NON-GEOMETRIC DISCRETE KERNEL FUNCTIONS FOR APPLIED DENSITY AND REGRESSION ESTIMATION

CHI-YANG CHU, DANIEL J. HENDERSON, AND CHRISTOPHER F. PARMETER

Abstract. Least-squares cross-validation is commonly used for selection of smoothing parameters in the discrete data setting; however, in many applied situations, it tends to select relatively small bandwidths, particularly when the data are sparse. This tendency to undersmooth can be a result of the geometric weighting scheme that many discrete kernels possess. This problem may be avoided by using alternative kernel functions. In this paper, we consider discrete kernel functions which do not have rapidly decaying weights. The analytic properties of these kernels are contrasted with commonly used kernel functions and their relative performance is compared using both simulated and real data. The simulation and empirical results show that the these kernel functions generally perform well. Further, we find substantial gains in terms of mean squared error in some cases.

Keywords: Least-Squares Cross-Validation, Discrete Data, Sparse Data, Panel Data, Geometric Weighting.

JEL Classification: C14

1. Introduction

An intuitive approach to estimate a univariate discrete probability function is to use the sample frequency of occurrence as the estimator of a cell probability (i.e., frequency approach). However, when the number of cells is close to or even greater than the sample size (when the data are sparse), the frequency approach does not work well due to many zero counts (Simonoff, 1996). In this case, applied researchers often resort to a smoothing approach, which introduces bias but can dramatically lower mean squared error (MSE). In this paper, we focus on the kernel smoothing approach where the underlying density, $p(x)$, is estimated by $\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} l(\cdot)$, with a kernel function $l(\cdot)$ for discrete data. Existing discrete kernel functions date back to Aitchison and Aitken

University of Alabama, University of Alabama, and University of Miami
Chi-Yang Chu, Department of Economics, Finance, and Legal Studies, University of Alabama, Tuscaloosa, AL 35487-0224, e-mail: cchu2@crimson.ua.edu. Daniel J. Henderson, Department of Economics, Finance, and Legal Studies, University of Alabama, Tuscaloosa, AL 35487-0224. Christopher F. Parmeter, Department of Economics, University of Miami, Coral Gables, FL 33124-6520.

(1976), Habbema et al. (1978), Titterington (1980), Wang and van Ryzin (1981), and Aitken (1983). More recently, Li and Racine (2003) develop kernel functions for unordered and ordered discrete data.

The bandwidth (i.e., smoothing parameter) is an important component of kernel estimation and least-squares cross-validation (LSCV) is a popular approach for its selection in the discrete data setting. However, in many applied situations, LSCV tends to select a relatively small bandwidth (undersmoothing), particularly when discrete data are sparse, see Coppejans (2003). One explanation for this problem is that many proposed ordered discrete kernel functions possess a geometrically decaying weighting, leading to a rapid decline in the weights used to smooth the data (Rajagopalan and Lall, 1995). Adding to this line of reasoning, Chu, Henderson, and Parmeter (2015) show that for an ordered discrete kernel function with geometric weighting structure, the optimal bandwidth, using the mean summed squared error (MSSE) criterion, is a real root from a polynomial, with the order of the polynomial being determined by the number of cells. Their main result is that the optimal bandwidth is inversely related to the order of the polynomial, compounding the small bandwidth problem.

These issues also occur in kernel regression estimation. For example, Henderson and Kumbhakar (2006) note that in their panel application, capturing unobserved heterogeneity through an unordered discrete variable (with respect to the cross-sectional dimension) results in a relatively small bandwidth. In this case, the regression estimator essentially uses only T observations for each cross-sectional unit. It is likely that this problem is pervasive in papers using nonparametric methods in the presence of panel data where the individual specific heterogeneity is treated as an unordered discrete variable. Although different methods have been proposed to resolve the issue of undersmoothing, most are modifications of existing error criterion and are typically designed for continuous variables (for example, see Härdle, Hall, and Marron 1988, Chiu 1990, Hart and Yi 1998, Hurvich, Simonoff, and Tsai 1998, and Hall and Robinson 2009). Unlike existing studies, we attempt to use alternative discrete kernel functions in conjunction with the LSCV criterion.

Rajagopalan and Lall (1995) develop an ordered discrete kernel function which does not possess a geometric weighting scheme, providing sufficient smoothing in the presence of sparse data.[1] Unfortunately, applied researchers who adopt kernel smoothing methods are largely unaware of this kernel function which motivates us to attempt to further its application. Specifically, we highlight

[1]Kokonendji, Senga Kiessé, and Zougab (2007) develop a so-called triangular probability mass function and use it as an ordered discrete kernel function. Their triangular kernel function also does not impose a geometric weighting structure, but is not relevant for our discussion here as this kernel function is designed for use with count data with excess zeros and the function consists of two parameters which adds additional complications to bandwidth selection.

Rajagopalan and Lall's (1995) non-geometrically weighted ordered kernel function and extend their work to the construction of an unordered discrete kernel function.

For both the unordered and ordered Rajagopalan and Lall kernel functions, we derive MSSE for the kernel density estimator. Further we demonstrate that the necessary condition for asymptotic normality of both the kernel density and regression estimators are satisfied by this new kernel, namely by establishing a second-order approximation of the discrete kernels proposed here, similar to that used by Li and Racine (2003).

As a comparison to the existing literature, we examine the Rajagopalan and Lall kernel functions versus Aitchison and Aitken's (1976), Wang and van Ryzin's (1981), and Li and Racine's (2003) kernel functions in simulations and empirical examples. For this set of kernel functions, the simulation results show the Rajagopalan and Lall kernel functions generally perform well. We find significant improvements in the univariate discrete density estimation setting with sparse data. Further, the relative performance of conditional density and cross-sectional regression estimators with the Rajagopalan and Lall kernel kernels perform at least as well as the existing kernels in both settings. In addition, the unordered Rajagopalan and Lall type kernel function performs admirably when an unordered discrete variable is used to capture unobservable individual effects (especially when an irrelevant discrete regressor exists) in panel data models. This feature alone is an important marker for applied use with this kernel as this setting represents a clear data sparse environment and the tendency to undersmooth can lead to poor inferences regarding the overall panel structure of the data.

The remainder of this paper is organized as follows: Section 2 presents the Rajagopalan and Lall ordered kernel function, develops the unordered Rajagopalan and Lall type kernel function and compares the analytic properties of these kernels with those commonly used in the literature. Section 3 shows the finite sample performance via simulations. Section 4 provides several empirical illustrations and Section 5 concludes.

2. The Rajagopalan and Lall Kernel Functions

For the case of a continuous random variable X, Epanechnikov (1969) shows that the MSE optimal second-order kernel function is

$$k(\psi_X) = \begin{cases} a\psi_X^2 + b \text{ if } |\psi_X| \leq 1 \\ 0 \qquad\quad \text{if } |\psi_X| > 1 \end{cases}, \tag{1}$$

where $-a = b = 0.75$, $\psi_X = \frac{X-x}{h}$ and h is the bandwidth. Rajagopalan and Lall (1995) extend this set-up to an ordered, discrete random variable X. The discrete version of the optimal

second-order kernel, $l(\cdot)$, is required to satisfy two conditions: (A) $\sum_{X=x-h}^{x+h} l(\psi_X) = 1$ and (B) $\sum_{X=x-h}^{x+h} l(\psi_X)\psi_X = 0$. Condition (A) is the discrete counterpart to requiring a kernel function to integrate to 1, while Condition (B) is the discrete counterpart of having a symmetric kernel with zero mean. The constants a and b in Equation (1) are solved to satisfy Conditions (A) and (B); consequently, Rajagopalan and Lall's (1995) kernel function, $l(\cdot)$, is

$$l(X,x,h) = \begin{cases} a\psi_X^2 + b \text{ if } |\psi_X| \leq 1 \\ 0 \qquad\quad \text{if } |\psi_X| > 1 \end{cases}, \tag{2}$$

where $-a = b = \frac{-3h}{(1-4h^2)}$ and the bandwidth h is defined as a positive integer. For the remainder of the paper we will refer to $k(\cdot)$ and $l(\cdot)$ as continuous and discrete kernel functions, respectively.

In order for us to extend this kernel function for smoothing an unordered, discrete random variable, we note that an unordered discrete kernel function does not take distance (i.e., $|X-x|$) or symmetry into consideration. In this case, Condition (A) becomes

$$\sum_{X=1}^{c} l(X,x,h) = a\left[\left(\frac{0}{h}\right)^2 + \left(\frac{1}{h}\right)^2 + \cdots + \left(\frac{1}{h}\right)^2\right] + cb = 1,$$

where c is the number of cells, which we assume is known (i.e., X with support $S = \{1, 2, \ldots, c\}$).[2] Note that Condition (B) is based on symmetry and hence is not needed for an unordered discrete kernel function. After solving for a and b, an unordered Rajagopalan and Lall type kernel function is given by

$$l(X,x,h) = \begin{cases} b \qquad\qquad\quad \text{if } X = x \\ a\left(\frac{1}{h}\right)^2 + b \text{ if } X \neq x \end{cases}, \tag{3}$$

where $-a = b = \frac{-h^2}{(c-1)-ch^2}$ with $h \in [1, \infty)$. Here h is not required to be an integer, as in the case of ordered categorical data. We set $\lambda = h - 1$ so that $\lambda \in [0, \infty)$.[3]

It is straightforward to show that (i) $\sum_{X=1}^{c} l(X,x,\lambda) = 1$, (ii) $l(X,x,\lambda) = 1\{X = x\}$ if $\lambda = 0$, where $1\{\mathcal{A}\}$ is the indicator function for the event $\mathcal{A}$, and (iii) $l(X,x,\lambda) \to \frac{1}{c}$ as $\lambda \to \infty$, $\forall X$. Property (i) guarantees that we have a proper probability density function, $\sum_x \widehat{p}(x) = 1$. Property (ii) implies that the kernel density estimator approaches the frequency estimator as the bandwidth goes to 0. Property (iii) suggests that if this kernel is used when conditioning, that irrelevant variables can be smoothed out through assigning upper-bound bandwidths to them or shrinking them toward the uniform distribution on their respective marginals (Hall, Racine, and Li, 2004).

[2]Li and Racine's (2003) kernel functions, to be discussed in more detail later, do not require the number of cells to be known.

[3]The upper bound of the bandwidth λ is infinity is because Equation (3) is derived from Equation (2), in which the upper bound of its bandwidth h is infinity.

2.1. **Smoothing Parameter Selection.** In general, the selection of the smoothing parameter λ is crucial for the finite sample performance of the kernel density estimator. It is commonly assumed theoretically that $\lambda \to 0$ as $n \to \infty$. Further, it is desirable for $\lambda \to \infty$ as $p(x) \to \frac{1}{c}$; that is, if our density is uniform, then the corresponding bandwidth should converge to ∞ rather than 0.

Bandwidths are commonly selected via LSCV. The cross-validation function in the discrete density setting (Ouyang, Li, and Racine, 2006) is given as

$$\begin{aligned} \mathrm{CV}(\lambda) &= \sum_{x=1}^{c}[\widehat{p}(x) - p(x)]^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{x=1}^{c} l(X_i, x, \lambda) l(X_j, x, \lambda) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} l(X_j, X_i, \lambda), \end{aligned}$$

where $\widehat{p}(x)$ is a kernel density estimator. Our simulation results (Section 4.2) show that as we take larger samples of X drawn from a multinomial distribution, the bandwidths chosen by LSCV approach zero, and for X randomly drawn from a uniform distribution, the bandwidth chosen by LSCV is relatively large. Therefore, when LSCV is applied for the selection of bandwidths using the unordered version of Rajagopalan and Lall's (1995) kernel function, the corresponding bandwidths possess these features.

2.2. **Mean Summed Squared Error.** The MSSE of the unordered univariate discrete kernel density estimator can be shown to be

$$\begin{aligned} MSSE\left[\widehat{p}\,(x)\right] &= \sum_{x=1}^{c}\left\{Bias\left[\widehat{p}\,(x)\right]\right\}^2 + \sum_{x=1}^{c} Var\left[\widehat{p}\,(x)\right] \\ &= \left[\frac{1-h^2}{(c-1)-ch^2}\right]^2 \sum_{x=1}^{c}\left[1-cp(x)\right]^2 + \frac{1}{n}\left[\frac{1}{(c-1)-ch^2}\right]^2 \sum_{x=1}^{c} p(x)\left[1-p(x)\right]. \end{aligned}$$

It is straightforward to show that the bias increases and the variance decreases as the bandwidth increases. Additionally, the variance decreases as the sample size grows. Note that if $h = 1$ ($\lambda = 0$), the MSSE will be the same as the frequency estimator ($\frac{1}{n}\sum_{x=1}^{c} p(x)\left[1-p(x)\right]$).

For the ordered Rajagopalan and Lall kernel function, the MSSE of the univariate discrete kernel density estimator (which we note was not given in Rajagopalan and Lall 1995) is

$$MSSE\left[\widehat{p}\,(x)\right] = \sum_{x=1}^{c}\left\{\left[\sum_{X=1}^{c} p(X)\left(a\left(\frac{X-x}{h}\right)^2 + b\right) 1\left\{\left|\frac{X-x}{h}\right| \leq 1\right\}\right] - p(x)\right\}^2 + O(n^{-1}).$$

It is difficult to derive an analytical form of the MSSE, but it is feasible to show the bandwidth's relation to the bias and variance using numerical methods.[4] As with the unordered kernel, if $h = 1$ ($\lambda = 0$), the MSSE will be the same as the frequency estimator.

2.3. **A Relationship Between Unordered Kernels.** An interesting relationship between our proposed unordered kernel and the Aitchison and Aitken kernel function exists. It turns out that this relationship is useful in demonstrating Theorem 2.1 in Racine and Li (2004), which holds for both the Li and Racine and Aitchison and Aitken kernel functions. Note that Equation (3) can be rearranged as

$$l(X, x, \lambda) = \begin{cases} 1 - d(h) \text{ if } X = x \\ \frac{1}{c-1} d(h) \text{ if } X \neq x \end{cases}, \tag{4}$$

where $d(h) = \frac{(c-1)(1-h^2)}{(c-1)-ch^2}$. If we think of $d(h)$ as λ, Equation (4) is the Aitchison and Aitken kernel function.[5].

Since the unordered Rajagopalan and Lall and Aitchison and Aitken kernel functions have a similar representation, it is natural to examine the relationship between their optimal bandwidths. The optimal bandwidth for use with each kernel can be determined via minimization of MSSE, which is asymptotically equal to those chosen by LSCV, see Chu, Henderson, and Parmeter (2015). As a result, the relationship between their MSSE optimal bandwidths is given by

$$\widehat{\lambda}^*_{URL} = \left[1 + \frac{1}{c} \left(\frac{c-1}{c\widehat{\lambda}^*_{AA}} - 1 \right)^{-1} \right]^{\frac{1}{2}} - 1, \tag{5}$$

where c is the number of cells and $\widehat{\lambda}^*_{URL}$ and $\widehat{\lambda}^*_{AA}$ are the MSSE optimal bandwidths for the unordered Rajagopalan and Lall and Aitchison and Aitken kernel functions, respectively.[6] This relationship is useful in determining the upper bound for the unordered Rajagopalan and Lall type kernel function, and this will be discussed in Section 3.2.[7] Moreover, this relationship can be used

[4]Kokonendji, Senga Kiessé, and Zougab (2007) use a discrete Taylor expansion to obtain an approximate and analytical expression for the MSE of a univariate discrete kernel density estimator. This technique is quite generic and hence is applicable to the ordered Rajagopalan and Lall kernel function.

[5]We may expect that the unordered Rajagopalan and Lall type kernel function performs similarly to the Aitchison and Aitken kernel function in a univariate discrete density setting. However, Li and Racine (2003) mention that the use of a mixed-data multivariate search algorithm naturally yields bandwidths (for each variable) which are different from bandwidths produced by the use of a discrete multivariate search algorithm.

[6]Following Chu, Henderson and Parmeter (2015), who do so for the Aitchison and Aitken kernel function, it can be shown that the MSSE optimal bandwidth for the unordered Rajagopalan and Lall kernel function is $\widehat{\lambda}^*_{URL} = \left\{ 1 + \frac{c \sum_{x=1}^{c} \widetilde{p}(x)[1-\widetilde{p}(x)]}{n \sum_{x=1}^{c} [1-c\widetilde{p}(x)]^2} \right\}^{\frac{1}{2}} - 1$, where $\widetilde{p}(x)$ is the frequency estimator. We do not believe a closed-form solution for the ordered Rajagopalan and Lall kernel exists.

[7]Given the number of cells c, $\widehat{\lambda}^*_{URL}$ increases as $\widehat{\lambda}^*_{AA}$ increases, but $\widehat{\lambda}^*_{URL}$ quickly passes unity once $\widehat{\lambda}^*_{AA}$ reaches its upper bound $\frac{c-1}{c}$.

to determine which kernel function has a relatively larger bandwidth and hence provides more smoothing, as we will see in the simulated and real examples.

2.3.1. *How does our Kernel Differ from Aitchison and Aitken's?* For ordered kernels with the geometric property, it is straightforward to show that the kernel weights drop off very quickly when X_i deviates from x. For example, with the Wang and van Ryzin kernel, $l(X_i = 1, x = 1, \lambda = 0.5) = 0.5$ and $l(X_i = 3, x = 1, \lambda = 0.5) = 0.0625$. However, this 'drop off' of the kernel weight is less intuitive in the unordered kernel setting, because no such geometric property exists. Precisely, there is no room for discussion of the geometric property because any unordered kernel is a binary-outcome function regardless of the number of cells. To observe relative changes in weights between unordered kernels, the rate of drop off in weights can be defined as a ratio of a weight from $X_i = x$ to a weight from $X_i \neq x$. When this ratio is closer to one (more equal weights), the kernel provides more smoothing. The ratios for the Aitchison and Aitken and our proposed kernels are given by

$$R_{AA}^{Weight} = \frac{1 - \lambda^{AA}}{\lambda^{AA}}(c - 1) \tag{6}$$

and

$$R_{URL}^{Weight} = \frac{h^2}{h^2 - 1}, \tag{7}$$

respectively, where c is the number of cells, $\lambda^{AA} \in [0, \frac{c-1}{c}]$, $h = \lambda^{URL} + 1$ with $\lambda^{URL} \in [0, \infty)$. Note that (7) is independent of c. As mentioned earlier, the bandwidths from the Aitchison and Aitken kernel and our proposed kernel have different upper bounds and hence we cannot plug the same bandwidth in these two ratios for comparison. A reasonable approach is to use Equation (5), which holds in the univariate setting.[8] For example, for $c = 3$ and $\lambda = 0.3$, $R_{AA}^{Weight} = 4.6667$ and $R_{URL}^{Weight} = 2.4493$, which implies that weights from the Aitchison and Aitken kernel drop off more quickly because the ratio is further from one. Figures 1 and 2 show the cases with $c = 3$ and $c = 10$, respectively. Note that the relative weight (i.e., the ratio) from our proposed kernel is the same in these two figures due to its independence of c. We only obtain more smoothing from the Aitchison and Aitken kernel when the bandwidth is close to its upper bound.

2.3.2. *Asymptotic Normality of the Density, Conditional Density and Conditional Mean Estimators.* The importance of the comparison to the unordered kernel of Aitchison and Aitken is subtle. The theoretical work of Li and Racine (2003) and Racine and Li (2004), establishing asymptotic normality of the kernel smoothed estimators for the density, conditional density and conditional mean (regression), does not make explicit assumptions on the discrete kernel function. However,

[8]It is not obvious whether this relationship holds in the multivariate setting.

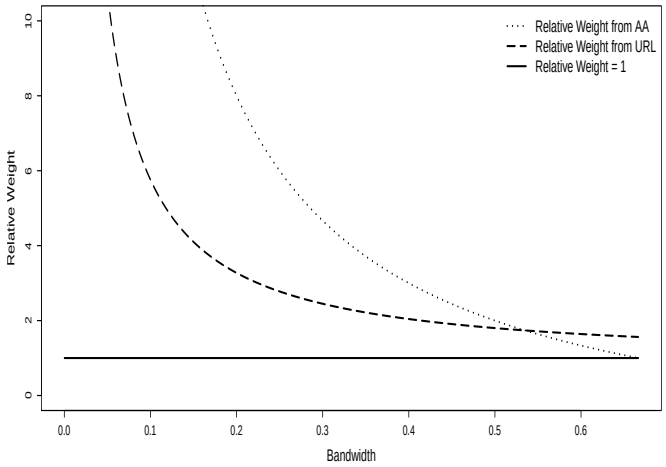


FIGURE 1. Change in Relative Weights: URL vs AA ($c = 3$, $\lambda_{AA}^{upper} = 0.6667$)

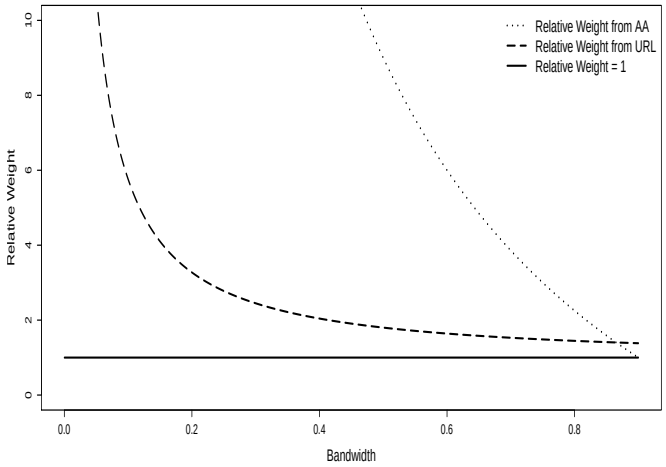


FIGURE 2. Change in Relative Weights: URL vs AA ($c = 10$, $\lambda_{AA}^{upper} = 0.9000$)

they do require (see A.3 in Li and Racine, 2003) that the discrete kernel function admits a power series expansion/approximation. The Aitchison and Aitken kernel satisfies this power series expansion by default. Thus, our connection here suggests that the unordered Rajagopalan and Lall kernel will also admit the same power series expansion,[9] which then implies that the asymptotic normality of the corresponding density, conditional density and conditional mean estimators remains.

The form of the asymptotic distribution of the density, conditional density and conditional mean estimators across various discrete kernels makes it difficult to assess which kernel is preferred. First, in the discrete only setting, Li and Racine (2003) have demonstrated that the smoothed discrete

[9]See Appendix B for a formal derivation of this expansion.

kernel density estimator does not possess any bias, and the variance is independent of the kernel selected. Second, in the mixed data setting, while a bias is introduced from smoothing the discrete component of the data, this bias depends on the unknown discrete marginals and not on the kernel function itself. The variance of the density estimator only depends on the variance of the continuous kernel deployed. This suggests the lack of clear theoretical evidence guiding choice of kernel. To remedy this, we turn to simulations to help provide some insight into the relative merits of the Rajagopalan and Lall kernels relative to their competitors.

3. Simulations

This section provides comprehensive simulation settings in order to find differences between kernel functions in their finite-sample performance. We will provide details on the design of the simulations and then present the results.

3.1. **Settings.** Here we consider three cases in the density and regression settings: standard data, sparse data, and inclusion of an irrelevant variable. Consequently, we develop nine (four density and five regression) settings: (a) univariate discrete density, (b) univariate discrete density with sparse data, (c) conditional density with a continuous random variable, (d) conditional density with a discrete random variable, (e) cross-sectional regression, (f) cross-sectional regression with sparse data, (g) cross-sectional regression with an irrelevant discrete regressor, (h) panel data regression, and (i) panel data regression with an irrelevant discrete regressor.[10]

3.1.1. *Univariate Discrete Density.* The data are generated (as shown in Tables A1 and A2 in Appendix A) by using the multinomial ($c = 3$) and beta-binomial ($c = 3$) distributions for unordered and ordered kernel functions, respectively. We consider four scenarios for each distribution. The beta-binomial distribution is the discrete version of the beta distribution and has two parameters (α and β) that determine the first four moments (see Dong and Simonoff, 1994). We use the value $r = \frac{p_{\max}}{p_{\min}}$ as a measure of the design. We have a uniform design as r goes to one and empty cells (few observations in almost all cells) as r goes to infinity.

We consider relative MSE to evaluate the finite sample performance of kernel functions and take the median over $1,000$ replications for each sample size $n = 50$, 100, and 200,

$$R^{MSE} = \frac{\sum_{i=1}^{n} \left[\widehat{p}\left(X_i\right) - p\left(X_i\right)\right]^2}{\sum_{i=1}^{n} \left[\widehat{p}^{*}\left(X_i\right) - p\left(X_i\right)\right]^2},$$

[10] The sparse data case is implicitly included in the panel data regression setting, see Section 3.1.8.

where $\widehat{p}^*(\cdot)$ is a kernel density estimator with the unordered or ordered Rajagopalan and Lall kernel function and $\widehat{p}(\cdot)$ is a kernel density estimator with one of the alternative unordered or ordered kernel functions.

3.1.2. *Univariate Discrete Density with Sparse Data.* Again, we consider the relative MSE to evaluate the finite sample performance of kernel functions and take the median over $1,000$ replications for each cell number $c = 10$, 20, and 50. We use the value $s = \frac{n}{c} = 1$ (very sparse) and 2 (moderate sparse) as a sparsity measure (Dong and Simonoff, 1994) and also use it to determine the sample size n. The data are generated by using Scenario (I) in Table A2 in Appendix A and hence we have a right-skewed distribution.

3.1.3. *Conditional Density with a Continuous Random Variable.* Inspired by Huynh and Jacho-Chávez (2009), we consider a modified version of their DGP

$$\text{DGP: } Y_i|X_i \sim N(X_i, 1),$$

where X is a discrete random variable taking values 0, 1, or 2 with equal probabilities from the multinomial distribution or with $p_0 = 0.40$, $p_1 = 0.20$, $p_2 = 0.40$ from the beta-binomial distribution. In other words, for the unordered discrete variable X, $f(Y)$ is an equal mixture of $N(0,1)$, $N(1,1)$, and $N(2,1)$ and hence Y_i is generated from $N(0,1)$, $N(1,1)$, or $N(2,1)$ with equal probabilities; for the ordered discrete variable X, $f(Y)$ is a 40/20/40 mixture of $N(0,1)$, $N(1,1)$, and $N(2,1)$ and hence Y_i is generated from $N(0,1)$, $N(1,1)$, or $N(2,1)$ with unequal probabilities.[11]

We consider the relative MSE to evaluate the finite sample performance of kernel functions and take the median over $1,000$ replications for each sample size $n = 50$, 100, and 200,

$$R^{MSE} = \frac{\sum_{i=1}^{n}\left[\widehat{f}(Y_i|X_i) - f(Y_i|X_i)\right]^2}{\sum_{i=1}^{n}\left[\widehat{f}^*(Y_i|X_i) - f(Y_i|X_i)\right]^2},$$

where $\widehat{f}^*(\cdot)$ is a kernel conditional density estimator with the unordered or ordered Rajagopalan and Lall kernel function and $\widehat{f}(\cdot)$ is a kernel conditional density estimator with alternative unordered or ordered kernel functions.

If the unordered discrete variable X takes values 0, 1, or 2 with $p_0 = 0.15$, $p_1 = 0.35$, and $p_2 = 0.50$, or the ordered discrete variable X takes values 0, 1, or 2 with $p_0 = 0.25$, $p_1 = 0.50$, $p_2 = 0.25$, the discrete variable X will become an irrelevant variable, which means $f(Y_i|X_i) = f(Y_i)$. Note that Y still follows its original distribution.

[11]A bivariate density with mixed data and a marginal density for the continuous variable are $f(Y,X) = f(Y|X)\cdot p(X)$ and $f(Y) = \sum_X f(Y,X)$, respectively.

3.1.4. *Conditional Density with a Discrete Random Variable.* Inspired by Hall, Racine, and Li (2004), we consider a modified version of their DGP

$$\text{DGP: } Y_i^* = Z_i + X_i + Z_i X_i + u_i$$

and

$$Y_i = \begin{cases} 1 \text{ if } Y_i^* > 0.5 \\ 0 \text{ otherwise} \end{cases},$$

where Y^* is a latent variable, $Z_i \sim U[0,1]$, $u_i \sim N(0,1)$, and X is a discrete random variable taking values -1, 0, or 1 with equal probabilities from the tri-nomial or beta-binomial distributions.

We consider the relative (out-of-sample) correct classification ratio (CCR) ($n_2 = 50$) to evaluate the finite sample performance of kernel functions and take the median over $1,000$ replications for each sample size $n_1 = 50$, 100, and 200,

$$R^{CCR} = \frac{CCR^*}{CCR},$$

where

$$CCR = \frac{\#\text{ of } \widehat{Y}_i = Y_i}{n_2},$$

and

$$\widehat{Y}_i = \begin{cases} 1 \text{ if } \widehat{p}(Y_i = 1 | Z_i = z, X_i = x) > 0.5 \\ 0 \text{ otherwise} \end{cases},$$

where CCR^* is obtained from a kernel conditional density estimator $\widehat{p}^*(\cdot)$ with the unordered or ordered Rajagopalan and Lall kernel function and CCR is obtained from a kernel conditional density estimator $\widehat{p}(\cdot)$ with different unordered or ordered kernel functions.[12] It is noted that if X is removed from the DGP, it will become an irrelevant discrete variable.

3.1.5. *Cross-Sectional Regression.* Here we consider the DGP

$$\text{DGP: } Y_i = \sin(\pi Z_i) + D_i + Z_i D_i + u_i,$$

where $Z_i \sim U[0,2]$, $u_i \sim N(0,1)$, and D is an effect related to a discrete random variable (X) taking values 0 or 1 with equal probabilities from the binomial distribution or taking values 0, 1, or 2 with equal probabilities from the beta-binomial distribution. D takes two or three different values which are randomly chosen from $U[0,6]$.

[12]Note that here our proposed kernels will be represented in the numerator. We do so to keep the tables easier to read. Values greater than one will represent improved performance.

We consider the relative (out-of-sample) MSE ($n_2 = 50$) to evaluate the finite sample performance of the kernel functions and take the median over $1,000$ replications for each sample size $n_1 = 50$, 100, and 200. Formally, this is calculated as

$$MSE = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \widehat{Y}_i)^2,$$

where $\widehat{Y}_i \equiv \widehat{g}(Z_i, X_i)$ and $\widehat{g}(\cdot)$ is estimated via local-constant least-squares (LCLS).

3.1.6. *Cross-Sectional Regression with Sparse Data.* We follow Section 3.1.5, but redefine X as a discrete regressor with $c = 50$, see Section 3.1.2. Specifically, D takes fifty different values, which are randomly chosen from $U[0, 6]$.

3.1.7. *Cross-Sectional Regression with an Irrelevant Discrete Regressor.* We follow Section 3.1.5, but redefine X as an irrelevant discrete regressor.

3.1.8. *Panel Data Regression.* Here we consider both one-way and two-way error component models under the so-called fixed effects assumption (the individual effects are allowed to be correlated with the regressors). For the one-way fixed effects model, the DGP is

$$\text{DGP: } Y_{it} = \sin(\pi Z_{it}) + x_i^u + w_i + Z_{it}x_i^u + Z_{it}w_i + u_{it},$$

where $i = 1 \dots n$, $t = 1 \dots T$, $Z_{it} \sim U[0, 2]$, $u_{it} \sim N(0, 1)$ and x^u is an individual-specific effect (e.g., country). For practical purposes, a second *discrete* and *time-invariant* effect w (e.g., type of political system) is generated in the DGP. x^u takes fifty, one-hundred, or two-hundred different values, which are randomly chosen from $U[0, 6]$, and w takes ten different values, which are randomly chosen from $U[0, 2]$. The nonparametric regression model is

$$Y_{it} = g(Z_{it}, X_i^u, W_i) + v_{it},$$

where $g(\cdot)$ is estimated via LCLS. As defined in Section 3.1.5, X^u and W correspond to x^u and w, respectively.

For the two-way fixed effects model, the DGP is

$$\text{DGP: } Y_{it} = \sin(\pi Z_{it}) + x_i^u + x_t^o + Z_{it}x_i^u + Z_{it}x_t^o + u_{it},$$

where x^o is a time-specific effect (e.g., year) and all others are the same as before. x^o takes three or five different values, which are randomly chosen from $U[0, 2]$. The nonparametric regression model is

$$Y_{it} = g(Z_{it}, X_i^u, X_t^o) + v_{it},$$

where $g(\cdot)$ is estimated via LCLS. Again, X^u and X^o correspond to x^u and x^o, respectively.

For each panel data setting, we consider the relative (out-of-sample) MSE with $n_2 = n$ ($T = 1$), to evaluate the finite sample performance of kernel functions and take the median over $1,000$ replications for each sample size $n_1 = 150$ ($n = 50, T = 3$), 250 ($n = 50, T = 5$), 300 ($n = 100, T = 3$), 500 ($n = 100, T = 5$), 600 ($n = 200, T = 3$), and 1000 ($n = 200, T = 5$). The (out-of-sample) MSE is then calculated as

$$MSE = \frac{1}{n_2} \sum_{t=1}^{T} \sum_{i=1}^{n} (Y_{it} - \widehat{Y}_{it})^2.$$

If we still use the sparsity measure $s = \frac{n}{c}$, we will always have $s = \frac{n_1}{nT} = 1$ for the two-way fixed effects model and $s = \frac{n_1}{10n} < 1$ for the one-way fixed effects model. Therefore, the sparse data case is implicitly included in the panel data regression setting.

3.1.9. *Panel Data Regression with an Irrelevant Discrete Regressor.* For the one-way fixed effects model, we follow Section 3.1.8, but redefine X^u as an irrelevant discrete regressor. For the two-way fixed effects model, we follow Section 3.1.8, but redefine X^o as an irrelevant discrete regressor.

3.2. **Estimation Specifics.** In each setting, LSCV is used as the bandwidth selector. For Settings 8-9, we choose LSCV with the leave-one-cross-section-out estimator in order to reduce the impact of an outlier of the individual level, see Chapter 11 of Henderson and Parmeter (2015). Additionally, a Gaussian kernel function is used for all continuous variables.

The upper bounds for bandwidths from the unordered and ordered Rajagopalan and Lall kernel functions are infinity. Theoretically, the importance of infinity is to enable the kernel functions to give precisely equal weights, see Property (iii) in Section 2. For a continuous variable, its standard deviation can be used to construct an upper bound for the bandwidth from the continuous kernel function, but this may be meaningless for discrete data. In Section 3.3, we will show that a value ensuring that the kernel functions can produce equal weights to the second or third decimal place can be used as an upper bound in settings with sparse data and irrelevant variables. In other words, a relatively small imposed upper bound is typically enough for most simulated and real data sets. Therefore, we set the upper bounds equal to one in most of the density settings and all of the regression settings. Typically, the optimal bandwidth chosen by LSCV occurs between zero and one (even with the higher theoretical upper bound). This can be further justified by Equation (5) for the unordered Rajagopalan and Lall kernel function.[13]

[13]In practice, we may first set the upper bound equal to one for the Rajagopalan and Lall kernel functions. If the cross-validated bandwidth reaches the upper bound, we may raise the upper bound up to some value ensuring that the kernel function can produce equal weights to say the second or third decimal place.

3.3. **Results.** Before presenting the results, we first explain the abbreviations used in the tables. The Aitchison and Aitken, Wang and van Ryzin, unordered and ordered Li and Racine, and unordered and ordered Rajagopalan and Lall kernel functions are listed as AA, WVR, ULR, OLR, URL, and ORL, respectively.[14] The frequency approach is listed as F. The Aitchison and Aitken kernel function is paired with the Wang and van Ryzin kernel function for the two-way fixed effects model and we abbreviate this to AW. Abbreviations for the other two pairs (i.e., unordered and ordered Li and Racine and unordered and ordered Rajagopalan and Lall kernel functions) are LR and RL.

To avoid too many tables, simulation results from those nine settings are sorted into: density estimation in the standard data case (Setting (a)), density estimation in the sparse data case (Setting (b)), density estimation in the irrelevant variable case (Settings (c)-(d)), cross-sectional regression (Settings (e)-(g)), and panel data regression (Settings (h)-(i)). We primarily use the median of the relative-MSE distribution to evaluate the performance of kernel functions and the median of the bandwidth distribution is given in Appendix A.[15]

3.3.1. *Density Estimation in the Standard Data Case.* Table 1 shows that the unordered Rajagopalan and Lall type kernel function performs similarly to the Aitchison and Aitken kernel function in all scenarios due to their similar representations, but better than the unordered Li and Racine kernel function and the frequency approach in Scenario (III) with $n = 50$ and 100, and Scenario (IV). The ordered Rajagopalan and Lall kernel function performs similarly to the Wang and van Ryzin kernel function, but better than the ordered Li and Racine kernel function and the frequency approach in Scenario (IV) with $n = 50$ and 100. Essentially the gains from the Rajagopalan and Lall kernel functions in the above scenarios decrease as the sample size increases except for the uniform design in the unordered case. The reason is that the bandwidth from the unordered Li and Racine kernel function does not converge to its upper bound as the underlying density approximates the uniform distribution.[16] As a result, its relatively small bandwidth makes its performance similar to the frequency approach.

[14]We would like to emphasize here that the Li and Racine kernels do not require the number of cells c to be known. We assume that they are known in our simulations and hence kernel functions which assume the value is known may exploit this relative advantage.

[15]Many of the results display a right-skewed distribution. Even though using the mean makes our kernels appear to perform even better, we believe that using the median is more objective. The results at the mean are available upon request.

[16]Table A3 shows that the bandwidth for the Li and Racine kernel function is relatively small in this case. However, in the conditional density case, Table A5, the kernel function produces a large bandwidth when the variable is irrelevant. This is not necessarily surprising as the theory in Hall, Racine and Li (2004) is for the case of a conditional density estimator.

TABLE 1. Median of Relative MSE (Density Estimation in the Standard Data Case)

	AA/URL	ULR/URL	F/URL	WVR/ORL	OLR/ORL	F/ORL
	(I) $p_1 = 0.15$, $p_2 = 0.35$, $p_3 = 0.50$			(I) $p_1 = 0.55$, $p_2 = 0.33$, $p_3 = 0.17$		
$n = 50$	1.0000	0.9588	1.0570	0.9428	0.9120	1.0279
$n = 100$	1.0000	1.0095	1.0176	0.9522	0.8946	0.9387
$n = 200$	1.0000	0.9737	0.9808	0.9439	0.8907	0.8955
	(II) $p_1 = 0.25$, $p_2 = 0.30$, $p_3 = 0.45$			(II) $p_1 = 0.25$, $p_2 = 0.50$, $p_3 = 0.25$		
$n = 50$	1.0000	1.1437	1.1694	1.0908	1.1575	1.2978
$n = 100$	1.0000	0.9321	0.9457	1.0060	1.0695	1.0798
$n = 200$	1.0000	0.8765	0.8852	1.0177	1.0144	1.0170
	(III) $p_1 = 0.25$, $p_2 = 0.35$, $p_3 = 0.40$			(III) $p_1 = 0.40$, $p_2 = 0.20$, $p_3 = 0.40$		
$n = 50$	1.0000	1.5150	1.5830	0.9867	1.0792	1.1117
$n = 100$	1.0000	1.2601	1.2905	0.9955	1.0279	1.0432
$n = 200$	1.0000	1.0754	1.0842	0.9292	0.9370	0.9477
	(IV) $p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$			(IV) $p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$		
$n = 50$	0.9868	4.0191	4.2563	1.0225	1.2904	1.3361
$n = 100$	0.9868	4.4062	4.5375	1.0054	1.2110	1.2398
$n = 200$	0.9868	4.4328	4.4990	0.9839	1.0411	1.0460

3.3.2. *Density Estimation in the Sparse Data Case.* Table 2 shows that the unordered Rajagopalan and Lall type kernel function has the same performance as the Aitchison and Aitken kernel function due to their similar representations, but better than the unordered Li and Racine kernel function and the frequency approach. The ordered Rajagopalan and Lall kernel function performs similarly to the Wang and van Ryzin kernel function, but better than the ordered Li and Racine kernel function and the frequency approach.[17] Essentially the gains from the unordered and ordered Rajagopalan and Lall kernel functions increase as the number of cells increases, but decrease as the sample size increases.

TABLE 2. Median of Relative MSE (Density Estimation in the Sparse Data Case)

	AA/URL	ULR/URL	F/URL	WVR/ORL	OLR/ORL	F/ORL
$c = 10$, $n = 10$	1.0000	4.0136	5.0549	1.1874	6.1655	7.1457
$c = 10$, $n = 20$	1.0000	2.6825	3.0176	1.0561	3.7734	4.2317
$c = 20$, $n = 20$	1.0000	5.8596	6.5554	1.0312	8.3315	9.0160
$c = 20$, $n = 40$	1.0000	3.4955	3.6931	0.9479	4.8224	5.0504
$c = 50$, $n = 50$	1.0000	8.0997	8.4792	1.0325	15.0093	15.4892
$c = 50$, $n = 100$	1.0000	4.2342	4.3376	0.9449	7.7759	7.9129

[17]For the ordered Rajagopalan and Lall kernel function, we set the upper bound to twenty in order to provide more smoothing for a relatively large numbers of cells in the sparse data case and this also ensures that Equation (2) can produce equal weights to the third decimal place.

3.3.3. *Density Estimation in the Irrelevant Variable Case.* Table 3 gives the results for the continuous random variable case. The unordered Rajagopalan and Lall type kernel function performs much better than the Aitchison and Aitken and unordered Li and Racine kernel functions, but similar to the frequency approach due to the relatively few number of cells with the relatively large sample size ($c = 3$ and $n = 50$, 100, or 200). If we consider $n = 25$, the unordered Rajagopalan and Lall kernel function can outperform the frequency approach. For ordered data, we find two cases whereby the Rajagopalan and Lall kernel does worse as compared to the Wang and van Ryzin kernel function. Table A5 in Appendix A shows that all of the bandwidths for the unordered and ordered cases reach their upper bounds.[18] We would like to note here that we find substantial differences in the bandwidths for the continuous variables even though we use a Gaussian kernel in each case.

For the discrete random variable, Table 3 shows that there is no significant difference between the kernel functions and between the kernel and frequency approaches.[19] Table A5 in Appendix A shows that all of the bandwidths reach their upper bounds. The main reason for the result here is that the finite sample performance is evaluated by CCR rather than MSE. Unless conditional density estimators with different kernel functions have quite different estimates ($\widehat{p}(\cdot)$), these estimators will give similar predicted values ($\widehat{Y}$) and hence similar CCR.

TABLE 3. Median of Relative MSE for Continuous Random Variable and Median of Relative CCR for Discrete Random Variable (Density Estimation in the Irrelevant Variable Case)

	AA/URL	ULR/URL	F/URL	WVR/ORL	OLR/ORL	F/ORL
	Continuous Random Variable			Continuous Random Variable		
$n = 50$	3.4638	3.4731	1.0459	0.1450	3.9198	0.4056
$n = 100$	3.4765	3.4828	0.9283	0.0932	3.9169	0.2431
$n = 200$	3.4718	3.4824	0.8646	0.0610	3.9109	0.1553
	Discrete Random Variable			Discrete Random Variable		
$n = 50$	1.0339	1.0392	1.0000	1.0000	1.0313	1.0000
$n = 100$	1.1035	1.0833	1.0290	0.9630	1.0714	1.0000
$n = 200$	1.1154	1.1429	1.0000	0.9667	1.1154	1.0000

[18]For the unordered Rajagopalan and Lall type kernel function, its upper bound is set to be six in order to make the discrete variable more thoroughly smoothed out via LSCV in the irrelevant variable case and this also ensures that Equation (3) can produce equal weights to the second decimal place. Similarly, the upper bound for ordered Rajagopalan and Lall kernel function is set to be ten.

[19] In Hall, Racine, and Li's (2004, p. 1023) simulation result for the irrelevant variable case, the median of the relative (out-of-sample) CCR of the unordered Li and Racine kernel function to the the frequency approach is 1.1164.

3.3.4. *Cross-Sectional Regression.* Table 4 gives the results for the standard data case. There is essentially no difference between the kernel functions, but for the unordered case, the kernel approach is slightly better than the frequency approach.

For the sparse data case, the unordered Rajagopalan and Lall type kernel function performs similarly to the Aitchison and Aitken and unordered Li and Racine kernel functions, but better than the frequency approach. The ordered Rajagopalan and Lall kernel function does not perform as well as the Wang and van Ryzin and ordered Li and Racine kernel functions, but better than the frequency approach. The reason is that the upper bound for the ordered Rajagopalan and Lall kernel function is set to be one here. If its upper bound is set to be twenty, its relative (out-of-sample) MSE to each of the other two kernel functions is around 0.85. Note that the ordered Rajagopalan and Lall kernel function performs closer to the other two kernel functions as the sample size increases.

For the irrelevant variable case, the results are similar to the standard data case. Table A6 in Appendix A shows that all of the bandwidths reach their upper bounds. The main reason why the results here are so similar is because the irrelevant variable is thoroughly smoothed out and regression estimates are solely determined by the continuous variable. Consequently, regression estimators give the same estimates and hence the same predicted values.

TABLE 4. Median of Relative Out-of-Sample MSE (Cross-Sectional Regression)

	AA/URL	ULR/URL	F/URL	WVR/ORL	OLR/ORL	F/ORL
	Standard Data Case			Standard Data Case		
$n = 50$	1.0251	1.0907	1.2485	1.0176	1.0010	1.0248
$n = 100$	1.0202	1.0700	1.1915	1.0098	1.0068	1.0210
$n = 200$	1.0095	1.0500	1.1341	1.0194	1.0140	1.0353
	Spare Data Case			Spare Data Case		
$n = 50$	1.0078	1.0000	2.0774	0.5562	0.5457	1.4149
$n = 100$	1.0025	1.0053	1.8981	0.6869	0.6722	1.5528
$n = 200$	1.0073	1.0248	1.5233	0.8250	0.8106	1.5469
	Irrelevant Variable Case			Irrelevant Variable Case		
$n = 50$	1.0000	1.0000	1.0359	0.9864	0.9833	1.0584
$n = 100$	1.0000	1.0000	1.0247	0.9920	0.9892	1.0397
$n = 200$	1.0000	1.0000	1.0221	0.9960	0.9924	1.0293

3.3.5. *Panel Data Regression.* For the one-way fixed effects model with the standard data case, Table 5 shows that the unordered Rajagopalan and Lall type kernel function performs similarly to the Aitchison and Aitken kernel function, but better than the unordered Li and Racine kernel function

and the frequency approach. Table A7 in Appendix A shows that the unordered Li and Racine kernel function undersmooths the individual-specific effect (i.e., relatively small bandwidths). For the two-way fixed effects model with the standard data case, there is essentially no difference between the kernel functions, but the frequency approach does not perform as well.

For the one-way fixed effects model with the irrelevant variable case, the unordered Rajagopalan and Lall type kernel function performs better than its competitors. Table A7 in Appendix A shows that compared to the Aitchison and Aitken kernel function, the unordered Rajagopalan and Lall kernel function not only smooths out the irrelevant regressor, but also provides sufficient smoothing for the relevant discrete regressor.[20] For the two-way fixed effect model with an irrelevant variable, the results are similar to the standard data case.

There are two points we want to emphasize for the one-way fixed effects model with an irrelevant variable. First, the relatively large bandwidths for the irrelevant variable (X^u) from the unordered Rajagopalan and Lall and Aitchison and Aitken kernel functions should be interpreted cautiously because they are caused by not only the irrelevance of the variable, but also the uniform distribution of the variable.[21] Second, the relative (out-of-sample) MSE from the kernel approach increases as the overall sample size increases, but this does not mean that the (out-of-sample) MSE of each kernel regression estimator increases. Actually, for each estimator, the (out-of-sample) MSE decreases as the overall sample size increases, but the reduced MSE is quite different.

4. Empirical Illustrations

Here we consider different types of real data to evaluate the empirical performance of the kernel functions. We consider a univariate unordered discrete density (Lindsey, 1995 and Greene, 2011), univariate ordered discrete density with sparse data (Simonoff, 1996), discrete conditional density (Li and Racine, 2004), and panel data regression (Cameron and Trivedi, 2005 and Henderson and Kumbhakar, 2006). We will introduce each of the datasets and then present the results.

4.1. **Data.**

[20]For comparison, we can use Equation (5), for the unordered regressor. For example, using Equation (5) for $n = 50$ and $T = 3$, for the relevant discrete regressor, we can obtain a new bandwidth for the unordered Rajagopalan and Lall kernel function on the basis of the value of the bandwidth from the Aitchison and Aitken kernel function, and it is 0.0011 (which is less than 0.0079).

[21]For the two-way fixed effects model with an irrelevant variable, the bandwidths for the (relevant) individual-specific effect from the unordered Rajagopalan and Lall and and Aitchison and Aitken kernel functions are relatively large due to their uniform distributions.

TABLE 5. Median of Relative Out-of-Sample MSE (Panel Data Regression)

	AA/URL	ULR/URL	F/URL	AW/RL	LR/RL	F/RL
	One-Way Fixed Effects			Two-Way Fixed Effects		
	Standard Data Case			Standard Data Case		
$n = 50$, $T = 3$	1.0040	1.3011	1.5249	1.0054	1.0783	1.6757
$n = 50$, $T = 5$	1.0084	1.3823	1.5134	0.9947	1.0703	1.6287
$n = 100$, $T = 3$	1.0327	1.3168	1.6311	1.0062	1.0693	1.6358
$n = 100$, $T = 5$	1.0417	1.3722	1.6376	0.9988	1.0571	1.5704
$n = 200$, $T = 3$	1.0361	1.2693	1.7041	1.0050	1.0490	1.5791
$n = 200$, $T = 5$	1.0407	1.3058	1.6972	1.0030	1.0397	1.5063
	Irrelevant Variable Case			Irrelevant Variable Case		
$n = 50$, $T = 3$	1.0159	1.0898	1.1566	1.0033	1.0534	1.7350
$n = 50$, $T = 5$	1.0795	1.1565	1.2071	1.0012	1.0635	1.6946
$n = 100$, $T = 3$	1.1834	1.4134	1.5470	1.0020	1.0525	1.6834
$n = 100$, $T = 5$	1.2700	1.5322	1.6169	1.0004	1.0596	1.6305
$n = 200$, $T = 3$	1.2132	1.5060	1.7073	1.0007	1.0338	1.6079
$n = 200$, $T = 5$	1.2557	1.4825	1.6432	1.0001	1.0445	1.5701

4.1.1. *Univariate Discrete Density Estimation.* Here we consider three univariate data sets. The first data set comes from Jarrett (1979). Jarrett (1979) tabulates explosions in British mines for the years 1851 to 1962 (see also, Lindsey 1995). Here we observe both the day of the week and the month in which the explosion occurred. We therefore have two data sets for which the former lists $n = 191$ such occurrences over $c = 7$ days of the week. The latter lists the same $n = 191$ such occurrences over $c = 12$ different months. Jarrett (1979) emphasizes that the frequency of accidents is related to days of the week but seems unrelated to months. No relationship between accidents and months would allow us to use unordered kernel functions in kernel density estimation. Given the nature of the data, we treat this data as unordered.

For our second example, we take the data from Greene (2011). Greene (2011) defines travel mode choice as transportation mode choice (i.e., air, train, bus, or car) for travel between Sydney and Melbourne, Australia ($n = 210$, $c = 4$) and this is an unordered discrete variable. We look at $n = 210$ individuals over the four modes of transportation $c = 4$.

Finally, we consider the sparse data case in Simonoff (1996). Simonoff (1996) also looks at British mine explosion data (1875-1951) and specifically looks at the time intervals in days between mine explosions which resulted in at least 10 people being killed ($n = 109$, $c = 55$). The sparsity measure measure $s = \frac{n}{c}$, for this particular ordered variable, displays moderate sparseness ($s = 1.9818$).

4.1.2. *Conditional Discrete Density Estimation.* Li and Racine (2004) use conditional density estimation to examine what factors could be relevant predictors of the propensity to engage in extramarital affairs. Eight discrete variables are used to explain the binary variable extramarital affairs (yes or no) and they are gender, children (yes or no), occupation, age, years married, how religious, education, and marriage rating. Note that the first three variables are unordered and the remaining five variables are ordered. The sample size n is 601. To avoid over-fitting, the relative (out-of-sample) correct classification ratio is considered. We randomly shuffle the data into an estimation sample of size $n_1 = 501$ and an evaluation sample of size $n_2 = 100$ and create $1,000$ such random splits.

4.1.3. *Panel Data Regression Estimation.* Cameron and Trivedi (2005) adopt several different nonlinear panel data models to examine which factors may determine the number of patents. They use firm-level panel data in the United States from 1975-1979 ($n = 346$, $T = 5$, $N = 1730$). Three variables are considered to explain the number of patents and they are a dummy for firms in the scientific sector, log of R&D spending, and log of R&D stock. We consider two additional variables: state-specific and time-specific effects. We adopt the two-way fixed effects model and use the last year as an evaluation sample ($n_2 = 346$).

Henderson and Kumbhakar (2006) adopt a nonparametric regression model, instead of Cobb-Douglas-based regression models (e.g., Munnell, 1990 and Baltagi and Pinnoi, 1995), to examine whether public expenditure affects output. They choose state-level panel data from 1970-1986 ($n = 48$, $T = 17$, $N = 816$). Six variables are used to explain gross state product and they are employment in non-agricultural payrolls, private capital stock, public capital, unemployment rate, state-specific, and time-specific effects. We adopt the two-way fixed effects model and use the last year as an evaluation sample ($n_2 = 48$).

4.2. **Results.**

4.2.1. *Univariate Discrete Density Estimation.* As expected, the unordered Rajagopalan and Lall type kernel function performs similarly to the Aitchison and Aitken kernel function in the data from Lindsey (1995) and Greene (2011). Figure 1 (a) shows that both of the kernel density estimates from the unordered Rajagopalan and Lall and the Aitchison and Aitken kernel functions are essentially the same and this is consistent with the result from the Pearson's chi-squared test. Figure 1 (b) shows that there is no meaningful difference between the kernel functions and between the kernel

and frequency approaches. The reason is that this dataset has relatively few cells and a relatively large sample size ($n = 210$, $c = 4$).

For the data from Simonoff (1996), Figure 1 (c) provides the frequency estimates and kernel density estimates from the ordered Li and Racine, Wang and van Ryzin, and ordered Rajagopalan and Lall kernel functions. The frequency estimates display two features: a rough structure around the 10th cell and many zero counts after the 20th cell. Although the kernel approach with the ordered Li and Racine or Wang and van Ryzin kernel function solves the problem with zero counts, they do not smooth the rough part. The ordered Rajagopalan and Lall kernel function not only fixes zero counts, but also smooths the rough part. However, the ordered Rajagopalan and Lall kernel function results in a boundary bias due to a downward fitted line in the first three cells.[22]

4.2.2. *Conditional Discrete Density Estimation.* Using the data from Li and Racine (2004), Table 6 shows that both of the Rajagopalan and Lall kernel functions perform similarly to their competitors and essentially all of the medians of relative (out-of-sample) CCR are unity.[23] This result is consistent with the simulations. In conditional density estimation with a discrete random variable, the kernel approach may be slightly better then the frequency approach (as shown in Hall, Racine, and Li, 2004), but if we restrict our attention to the kernel approach, there is no significant difference between the kernel functions.

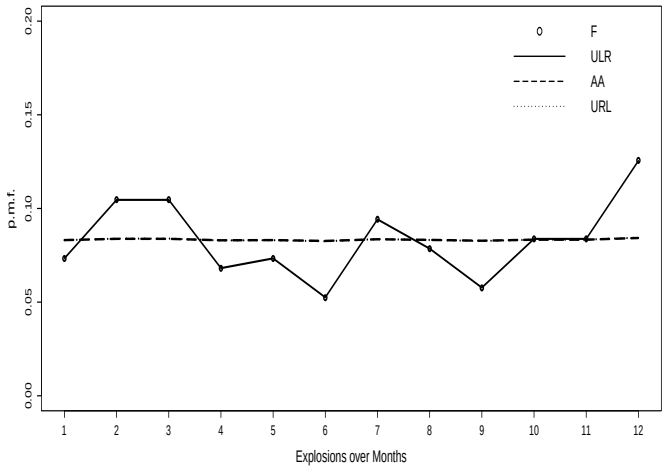
TABLE 6. Median of Relative Out-of-Sample CCR for Li and Racine (2004) and Relative Out-of-Sample MSE for Cameron and Trivedi (2005) and Henderson and Kumbhakar (2006)

	AW/RL	LR/RL	F/RL
Li and Racine (2004)	1.0308	1.0400	1.0435
Cameron and Trivedi (2005)	0.2291	0.3174	0.3222
Henderson and Kumbhakar (2006)	1.7997	1.8494	3.1872

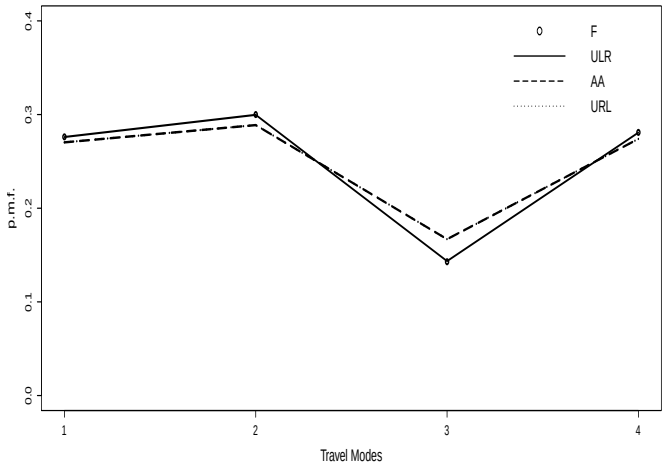
4.2.3. *Panel Data Regression Estimation.* Using the data from Cameron and Trivedi (2005), Table 6 shows that both of the Rajagopalan and Lall kernel functions do not perform as well as their

[22]The results here for the ordered Rajagopalan and Lall kernel function is similar to Dong and Simonoff (1994), Rajagopalan and Lall (1995), and Simonoff (1996). It is noted that Dong and Simonoff (1994) and Rajagopalan and Lall (1995) use Hall and Titterington's (1989) kernel density estimator, which is different from what we use (i.e., Rosenblatt, 1956), and Simonoff (1996) uses the local-polynomial smooth approach.

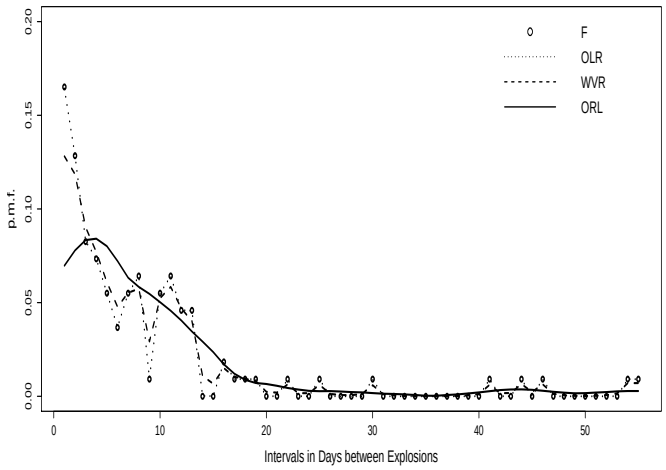
[23]The 10th and 90th percentiles of the relative (out-of-sample) CCR distribution for AW/RL is 0.9857 and 1.0975, respectively. The 10th and 90th percentiles of the relative (out-of-sample) CCR distribution for LR/RL is 0.9733 and 1.1077, respectively. The 10th and 90th percentiles of the relative (out-of-sample) CCR distribution for F/RL is 0.9726 and 1.1143, respectively.



(a)



(b)



(c)

FIGURE 3. Univariate Discrete Density Estimation: (a) Mine Explosion (Jarrett, 1979), (b) Travel Mode Choice (Greene, 2011), and (c) Mine Explosion (Simonoff, 1996)

competitors. The main reason is that for the state-specific effect, the unordered Rajagopalan and Lall type kernel function now oversmooths.[24]

Using the data from Henderson and Kumbhakar (2006), Table 6 shows that both of the Rajagopalan and Lall kernel functions significantly outperform their competitors. It turns out that an additional amount of smoothing leads to the improved performance.[25]

5. Conclusion

In this paper we consider non-geometric kernel functions for use with discrete data. Specifically, we start with Rajagopalan and Lall's (1995) kernel function for ordered data and propose an unordered version. For each of these kernel functions, MSSE for the kernel density estimator is derived and we demonstrate conditions for asymptotic normality for the respective kernel density and regression estimators.

We compare these kernels with existing kernel functions via simulated and real data. The simulation results show our kernel functions are typically as good as or better than alternative kernel functions. That being said, we find several cases of substantial improvement. For example, we find improvements when the data are sparse, the original motivation for such kernels. One case of specific interest is in panel data estimation when an unordered discrete variable is used to capture unobservable individual effects. The empirical results essentially mimic the simulation results and hence we believe that these kernel functions will perform well in practice.

[24]For the dummy for firms in the scientific sector, the bandwidths for URL, AA, and ULR are 1.0000, 0.4417, and 0.4064, respectively. For the state-specific effect, the bandwidths for URL, AA, and ULR are 0.9938, 0.6075, and 0.0841, respectively. For the time-specific effect, the bandwidths for ORL, WVR, and OLR are 0.1838, 0.5374, and 0.0972, respectively. However, these numbers are not directly comparable. To try to make such a comparison, we can use Equation (5), for the unordered regressor. Using Equation (5), for both of the unordered discrete regressors, we can obtain new bandwidths for the unordered Rajagopalan and Lall kernel function on the basis of the values of the bandwidths from the Aitchison and Aitken kernel function, and they are 0.0012 (< 1.0000) and 0.0023 (< 0.9938) for the dummy for firms in the scientific sector and the state-specific effect, respectively. All else equal, the relative (out-of-sample) MSE would increase from 0.2291 to 0.8648.

[25] For the state-specific effect, the bandwidths for URL, AA, and ULR are 0.2062, 0.7478, and 0.1438, respectively. For the time-specific effect, the bandwidths for ORL, WVR, and OLR are 1.0000, 0.7616, and 0.7676, respectively. However, these numbers are not directly comparable. To try to make such a comparison, we can use Equation (5), for the unordered regressor. Using this formula, the bandwidth for the unordered Rajagopalan and Lall type kernel function on the basis of the value of the bandwidth from the Aitchison and Aitken kernel function is 0.0331 (< 0.2062) for the state-specific effect. In other words, this implies that the 0.0331 bandwidth for the RL kernel corresponds to 0.7478 from the Aitchison and Aitken kernel. From this point of view, our bandwidth of 0.2062 is greater than 0.0331 and hence additional smoothing.

References

[1] Aitken CGG. 1983. Kernel Methods for the Estimation of Discrete Distributions. *Journal of Statistical Computation and Simulation* 16: 189-200.

[2] Aitchison J, Aitken CGG. 1976. Multivariate Binary Discrimination by the Kernel Method. *Biometrika* 63: 413-420.

[3] Baltagi BH, Pinnoi N. 1995. Public Capital Stock and State Productivity Growth: Further Evidence from An Error Components Model. *Empirical Economics* 20: 351-359.

[4] Cameron AC, Trivedi PK. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press: New York, NY, USA.

[5] Chiu S-T. 1990. Why Bandwidth Selectors Tend to Choose Smaller Bandwidths, and A Remedy. *Biometrika* 77: 222-226.

[6] Chu C-Y, Henderson DJ, Parmeter CF. 2015. Plug-In Bandwidth Selection for Kernel Density Estimation with Discrete Data. *Econometrics* 3: 199-214.

[7] Coppejans M. 2003. Effective Nonparametric Estimation in the Case of Severely Discretized Data. *Journal of Econometrics* 117: 331-367.

[8] Dong J, Simonoff JS. 1994. The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials. *Journal of Computational and Graphical Statistics* 3: 57-66.

[9] Epanechnikov VA. 1969. Nonparametric Estimations of A Multivariate Probability Density. *Theory of Probability and Its Applications* 14: 153-158.

[10] Greene W. 2011. *Econometric Analysis*. Prentice Hall: Upper Saddle River, NJ, USA.

[11] Habbema JDF, Hermans J, Remme J. 1978. Variable Kernel Density Estimation in Discriminant Analysis. *Compstat*, LCA Corster, J Hermans (Ed.), Vienna, Austria: Physica-Verlag, 178-185.

[12] Hall P, Li Q, Racine J. 2007. Estimation of Regression Functions in the Presence of Irrelevant Regressors. *Review of Economics and Statistics* 89: 784-789.

[13] Hall P, Racine J, Li Q. 2004. Cross-Validation and the Estimation of Conditional Probability Densities. *Journal of the American Statistical Association* 99: 1015-1026.

[14] Hall P, Robinson AP. 2009. Reducing Variability of Crossvalidation for Smoothing-Parameter Choice. *Biometrika* 96: 175-786.

[15] Hall P, Titterington DM. 1987. On Smoothing Sparse Multinomial Data. *Australian Journal of Statistics* 29: 19-37.

[16] Hart JD, Yi S. 1998. One-Sided Cross-Validation. *Journal of the American Statistical Association* 93: 620-631.

[17] Härdle W, Hall P, Marron JS. 1988. How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum? *Journal of the American Statistical Association* 83: 86-101.

[18] Henderson DJ, Kumbhakar SC. 2006. Public and Private Capital Productivity Puzzle: A Nonparametric Approach. *Southern Economic Journal* 73: 219-232.

[19] Henderson DJ, Parmeter CF. 2015. *Applied Nonparametric Econometrics*. Cambridge University Press: New York, NY, USA.

[20] Hurvich C, Simonoff JS, Tsai C-L.1998. Smoothing Parameter Selection in Nonparametric Regression Using An Improved Akaike Information Criterion. *Journal of the Royal Statistical Society, Series B*: 271-293.

[21] Huynh KP, Jacho-Chávez DT. 2009. Internally-Corrected Conditional Density Estimation. *Journal of Comparative Economics* 37: 122-143.

[22] Jarrett RG. 1979. A Note on the Intervals Between Coal-Mining Disasters. *Biometrika* 66: 191-193.

[23] Kokonendji CC, Senga Kiessé T, Zocchi SS. 2007. Discrete Triangular Distributions and Non-Parametric Estimation for Probability Mass Function. *Journal of Nonparametric Statistics* 19: 241-254.

[24] Li Q, Racine J. 2003. Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis* 86: 266-292.

[25] Li Q, Racine J. 2004. Predictor Relevance and Extramarital Affairs. *Journal of Applied Econometrics* 19: 533-535.

[26] Li Q, Racine J. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press: Princeton, NJ, USA.

[27] Lindsey JK. 1995. *Modelling Frequency and Count Data*. Clarendon Press: Oxford, UK.

[28] Munnell A. 1990. How Does Public Infrastructure Affect Regional Economic Performance? *New England Economic Review* (Spetember): 11-32.

[29] Ouyang D, Li Q, Racine J. 2006. Cross-Validation and the Estimation of Probability Distributions with Categorical Data. *Journal of Nonparametric Statistics* 18: 69-100.

[30] Racine J, Li Q. 2004. Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data. *Journal of Econometrics* 119: 99-130.
[31] Rajagopalan B, Lall U. 1995. A Kernel Estimator for Discrete Distribution. *Nonparametric Statistics* 4: 409-426.
[32] Rosenblatt M. 1956. Remarks on Some Nonparametric Estimates of A Density Function. *Annals of Mathematical Statistics* 27: 832-837.
[33] Simonoff JS. 1996. *Smoothing Methods in Statistics*. Springer-Verlag: New York, NY, USA.
[34] Simonoff JS. 2003. *Analyzing Categorical Data*. Springer-Verlag: New York, NY, USA.
[35] Titterington DM. 1980. A Comparative Study of Kernel-Based Density Estimates for Categorical Data. *Technometrics* 22: 259-268.
[36] Wang M-C, van Ryzin J. 1981. A Class of Smooth Estimators for Discrete Distributions. *Biometrika* 68: 301-309.

Appendix A

Here we provide four scenarios for the multinomial and beta-binomial distributions (Tables A1-A2) and these are primarily used in Section 3.1. In Table A2, Scenario (I) with $c = 10$, 20, and 50 is used in Section 3.1.2 and their probability vectors are (1) $P_{c=10}$ = (0.1818, 0.1636, 0.1455, 0.1273, 0.1091, 0.0909, 0.0727, 0.0545, 0.0364, 0.0182), (2) $P_{c=20}$ = (0.0952, 0.0905, 0.0857, 0.0810, 0.0762, 0.0714, 0.0667, 0.0619, 0.0571, 0.0524, 0.0476, 0.0429, 0.0381, 0.0333, 0.0286, 0.0238, 0.0190, 0.0143, 0.0095, 0.0048), and (3) $P_{c=50}$ = (0.0392, 0.0384, 0.0376, 0.0369, 0.0361, 0.0353, 0.0345, 0.0337, 0.0329, 0.0322, 0.0314, 0.0306, 0.0298, 0.0290, 0.0282, 0.0275, 0.0267, 0.0259, 0.0251, 0.0243, 0.0235, 0.0227, 0.0220, 0.0212, 0.0204, 0.0196, 0.0188, 0.0180, 0.0173, 0.0165, 0.0157, 0.0149, 0.0141, 0.0133, 0.0125, 0.0118, 0.0110, 0.0102, 0.0094, 0.0086, 0.0078, 0.0071, 0.0063, 0.0055, 0.0047, 0.0039, 0.0031, 0.0024, 0.0016, 0.0008). Additionally, we provide the bandwidths used to calculate the results in Tables 1-5 in Section 3.3 (Tables A3-A7).

Table A1. Multinomial Distribution for Unordered Discrete Kernel Functions

	$c = 3$	
Scenario	Probability	r
(I)	$p_1 = 0.15$, $p_2 = 0.35$, $p_3 = 0.50$	$3.\overline{3}$
(II)	$p_1 = 0.25$, $p_2 = 0.30$, $p_3 = 0.45$	2.5
(III)	$p_1 = 0.25$, $p_2 = 0.35$, $p_3 = 0.40$	1.6
(IV)	$p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$	1.0

Table A2. Beta-Binomial Distribution for Ordered Discrete Kernel Functions

	$c = 3$		
Scenario	Probability	Shape Parameters	r
(I)	$p_1 = 0.50$, $p_2 = 0.33$, $p_3 = 0.17$	$\alpha = 1$, $\beta = 2$	2.9
(II)	$p_1 = 0.25$, $p_2 = 0.50$, $p_3 = 0.25$	$\alpha = 50$, $\beta = 50$	2.0
(III)	$p_1 = 0.40$, $p_2 = 0.20$, $p_3 = 0.40$	$\alpha = 0.33$, $\beta = 0.33$	2.0
(IV)	$p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$	$\alpha = 1$, $\beta = 1$	1.0

TABLE A3. Median of bandwidth (Density Estimation in the Standard Data Case)

	URL	AA	ULR	ORL	WVR	OLR
	(I) $p_1 = 0.15$, $p_2 = 0.35$, $p_3 = 0.50$			(I) $p_1 = 0.55$, $p_2 = 0.33$, $p_3 = 0.17$		
$n = 50$	0.0366	0.1217	0.0087	0.0726	0.1129	0.0133
$n = 100$	0.0171	0.0625	0.0043	0.0284	0.0674	0.0067
$n = 200$	0.0084	0.0320	0.0022	0.0196	0.0353	0.0033
	(II) $p_1 = 0.25$, $p_2 = 0.30$, $p_3 = 0.45$			(II) $p_1 = 0.25$, $p_2 = 0.50$, $p_3 = 0.25$		
$n = 50$	0.1270	0.2985	0.0096	0.5291	0.1654	0.0135
$n = 100$	0.0591	0.1783	0.0048	0.0456	0.1011	0.0068
$n = 200$	0.0272	0.0945	0.0024	0.0196	0.0582	0.0034
	(III) $p_1 = 0.25$, $p_2 = 0.35$, $p_3 = 0.40$			(III) $p_1 = 0.40$, $p_2 = 0.20$, $p_3 = 0.40$		
$n = 50$	0.3052	0.4523	0.0098	0.0219	0.0807	0.0092
$n = 100$	0.1247	0.2952	0.0049	0.0196	0.0376	0.0045
$n = 200$	0.0542	0.1669	0.0025	0.0196	0.0182	0.0022
	(IV) $p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$			(IV) $p_1 = 1/3$, $p_2 = 1/3$, $p_3 = 1/3$		
$n = 50$	1.0000	0.6667	0.0100	0.5199	0.1955	0.0135
$n = 100$	1.0000	0.6667	0.0050	0.0439	0.1251	0.0067
$n = 200$	1.0000	0.6667	0.0025	0.0196	0.0732	0.0034

TABLE A4. Median of bandwidth (Density Estimation in the Sparse Data Case)

	URL	AA	ULR	ORL	WVR	OLR
$c = 10$, $n = 10$	0.3416	0.8000	0.0108	4.0000	0.6698	0.0433
$c = 10$, $n = 20$	0.0994	0.6085	0.0054	2.6111	0.5819	0.0234
$c = 20$, $n = 20$	0.0854	0.7417	0.0026	7.3764	0.7478	0.0199
$c = 20$, $n = 40$	0.0427	0.6034	0.0013	5.0000	0.6800	0.0112
$c = 50$, $n = 50$	0.0313	0.7456	0.0004	9.8433	0.8268	0.0076
$c = 50$, $n = 100$	0.0154	0.5964	0.0002	8.4442	0.7828	0.0043

TABLE A5. Median of bandwidth (Density Estimation in the Irrelevant Variable Case)

	URL	AA	ULR	ORL	WVR	OLR
	Continuous Random Variable			Continuous Random Variable		
$n = 50$	6.0000	0.6667	1.0000	10.0000	0.7952	1.0000
$n = 100$	6.0000	0.6667	1.0000	10.0000	0.7952	1.0000
$n = 200$	6.0000	0.6667	1.0000	10.0000	0.7952	1.0000
	Discrete Random Variable			Discrete Random Variable		
$n = 50$	1.0000	0.6667	1.0000	1.0000	0.7859	1.0000
$n = 100$	1.0000	0.6667	1.0000	1.0000	0.7859	1.0000
$n = 200$	1.0000	0.6667	1.0000	1.0000	0.7859	1.0000

TABLE A6. Median of bandwidth (Cross-Sectional Regression)

	URL	AA	ULR	ORL	WVR	OLR
	Standard Data Case			Standard Data Case		
$n = 50$	0.7817	0.1234	0.0452	0.0425	0.0461	0.0434
$n = 100$	0.5859	0.1213	0.0359	0.0466	0.0466	0.0450
$n = 200$	0.2287	0.1161	0.0301	0.0525	0.0355	0.0483
	Spare Data Case			Spare Data Case		
$n = 50$	0.0170	0.4374	0.0254	1.0000	0.6410	0.6471
$n = 100$	0.0137	0.3528	0.0081	1.0000	0.6592	0.6318
$n = 200$	0.0041	0.2400	0.0066	1.0000	0.6534	0.6129
	Irrelevant Variable Case			Irrelevant Variable Case		
$n = 50$	1.0000	0.5000	1.0000	1.0000	0.8619	1.0000
$n = 100$	1.0000	0.5000	1.0000	1.0000	0.8764	1.0000
$n = 200$	1.0000	0.5000	1.0000	1.0000	0.7916	1.0000

TABLE A7. Median of bandwidth: Upper for X^u and Bottom for W in One-Way Fixed Effects; Upper for X^u and Bottom for X^o in Two-Way Fixed Effects (Panel Data Regression)

	URL	AA	ULR	RL	AW	LR
	One-Way Fixed Effects			Two-Way Fixed Effects		
	Standard Data Case			Standard Data Case		
$n = 50,\ T = 3$	0.9810	0.9530	0.0104	0.9734	0.9572	0.0406
	0.1554	0.9028	0.1493	0.2238	0.3650	0.2130
$n = 50,\ T = 5$	0.9968	0.9678	0.0086	0.9852	0.9628	0.0353
	0.1977	0.5953	0.1336	0.4191	0.4257	0.2301
$n = 100,\ T = 3$	1.0000	0.9895	0.0099	1.0000	0.9883	0.0352
	0.2000	0.4459	0.1252	0.2000	0.2459	0.1724
$n = 100,\ T = 5$	1.0000	0.9895	0.0084	1.0000	0.9881	0.0299
	0.1977	0.4326	0.1131	0.3807	0.3782	0.2078
$n = 200,\ T = 3$	1.0000	0.9945	0.0147	1.0000	0.9950	0.0303
	0.1976	0.4222	0.0826	0.1978	0.1990	0.1213
$n = 200,\ T = 5$	1.0000	0.9941	0.0144	1.0000	0.9950	0.0250
	0.1976	0.3883	0.0724	0.1980	0.1990	0.1882
	Irrelevant Variable Case			Irrelevant Variable Case		
$n = 50,\ T = 3$	0.9985	0.9657	0.0452	0.9322	0.9583	0.0429
	0.0079	0.0960	0.0053	1.0000	0.8842	0.3367
$n = 50,\ T = 5$	0.9986	0.9664	0.0407	0.9897	0.9608	0.0348
	0.0117	0.0695	0.0036	0.4370	0.8712	0.3345
$n = 100,\ T = 3$	1.0000	0.9900	0.0394	0.9991	0.9891	0.0327
	0.1976	0.1980	0.0034	0.4189	0.4304	0.3639
$n = 100,\ T = 5$	1.0000	0.9900	0.0354	0.9995	0.9895	0.0277
	0.1976	0.1957	0.0030	0.4189	0.4326	0.3433
$n = 200,\ T = 5$	1.0000	0.9950	0.0342	0.9991	0.9941	0.0271
	0.1976	0.2379	0.0028	0.4189	0.4213	0.3380
$n = 200,\ T = 5$	1.0000	0.9950	0.0308	0.9981	0.9941	0.0217
	0.1976	0.1967	0.0025	0.4189	0.4213	0.3138

Appendix B

Following Equation (A.3) in Li and Racine (2003, p. 278), a power series expansion for our proposed kernel can be represented as

$$\begin{aligned}
L(X_i,x,h) &= \mathbf{1}_{d_{ix}=0}\left[\frac{-h^2}{(c-1)-ch^2}\right]^k + \\
&\quad \mathbf{1}_{d_{ix}=1}\left[\frac{-h^2}{(c-1)-ch^2}\right]^{k-1}\left[\frac{1-h^2}{(c-1)-ch^2}\right] + \\
&\quad \mathbf{1}_{d_{ix}=2}\left[\frac{-h^2}{(c-1)-ch^2}\right]^{k-2}\left[\frac{1-h^2}{(c-1)-ch^2}\right]^2 + \\
&\quad \mathbf{1}_{d_{ix}=3}\left[\frac{-h^2}{(c-1)-ch^2}\right]^{k-3}\left[\frac{1-h^2}{(c-1)-ch^2}\right]^3 + \cdots \\
&= \text{the 1st term} + \text{the 2nd term} + \text{the 3rd term} + \\
&\quad \mathbf{1}_{d_{ix}=3}\left[\frac{-h^2}{(c-1)-ch^2}\right]^{k-3}\left[\frac{1}{ch^2-(c-1)}\right]^3\left[(\lambda+1)^2-1\right]^3 + \cdots \\
&= \text{the 1st term} + \text{the 2nd term} + \text{the 3rd term} + \\
&\quad \mathbf{1}_{d_{ix}=3}\left[\frac{-h^2}{(c-1)-ch^2}\right]^{k-3}\left[\frac{1}{ch^2-(c-1)}\right]^3(\lambda+2)^3\lambda^3 + \cdots \\
&\approx \text{the 1st term} + \text{the 2nd term} + \text{the 3rd term} + O_p(\lambda^3)
\end{aligned}$$

where $h = \lambda + 1$ with $\lambda \in [0, \infty)$. Note that when $\lambda = 0$, $\left[\frac{-h^2}{(c-1)-ch^2}\right] = 1$ and $\left[\frac{1-h^2}{(c-1)-ch^2}\right] = 0$, which correspond to $(1-\lambda) = 1$ and $\lambda = 0$ for the unordered Li and Racine or Aitchison and Aitken kernels. Therefore, although c cannot be moved out of the square brackets and h appears in the denominators, these do not seem to affect any subsequent derivations and proof (e.g., Lemmas A.1-A.5).[26]

[26]Although c can be rearranged as an extra multiplicative constant for the the Aitchison and Aitken kernel, its value still can affect the power series expansion of the product kernel.