad (A.1)
\end{aligned}
$$

where the definition of I_{1n} and I_{2n} should be apparent. Define the shorthand notation $W_{h_0,ix}=h_0^{-p}W((X_i^c-x^c)/h_0)$ and $L_{\lambda_0,ix}=L(X_i^d,x^d,\lambda_0)$. It is straightforward to show that

$$
\begin{aligned}
E(I_{1n}) &= E[(g(X_i)-g(x))W_{h_0,ix}L_{\lambda_0,ix}]\\
&= E[(g(X_i)-g(x))W_{h_0,ix}L_{\lambda_0,ix}|d_{ix}=0]P(d_{ix}=0)\\
&\quad + E[(g(X_i)-g(x))W_{h_0,ix}L_{\lambda_0,ix}|d_{ix}=1]P(d_{ix}=1)+O(\lambda_0^2)\\
&= E[(g(X_i)-g(x))W_{h_0,ix}|d_{ix}=0]P(x^d)\\
&\quad + \lambda_0 E[(g(X_i)-g(x))W_{h_0,ix}|d_{ix}=1]
\end{aligned}
$$

$$
\begin{aligned}
&\times \mathrm{P}(d_{ix}=1)+\mathrm{O}(\lambda_0^2)\\
&=\int f(x^c+h_0v,x^d)(g(x+h_0v,x^d)-g(x^c,x^d))W(v)\,\mathrm{d}v+\mathrm{O}(\lambda_0^2)\\
&\quad+\lambda_0\sum_{\tilde{x}^d,d_{\tilde{x}x}=1}\left[\int f(x^c+hv|\tilde{x}^d)[g(x^c+h_0v,\tilde{x}^d)-g(x)]W(v)\,\mathrm{d}v\right]\\
&\quad\times p(d_{\tilde{x},x}=1)+\mathrm{O}(\lambda_0^2)\\
&=h_0^2\{\nabla f(x)'\nabla g(x)+f(x)\,tr[\nabla^2 g(x)]/2\}\left[\int w(v)v^2\,\mathrm{d}v\right]+\mathrm{O}(\lambda_0h_0^2+h_0^4)\\
&\quad+\lambda_0\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}[g(x^c,\tilde{x}^d)-g(x)]f(x^c,\tilde{x}^d)+\mathrm{O}(\lambda_0h_0^2)+\mathrm{O}(\lambda_0^2)\\
&=f(x)B(h_0,\lambda_0)+\mathrm{O}(h_0^4+\lambda_0h_0^2+\lambda_0^2),
\end{aligned}
$$

where $B(h_0,\lambda_0)=h_0^2\{\nabla f(x)'\nabla g(x)/f(x)+tr[\nabla^2 g(x)]/2\}[\int w(v)v^2\,\mathrm{d}v]+\lambda_0\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}[g(x^c,\tilde{x}^d)-g(x)]f(x^c,\tilde{x}^d)/f(x)$. Similarly, one can easily show that $var(I_{1n})=\mathrm{o}((h_0^2+\lambda_0)^2)$, which implies that

$$
I_{1n}=\mathrm{E}[I_{1n}]+(s.o.)=B(h_0,\lambda_0)+\mathrm{o_p}(h_0^2+\lambda_0). \tag{A.2}
$$

Also, $\mathrm{E}(I_{2n})=0$ and

$$
\begin{aligned}
Var(I_{2n})&=\mathrm{E}[(I_{2n})^2]\\
&=n^{-1}\mathrm{E}[\sigma^2(X_i)W_{h_0,ix}^2L_{\lambda_0,ix}^2]\\
&=n^{-1}\{\mathrm{E}[\sigma^2(X_i)W_{h_0,ix}^2|d_{ix}=0]\,\mathrm{P}(x^d)+\mathrm{O}(\lambda_0)\}\\
&=(nh_0^p)^{-1}\left\{\sigma^2(x)f(x)\left[\int W^2(v)\,\mathrm{d}v\right]+\mathrm{O}(\lambda_0+h_0^2)\right\}\\
&=(nh_0^p)^{-1}\{\Omega(x)f^2(x)+\mathrm{o}(1)\}.
\end{aligned}
$$

By a standard triangular array central limit theorem argument, we have

$$
\sqrt{nh_0^p}I_{2n}\to \mathrm{N}(0,\Omega(x)f^2(x)) \text{ in distribution.} \tag{A.3}
$$

Finally, it is easy to show that

$$
\tilde{f}(x)=f(x)+\mathrm{o_p}(1). \tag{A.4}
$$

Combining Eqs. (A.1)–(A.4), we have

$$\sqrt{nh_0}(\tilde{g}(x) - g(x) - B(h_0, \lambda_0))$$
$$= \frac{\sqrt{nh_0^p}(\tilde{g}(x) - g(x) - B(h_0, \lambda_0))\tilde{f}(x)}{\tilde{f}(x)}$$
$$= \frac{\sqrt{nh_0} I_{2n}}{f(x)} + o_p(1) \to N(0, \Omega(x)) \text{ in distribution.} \tag{A.5}$$

A.2. Proof of Theorem 2.2(i)

From Eq. (2.8) we have

$$CV(\lambda, h) \stackrel{\text{def}}{=} n^{-1} \sum_i [Y_i - \hat{g}(X_i)]^2 M(X_i) = n^{-1} \sum_i (g_i + u_i - \hat{g}_i)^2 M_i$$
$$= n^{-1} \sum_i (g_i - \hat{g}_i)^2 M_i + 2n^{-1} \sum_i u_i (g_i - \hat{g}_i) M_i + n^{-1} \sum_i u_i^2 M_i, \tag{A.6}$$

where $g_i = g(X_i)$, $\hat{g}_i = \hat{g}_{-i}(X_i)$ and $M_i = M(X_i)$.

Write $g_i - \hat{g}_i = (g_i - \hat{g}_i)\hat{f}_i/f_i + (g_i - \hat{g}_i)(f_i - \hat{f}_i)/f_i$ $(\hat{f}_i = \hat{f}_{-i}(X_i))$. By similar arguments as in the proof of Lemma 1 of Härdle and Marron (1985), one can establish the uniform consistency of $\hat{f}(x)$ to $f(x)$ and $\hat{g}(x)$ to $g(x)$. Therefore, the second term is of smaller order than the first term. Replacing $(g_i - \hat{g}_i)$ by $(g_i - \hat{g}_i)\hat{f}_i/f_i$ in Eq. (A.6), we obtain the leading term of $CV(\lambda, h)$ (ignoring $n^{-1}\sum_i u_i^2 M_i$ since it is independent of λ) and we denote this by $CV_1(\lambda, h)$:

$$CV_1(\lambda, h) = n^{-1} \sum_i (g_i - \hat{g}_i)^2 \hat{f}_i^2 M_i / f_i^2 + 2n^{-1} \sum_i u_i (g_i - \hat{g}_i) \hat{f}_i M_i / f_i. \tag{A.7}$$

To simplify notation and to save space, we will omit the trimming function M_i below. Substituting Eqs. (2.9) and (2.10) into Eq. (A.7), and noting that $Y_j = g_j + u_j$, we have (omitting M_i)

$$CV_1 = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} (g_i - Y_j)(g_i - Y_l) K_{h,ij} K_{h,il}^2 / f_i^2 + 2n^{-1} \sum_i u_i (g_i - \hat{g}_i) \hat{f}_i / f_i.$$
$$= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} (g_i - g_j - u_j)(g_i - g_l - u_l) K_{h,ij} K_{h,il} / f_i^2$$
$$+ 2n^{-1} \sum_i u_i (g_i - \hat{g}_i) \hat{f}_i / f_i.$$
$$= n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i} (g_i - g_j)(g_i - g_l) K_{h,ij} K_{h,il} / f_i^2$$

$$
\begin{aligned}
&+ n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l K_{h,ij}K_{h,il}/f_i^2\\
&- 2n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i-g_j)u_l K_{h,ij}K_{h,il}/f_i^2\\
&+ 2n^{-2}\sum_i\sum_{j\neq i} u_i(g_i-Y_j)K_{h,ij}/f_i\\
=& \left\{ n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i-g_j)(g_i-g_l)K_{h,ij}K_{h,il}/f_i^2 \right\}\\
&+ \left\{ n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l K_{h,ij}K_{h,il}/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i} u_i u_j K_{h,ij}/f_i \right\}\\
&+ 2\left\{ n^{-2}\sum_i\sum_{j\neq i} u_i(g_i-g_j)K_{h,ij}/f_i - n^{-3}\right.\\
&\left.\times \sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i-g_j)u_l K_{h,ij}K_{h,il}/f_i^2 \right\}\\
\equiv& \{S_1\} + \{S_2\} + 2\{S_3\},
\end{aligned}
$$

where the definition of S_j $(j=1,2,3)$ should be apparent.

By lemmas B.1–B.3 we know that

$$
\begin{aligned}
S_1 &= B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + \bar{B}_5h^2(nh^p)^{-1}\\
&\quad + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + (s.o.),\\
S_2 &= (nh^p)^{-1}[B_4 + h^2\tilde{B}_5] + (nh^{p/2})^{-1}\mathscr{Z}_{1n} + (s.o.),\\
S_3 &= h^2n^{-1/2}\mathscr{Z}_{2n} + \lambda n^{-1/2}\mathscr{Z}_{3n} + (s.o.), \qquad (A.8)
\end{aligned}
$$

where the $\mathscr{Z}_{jn}$'s are mean-zero $O_p(1)$ random variables, while the B_j's, $\tilde{B}_j$'s and $\bar{B}_5$ are some constants defined in the proofs of the lemmas in Appendix B.

Define $CV_2 = CV - CV_1$. By Lemma B.4 we know that

$$
CV_2 = \tilde{C}_1h^6 + \tilde{C}_2h^4\lambda + \tilde{C}_3h^2\lambda^2 + \tilde{C}_4\lambda^3 + \tilde{C}_5h^2(nh^p)^{-1} + (s.o.). \qquad (A.9)
$$

Using Eqs. (A.8) and (A.9), we have

$$
\begin{aligned}
CV &= CV_1 + CV_2 = S_1 + S_2 + 2S_3 + CV_2\\
&= \{B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + B_4(nh^p)^{-1}\}
\end{aligned}
$$

$$
\begin{aligned}
&+\{(nh^{p/2})^{-1}\mathscr{Z}_{1n}+h^2n^{-1/2}\mathscr{Z}_{2n}+\lambda n^{-1/2}\mathscr{Z}_{3n}\}\\
&+\{C_1h^6+C_2h^4\lambda+C_3h^2\lambda^2+C_4\lambda^3+C_5h^2(nh^p)^{-1}\}+(s.o.),\\
\equiv\ &\{A_{1n}\}+\{A_{2n}\}+\{A_{3n}\}+(s.o.),
\end{aligned}
\tag{A.10}
$$

where $A_{1n}=B_1h^4-B_2h^2\lambda+B_3\lambda^2+B_4(nh^p)^{-1}$, $A_{2n}=(nh^{p/2})^{-1}\mathscr{Z}_{1n}+h^2n^{-1/2}\mathscr{Z}_{2n}+\lambda n^{-1/2}\mathscr{Z}_{3n}$, and $A_{3n}=C_1h^6+C_2h^4\lambda+C_3h^2\lambda^2+C_4\lambda^3+C_5h^2(nh^p)^{-1}$, $C_j=\tilde{B}_j+\tilde{C}_j$ $(j=1,2,3,4)$ and $C_5=\bar{B}_5+\tilde{B}_5+\tilde{C}_5$.

Given that $h=o(1)$ and $\lambda=o(1)$, we have $CV=A_{1n}+(s.o.)$. That is, A_{1n} is the leading term of CV. We rewrite A_{1n} as

$$
A_{1n}=B_3[\lambda-B_2h^2/(2B_3)]^2+[B_1-B_2^2/(4B_3)]h^4+B_4(nh^p)^{-1}. \tag{A.11}
$$

Let h_0 and λ_0 denote the values of h and λ that minimize $A_{1n}=A_{1n}(h,\lambda)$. Then from Eq. (A.11) it is easy to see that λ_0 and h_0 satisfy the following equations:

$$
\lambda_0=B_2h_0^2/(2B_3)\quad\text{and}\quad 4[B_1-B_2^2/(4B_3)]h_0^{p+4}=\frac{pB_4}{n}. \tag{A.12}
$$

Solving Eq. (A.12) leads to

$$
h_0=c_1n^{-1/(4+p)}\quad\text{and}\quad \lambda_0=c_2n^{-2/(4+p)}, \tag{A.13}
$$

where $c_1=\{pB_4/(4[B_1-B_2^2/(4B_3)])\}^{1/(4+p)}$ and $c_2=B_2c_1^2/(2B_3)$.

From $CV=A_{1n}+o_p(A_{1n})$, we know that $\hat{h}=h_0+o_p(h_0)$ and $\hat{\lambda}=\lambda_0+o_p(\lambda_0)$. Therefore, we have $(\hat{h}-h_0)/h_0\xrightarrow{p}0$ and $(\hat{\lambda}-\lambda_0)/\lambda_0\xrightarrow{p}0$.

Next we derive the rates of convergence of $(\hat{h}-h_0)/h_0$ and $(\hat{\lambda}-\lambda_0)/\lambda_0$.

A.2.1. The case of $p\leqslant 3$

In the case of $p\leqslant 3$, we have $CV=A_{1n}+A_{2n}+(s.o.)$, and we rewrite $A_{1n}+A_{2n}$ as

$$
\begin{aligned}
A_{1n}+A_{2n}=\ &B_3[\lambda-(h^2B_2-n^{-1/2}\mathscr{Z}_{3n})/(2B_3)]^2+[B_1-B_2^2/(4B_3)]h^4\\
&+h^2n^{-1/2}[\mathscr{Z}_{2n}+B_2\mathscr{Z}_{3n}/(2B_3)]\\
&+(nh^p)^{-1}[B_4+h^{p/2}\mathscr{Z}_{1n}]-n^{-1}\mathscr{Z}_{3n}^2/(4B_3).
\end{aligned}
\tag{A.14}
$$

Using Eq. (A.14), we minimize $CV=A_{1n}+A_{2n}+(s.o.)$ over λ and h, and we obtain

$$
\hat{\lambda}=\hat{h}^2B_2/(2B_3)-n^{-1/2}\mathscr{Z}_{3n}/(2B_3)+(s.o.) \tag{A.15}
$$

and

$$
\begin{aligned}
&B_0\hat{h}^{p+4}-\frac{p}{n}B_4+2\hat{h}^{p+2}n^{-1/2}[\mathscr{Z}_{2n}+B_2\mathscr{Z}_{3n}/(2B_3)]-\frac{p\hat{h}^{p/2}}{2}\mathscr{Z}_{1n}+(s.o.)\\
&\quad=0,
\end{aligned}
\tag{A.16}
$$

where $B_0=4[B_1-B_2^2/(4B_3)]$. Writing $\hat{h}=h_0+h_1$ and noting that h_1 has an order smaller than that of h_0 because $(\hat{h}-h_0)/h_0=o_p(1)$ which implies $h_1/h_0=o_p(1)$, we have

$$\hat{h}^4 \equiv (h_0 + h_1)^{4+p} = h_0^{4+p} + (4+p)h_0^{p+3}h_1 + (s.o.). \tag{A.17}$$

Using Eqs. (A.12) and (A.17), then from Eq. (A.16), we obtain

$$B_0(p+4)h_0^{p+3}h_1 + 2h_0^{p+2}n^{-1/2}[\mathscr{Z}_{2n} + B_2\mathscr{Z}_{3n}/(2B_3)] - \frac{ph_0^{p/2}}{2n}\mathscr{Z}_{1n} + (s.o.) = 0. \tag{A.18}$$

Eq. (A.18) gives

$$h_1 = \frac{p(2nh_0^{3+p/2})^{-1}\mathscr{Z}_{1n} - 2(n^{1/2}h_0)^{-1}[\mathscr{Z}_{2n} + B_2\mathscr{Z}_{3n}/(2B_3)]}{B_0(4+p)} + (s.o.). \tag{A.19}$$

By noting that $h_1 = \hat{h} - h_0$, we have from Eq. (A.19) that

$$\begin{aligned}(\hat{h} - h_0)/h_0 &= \frac{1}{B_0(4+p)}\{p(2nh_0^{4+p/2})^{-1}\mathscr{Z}_{1n} \\ &\quad - 2(n^{1/2}h_0^2)^{-1}[\mathscr{Z}_{2n} + B_2\mathscr{Z}_{3n}/(2B_3)]\} + (s.o.) \\ &= O_p(h_0^{p/2}) = O_p(n^{-p/[2(4+p)]}).\end{aligned} \tag{A.20}$$

Using $\hat{h} = h_0 + h_1$ in Eq. (A.15) gives us

$$\begin{aligned}\hat{\lambda} &= (h_0 + h_1)^2 B_2/(2B_3) + n^{-1/2}\mathscr{Z}_{3n}/(2B_3) + (s.o.) \\ &= \lambda_0 + 2h_0h_1B_2/(2B_3) + n^{-1/2}\mathscr{Z}_{3n}/(2B_3) + (s.o.) = \lambda_0 + O_p(n^{-1/2}),\end{aligned} \tag{A.21}$$

because $h_0h_1 = O(n^{-1/2})$ by Eq. (A.19).

Eqs. (A.20) and (A.21) complete the proof for $p \leqslant 3$, part (i) of Theorem 2.2.

A.2.2. The case of $p \geqslant 4$

We now consider the case of $p \geqslant 4$. When $p = 4$, A_{3n} has the same order as A_{2n} and when $p \geqslant 5$, A_{3n} has an order larger than that of A_{2n}. We first consider the case of $p \geqslant 5$ below. In this case we have $CV = A_{1n} + A_{3n} + (s.o.)$ since $A_{2n} = o_p(A_{3n})$ in this case. Therefore, we have

$$\begin{aligned}CV &= A_{1n} + A_{3n} + (s.o.) = B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + B_4(nh^p)^{-1} \\ &\quad + C_1h^6 + C_2h^4\lambda + C_3h^2\lambda^2 + C_4\lambda^3 + C_6h^2(nh^p)^{-1} + (s.o.).\end{aligned} \tag{A.22}$$

Taking derivatives of Eq. (A.22) with respect to λ and h and setting them to zero will give us two equations. We then replace $\hat{h}$ by $h_0 + h_1$ and $\hat{\lambda}$ by $\lambda_0 + \lambda_1$. Noting that h_1 has an order smaller than h_0 and that λ_1 has an order smaller than λ_0, then using expansions of

$$\begin{aligned}\hat{h}^s &= (h_0 + h_1)^s = h_0^s + sh_0^{s-1}h_1 + (s.o.), \\ \hat{\lambda}^t &= (\lambda_0 + \lambda_1)^t = \lambda_0^t + t\lambda_0^{t-1}\lambda_1 + (s.o.),\end{aligned} \tag{A.23}$$

for some positive integers s and t, we obtain two equations that are linear in h_1 and λ_1 (i.e., we only retain up to the linear terms in h_1 and λ_1). It is easy to see that solving these two linear equations for h_1 and λ_1 leads to

$$\lambda_1 = (\hat{\lambda} - \lambda_0) = O_p(h_0^4) = O_p(n^{-4/(4+p)}),$$
$$h_1/h_0 = (\hat{h} - h_0)/h_0 = O_p(h_0^2) = O_p(n^{-2/(4+p)}). \tag{A.24}$$

Finally, when $p=4$, A_{2n} has the same order as A_{3n}, but this only amounts to adding some extra terms having the same order as A_{3n}, while the above arguments leading to Eq. (A.24) remain unchanged. Hence, Eq. (A.24) holds true for the case of $p=4$. This completes the proof of Theorem 2.2(i).

A.3. Proof of Theorem 2.3(i)

One can prove Theorem 2.3 using the stochastic equicontinuity arguments as in Ichimura (2000). However, below we use a simple Taylor series expansion argument to prove Theorem 2.3.

From $(\hat{h} - h_0)/h_0 = o_p(1)$ we have

$$\frac{1}{\hat{h}^p} = \frac{1}{h_0^p} + \frac{1}{h_0^p} O_p\left(\frac{\hat{h} - h_0}{h_0}\right) = \frac{1}{h_0^p}(1 + o_p(1)). \tag{A.25}$$

Using Eq. (A.25) and $\hat{\lambda} - \lambda_0 = o_p(1)$, it is easy to see that

$$\begin{aligned}
&(\hat{g}(x) - g(x))\hat{f}(x) \\
&= \frac{1}{n\hat{h}^p} \sum_i [g(X_i) - g(x) + u_i] W\left(\frac{X_i^c - x^c}{\hat{h}}\right) L(X_i^d, x^d, \hat{\lambda}) \\
&= \frac{1}{nh_0^p} \sum_i [g(X_i) - g(x) + u_i] W\left(\frac{X_i^c - x^c}{\hat{h}}\right) L(X_i^d, x^d, \lambda_0) + (s.o.) \\
&= \frac{1}{nh_0^p} \sum_i [g(X_i) - g(x) + u_i] W\left(\frac{X_i^c - x^c}{\hat{h}}\right) L(X_i^d, x^d, \lambda_0) + (s.o.) \\
&\equiv J_n + (s.o.),
\end{aligned} \tag{A.26}$$

where $J_n = (nh_0^p)^{-1} \sum_i [g(X_i) - g(x) + u_i] W((X_i^c - x^c)/\hat{h}) L(X_i^d, x^d, \lambda_0)$. Define $J_{n,0}$ by replacing $\hat{h}$ by h_0 in J_n:

$$\begin{aligned}
J_{n,0} &\stackrel{\text{def}}{=} \frac{1}{nh_0^p} \sum_i [g(X_i) - g(x) + u_i] W\left(\frac{X_i^c - x^c}{h_0}\right) L(X_i^d, x^d, \lambda_0) \\
&\equiv (\tilde{g}(x) - g(x))\tilde{f}(x).
\end{aligned} \tag{A.27}$$

Then by the proof of Theorem 2.1(i) we know that $J_{n,0} = O_p((nh_0^p)^{1/2}) = O_p(h_0^2)$. Next, applying a Taylor expansion to $W((X_i^c - x^c)/\hat{h})$ at $\hat{h} = h_0$, we have

$$W\left(\frac{X_i^c - x^c}{\hat{h}}\right) = W\left(\frac{X_i^c - x^c}{h_0}\right) + \sum_{1 \leqslant s \leqslant m-1} \frac{1}{s!}\tilde{W}^{(s)}\left(\frac{X_i^c - x^c}{h_0}\right)\left(\frac{\hat{h} - h_0}{h_0}\right)^s$$

$$+ \frac{1}{m!}\tilde{W}^{(m)}\left(\frac{X_i^c - x^c}{\tilde{h}}\right)\left(\frac{\hat{h} - h_0}{\tilde{h}}\right)^m, \tag{A.28}$$

where $\tilde{W}^{(s)}((X_i^c - x^c)/h) \stackrel{\text{def}}{=} h^s(\partial^s/\partial h^s)W((X_i^c - x^c)/h)$, and $\tilde{h}$ is between $\hat{h}$ and h_0. It is easy to see that $\tilde{W}^{(s)}(v)$ contains terms of $W^{(t)}(v) = (\partial^t/\partial v_1^{s_1} \ldots \partial v_p^{s_p})W(v)$ $(s_1 + \cdots + s_p = t)$ times a tth-order polynomial in v for $1 \leqslant t \leqslant s$. Also, $\tilde{W}^{(s)}(v)$ is an even function and thus can be viewed as a second-order kernel function (though it may take negative values).

Substituting Eq. (A.28) into Eq. (A.26) we obtain

$$J_n = \frac{1}{nh_0^p}\sum_i [g(X_i) - g(x) + u_i]W\left(\frac{X_i^c - x^c}{h_0}\right)L(X_i^d, x^d, \lambda_0)$$

$$+ O_p(J_{n,0})O_p\left(\frac{\hat{h} - h_0}{h_0}\right) + h_0^{-p}O_p\left(\left(\frac{\hat{h} - h_0}{h_0}\right)^m\right)$$

$$= J_{n,0} + o_p(J_{n,0}) + o_p(h_0^2) = J_{n,0} + o_p(h_0^2), \tag{A.29}$$

since $J_{n,0} = O_p(h_0^2)$ and $((\hat{h} - h_0)/\tilde{h})^m/\tilde{h}^p = h_0^{-p}O_p([(\hat{h} - h_0)/h_0]^m) + (s.o.) = o_p(h_0^2)$ by Theorem 2.2 and Assumption (A1) (ii).

Similarly, it is straightforward to show that

$$\hat{f}(x) = f(x) + o_p(1). \tag{A.30}$$

Summarizing the results in Eqs. (A.26), (A.27), (A.29) and (A.30) we have

$$\sqrt{n\hat{h}^p}(\hat{g}(x) - g(x) - B(h_0, \lambda_0))$$

$$= \frac{\sqrt{n\hat{h}^p}(\hat{g}(x) - g(x) - B(h_0, \lambda_0))\hat{f}(x)}{\hat{f}(x)}$$

$$= \frac{\sqrt{nh_0^p}(J_{n,0} - B(h_0, \lambda_0))}{f(x)} + o_p(1) \to N(0, \Omega(x)) \text{ in distribution}, \tag{A.31}$$

where the last convergence result follows from the proof of Theorem 2.1.

A.4. Proof of Theorem 2.3(ii)

Using the results of Theorem 2.3(i), it is obvious that $\hat{B}(h_0, \lambda_0) = B(h_0, \lambda_0) + o_p(h_0^2 + \lambda_0)$ and $\hat{\Omega}(x) = \Omega(x) + o_p(1)$. Hence, Theorem 2.3(ii) follows from these results and from Theorem 2.3(i).

A.5. Proof of Theorem 2.4

From Eqs. (A.19)–(A.21) we know that both $n^{p/[2(4+p)]}(\hat{h} - h_0)/h_0$ and $\sqrt{n}(\hat{\lambda} - \lambda_0)$ can be written as linear combinations of $\mathscr{Z}_{1n}$, $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$ (plus some $o_p(1)$ terms), where $\mathscr{Z}_{1n}$ is defined in Lemma B.2, $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$ are defined in Lemma B.3. For example, $\mathscr{Z}_{1n} = (nh^{p/2})\{n^{-2}\sum\sum_{j\neq i, d_{ij}=0} u_i u_j (W_{h,ij}^{(2)} - 2W_{h,ij})/f_i\}$ ($W^{(2)}(.)$ is the two-fold convolution kernel defined from $W(.)$). Obviously $\mathscr{Z}_{1n}$ is a second-order degenerate U-statistic, thus using the central limit theorem for degenerate U-statistics of Hall (1984), it is straightforward to show that $\mathscr{Z}_{1n}$ converges in distribution to a mean-zero finite-variance normal random variable. Similarly, for $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$, using H-decomposition, it is easy to see that the leading terms of both $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$ are partial sums of the form of $n^{-1/2}\sum_i u_i \mathscr{C}(X_i) + (s.o.)$ for some function $\mathscr{C}(\cdot)$. Therefore, $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$ are asymptotically normally distributed with mean zero and finite variance. Note that $\mathscr{Z}_{1n}$ is uncorrelated with either $\mathscr{Z}_{2n}$ or $\mathscr{Z}_{3n}$. It is easy to show that a linear combination of $\mathscr{Z}_{1n}$, $\mathscr{Z}_{2n}$ and $\mathscr{Z}_{3n}$ has an asymptotic normal distribution with mean zero and finite variance which results in Theorem 2.4.

Appendix B.

Lemmas B.1–B.3 below utilize the U-statistics H-decomposition with variable kernels. Here we provide an intuitive explanation of H-decomposition for a second-order U-statistic. A second-order U-statistic has the form

$$\mathscr{U}_n = \frac{2}{n(n-1)} \sum \sum_{1 \leqslant i < j \leqslant n} H_n(X_i, X_j), \tag{B.1}$$

where $H_n(.,.)$ is a symmetric function. The H-decomposition involves rewriting $\mathscr{U}_n$ in the form of uncorrelated terms of differing order:

$$\begin{aligned}\mathscr{U}_n &= \mathrm{E}[H_n(X_i, X_j)] + \frac{2}{n}\sum_i \{\mathrm{E}[H_n(X_i, X_j)|X_i] - \mathrm{E}[H_n(X_i, X_j)]\} \\ &\quad + \frac{2}{n(n-1)} \sum \sum_{1 \leqslant i < j \leqslant n} \{H_n(X_i, X_j) - \mathrm{E}[H_n(X_i, X_j)|X_i] \\ &\quad - \mathrm{E}[H_n(X_i, X_j)|X_j] + \mathrm{E}[H_n(X_i, X_j)]\}.\end{aligned} \tag{B.2}$$

If $\mathrm{E}[H_n^4(X_i, X_j)] = \mathrm{O}(1)$, then it is easy to see that the three terms in Eq. (B.2) are of the orders $\mathrm{O_p}(1)$, $\mathrm{O_p}(n^{-1/2})$ and $\mathrm{O_p}(n^{-1})$, respectively. Moreover, the three terms are uncorrelated with each other. In our application of the H-decomposition below, usually $\mathrm{E}[H_n(X_i, X_j)] = \mathrm{O}(a_n)$ (say $a_n = \mathrm{O}((h^2 + \lambda)^2)$, the second term in the decomposition is of the order of $\mathrm{O_p}(n^{-1/2} a_n)$, and the third term is of even smaller order. We also use the H-decomposition of a third-order U-statistic, while Lee (1990, Section 1.6) provides a detailed result of H-decomposition for a general kth-order U-statistic. For U-statistics with variable kernels, see Powell et al. (1989).

Lemma B.1. *$S_1 = B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + \bar{B}_5h^2(nh^p)^{-1} + (s.o.)$, where B_j's, $\tilde{B}_j$'s and $\bar{B}_5$ are some constants defined in the proof below.*

Proof. $S_1 = n^{-3}\sum\sum\sum_{i\neq j\neq l}(g_i - g_j)(g_i - g_l)K_{h,ij}K_{h,il}/f_i^2 + n^{-3}\sum_i\sum_{j\neq i}(g_i - g_j)^2 K_{h,ij}^2/f_i^2 = S_{1a} + S_{1b}$. We first consider S_{1a}. $S_{1a} = [n^{-3}\sum\sum\sum_{i\neq j\neq l} H_{1a}(X_i,X_j,X_l)]$, where $H_{1a}(X_i,X_j,X_l)$ is a symmetrized version of $(g_i - g_j)(g_i - g_l)K_{h,ij}K_{h,il}/f_i^2$ given by $H_{1a}(X_i,X_j,X_l) = (1/3)\{(g_i - g_j)(g_i - g_l)K_{h,ij}K_{h,il}/f_i^2 + (g_j - g_i)(g_j - g_l)K_{h,ij}K_{jl}/f_j^2 + (g_l - g_j)(g_l - g_i)K_{lj}K_{h,il}/f_l^2\}$.

We first compute $\mathrm{E}[(g_i - g_j)K_{h,ij}|X_i]$ (note that $d_{ij} = d_{x_i,x_j}$):

$$\begin{aligned}
&\mathrm{E}[(g_i - g_j)K_{h,ij}|X_i]\\
&\quad = \mathrm{E}[(g_i - g_j)W_{h,ij}|X_i, d_{ij} = 0]\mathrm{P}(d_{ij} = 0|X_i)\\
&\qquad + \mathrm{E}[(g_i - g_j)W_{h,ij}|X_i, d_{ij} = 1]\mathrm{P}(d_{ij} = 1|X_i)\lambda\\
&\qquad + \sum_{l=2}^{k}\mathrm{E}[(g_i - g_j)W_{h,ij}|X_i, d_{ij} = l]\mathrm{P}(d_{ij} = l|X_i)\lambda^l\\
&\quad = \{B_{1,1}(X_i)h^2 + \mathrm{O}(h^4)\} + \{-B_{1,2}(X_i)\lambda + \mathrm{O}(\lambda h^2)\} + \{\mathrm{O}(\lambda^2)\},
\end{aligned} \tag{B.3}$$

where $B_{1,1}(X_i) = \{\nabla f_i'\nabla g_i + (1/2)f(X_i)\int tr(\nabla^2 g_i)\}[\int w(v)v^2\,\mathrm{d}v]$, and $B_{1,2}(X_i) = \mathrm{E}[g(X_i^c,X_j^d) - g(X_i^c,X_i^d)|X_i, d_{ij} = 1]\mathrm{P}(d_{ij} = 1|X_i)$ are as defined in Assumption (A1).

Note that in the above calculation, $g_i - g_j \neq 0$ even when $d_{ij} = 0$. This arises because $d_{ij} = 0$ only restricts $X_i^d = X_j^d$, while $[(g_i - g_j)|d_{ij} = 0] = g(X_i^c,X_i^d) - g(X_j^c,X_i^d) \neq 0$ because $X_i^c \neq X_j^c$.

Using Eq. (B.3) we have

$$\begin{aligned}
&\mathrm{E}[H_{1a}(X_i,X_j,X_l)]\\
&\quad = \mathrm{E}[(g_i - g_j)(g_i - g_l)K_{h,ij}K_{h,il}/f_i^2]\\
&\quad = \mathrm{E}\{\mathrm{E}[(g_i - g_j)K_{h,ij}|X_i]\mathrm{E}[(g_i - g_l)K_{h,il}|X_i]/f_i^2\}\\
&\quad = \mathrm{E}\{\mathrm{E}[(g_i - g_j)K_{h,ij}|X_i]/f_i\}^2\\
&\quad = \mathrm{E}[(B_{1,1}(X_i)/f_i)^2h^4 - 2f_i^{-2}B_{1,1}(X_i)B_{1,2}(X_i)h^2\lambda + (B_{1,2}(X_i)/f_i)^2\lambda^2]\\
&\qquad + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + (s.o.)\\
&\quad \equiv [B_1h^4 - B_2h^2\lambda + B_3\lambda^2] + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + (s.o.),
\end{aligned} \tag{B.4}$$

where $B_1 = \mathrm{E}[(B_{1,1}(X_i)/f_i)^2]$, $B_2 = \mathrm{E}[2f_i^{-2}B_{1,1}(X_i)B_{1,2}(X_i)]$ and $B_3 = \mathrm{E}[(B_{1,2}(X_i)/f_i)^2]$. Similarly, the $\tilde{B}_j$'s correspond to terms with higher-order derivatives (with respect to the continuous variables) and/or terms where d_{x_i,x_j} assumes values larger than 1 (which results in higher-order polynomials in λ). We do not give the explicit definitions of the $\tilde{B}_j$'s here to save space and because we do not use their specific expressions in the paper.

Therefore, by Eqs. (B.3) and (B.4) and the H-decomposition, we have

$$
\begin{aligned}
S_{1a} &= \mathrm{E}[H_{1a}(X_i,X_j,X_l)] \\
&\quad + 3n^{-1}\sum_i \{\mathrm{E}[H_{1a}(X_i,X_j,X_l)|X_i] - \mathrm{E}[H_{1a}(X_i,X_j,X_l)]\} + (s.o.) \\
&= \mathrm{E}[H_{1a}(X_i,X_j,X_l)] + n^{-1/2}\mathrm{O_p}(h^4 + h^2\lambda + \lambda^2) + (s.o.) \\
&= B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + (s.o.).
\end{aligned}
$$

Next, we consider S_{1b}. Defining $H_{1b}(X_i,X_j) = (g_i - g_j)^2K_{h,ij}^2(1/f_i^2 + 1/f_j^2)/2$, then $S_{1b} = n^{-1}[n^{-2}\sum_i\sum_{j\neq i}H_{1b}(X_i,X_j)]$, and it is easy to see that

$$
\begin{aligned}
\mathrm{E}[H_{1b}(X_i,X_j)] &= \mathrm{E}[(g_i-g_j)^2K_{h,ij}^2/f_i^2] \\
&= \mathrm{E}[(g_i-g_j)^2K_{h,ij}^2/f_i^2|d_{ij}=0]\,p(d_{ij}=0) \\
&\quad + \mathrm{E}[(g_i-g_j)^2K_{h,ij}^2/f_i^2|d_{ij}\geqslant 1]\,p(d_{ij}\geqslant 1) \\
&= \mathrm{E}[(g_i-g_j)^2W_{h,ij}^2/f_i^2|d_{ij}=0]\,p(d_{ij}=0) + h^{-p}\mathrm{O}(\lambda^2) \\
&= \bar{B}_5h^2h^{-p} + \mathrm{O}(h^{-p}(h^4+\lambda^2)),
\end{aligned}
$$

where $\bar{B}_5 = \mathrm{E}[(\nabla g_i)'\nabla g_i/f_i][\int w^2(v)v^2\,\mathrm{d}v]\,p(d_{ij}=0)$.

Similarly one can easily show that $\mathrm{E}[H_{1b}(X_i,X_j)|X_i] = \mathrm{O}(h^2h^{-p})$. Hence,

$$
\begin{aligned}
S_{1b} &= n^{-1}[\mathrm{E}[H_{1b}(X_i,X_j)] \\
&\quad + 2n^{-1}\sum_i \{\mathrm{E}[H_{1b}(X_i,X_j)|X_i] - \mathrm{E}[H_{1b}(X_i,X_j)]\} + (s.o.)] \\
&= h^2(nh^p)^{-1}\bar{B}_5 + n^{-1/2}\mathrm{O}((nh^p)^{-1}h^2).
\end{aligned}
$$

Summarizing the above we have shown that

$$
\begin{aligned}
S_1 = S_{1a} + S_{1b} &= B_1h^4 - B_2h^2\lambda + B_3\lambda^2 + \tilde{B}_1h^6 + \tilde{B}_2h^4\lambda \\
&\quad + \tilde{B}_3h^2\lambda^2 + \tilde{B}_4\lambda^3 + \bar{B}_5h^2(nh^p)^{-1} + (s.o.) \qquad \text{(B.5)}
\end{aligned}
$$

Lemma B.2. *$S_2 = (nh^p)^{-1}[B_4 + \tilde{B}_5h^2] + (nh^{p/2})^{-1}\mathscr{Z}_{1n} + (s.o.)$, where B_4 and $\tilde{B}_j$ ($j=5,6$) are some constants and $\mathscr{Z}_{1n}$ is a $\mathrm{O_p}(1)$ random variable.*

Proof.

$$
\begin{aligned}
S_2 &= n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_ju_lK_{h,ij}K_{h,il}/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i}u_iu_jK_{h,ij}/f_i \\
&= n^{-3}\sum_i\sum_{j\neq i}u_j^2K_{h,ij}^2/f_i^2 + n^{-3}\sum\sum\sum_{i\neq j\neq l}u_ju_lK_{h,ij}K_{h,il} \\
&\quad - 2n^{-2}\sum_i\sum_{j\neq i}u_iu_jK_{h,ij}/f_i \equiv S_{2a} + S_{2b} - 2S_{2c}.
\end{aligned}
$$

Define $H_{2a}(Z_i,Z_j)=(1/2)(u_i^2/f_i^2+u_j^2/f_j^2)K_{h,ij}^2$, then $S_{2a}=n^{-1}[n^{-2}\sum\sum_{i\neq j}H_{2a}(Z_i,Z_j)]$.

$$\begin{aligned}
&\mathrm{E}[H_{2a}(Z_i,Z_j)]\\
&\quad=\mathrm{E}[u_i^2K_{h,ij}^2/f_i^2]=\mathrm{E}[\sigma^2(X_i)K_{h,ij}^2/f_i^2]\\
&\quad=\mathrm{E}[\sigma^2(X_i)K_{h,ij}^2/f_i^2|d_{ij}=0]\,p(d_{ij}=0)+\mathrm{E}[\sigma^2(X_i)K_{h,ij}^2/f_i^2|d_{ij}\geqslant 1]\,p(d_{ij}\geqslant 1)\\
&\quad=\mathrm{E}[\sigma^2(X_i)W_{h,ij}^2/f_i^2|d_{ij}=0]\,p(d_{ij}=0)+\mathrm{O}(\lambda^2h^{-p})\\
&\quad=h^{-p}\left[\sum_{x^d}\int f^{-1}(x^c,x^d)\sigma^2(x^c,x^d)f(x^c+hv,x^d)W^2(v)\,\mathrm{d}x^c\,\mathrm{d}v\right]\\
&\qquad\times p(d_{ij}=0)+\mathrm{O}(\lambda^2h^{-p})\\
&\quad=h^{-p}[B_4+\tilde{B}_5h^2+\mathrm{O}(h^4)]+\mathrm{O}(\lambda^2h^{-p}),
\end{aligned}$$

where $B_4=\mathrm{E}[\sigma^2(X_i)/f(X_i)][\int W^2(v)\,\mathrm{d}v]\,p(d_{ij}=0)$ and
$\tilde{B}_5=(1/2)\mathrm{E}[\sigma^2(X_i)\,tr(\nabla^2f(X_i))/f^2(X_i)][\int w^2(v)v^2\,\mathrm{d}v]\,p(d_{ij}=0)$. Next,

$$\begin{aligned}
&\mathrm{E}[H_{2a}(Z_i,Z_j)|Z_i]=(1/2)\{(u_i^2/f_i^2)\mathrm{E}[K_{h,ij}^2|Z_i]+\mathrm{E}[(\sigma^2(X_j)/f_j^2)K_{h,ij}^2|Z_i]\}\\
&\quad=(1/2)u_i^2f_i^{-2}\{\mathrm{E}[K_{h,ij}^2|X_i,d_{ij}=0]\,p(d_{ij}=0|X_i)\\
&\qquad+\sum_{l=1}^{k}\mathrm{E}[K_{h,ij}^2/f_i^2|X_i,d_{ij}=l]\,p(d_{ij}=l|X_i)\}\\
&\qquad+(1/2)\mathrm{E}[\sigma^2(X_j)K_{h,ij}^2/f_j^2|X_i,d_{ij}=0]\,p(d_{ij}=0|X_i)\\
&\qquad+\sum_{l=1}^{k}\mathrm{E}[\sigma^2(X_j)K_{h,ij}^2/f_j^2|X_i,d_{ij}=l]\,p(d_{ij}=l|X_i)\\
&\quad=(1/2)h^{-p}f_i^{-1}\left\{[u_i^2+\sigma^2(X_i)]\left[\int W^2(v)\,\mathrm{d}v\right]+\mathrm{O}(h^2+\lambda^2)\right\}\\
&\quad=(1/2)h^{-p}f^{-1}(X_i)\left\{[u_i^2+\sigma^2(X_i)]\left[\int W^2(v)\,\mathrm{d}v\right]+\mathrm{O_p}(h^2+\lambda)\right\}\\
&\quad=\mathscr{B}_4(Z_i)h^{-p}+\mathrm{O_p}(h^{-p}(h^2+\lambda)),
\end{aligned}$$

where $B_4(Z_i)=(1/2)f_i^{-1}[u_i^2+\sigma^2(X_i)][\int W^2(v)\,\mathrm{d}v]$. It is easy to check that $B_4=\mathrm{E}[\mathscr{B}_4(Z_i)]$.

Hence, by the H-decomposition we have

$$S_{2a} = n^{-1}\Bigg\{\mathrm{E}[H_{2a}(Z_i,Z_j)]$$

$$+\,2n^{-1}\sum_i \{\mathrm{E}[H_{2a}(Z_i,Z_j)|Z_i] - \mathrm{E}[H_{2a}(Z_i,Z_j)]\} + (s.o.)\Bigg\}$$

$$= (nh^p)^{-1}[B_4 + \tilde{B}_5 h^2 + \mathrm{O}(h^4 + \lambda^2 + h^2\lambda)]$$

$$+\,(nh^p)^{-1}n^{-1/2}[\mathscr{Z}_{2a,n} + \mathrm{O_p}(h^2+\lambda)] + (s.o.),$$

where $\mathscr{Z}_{2a,n} = n^{-1/2}\sum_i [\mathscr{B}_4(Z_i) - \mathrm{E}(\mathscr{B}_4(Z_i))]$.

Next, S_{2b} can be written as a third-order U-statistic. $S_{2b} = [n^{-3}\sum\sum\sum_{i\neq j\neq l} H_{2b}(Z_i,Z_j,Z_l)]$, where $H_{2b}(Z_i,Z_j,Z_l)]$ is a symmetrized version of $u_j u_l K_{h,ij}K_{h,il}/f_i^2$ given by

$$H_{2b}(Z_i,Z_j,Z_l) = (1/3)[u_j u_l K_{h,ij}K_{h,il}/f_i^2 + u_i u_l K_{h,ij}K_{h,jl}/f_j^2 + u_j u_i K_{h,lj}K_{h,il}/f_l^2].$$

Note that $\mathrm{E}[H_{2b}(Z_i,Z_j,Z_l)|Z_j] = 0$ because $\mathrm{E}(u_l|Z_j) = 0$. Hence the leading term of S_{2b} is a second-order degenerate U-statistic:

$$\mathrm{E}[H_{2b}(Z_i,Z_j,Z_l)|Z_i,Z_j]$$

$$=(1/3)u_i u_j \mathrm{E}[K_{h,lj}K_{h,il}/f_l^2|X_i,X_j]$$

$$=u_i u_j \mathrm{E}[K_{h,lj}K_{h,il}/f_l^2|X_i,X_j,d_{lj}+d_{il}=0]\mathrm{P}(d_{lj}+d_{il}=0|X_i,X_j) + (s.o.).$$

Note that

$$\mathrm{E}[K_{h,lj}K_{il}/f_l^2|X_i,X_j,d_{lj}+d_{il}=0] = \mathrm{E}[W_{h,lj}W_{h,il}/f_l^2|X_i,X_j,d_{lj}+d_{il}=0]$$

$$= W_{h,ij}^{(2)}\mathbf{1}(d_{ij}=0)/f_i + \mathrm{O}(h^2),$$

where $W_{h,ij}^{(2)} = h^{-p}W^{(2)}((X_i^c - X_j^c)/h)$ with $W^{(2)}(v) \stackrel{\text{def}}{=} \int W(u)W(v+u)\,\mathrm{d}u$ is the two-fold convolution kernel derived from $W(\cdot)$. Hence,

$$S_{2b} = 3\left\{n^{-2}\sum\sum_{j\neq i}\mathrm{E}[H_{2b}(Z_i,Z_j,Z_l)|Z_i,Z_j] + (s.o)\right\}$$

$$= \left\{n^{-2}\sum\sum_{j\neq i}u_i u_j \mathrm{E}[K_{h,lj}K_{h,il}/f_l^2|Z_i,Z_j] + (s.o.)\right\}$$

$$= \left[n^{-2}h^p\sum\sum_{j\neq i, d_{ij}=0}u_i u_j W_{h,ij}^{(2)}/f_i + (s.o.)\right]$$

$$= (nh^{p/2})^{-1}\mathscr{Z}_{2b,n} + \mathrm{o_p}((nh^{p/2})^{-1}),$$

where $\mathscr{Z}_{2b,n} = (nh^{p/2})\{n^{-2}\sum\sum_{j\neq i,d_{ij}=0} u_i u_j W_{h,ij}^{(2)}/f_i\}$.

Finally,

$$
\begin{aligned}
S_{2c} &= n^{-2}\sum_i\sum_{j\neq i} u_i u_j K_{h,ij}/f_i \\
&= n^{-2}\sum_i\sum_{j\neq i,d_{ij}=0} u_i u_j W_{h,ij}/f_i + n^{-2}\sum_i\sum_{j\neq i,d_{ij}\geqslant 1} \lambda^{d_{ij}} u_i u_j W_{h,ij}/f_i \\
&= n^{-2}\sum_i\sum_{j\neq i,d_{ij}=0} u_i u_j W_{h,ij}/f_i + (s.o.) \\
&= (nh^{p/2})^{-1}\mathscr{Z}_{2c,n} + (s.o.),
\end{aligned}
$$

where $\mathscr{Z}_{2c,n} = (nh^{p/2})[n^{-2}\sum_i\sum_{j\neq i,d_{ij}=0} u_i u_j W_{h,ij}/f_i]$.

Summarizing the above we have shown that

$$
S_2 = S_{2a} + S_{2b} - 2S_{2c} = (nh^p)^{-1}[B_4 + \tilde{B}_5 h^2] + (nh^{p/2})^{-1}\mathscr{Z}_{1n} + (s.o.), \tag{B.6}
$$

where $\mathscr{Z}_{1n} = \mathscr{Z}_{2b,n} - 2\mathscr{Z}_{2c,n}$. Note that $\mathscr{Z}_{1n}$ is a second-order generate U-statistic. Using Theorem 1 of Hall (1984), it is easy to see that $\mathscr{Z}_{1n}$ has an asymptotic mean-zero finite-variance normal distribution. Hence, $\mathscr{Z}_{1n} = \mathrm{O_p}(1)$.

Lemma B.3. $S_3 = h^2 n^{-1/2}\mathscr{Z}_{2n} + \lambda n^{-1/2}\mathscr{Z}_{3n} + \mathrm{o_p}(n^{-1/2}(h^2+\lambda))$, *where both* $\mathscr{Z}_{2n}$ *and* $\mathscr{Z}_{3n}$ *are mean-zero* $\mathrm{O_p}(1)$ *random variables.*

Proof.

$$
\begin{aligned}
S_3 &= n^{-2}\sum_i\sum_{j\neq i} u_i(g_i - g_j)K_{h,ij}/f_i - n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i - g_j)u_l K_{h,ij}K_{h,il}/f_i^2 \\
&= n^{-2}\sum_i\sum_{j\neq i} u_i(g_i - g_j)K_{h,ij}/f_i - n^{-3}\sum_i\sum_{j\neq i}(g_i - g_j)u_j K_{h,ij}^2/f_i^2 \\
&\quad - n^{-3}\sum\sum\sum_{i\neq j\neq l}(g_i - g_j)u_l K_{h,ij}K_{h,il}/f_i^2 \\
&\equiv S_{3a} - S_{3b} - S_{3c}.
\end{aligned}
$$

We first consider S_{3a}. The leading terms of S_{3a} are the cases (i) $d_{ij}=0$ and (ii) $d_{ij}=1$. We use $S_{3a,(\mathrm{i})}$ and $S_{3a,(\mathrm{ii})}$ to denote these two cases. For case (i), we have $S_{3a,(\mathrm{i})} = n^{-2}\sum_i\sum_{j\neq i,d_{ij}=0} H_{3a}(Z_i,Z_j)$, where $H_{3a}(Z_i,Z_j)] = (1/2)[u_i(g_i-g_j)/f_i + u_j(g_j - g_i)/f_j]W_{h,ij}$:

$$
\begin{aligned}
[H_{3a}(Z_i,Z_j)|Z_i] &= (1/2)(u_i/f_i)\mathrm{E}[(g_i-g_j)W_{h,ij}|X_i^c] \\
&= -(1/4)h^2(u_i/f_i)\,tr[\nabla^2 g_i]\left[\int w^2(v)v^2\,\mathrm{d}v\right] + \mathrm{O_p}(h^4) \\
&\equiv h^2\mathscr{B}_{3a,(\mathrm{i})}(Z_i) + \mathrm{O_p}(h^4),
\end{aligned}
$$

where $\mathscr{B}_{3a,(\mathrm{i})}(Z_i) = -(1/4)(u_i/f_i)\,tr[\nabla^2 g_i][\int w^2(v)v^2\,\mathrm{d}v]$.

Using H-decomposition and noting that $\mathrm{E}[H_{3a}(Z_i, Z_j)] = 0$, we have

$$S_{3a,(\mathrm{i})} = \left\{ 2n^{-1} \sum_i \mathrm{E}[H_{3a}(Z_i, Z_j)|Z_i] + (s.o.) \right\}$$

$$= 2h^2 n^{-1} \sum_i \mathscr{B}_{3a,(i)}(Z_i) + (s.o.)$$

$$\equiv n^{-1/2} h^2 \mathscr{Z}_{3a,(\mathrm{i})} + (s.o.),$$

where $\mathscr{Z}_{3a,(i)} = n^{-1/2} \sum_i \mathscr{B}_{3a,(i)}(Z_i)$.

Now consider $S_{3a,(\mathrm{ii})}$:

$$S_{3a,(\mathrm{ii})} = \left\{ \lambda n^{-2} \sum_i \sum_{j \neq i, d_{ij}=1} u_i(g_i - g_j) W_{h,ij}/f_i \right\} = \lambda n^{-1/2} \mathscr{Z}_{3a,(\mathrm{ii})}, \tag{B.7}$$

where $\mathscr{Z}_{3a,(\mathrm{ii})} = n^{-3/2} \sum_i \sum_{j \neq i, d_{ij}=1} u_i(g_i - g_j) W_{h,ij}/f_i$. Obviously, $\mathscr{Z}_{3n}$ is $\mathrm{O_p}(1)$.

It is easy to see that when $d_{ij} \geqslant 2$, we have $S_{3a} = \mathrm{O_p}(\lambda^2 n^{-1/2})$. Thus, we have shown that

$$S_{3a} = h^2 n^{-1/2} \mathscr{Z}_{3a,(\mathrm{i})} + \lambda n^{-1/2} \mathscr{Z}_{3a,(\mathrm{ii})} + \mathrm{O_p}(\lambda^2 n^{-1/2}). \tag{B.8}$$

Next, for S_{3b} it is easy to see that

$$S_{3b} = (nh^p)^{-1} \mathrm{O_p}(S_{3a}) = \mathrm{O_p}((nh^p)^{-1}(h^2 + \lambda) n^{-1/2}) = \mathrm{o_p}(n^{-1/2}(h^2 + \lambda)). \tag{B.9}$$

Finally, we consider S_{3c}. The leading terms should have $d_{il} = 0$ (since there is no $(g_i - g_l)$ term in S_{3c}). The two leading cases are (i) $d_{il} = 0$ and $d_{ij} = 0$, (ii) $d_{il} = 0$ and $d_{ij} = 1$. We use $S_{3c,(\mathrm{i})}$ and $S_{3c,(\mathrm{ii})}$ to denote these two cases.

$S_{3c,(\mathrm{i})}$ can be written as a third-order U-statistic $S_{3c,(\mathrm{i})} = n^{-3} \sum \sum \sum_{i \neq j \neq l, d_{ij}=0, d_{il}=0} H_{3c,(\mathrm{i})}(Z_i, Z_j, Z_l)$, where $H_{3c,(\mathrm{i})}(Z_i, Z_j, Z_l)$ is a symmetrized version of $u_l(g_i - g_j) K_{h,ij} K_{h,il}/f_i^2$. Obviously $\mathrm{E}[H_{3c,(\mathrm{i})}(Z_i, Z_j, Z_l)] = 0$ and it can easily be verified that

$$\mathrm{E}[H_{3c,(\mathrm{i})}(Z_i, Z_j, Z_l)|Z_i] = (1/3) h^2 u_i B_{3c,(\mathrm{i})}(Z_i),$$

where

$$\mathscr{B}_{3c,(\mathrm{i})}(Z_i) = \{(\nabla g_i)' \nabla f_i / f_i + (1/2)\, tr[\nabla^2 g_i]\} \left[\int w(v) v^2 \, \mathrm{d}v \right]. \tag{B.10}$$

Therefore, by H-decomposition we have

$$S_{3c,(\mathrm{i})} = 3n^{-1} \sum_i \mathrm{E}[H_{3c,(\mathrm{i})}(Z_i, Z_j, Z_l)|Z_i] + (s.o.)$$

$$= h^2 n^{-1/2} \left[n^{-1/2} \sum_i u_i B_{3c,(\mathrm{i})}(X_i) \right] + (s.o.)$$

$$\equiv h^2 n^{-1/2} \mathscr{Z}_{3c,(\mathrm{i})} + (s.o.),$$

where $\mathscr{Z}_{3c,(\mathrm{i})} = [n^{-1/2} \sum_i u_i \mathscr{B}_{3c,(\mathrm{i})}(X_i)]$ with $\mathscr{B}_{3c,(\mathrm{i})}$ defined in Eq. (B.10).

Next,

$$S_{3c,(\mathrm{ii})} = \lambda n^{-3} \sum\sum \sum_{\{i \neq j \neq l, d_{il}=0, d_{ij}=1\}} (g_i - g_j) u_l W_{h,ij} W_{h,il} / f_i^2$$

$$\equiv \lambda n^{-1/2} \{\mathscr{Z}_{3c,(\mathrm{ii})}\}, \tag{B.11}$$

where $\mathscr{Z}_{3c,(\mathrm{ii})} = n^{-5/2} \sum\sum\sum_{\{i \neq j \neq l, d_{il=0}, d_{ij}=1\}} (g_i - g_j) u_l W_{h,ij} W_{h,il} / f_i^2\}$. It is straightforward to show that $\mathrm{E}\{[\mathscr{Z}_{3c,(\mathrm{ii})}]^2\} = \mathrm{O}(1)$. Hence, $\mathscr{Z}_{3c,(\mathrm{ii})} = \mathrm{O_p}(1)$.

It is easy to see that when $d_{ij} + d_{il} \geqslant 2$, we will have a factor of λ^2, and $S_{3c} = \mathrm{O_p}(\lambda^2 n^{-1/2})$ in such cases. Hence, we have

$$S_{3c} = h^2 n^{-1/2} \mathscr{Z}_{3c,(\mathrm{i})} + n^{-1/2} \lambda \mathscr{Z}_{3c,(\mathrm{ii})} + \mathrm{O_p}(\lambda^2 n^{-1/2}). \tag{B.12}$$

Summarizing Eqs. (B.8), (B.9), and (B.12), we have shown that

$$S_3 = S_{3a} - S_{3b} - S_{3c} = h^2 n^{-1/2} \mathscr{Z}_{2n} + \lambda n^{-1/2} \mathscr{Z}_{3n} + \mathrm{o_p}(n^{-1/2}(h^2 + \lambda)), \tag{B.13}$$

where $\mathscr{Z}_{2n} = \mathscr{Z}_{3a,(\mathrm{i})} - \mathscr{Z}_{3c,(\mathrm{i})}$ and $\mathscr{Z}_{3n} = \mathscr{Z}_{3a,(\mathrm{ii})} - \mathscr{Z}_{3c,(\mathrm{ii})}$, both are mean-zero $\mathrm{O_p}(1)$ random variables.

Lemma B.4. *$CV_2(h,\lambda) = \tilde{C}_1 h^6 + \tilde{C}_2 h^4 \lambda + \tilde{C}_3 h^2 \lambda^2 + \tilde{C}_4 \lambda^3 + \tilde{C}_5 h^2 (nh^p)^{-1} + (s.o.)$, where $\tilde{C}_j$'s are some finite constants.*

Proof. Since the details of the proof are very similar to the proofs of Lemmas B.1 and B.3, we only sketch a proof here.

$$CV_2 = n^{-1} \sum_i (\hat{g}_i - g_i)^2 (\hat{f}_i - f_i)^2 / f_i^2 + 2n^{-1} \sum_i (\hat{f}_i - f_i)(\hat{g}_i - g_i)^2 \hat{f}_i / f_i^2$$

$$+ 2n^{-1} \sum_i u_i (\hat{g}_i - g_i)(\hat{f}_i - f_i) / f_i. \tag{B.14}$$

It is easy to see that the first term on the right-hand side of Eq. (B.14) has an order smaller than the second and third terms. Let $CV_{2,L}$ denote the leading term of CV_2, i.e., $CV_2 = CV_{2,L} + (s.o.)$. Replacing $(\hat{g}_i - g_i)$ by $(\hat{g}_i - g_i)\hat{f}_i / f_i$ in the second and third terms of Eq. (B.14), we obtain the leading term of CV_2:

$$CV_{2,L} = 2n^{-1} \sum_i (\hat{f}_i - f_i)(\hat{g}_i - g_i)^2 \hat{f}_i^2 / f_i^3 + 2n^{-1} \sum_i u_i (\hat{f}_i - f_i)(\hat{g}_i - g_i) \hat{f}_i / f_i^2$$

$$= \mathrm{O_p}(h^6 + h^4 \lambda + h^2 \lambda^2 + \lambda^3 + h^2 (nh^p)^{-1}) + (s.o.), \tag{B.15}$$

where the order calculations follow the same arguments as in the proofs of Lemmas B.1–B.3. It is not hard to see that a detailed calculation would reveal that

$$CV_{2,L} = \tilde{C}_1 h^6 + \tilde{C}_2 h^4 \lambda + \tilde{C}_3 h^2 \lambda^2 + \tilde{C}_4 \lambda^3 + \tilde{C}_5 h^2 (nh^p)^{-1} + (s.o.), \tag{B.16}$$

for some constants $\tilde{C}_j$. We will not give the explicit definitions of C_j's to save space.

References

Ahmad, I.A., Cerrito, P.B., 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.

Bierens, H., 1983. Uniform consistency of kernel estimators of a regression function under generalized conditions. Journal of the American Statistical Association 78, 699–707.

Bierens, H., 1987. Kernel estimation of regression functions. In: Bewley, T.F. (Ed.), Advances in Econometrics. Cambridge University Press, Cambridge.

Delgado, M.A., Mora, J., 1995. Nonparametric and semiparametric estimation with discrete regressors. Econometrica 63, 1477–1484.

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

Fan, J., Härdle, W., Mammen, E., 1998. Direct estimation of low dimensional components in additive models. Annals of Statistics 26, 943–971.

Grund, B., Hall, P., 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., 1984. Central limit theorem for integrated square error of multivariate nonparametric density estimators. Journal of Multivariate Analysis 14, 1–16.

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

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

Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83, 86–95.

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

Hausman, J., Hall, B.H., Griliches, Z., 1984. Econometric models for count data with an application to the patents–R&D relationship. Econometrica 52, 909–938.

Horowitz, J.L., 1998. Semiparametric Methods in Econometrics. Lecture Notes in Statistics. Springer, Berlin.

Hsiao, C., 1983. Regression analysis with categorized explanatory variables. In: Karlin, S., Amemiya, T., Goodman, L. (Eds.), Studies in Econometrics, Time series and Multivariate Statistics. Academic Press, New York, pp. 93–129.

Hsiao, C., Mountain, D., 1985. Estimating the short-run income elasticity of demand for electricity using cross-sectional categorical data. Journal of the American Statistical Association 80, 259–265.

Ichimura, H., 2000. Asymptotic distribution of nonparametric and semiparametric estimators with data dependent smoothing parameters. Unpublished manuscript.

Insik, M., Fang, C., Li, Q., 2002. Investigation of patterns in food-away-from-home expenditure for China: a nonparametric approach. Unpublished manuscript.

Ker, A., Racine, J., 2001. Nonparametric rating of crop insurance policies: adverse selecting against the FCIC/RMA. Unpublished manuscript.

Lee, J., 1990. U-Statistics: Theory and Practice. Marcel Dekker, Inc., New York, Basel.

Li, Q., Racine, J. 2001. Empirical applications of smoothing categorical variables. Unpublished manuscript.

Li, Q., Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis, forthcoming.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 1403–1430.

Robinson, P., 1988. Root-N consistent semiparametric regression. Econometrica 56, 931–954.

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

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

Stock, J.H., 1989. Nonparametric policy analysis. Journal of the American Statistical Association 84, 567–575.

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