DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2017-14

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

<http://www.mcmaster.ca/economics/>

Kernel Smoothed Probability Mass Functions for Ordered Datatypes

Jeffrey S. Racine

*Department of Economics and Graduate Program in Statistics, McMaster University,
racinej@mcmaster.ca*

Qi Li

Department of Economics, Texas A&M University

Karen X. Yan

Department of Economics, Texas A&M University

November 03, 2017

Abstract

We propose a kernel function for ordered categorical data that overcomes certain limitations present in ordered kernel functions that have appeared in the literature on the estimation of probability mass functions for multinomial ordered data. Some of these limitations arise from assumptions made about the support of the random variable that may be at odds with the data at hand. Furthermore, many existing ordered kernel functions lack a particularly appealing property, namely the ability to deliver discrete uniform probability estimates for some value of the smoothing parameter. To overcome these limitations, we propose an asymmetric *empirical support* kernel function that adapts to the data at hand and possesses certain desirable features. In particular, there are no difficulties arising from zero counts caused by gaps in the data while it encompasses both the empirical proportions and the discrete uniform probabilities at the lower and upper boundaries of the smoothing parameter. We propose using likelihood and least squares cross-validation for smoothing parameter selection, and study the asymptotic behaviour of these data-driven methods. We use Monte Carlo simulations to examine the finite sample performance of the proposed estimator and we also provide a simple empirical example to illustrate the usefulness of the proposed estimator in applied settings.

1 Introduction

In multinomial discrete support random variable settings, it is common to encounter situations in which the support contains only a handful of values, and such values may contain gaps (e.g., $\{0, 1, 2, 5\}$). When such data are of the *ordered* type, using a kernel function that recognizes order present in data can lead to improved accuracy relative to kernel functions that ignore order (e.g., binary unordered *counting* kernel functions). Though a range of ordered kernel functions for multinomial data have been proposed in the literature, many make assumptions about the support that, when unwarranted, may lead to improper probability estimates and thereby require special treatment. The presence of zero counts in the assumed support can sometimes lead to difficulties. Furthermore, it turns out that a kernel function that can span the empirical (nonsmooth) estimator (i.e., the sample proportions) and the discrete uniform estimator for some values of the smoothing parameter is an asset; virtually all existing ordered kernel functions possess the former property, while many lack the latter. We propose an asymmetric *empirical support* kernel function that encompasses the empirical proportions and the discrete uniform distribution at the extremes of the admissible smoothing parameter. The resulting estimator automatically adapts to gaps in the data and makes no assumptions about the support other than that it is discrete, ordered, and finite.

We also provide theoretical support for two methods of smoothing parameter selection, namely likelihood cross-validation and least squares cross-validation, and compare their finite-sample performance via Monte Carlo simulation. The proposed approach exhibits better finite sample performance than estimators based on kernel functions that ignore order present in the data and than the empirical proportions themselves. Unlike Hall (1987), who considered likelihood cross-validation in a density estimation context and demonstrated how its asymptotic properties are profoundly influenced by tail properties of the kernel function and of the unknown density function, our approach is immune to this phenomenon because we explicitly treat our problem as one having finite support, hence there is no “tail” in the sense of Hall (1987).

2 Background

Suppose we are interested in modeling a univariate probability function for some random variable X which is discrete in nature. We denote the probability function $p(x) = \Pr(X = x)$, $x \in \mathcal{D}$, $0 \leq p(x) \leq 1$, $\sum_{x \in \mathcal{D}} p(x) = 1$. In general, $p(x)$ is unknown and must be estimated. When in addition X is ordered (i.e., distance between elements of the support is defined) we can define the cumulative probability function, $F(x) = \sum_{z \in \mathcal{D}, z \leq x} p(z)$. Nonparametric estimates of $p(x)$ have been studied for decades and include the approaches of Fienberg and Holland (1973), Aitchison and Aitken (1976), van Ryzin and Wang (1978), Wang and van Ryzin (1981), Titterton and Bowman (1985), Hall and Titterton (1987), Burman (1987), Dong and Simonoff (1994), Ahmad and Cerrito (1994), Rajagopalan and Lall (1995), Ouyang et al. (2006), and

more recently Chen et al. (2014), Chu et al. (2017), and Harfouche et al. (2017), among others.

It is widely appreciated that by adopting a kernel function that is capable of reflecting the order present in data, we can improve the estimator's performance relative to one that ignores order. An *ordered kernel function* is one that can assign higher weight to neighbouring values of x than to those lying farther away.¹

We assume that X takes c different *ordered* discrete values, e.g., $X \in \mathcal{D}$, where $\mathcal{D} = \{a_1, a_2, \dots, a_c\}$ with $a_1 < a_2 < \dots < a_c$. Without loss of generality, we assume that $\min_{1 \leq i, j \leq c; j \neq i} |a_i - a_j| = 1$ (i.e., the minimum distance is 1) and that $a_1, \dots, a_c$ are c distinct non-negative integers. By way of illustration, Aitchison and Aitken (1976) (Page 419) considered the case of $\mathcal{D} = \{0, 1, \dots, c-1\}$ ($\mathcal{D}$ contains c consecutive non-negative integers) and suggested using the binomial kernel function given by

$$l_{AA}(X_i, x, \lambda) = \binom{c}{|x - X_i|} \lambda^{|x - X_i|} (1 - \lambda)^{c - |x - X_i|},$$

where $0 \leq |x - X_i| \leq c$ and where

$$\binom{c}{|x - X_i|} = \frac{c!}{(c - |x - X_i|)! |x - X_i|!}. \quad (1)$$

The resulting smoothed probability estimator is given by

$$\tilde{p}(x) = \frac{1}{n} \sum_{i=1}^n l_{AA}(X_i, x, \lambda).$$

It turns out that the kernel function defined in Equation (1) is lacking an important property, namely the ability to place equal weight on all outcomes for some value of its smoothing parameter (there is no value of λ for which the binomial probability function is *flat* for $c > 2$). When estimating conditional distributions or moments in the presence of ordered discrete predictors, the ability to automatically *smooth out* a variable is particularly important and can only occur if the kernel function can assume the discrete uniform distribution for some value of its smoothing parameter; see Hall et al. (2004) and Hall et al. (2007) by way of illustration. Aitchison and Aitken's unordered kernel function possesses this important property since for $\lambda = (c-1)/c$ (its upper bound) their unordered kernel function is $1/c$ for all $x \in \mathcal{D}$, but this property is not shared by its ordered counterpart. Furthermore, zero counts in the assumed support (i.e., gaps) will render the estimated probabilities improper as the estimated probabilities will not sum to 1 (the empirical proportions, on the other hand, will always sum to 1). Other ordered kernel functions have been proposed such as Wang and van

¹Unordered kernel functions place a binary *counting* weight on each observation, for example, $1 - \lambda$ when $X_i = x$ and $\lambda/(c-1)$ when $X_i \neq x$, $\lambda \in [0, (c-1)/c]$, but cannot assess distance (see Aitchison and Aitken (1976) Page 419).

Ryzin (1981) and Ahmad and Cerrito (1994), but they presume that the support is the set of all consecutive integers which may not be the case for the data at hand, while there is no value of the smoothing parameter for which the kernel function is the discrete uniform (the same goes for the ordered kernels proposed by Rajagopalan and Lall (1995), Chu et al. (2017) and others). Furthermore, existing kernel functions can exhibit undesirable behaviour when there are *gaps* in the support of a discrete random variable and the resulting estimator (say, $\hat{p}(\cdot)$) may fail to satisfy the condition $\sum_{x \in \mathcal{D}} \hat{p}(x) = 1$, rendering the estimate an improper probability estimator; see Titterton and Bowman (1985) for a discussion on difficulties arising from zero counts. It turns out that constructing a bounded support discrete kernel function that is *also* capable of applying uniform weight, can adapt to gaps, and is always proper simply requires a bit of kernel carpentry.

3 Estimation of $p(x)$ and $F(x)$ with Ordered Data

3.1 An Estimator of $p(x)$ for Ordered Data

In this paper we propose using the following asymmetric empirical support kernel function,

$$l(z, x, \lambda) = \frac{\lambda^{d_{zx}}}{\Lambda_z}, \quad (2)$$

where $0 \leq \lambda \leq 1$, $d_{zx} = |z - x|$ and $\Lambda_z = \sum_{x \in \mathcal{D}} \lambda^{d_{zx}}$. Note that, $d_{zx} = d_{xz}$, but due to the normalization factor $1/\Lambda_z$, $l(z, x, \lambda) = d_{xz}/\Lambda_z \neq d_{xz}/\Lambda_x = l(x, z, \lambda)$, i.e., the kernel function $l(x, z, \cdot)$ is *not* symmetric in (x, z) .

The resulting kernel estimator of $p(x)$ is given by

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n l(X_i, x, \lambda) = \frac{1}{n} \sum_{i=1}^n \frac{\lambda^{d_{ix}}}{\Lambda_i}, \quad (3)$$

where $d_{ix} = |X_i - x|$ and $\Lambda_i = \sum_{x \in \mathcal{D}} d_{ix}$.

We begin by outlining the properties of the kernel function $l(z, x, \lambda)$, and then consider the properties of $\hat{p}(x)$ and its cumulative counterpart $\hat{F}(x)$. We also consider two data-driven smoothing parameter selection methods and the multivariate extensions of $\hat{p}(x)$ and $\hat{F}(x)$, while a Monte Carlo simulation compares the performance of $\hat{p}(x)$ with the nonsmooth sample proportion $p_n(x)$ and the unordered counterpart to $\hat{p}(x)$.

3.1.1 Properties of $l(z, x, \lambda)$

The kernel function $l(z, x, \lambda)$ defined in Equation (2) has the following properties:

1. $l(z, x, \lambda)$ is non-negative, and for all $z \in \mathcal{D}$, $\lambda \in [0, 1]$, it sums to one (over $x \in \mathcal{D}$), i.e.,

$$\sum_{x \in \mathcal{D}} l(z, x, \lambda) = \sum_{x \in \mathcal{D}} \frac{\lambda^{d_{zx}}}{\Lambda_z} = \sum_{x \in \mathcal{D}} \frac{\lambda^{d_{zx}}}{\sum_{x \in \mathcal{D}} \lambda^{d_{zx}}} = 1.$$

2. $l(z, x, \lambda)$ adapts to the presence of non-consecutive outcomes such as $\mathcal{D} = \{a_1, a_2, a_3\} = \{0, 1, 4\}$.
3. When $\lambda = 0$, $\Lambda_z = 1$ for all $z \in \mathcal{D}$ (i.e., $\Lambda_z = \sum_{x \in \mathcal{D}} \lambda^{d_{zx}} = 0^0 + \sum_{x \in \mathcal{D}, x \neq z} 0^{|z-x|} = 1$, where we define $0^0 = 1$). Hence, when $\lambda = 0$, $l(z, x, \lambda) = \mathbf{1}(z = x)$ for all $z \in \mathcal{D}$, where $\mathbf{1}(A)$ is an indicator function which takes value 1 if event A holds true and 0 otherwise.
4. For $\lambda = 1$, the kernel function applies weight $1/c$ to all observations since $\Lambda_z = c$ (i.e., $\Lambda_z = \sum_{x \in \mathcal{D}} \lambda^{d_{zx}} = 1^{|a_1-z|} + 1^{|a_2-z|} + \dots = c$). Hence $l(z, x, \lambda = 1) = 1/c$ for all $z, x \in \mathcal{D}$.
5. We can express $\Lambda_z = \sum_{x \in \mathcal{D}} \lambda^{d_{zx}} = 1 + \sum_{x \in \mathcal{D}, x \neq z} \lambda^{d_{zx}} = 1 + \Lambda_{-z}$ (where $\Lambda_{-z} = \sum_{x \in \mathcal{D}, x \neq z} \lambda^{d_{zx}}$) since when $z = x$, $d_{zx} = 0$ and $\lambda^0 = 1$ for all $\lambda \in [0, 1]$. In Section 4, we will show that, when X is not uniformly distributed, the data-driven smoothing parameter $\hat{\lambda}$ has an order $O_p(1/n)$. Therefore, the kernel function $l(z, x, \lambda)$ can be written as $l(z, x, \lambda) = \frac{\lambda^{d_{zx}}}{1 + \Lambda_{-z}} = \lambda^{d_{zx}}(1 - \Lambda_{-z} + \Lambda_{-z}^2 - \dots) = \lambda^{d_{zx}}(1 - b_{1z}\lambda + O(\lambda^2))$. For $d_{zx} = |x - z| = 0$, $l(z, x, \lambda) = 1 - b_{1z}\lambda + O(\lambda^2)$, and for $d_{zx} = |x - z| = 1$, $l(z, x, \lambda) = \lambda(1 - b_{1z}\lambda + O(\lambda^2)) = \lambda - b_{1z}\lambda^2 + O(\lambda^3)$, where b_{1z} takes values in $\{0, 1, 2\}$.

Remark. For any $z \in \mathcal{D}$, we can write $\Lambda_{-z} = b_{1z}\lambda + b_{2z}\lambda^2 + \dots$, where $b_{jz} \in \{0, 1, 2\}$ for $j = 1, \dots, a_c$, by simply rearranging the order of the elements $\lambda^{d_{zx}}$ as they appear in Λ_{-z} and, instead, ordering elements from the smallest to the largest exponent. Consider the simple case where $\mathcal{D} = \{0, 1, 2, 6\}$. Then we have $\Lambda_{-0} = \sum_{x \in \mathcal{D}, x \neq 0} \lambda^{d_{0x}} = \lambda + \lambda^2 + \lambda^6$ and $\Lambda_{-1} = \sum_{x \in \mathcal{D}, x \neq 1} \lambda^{d_{1x}} = 2\lambda + \lambda^5$, where the 2λ term arises when $x = 0$ and $x = 2$ as $d_{x1} = 1$ for both $x = 0$ and $x = 2$. It is straightforward to see that for any $z \in \{0, 1, 2, 6\}$, we can write $\Lambda_{-z} = \sum_{j=1}^6 b_{jz}\lambda^j$, where the coefficients b_{jz} assume values in $\{0, 1, 2\}$.

3.1.2 Properties of $\hat{p}(x)$

It is straightforward to check that $\hat{p}(x)$ defined in Equation (3) satisfies $\sum_{x \in \mathcal{D}} \hat{p}(x) = 1$, therefore the estimate is a proper probability estimator. Next, we analyze the mean and variance of $\hat{p}(x)$. Assuming i.i.d. draws and assuming that $\lambda = o(1)$,

$$E\hat{p}(x) = p(x) + c_{1x}\lambda + O(\lambda^2), \quad (4)$$

where b_{1x} takes values in $\{0, 1, 2\}$ and $c_{1x} = \sum_{z \in \mathcal{D}, d_{zx}=1} p(z) - b_{1x}p(x)$.

$$\text{Var } \hat{p}(x) = \frac{p(x)(1-p(x))}{n} + \frac{d_{1x}\lambda}{n} + O\left(\frac{\lambda^2}{n}\right),$$

where $d_{1x} = -2(b_{1x} + c_{1x})p(x)$ is a constant, b_{1x} and c_{1x} are defined in Equation (4).

We propose using data-driven methods for selecting the smoothing parameter λ . Specifically, we consider using likelihood cross-validation (LCV) and least squares cross-validation (LSCV) methods for choosing λ . In Theorem 4.1 and Proposition 4.1, we show that the data-driven selected $\hat{\lambda}$ is of order $O_p(1/n)$. Hence, by the Liapounov central limit theorem (CLT, see White (1984), Page 112), we have, as $n \rightarrow \infty$

$$\sqrt{n}(\hat{p}(x) - p(x)) \xrightarrow{d} N(0, p(x)(1-p(x))).$$

3.2 An Estimator of $F(x)$ for Ordered Data

The univariate kernel estimator of $F(x)$ is given by

$$\hat{F}(x) = \sum_{z \in \mathcal{D}, z \leq x} \hat{p}(z) = \sum_{z \in \mathcal{D}, z \leq x} \frac{1}{n} \sum_{i=1}^n l(X_i, z, \lambda) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(X_i, z, \lambda), \quad (5)$$

where the ordered kernel $\mathcal{L}(X_i, z, \lambda)$ is defined by

$$\mathcal{L}(X_i, x, \lambda) = \sum_{z \in \mathcal{D}, z \leq x} l(X_i, z, \lambda) = \sum_{z \in \mathcal{D}, z \leq x} \frac{\lambda^{d_{iz}}}{\Lambda_i},$$

where $0 \leq \lambda \leq 1$, $d_{iz} = |X_i - z|$ and where $\Lambda_i = \sum_{z \in \mathcal{D}} \lambda^{d_{iz}}$ is a normalizing factor that is tailored to the particular placement of $X_i \in \mathcal{D}$. When $\lambda = 0$, $\mathcal{L}(X_i, x, \lambda) = \mathbf{1}(X_i \leq x) = \sum_{z \in \mathcal{D}, z \leq x} \mathbf{1}(X_i = z)$.² The $\hat{F}(x)$ defined in Equation (5) collapses to the empirical CDF $F_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$ when $\lambda = 0$.

²To see this, note that $d_{zi} = 0$ when $z = X_i$, and when $\lambda = 0$, $l(X_i, z, 0) = \lambda^{d_{iz}}/\Lambda_i = 0^0/(0^0 + \sum_{y \in \mathcal{D}, y \neq X_i} 0^{|X_i - y|}) = 1$ when $z = X_i$ and 0 otherwise. So, for any X_i and $z \leq x$, $l(X_i, z, 0)$ will equal 1 if X_i equals any value $z \leq x$ and zero otherwise, hence $\mathcal{L}(X_i, x, \lambda) = \mathbf{1}(X_i \leq x) = \sum_{z \in \mathcal{D}, z \leq x} \mathbf{1}(X_i = z)$ when $\lambda = 0$.

Intuitively, if $\sqrt{n}\lambda \rightarrow 0$ as $n \rightarrow \infty$ then $\hat{F}(x) - F_n(x) = O_p(\lambda) = o_p(n^{-1/2})$. Then $\sqrt{n}(\hat{F}(x) - F(x))$ and $\sqrt{n}(F_n(x) - F(x))$ will have the same asymptotic distribution. Note that we require $\sqrt{n}\lambda \rightarrow 0$ as $n \rightarrow \infty$, i.e., $\lambda = o(1/\sqrt{n})$, which is true when using either LCV or LSCV to select λ since both $\hat{\lambda}_{LCV}$ and $\hat{\lambda}_{LSCV}$ are of order $O_p(n^{-1})$.

It can be shown that the expected value of $\hat{F}(x)$ is given by

$$E\hat{F}(x) = F(x) + d_{1x}\lambda + O(\lambda^2),$$

where $d_{1x} = \sum_{z \in \mathcal{D}, z \leq x} c_{1z}$.

Similar to the calculation for $\text{Var } \hat{p}(x)$, one can show that the variance of $\hat{F}(x)$ can be written as

$$\text{Var } \hat{F}(x) = \frac{F(x)(1 - F(x))}{n} + e_{1x}\lambda/n + O(\lambda^2/n),$$

where e_{1x} is a constant. The detailed derivation to obtain an explicit expression for e_{1x} is tedious, but this result should be quite intuitive. Therefore, we omit the derivation because the explicit expression of e_{1x} is immaterial.

Furthermore, under the condition $\sqrt{n}\lambda \rightarrow 0$ as $n \rightarrow \infty$ at its optimal rate (i.e., $\lambda = O(n^{-1}) = o(1/\sqrt{n})$), by the Liapounov CLT we have

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))).$$

4 Data-driven Smoothing Parameter Selection Methods

We consider two data-driven methods to select the smoothing parameter λ : LCV and LSCV.

4.1 The Likelihood Cross-Validation (LCV) Method

The LCV method selects λ to maximize the leave-one-out log-likelihood function given by

$$\hat{\lambda}_{LCV} = \operatorname{argmax}_{\lambda} \sum_{i=1}^n \log \hat{p}_{-i}(X_i),$$

where $\hat{p}_{-i}(X_i) = n^{-1} \sum_{j \neq i} l(X_j, X_i, \lambda)$ is the leave-one-out estimator of $p(X_i)$.

Theorem 4.1. *Let $\hat{\lambda}_{LCV}$ denote the value of λ that maximizes the likelihood cross-validation function, assume that $p(x)$ is not uniformly distributed over its support $\mathcal{D}$ (with $c \geq 2$), and assume that $\lambda \rightarrow 0$ as $n \rightarrow \infty$. Then we have*

$$\hat{\lambda}_{LCV} = O_p(n^{-1}).$$

The proof is given in the appendix.

4.2 The Least Squares Cross-Validation (LSCV) Method

An alternative method is to use the least squares cross-validation (LSCV) method to select λ . The estimation sum of squared errors (SSE) is given by

$$\begin{aligned} \text{SSE } \hat{p}(x) &= \sum_{x \in \mathcal{D}} (\hat{p}(x) - p(x))^2 \\ &= \sum_{x \in \mathcal{D}} \hat{p}^2(x) - 2 \sum_{x \in \mathcal{D}} p(x) \hat{p}(x) + \sum_{x \in \mathcal{D}} p^2(x) \\ &= J_{1n} + J_{2n} + J_3 \end{aligned} \tag{6}$$

where $J_3 = \sum_{x \in \mathcal{D}} p^2(x)$ is a constant that is unrelated to λ .

Therefore, minimizing SSE is equivalent to minimizing the sum of the first two terms $J_{1n} + J_{2n}$. Note that J_{2n} contains an unknown quantity $p(x)$. We estimate J_{2n} by $\hat{J}_{2n} = -2n^{-1} \sum_{i=1}^n \hat{p}_{-i}(X_i)$. Hence, we recommend choosing λ to minimize the following cross-validation function:

$$J_n(\lambda) = J_{1n} + \hat{J}_{2n}.$$

Proposition 4.1. Let $\hat{\lambda}_{LSCV}$ denote the value of λ that minimizes $J_n(\lambda) = J_{1n} + \hat{J}_{2n}$. Assume that $p(\cdot)$ is not uniformly distributed over its support $\mathcal{D}$, $\lambda \rightarrow 0$ as $n \rightarrow \infty$. Then

$$\hat{\lambda}_{LSCV} = O_p(n^{-1}).$$

The proof of Proposition 4.1 is similar to the proof of Theorem 2.1 in Ouyang et al. (2006).

Remark. Up to now we have restricted attention to the case for which $\lambda = o(1)$ which, as stated in Proposition 4.1, requires that $p(\cdot)$ is not uniformly distributed, i.e., that there exists $x, z \in \mathcal{D} = \{a_1, \dots, a_c\}$ such that $p(x) \neq p(z)$. If $p(\cdot)$ is uniformly distributed in the sense that $p(X = a_i) = 1/c$ for all $i = 1, \dots, c$, then using a proof similar to the proof of Theorem 2.4 in Ouyang et al. (2006), one can prove that, asymptotically, $CV(\lambda)$ is minimized at $\hat{\lambda} = 1$ with probability δ for some $\delta \in (0, 1)$, i.e., $\lim_{n \rightarrow \infty} P(\hat{\lambda} = 1) = \delta$ for some $\delta \in (0, 1)$. That is, the smoothing parameter will not converge to zero, rather it tends to assume its upper extreme value with a high probability, so that the estimated probability function satisfies the uniform distribution characteristics. In this case, $\hat{p}(x)$ is more efficient than the frequency estimator because the frequency method does not exploit this information.

5 Monte Carlo Simulation

Tables 1 and 2 present summaries of a Monte Carlo simulation that compares the nonsmooth estimator $p_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i = x)$ with the proposed smooth estimator. We include the unordered kernel estimator as an additional comparator to assess whether incorporating the information that the data is ordered has a perceptible impact. We simulate data having discrete ordered support $\mathcal{D} = \{0, 1, 2, 9, 10\}$ that contains a gap. The data is drawn from a continuum of distributions, at one extreme the discrete uniform, $p_u = (0.20, 0.20, 0.20, 0.20, 0.20)$, and at the other extreme a case where the probabilities are starkly unequal, $p_s = (0.04, 0.12, 0.20, 0.28, 0.36)$. We take a convex combination of these extremes given by $\delta p_s + (1 - \delta)p_u$ and let $\delta \in [0, 1]$. The sample size is $n = 100$ and $M = 1000$ Monte Carlo replications are conducted. In Table 1 (Table 2), for each replication we maximize (minimize) the cross-validation function for the ordered and unordered estimators via search over a grid of 25 equally spaced points in the intervals $[0, 1]$ and $[0, (c - 1)/c] = [0, 0.8]$, respectively. For each replication we compute the SSE for each estimator, e.g., $SSE = \sum_{x \in \mathcal{D}} (\hat{p}(x) - p(x))^2$, and report their median SSE over all $M = 1000$ replications.

One would expect that when the probabilities differ substantially then cross-validation will tend to select small smoothing parameters, while when they are similar cross-validation will tend to select large smoothing parameters. Furthermore, for the discrete uniform case the distribution of the cross-validated smoothing

Table 1: Likelihood cross-validation median SSEs ($\times 100$), relative median SSEs and median cross-validated smoothing parameters λ_{LCV} . For columns 5 and 6 (relative median SSEs of the ordered versus empirical and ordered versus unordered, respectively) numbers < 1 indicate superior performance of the proposed ordered kernel approach. The superscripts o and u denote the estimator using the proposed ordered kernel function and Aitchison and Aitken's unordered kernel function, respectively. The discrete uniform case corresponds to $\delta = 0$.

δ	$\text{SSE } p_n(x)$	$\text{SSE } \hat{p}^o(x)$	$\text{SSE } \hat{p}^u(x)$	$\frac{\text{SSE } \hat{p}^o(x)}{\text{SSE } p_n(x)}$	$\frac{\text{SSE } \hat{p}^o(x)}{\text{SSE } \hat{p}^u(x)}$	λ_{LCV}^o	λ_{LCV}^u
0.000	0.136	0.000	0.000	0.000	NaN	1.000	0.800
0.250	0.132	0.080	0.080	0.606	1.000	0.792	0.600
0.500	0.124	0.083	0.121	0.672	0.687	0.667	0.300
0.750	0.124	0.111	0.121	0.892	0.915	0.292	0.133
1.000	0.116	0.097	0.116	0.840	0.839	0.167	0.067

Table 2: Least squares cross-validation median SSEs ($\times 100$), relative median SSEs and median cross-validated smoothing parameters λ_{LSCV} . For columns 5 and 6 (relative median SSEs of the ordered versus empirical and ordered versus unordered, respectively) numbers < 1 indicate superior performance of the proposed ordered kernel approach. The superscripts o and u denote the estimator using the proposed ordered kernel function and Aitchison and Aitken's unordered kernel function, respectively. The discrete uniform case corresponds to $\delta = 0$.

δ	$\text{SSE } p_n(x)$	$\text{SSE } \hat{p}^o(x)$	$\text{SSE } \hat{p}^u(x)$	$\frac{\text{SSE } \hat{p}^o(x)}{\text{SSE } p_n(x)}$	$\frac{\text{SSE } \hat{p}^o(x)}{\text{SSE } \hat{p}^u(x)}$	λ_{LSCV}^o	λ_{LSCV}^u
0.000	0.136	0.000	0.000	0.000	NaN	1.000	0.800
0.250	0.132	0.080	0.080	0.606	1.000	0.792	0.600
0.500	0.124	0.084	0.123	0.680	0.687	0.667	0.300
0.750	0.124	0.118	0.122	0.949	0.963	0.292	0.133
1.000	0.116	0.111	0.120	0.961	0.926	0.208	0.100

parameters would be expected to be more concentrated at their upper bounds (1 and 0.8 for the ordered and unordered kernels, respectively) as n increases. The opposite would be expected when the probabilities differ substantially, with the cross-validated smoothing parameters becoming more concentrated at $\lambda = 0$.

All three estimators are consistent, but the nonsmooth estimator $p_n(x)$ is the least efficient. Both the unordered and ordered smooth estimators shrink the empirical probabilities towards the discrete uniform distribution, and are most efficient when the underlying probabilities are those for the discrete uniform. Across the range of distributions, the ordered estimator is the best performer, and it is evident that cross-validated smoothing parameter selection is performing as expected. The LCV approach is more efficient than the LSCV approach when the probabilities are starkly unequal (this corresponds to the cases for which $\delta = 0.5, 0.75$, and 1.).

6 The Multivariate Case

Let $x = (x_1, \dots, x_r)$ be an r -dimensional ordered discrete variable with $x_s \in \mathcal{D}_s = \{a_{s,1}, \dots, a_{s,c_s}\}$ for $s = 1, \dots, r$, where $0 \leq a_{s,1} < a_{s,2} < \dots < a_{s,c_s}$. Define $d_{x_s, z_s} = |x_s - z_s|$ for $x_s, z_s \in \mathcal{D}_s$. We assume that when $x_s \neq z_s$, we have $d_{x_s, z_s} \geq 1$. Also, we use X_{is} to denote the s^{th} component of X_i , i.e., $X_i = (X_{i1}, \dots, X_{ir})$. Letting $\Lambda_{X_{is}} = \sum_{z_s \in \mathcal{D}_s} \lambda_s^{d_{X_{is}, z_s}}$, we estimate $p(x)$ by

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

where $L(X_i, x, \lambda) = \prod_{s=1}^r l(X_{is}, x_s, \lambda_s)$ is a product kernel function, and where $l(X_{is}, x_s, \lambda_s) = \lambda_s^{d_{X_{is}, x_s}} / \Lambda_{X_{is}}$.

6.1 Multivariate Data-Driven Smoothing Parameter Selection

Similar to Section 4, we consider using LCV and LSCV methods for selecting the vector of smoothing parameters, $\lambda = (\lambda_1, \dots, \lambda_r)$.

The LCV method selects the vector λ by maximizing

$$CV_n(\lambda) = \sum_{i=1}^n \log \hat{p}_{-i}(X_i), \tag{7}$$

where $\hat{p}_{-i}(X_i) = n^{-1} \sum_{j \neq i} L(X_j, X_i, \lambda)$ is the leave-one-out probability estimator of $p_i = p(X_i)$.

The LSCV method selects the vector λ by minimizing

$$J_n(\lambda) = \sum_{x \in \mathcal{D}} \hat{p}^2(x) - \frac{2}{n} \sum_{i=1}^n \hat{p}_{-i}(X_i), \quad (8)$$

where $\sum_{x \in \mathcal{D}} = \sum_{x_1 \in \mathcal{D}_1} \cdots \sum_{x_r \in \mathcal{D}_r}$, and $\hat{p}_{-i}(X_i)$ is defined in Equation (7).

Proposition 6.1. Let $\hat{\lambda}_{LCV} = (\hat{\lambda}_1, \dots, \hat{\lambda}_r)$ denote the value of the vector λ that maximizes $CV_n(\lambda)$ defined in Equation (7). If $p(\cdot)$ is not uniformly distributed in any of its components, $\lambda_s \rightarrow 0$ as $n \rightarrow \infty$ for $s = 1, \dots, r$, then we have

$$\hat{\lambda}_s = O_p(n^{-1}),$$

for $s = 1, \dots, r$.

Proposition 6.2. Let $\hat{\lambda}_{LSCV} = (\hat{\lambda}_1, \dots, \hat{\lambda}_r)$ denote the value of the vector λ that minimizes $J_n(\lambda)$ defined in Equation (8). If $p(\cdot)$ is not uniformly distributed in any of its components, $\lambda_s \rightarrow 0$ as $n \rightarrow \infty$ for $s = 1, \dots, r$, and we have

$$\hat{\lambda}_s = O_p(n^{-1}),$$

for $s = 1, \dots, r$.

The proofs of propositions 6.1 and 6.2 are similar to the univariate case and are therefore omitted.

With the fast convergence rate $\hat{\lambda} = O_p(n^{-1})$, we have, for any $x \in \mathcal{D} = \prod_{s=1}^r \mathcal{D}_s$,

$$\sqrt{n}(\hat{p}(x) - p(x)) \xrightarrow{d} N(0, p(x)(1 - p(x))),$$

and

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))),$$

where $\hat{F}(x) = \sum_{z \in \mathcal{D}, z \leq x} \hat{p}(z)$, and where $z \leq x$ means $z_s \leq x_s$ for all $s = 1, \dots, r$.

Remark. Suppose that $p(\cdot)$ is uniformly distributed with respect to some component, e.g., $p(\cdot)$ is uniformly distributed only with respect to its last component, i.e., $p(x_{-r}, x_r) = p(x_{-r}, z_r)$ for all $x_{-r} \in \mathcal{D}_{-r}$, $x_r, z_r \in \mathcal{D}_r$, where $x_{-r} = (x_1, \dots, x_{r-1})$, and where $\mathcal{D}_{-r}$ is the support of x_{-r} . Then, using a proof similar to the proof for Theorem 2.4 in Ouyang et al. (2006), one can prove that $\hat{\lambda}_s = O_p(n^{-1})$ for $s = 1, \dots, r-1$ and that $\lim_{n \rightarrow \infty} P(\hat{\lambda}_r = 1) = \delta$ for some $\delta \in (0, 1)$, where $\hat{\lambda}_s$ could be obtained by either the LCV or LSCV methods described above.

7 Empirical Illustration

We consider a dataset collected by Hausman et al. (1984) which records the number of successful patent applications by 128 U.S. firms across a seven-year period (1968-1974). First we considered three popular parametric probability models for ordered data, the binomial, Poisson, and negative binomial. Goodness of fit tests are conducted for each parametric model and they are all rejected by the data, therefore we consider instead consistent nonparametric estimates.³ We model the probability function for the number of successful patent applications using the two methods of smoothing parameter selection outlined above and present results in Figure 1. The nonsmooth estimate is quite noisy, while both smooth estimates are much less so. Of the two methods of smoothing parameter selection considered, LCV produces the least noisy estimates. The smoothing parameters chosen by LCV and LSCV are 0.8686869 and 0.4545455, respectively. Note that this dataset is characterized by numerous gaps in the data (i.e., outcomes with zero counts). Like its empirical counterpart, the smooth estimate delivers probability estimates that sum to 1, but the smooth estimate is expected to be more efficient from the square error perspective.

8 Summary

We propose an asymmetric empirical support ordered kernel function for the nonparametric estimation of probability mass functions, provide theoretical underpinnings, data-driven smoothing parameter selection procedures, and compare the proposed estimator with the nonsmooth empirical estimator and an estimator that ignores order present in the data. The proposed estimator shrinks the unbiased empirical probabilities towards the discrete uniform probabilities. The estimator is always proper, automatically adapts to gaps in data if present, contains the conventional frequency estimator as a special case when the smoothing parameter takes its lower bound value, and can assume the discrete uniform probabilities at the upper bound of its

³The negative binomial provided the best fit overall, but its tail is too thin, i.e., as the number of successful patent applications increases, the probability estimates approach zero too quickly yet there is still a substantial amount of empirical probability mass remaining in the tail.

Figure 1: Probability function estimates for the patent data. The smooth estimate appears as a solid line, the nonsmooth estimate as a dashed line. The LCV smoothed estimate appears on the left, the LSCV smoothed estimate appears on the right.

smoothing parameter, properties that are collectively lacking from existing ordered kernel functions that appear in the literature on the estimation of probability mass functions for multinomial ordered data.

9 Appendix: Proof of the Main Results

9.1 Calculation of the Mean and Variance of $\hat{p}(x)$ and $\hat{F}(x)$

Assuming i.i.d. draws and assuming that $\lambda = o(1)$,

$$\begin{aligned}
E\hat{p}(x) &= \frac{1}{n} \sum_{i=1}^n El(X_i, x, \lambda) \\
&= El(X_i, x, \lambda) \\
&= \sum_{z \in \mathcal{D}} l(z, x, \lambda) p(z) \\
&= l(x, x, \lambda) p(x) + \sum_{z \in \mathcal{D}, z \neq x} l(z, x, \lambda) p(z) \\
&= \frac{1}{\Lambda_x} p(x) + \sum_{z \in \mathcal{D}, z \neq x} \frac{\lambda^{d_{zx}}}{\Lambda_z} p(z) \\
&= (1 - b_{1x}\lambda - b_{2x}\lambda^2 - \dots) p(x) + \sum_{z \in \mathcal{D}, z \neq x} \frac{\lambda^{d_{zx}}}{\Lambda_z} p(z) \\
&= p(x) + c_{1x}\lambda + O(\lambda^2),
\end{aligned} \tag{9}$$

where b_{1x} takes values in $\{0, 1, 2\}$ and $c_{1x} = \sum_{z \in \mathcal{D}, d_{zx}=1} p(z) - b_{1x}p(x)$. As well,

$$\text{Var } \hat{p}(x) = \frac{1}{n} \left(El^2(X_i, x, \lambda) - (El(X_i, x, \lambda))^2 \right).$$

We already obtained the second term on the right hand side of the above equation when we obtained $E\hat{p}(x) = El(X_i, x, \lambda)$ previously. Therefore, we only need to consider $El^2(X_i, x, \lambda)$. We observe that

$$\begin{aligned} El^2(X_i, x, \lambda) &= \sum_{z \in \mathcal{D}} l^2(z, x, \lambda) p(z) \\ &= l^2(x, x, \lambda) p(x) + \sum_{z \in \mathcal{D}, z \neq x} l^2(z, x, \lambda) p(z) \\ &= \left(\frac{1}{\Lambda_x} \right)^2 p(x) + \sum_{z \in \mathcal{D}, z \neq x} \left(\frac{\lambda^{d_{zx}}}{\Lambda_z} \right)^2 p(z) \\ &= (1 - b_{1x}\lambda - \dots)^2 p(x) + \sum_{z \in \mathcal{D}, z \neq x} \lambda^{2d_{zx}} (1 + O(\lambda)) p(z) \\ &= p(x) - 2\lambda b_{1x} p(x) + O(\lambda^2), \end{aligned} \tag{10}$$

where the leading term in the expression $\sum_{z \in \mathcal{D}, z \neq x} (\lambda^{d_{zx}} / \Lambda_z)^2 p(z)$ is $O(\lambda^2)$ because $d_{zx} \geq 1$ (when $x \neq z$). Therefore, using equations (9) and (10), we can write

$$\begin{aligned} \text{Var } \hat{p}(x) &= \frac{1}{n} \left(El^2(X_i, x, \lambda) - (El(X_i, x, \lambda))^2 \right) \\ &= \frac{1}{n} \left(p(x) - 2\lambda b_{1x} p(x) - (p^2(x) + 2\lambda c_{1x} p(x)) + O(\lambda^2) \right) \\ &= \frac{p(x)(1 - p(x))}{n} + \frac{d_{1x}\lambda}{n} + O\left(\frac{\lambda^2}{n}\right), \end{aligned}$$

where $d_{1x} = -2(b_{1x} + c_{1x})p(x)$ is a constant.

It is easy to show that the expected value of $\hat{F}(x)$ is given by

$$\begin{aligned}
E\hat{F}(x) &= \frac{1}{n} \sum_{i=1}^n E\mathcal{L}(X_i, x, \lambda) \\
&= E\mathcal{L}(X_i, x, \lambda) \\
&= \sum_{z \in \mathcal{D}, z \leq x} El(X_i, z, \lambda) \\
&= \sum_{z \in \mathcal{D}, z \leq x} (p(z) + c_{1z}\lambda + O(\lambda^2)) \\
&= F(x) + d_{1x}\lambda + O(\lambda^2),
\end{aligned}$$

where the fourth equality follows from Equation (9), and where $d_{1x} = \sum_{z \in \mathcal{D}, z \leq x} c_{1z}$. We retain the leading bias term, which is of order $O(\lambda)$, and we square this when computing the mean square error (MSE) so that this term's contribution to the MSE expression is of order $O(\lambda^2)$.

Similar to the calculation for $\text{Var } \hat{p}(x)$, one can show that the variance of $\hat{F}(x)$ can be written as

$$\begin{aligned}
\text{Var } \hat{F}(x) &= \frac{1}{n} (E\mathcal{L}^2(X_i, x, \lambda) - (E\mathcal{L}(X_i, x, \lambda))^2) \\
&= \frac{F(x)(1 - F(x))}{n} + e_{1x}\lambda/n + O(\lambda^2/n),
\end{aligned}$$

where e_{1x} is a constant. The detailed derivation to obtain an explicit expression for e_{1x} is tedious, but this result should be quite intuitive. Therefore, we omit the derivation because the explicit expression of e_{1x} is immaterial.

9.2 Proof of Theorem 4.1

Proof. Let $\hat{\lambda}_{LCV}$ denote the LCV selected λ which maximizes

$$\frac{1}{n} \sum_{i=1}^n \log \hat{p}_{-i}(X_i). \tag{11}$$

Since $x \in \mathcal{D}$ takes a finite number of possible values, say c , it is straightforward to see that (because $\lambda = o(1)$) $\max_{x \in \mathcal{D}} |\hat{p}(x) - p(x)| = o_p(1)$, and that

$$\max_{1 \leq i \leq n} |\hat{p}_{-i} - p_i| = o_p(1), \quad (12)$$

where $\hat{p}_{-i} = n^{-1} \sum_{j \neq i}^n \lambda^{d_{ji}} / \Lambda_j$ is the leave-one-out probability estimator of $p_i = p(X_i)$, and where $d_{ji} = |X_j - X_i|$.

Note that maximizing Equation (11) is equivalent to maximizing

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \log \hat{p}_{-i}(X_i) - \frac{1}{n} \sum_{i=1}^n \log p_i, \quad (13)$$

because the second term is unrelated to λ .

Letting $\Delta_i = (\hat{p}_{-i} - p_i)/p_i$, then Equation (13) can be written as

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \log \frac{\hat{p}_{-i}}{p_i} &= \frac{1}{n} \sum_{i=1}^n \log \left[1 + \frac{\hat{p}_{-i} - p_i}{p_i} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \log(1 + \Delta_i) \\ &= \frac{1}{n} \sum_{i=1}^n \Delta_i - \frac{1}{2}(1 + o_p(1)) \frac{1}{n} \sum_{i=1}^n \Delta_i^2, \end{aligned}$$

where in the last equality we use the mean-value theorem.⁴

Hence, choosing λ to maximize Equation (11) is asymptotically equivalent to choosing λ to maximize the leading term of $CV(\lambda)$ which is given by

$$\begin{aligned} CV_0(\lambda) &= \frac{1}{n} \sum_{i=1}^n \Delta_i - \frac{1}{2n} \sum_{i=1}^n \Delta_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\hat{p}_{-i}}{p_i} - 1 - \frac{1}{2n} \sum_{i=1}^n \left[\frac{\hat{p}_{-i}}{p_i} - 1 \right]^2 \\ &= \frac{2}{n} \sum_{i=1}^n \frac{\hat{p}_{-i}}{p_i} - \frac{1}{2n} \sum_{i=1}^n \frac{\hat{p}_{-i}^2}{p_i^2} - \frac{3}{2} \\ &\equiv 2A_1 - \frac{1}{2}A_2 - \frac{3}{2}, \end{aligned} \quad (14)$$

⁴Using a Taylor expansion, $f(\Delta_i) \equiv \log(1 + \Delta_i) = \Delta_i - \frac{1}{2(1+\bar{\Delta}_i)^2} \Delta_i^2$, where $\bar{\Delta}_i \in (0, \Delta_i)$. Note that $\frac{1}{(1+\bar{\Delta}_i)^2} = 1 - 2\bar{\Delta}_i + O(\bar{\Delta}_i^2) = 1 + o_p(1)$ because $|\bar{\Delta}_i| \leq \max_{1 \leq j \leq n} |\bar{\Delta}_j| = o_p(1)$.

where $A_1 = n^{-1} \sum_{i=1}^n \hat{p}_{-i}/p_i$ and $A_2 = n^{-1} \sum_{i=1}^n \hat{p}_{-i}^2/p_i^2$.

Replacing A_1 and A_2 in Equation (14) by their leading terms given in lemmas 9.1 and 9.2, we obtain that

$$CV(\lambda) = -\left[2\bar{B}_2 + \frac{1}{2}\bar{H}_2\right]\lambda^2 + \frac{\lambda}{n}C_n + o_p(\lambda^2 + \lambda/n) + \text{terms unrelated to } \lambda,$$

where $\bar{B}_2$ and $\bar{H}_2$ are positive constants (by Lemma 9.3), $C_n = C_{n1} + C_{n2}$ is an $O_p(1)$ random variable, with C_{n1} and C_{n2} defined in lemmas 9.1 and 9.2, respectively.

Taking the derivative of $CV(\lambda)$ with respect to λ , setting it equal to zero, and solving for λ yields

$$\hat{\lambda} = \begin{cases} \frac{1}{n} \left[\frac{C_n}{4\bar{B}_2 + \bar{H}_2} \right] + o_p(n^{-1}), & \text{if } C_n \geq 0 \\ o_p(n^{-1}) & \text{if } C_n < 0 \end{cases}.$$

This implies that $\hat{\lambda} = O_p(n^{-1})$, completing the proof of Theorem 4.1. $\square$

9.3 Some Useful Lemmas

Lemma 9.1. *Letting A_1 be defined as in Equation (14), then under the same conditions as in Theorem 4.1, we have*

$$A_1 = -\bar{B}_2\lambda^2 + \frac{2\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{\lambda}{n}C_{n1} + o_p(\lambda^2 + \lambda/n) + \text{terms unrelated to } \lambda,$$

where $\bar{B}_2 = \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} p(x) \mathbf{1}_{1,x,y} \mathbf{1}_{1,x,z}$, $C_{n1} = \frac{2}{n-1} \sum_{i=1}^n \sum_{j>i}^n [B_{2,ij} - B_{1,i} - B_{1,j}]$, $B_{1,i} = \frac{1}{2} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right]$, and $B_{2,ij} = \frac{1}{2} \left\{ \mathbf{1}_{1,i,j} \left[\frac{1}{p_i} + \frac{1}{p_j} \right] - \mathbf{1}_{0,i,j} \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{1,j,z} + \frac{1}{p_j} \mathbf{1}_{1,i,z} \right] \right\}$.

Proof.

$$\begin{aligned} A_1 &= \frac{n-1}{n} \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n \frac{1}{2} \left[\frac{\lambda^{d_{ji}}}{p_i \Lambda_j} + \frac{\lambda^{d_{ij}}}{p_j \Lambda_i} \right] \\ &\equiv \frac{n-1}{n} U_2, \end{aligned} \tag{15}$$

where $U_2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n \frac{1}{2} \left[\frac{\lambda^{d_{ji}}}{p_i \Lambda_j} + \frac{\lambda^{d_{ij}}}{p_j \Lambda_i} \right]$ is a second order U -statistic.

Let

$$K_{ij} = \frac{1}{2} \left[\frac{\lambda^{d_{ji}}}{p_i \Lambda_j} + \frac{\lambda^{d_{ij}}}{p_j \Lambda_i} \right] = \frac{1}{2} \lambda^{d_{ji}} \left[\frac{1}{p_i \Lambda_j} + \frac{1}{p_j \Lambda_i} \right], \quad (16)$$

where the last equality follows from $d_{ij} = d_{ji}$.

Using U-statistic H-decomposition,

$$U_2 = E(K_{ij}) + U_{2,1} + U_{2,2}, \quad (17)$$

where $U_{2,1} = \frac{2}{n} \sum_{i=1}^n \left[E(K_{ij}|X_i) - E(K_{ij}) \right]$ and $U_{2,2} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [K_{ij} - E(K_{ij}|X_i) - E(K_{ij}|X_j) + E(K_{ij})]$.

To simplify the exposition we will only consider the case where $\mathcal{D}$ contains c consecutive non-negative integers.

The proof for the general case is similar but more tedious. Define

$$\mathbf{1}_{s,x,z} = \mathbf{1}(|x - z| = s), \quad \mathbf{1}_{s,i,j} = \mathbf{1}(|X_i - X_j| = s), \quad \mathbf{1}_{s,i,z} = \mathbf{1}(|X_i - z| = s),$$

for $s = 0, 1, 2$.

Note that $\mathbf{1}_{0,x,z} = \mathbf{1}(x = z)$, $\mathbf{1}_{0,i,j} = \mathbf{1}(X_i = X_j)$ and $\mathbf{1}_{0,i,z} = \mathbf{1}(X_i = z)$.

From Equation (16), we have

$$\begin{aligned} K_{ij} &= \frac{1}{2} \left[\mathbf{1}_{0,i,j} + \mathbf{1}_{1,i,j} \lambda + \mathbf{1}_{2,i,j} \lambda^2 + O_p(\lambda^3) \right] \left\{ \frac{1}{p_i} \left[1 - \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,j,z} \lambda + \mathbf{1}_{2,j,z} \lambda^2 + O_p(\lambda^3)] \right] \right. \\ &\quad \left. + \frac{1}{p_j} \left[1 - \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,i,z} \lambda + \mathbf{1}_{2,i,z} \lambda^2 + O_p(\lambda^3)] \right] \right\} \\ &= \frac{1}{2} \mathbf{1}_{0,i,j} \left[\frac{1}{p_i} + \frac{1}{p_j} \right] + \frac{\lambda}{2} \left\{ \mathbf{1}_{1,i,j} \left[\frac{1}{p_i} + \frac{1}{p_j} \right] - \mathbf{1}_{0,i,j} \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{1,j,z} + \frac{1}{p_j} \mathbf{1}_{1,i,z} \right] \right\} \\ &\quad + \frac{\lambda^2}{2} \left\{ \mathbf{1}_{2,i,j} \left[\frac{1}{p_i} + \frac{1}{p_j} \right] - \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{2,j,z} + \frac{1}{p_j} \mathbf{1}_{2,i,z} + \mathbf{1}_{1,i,j} \left(\frac{1}{p_i} \mathbf{1}_{1,j,z} + \frac{1}{p_j} \mathbf{1}_{1,i,z} \right) \right] \right\} + O_p(\lambda^3) \\ &\equiv B_{1,ij} + B_{2,ij} \lambda + B_{3,ij} \lambda^2 + O_p(\lambda^3), \end{aligned} \quad (18)$$

where the definition of $B_{k,ij}$, $k = 1, 2, 3$ should be apparent.

$$\begin{aligned}
E(K_{ij}|X_i) &= \frac{1}{2} \sum_{x \in \mathcal{D}} p(x) \mathbf{1}_{0,i,x} \left[\frac{1}{p_i} + \frac{1}{p(x)} \right] \\
&\quad + \frac{\lambda}{2} \sum_{x \in \mathcal{D}} p(x) \left\{ \mathbf{1}_{1,i,x} \left[\frac{1}{p_i} + \frac{1}{p(x)} \right] - \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{0,i,x} \mathbf{1}_{1,x,z} + \frac{1}{p(x)} \mathbf{1}_{0,i,x} \mathbf{1}_{1,i,z} \right] \right\} \\
&\quad + \frac{\lambda^2}{2} \sum_{x \in \mathcal{D}} p(x) \left\{ \mathbf{1}_{2,i,x} \left[\frac{1}{p_i} + \frac{1}{p(x)} \right] - \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{2,x,z} + \frac{1}{p(x)} \mathbf{1}_{2,i,x} \right. \right. \\
&\quad \left. \left. + \mathbf{1}_{1,i,x} \left(\frac{1}{p_i} \mathbf{1}_{1,x,z} + \frac{1}{p(x)} \mathbf{1}_{1,i,z} \right) \right] \right\} + O_p(\lambda^3) \\
&= 1 + \frac{\lambda}{2} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] \\
&\quad + \frac{\lambda^2}{2} \sum_{x \in \mathcal{D}} p(x) \left\{ \mathbf{1}_{2,i,x} \left[\frac{1}{p_i} + \frac{1}{p(x)} \right] - \sum_{z \in \mathcal{D}} \left[\frac{1}{p_i} \mathbf{1}_{2,x,z} + \frac{1}{p(x)} \mathbf{1}_{2,i,x} \right. \right. \\
&\quad \left. \left. + \mathbf{1}_{1,i,x} \left(\frac{1}{p_i} \mathbf{1}_{1,x,z} + \frac{1}{p(x)} \mathbf{1}_{1,i,z} \right) \right] \right\} + O_p(\lambda^3) \\
&\equiv 1 + B_{1,i} \lambda + B_{2,i} \lambda^2 + O_p(\lambda^3), \tag{19}
\end{aligned}$$

where $B_{1,i}$ and $B_{2,i}$ should be apparent.

In deriving the second equality in Equation (19), when simplifying the first term that is unrelated to λ we used the fact that $\sum_{x \in \mathcal{D}} p(x) \mathbf{1}_{0,i,x} \left[\frac{1}{p_i} + \frac{1}{p(x)} \right] = p_i \left[\frac{1}{p_i} + \frac{1}{p_i} \right] = 2$, because $\mathbf{1}_{0,i,x} = \mathbf{1}(X_i = x)$ which equals 1 if and only if $X_i = x$; for the second term, $\sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{D}} p(x) \left[\frac{1}{p_i} \mathbf{1}_{0,i,x} \mathbf{1}_{1,x,z} + \frac{1}{p(x)} \mathbf{1}_{0,i,x} \mathbf{1}_{1,i,z} \right] = 2 \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} = 2 \sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} = 2 \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x}$.

Using Equation (19), the law of iterated expectation and noting that $E(B_{1,i}) = 0$, we have

$$\begin{aligned}
E(K_{ij}) &= E[E(K_{ij}|X_i)] \\
&= 1 - \lambda^2 \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} p(x) \mathbf{1}_{1,x,y} \mathbf{1}_{1,x,z} \\
&\equiv 1 - \bar{B}_2 \lambda^2 + O(\lambda^3), \tag{20}
\end{aligned}$$

where $\bar{B}_2 = \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} p(x) \mathbf{1}_{1,x,y} \mathbf{1}_{1,x,z}$.

The leading term of $U_{2,1}$ in Equation (17) is a zero mean partial sum.

$$\begin{aligned}
U_{2,1} &= \frac{2\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{2\lambda^2}{n} \sum_{i=1}^n [B_{2,i} + \bar{B}_2] + O_p(\lambda^3) \\
&= \frac{2\lambda}{n} \sum_{i=1}^n B_{1,i} + O_p(n^{-1/2}\lambda^2 + \lambda^3).
\end{aligned} \tag{21}$$

Note that the term $U_{2,2}$ in Equation (17) is a degenerate U-statistic. Combining equations (18), (19) and (20), we obtain

$$\begin{aligned}
U_{2,2} &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [B_{1,ij} - 1] \\
&\quad + \frac{2\lambda}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n [B_{2,ij} - B_{1,i} - B_{1,j}] + O_p(\lambda^3 + n^{-1/2}\lambda^2).
\end{aligned} \tag{22}$$

Combining equations (15), (17), (20), (21), and (22), we have that

$$\begin{aligned}
A_1 &= 1 + \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [B_{1,ij} - 1] - \bar{B}_2\lambda^2 + \frac{2\lambda}{n} \sum_{i=1}^n B_{1,i} \\
&\quad + \frac{2\lambda}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [B_{2,ij} - B_{1,i} - B_{1,j}] + O_p(\lambda^3 + n^{-1/2}\lambda^2) \\
&= -\bar{B}_2\lambda^2 + \frac{2\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{\lambda}{n} C_{n1} + O_p(\lambda^3 + n^{-1/2}\lambda^2) + \text{terms unrelated to } \lambda,
\end{aligned}$$

where $C_{n1} \equiv \frac{2}{n-1} \sum_{i=1}^{n-1} \sum_{j>i}^n [B_{2,ij} - B_{1,i} - B_{1,j}]$. It is easy to see that C_{n1} is a zero mean $O_p(1)$ random variable. $\square$

Lemma 9.2. *Let A_2 be defined as in Equation (14). Then under the same conditions as in Theorem 4.1, we have*

$$A_2 = \lambda \bar{H}_2 + \frac{8\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{\lambda}{n} C_{n2} + o_p(\lambda^2 + \lambda/n),$$

where $\bar{H}_2 = \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - p(y) \mathbf{1}_{1,w,x} \left[\mathbf{1}_{1,w,y} + \mathbf{1}_{1,x,y} \right] + \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} \right\},$

$C_{n2} = \overline{H}_1 + \frac{6}{n-1} \sum_{i=1}^{n-1} \sum_{j>i}^n \left[D_{ij,1,2} + D_{ij,2,2} + D_{ij,3,2} \right]$ with $\overline{H}_1$ defined in Equation (24), and $D_{ij,k,2}$, $k = 1, 2, 3$ defined in equations (28), (31), and (33), respectively.

Proof. Note that

$$\begin{aligned} A_2 &= \frac{1}{n^3} \sum_{i=1}^n \sum_{j \neq i}^n \sum_{k \neq i}^n \frac{\lambda^{d_{ji}+d_{ki}}}{p_i p_j \Lambda_j \Lambda_k} \\ &= \frac{1}{n^3} \sum_{i=1}^n \sum_{j \neq i}^n \frac{\lambda^{2d_{ji}}}{p_i^2 \Lambda_j^2} + \frac{1}{n^3} \sum_{i=1}^n \sum_{j \neq i}^n \sum_{k \neq i, k \neq j}^n \frac{\lambda^{d_{ji}+d_{ki}}}{p_i^2 \Lambda_j \Lambda_k} \\ &\equiv A_{2,1} + A_{2,2}, \end{aligned} \tag{23}$$

where $A_{2,1} = n^{-3} \sum_{i=1}^n \sum_{j \neq i}^n \lambda^{2d_{ji}} / (p_i^2 \Lambda_j^2)$ and $A_{2,2} = n^{-3} \sum_{i=1}^n \sum_{j \neq i}^n \sum_{k \neq i, k \neq j}^n \lambda^{d_{ji}+d_{ki}} / (p_i p_j \Lambda_j \Lambda_k)$.

It is straightforward to see that

$$\begin{aligned} A_{2,1} &= \frac{2\lambda}{n} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} \frac{p(x)}{p(y)} \mathbf{1}_{1,z,x} + O_p(n^{-1} \lambda^2) \\ &= \frac{\lambda}{n} \overline{H}_1 + O_p(n^{-1} \lambda^2), \end{aligned} \tag{24}$$

where $\overline{H}_1 = 2 \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} \frac{p(x)}{p(y)} \mathbf{1}_{1,z,x}$.

Next we examine term $A_{2,2}$.

$$\begin{aligned} A_{2,2} &= \frac{(n-1)(n-2)}{n^2} \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{k>j}^n \frac{1}{3} \left[\frac{\lambda^{d_{ji}+d_{ki}}}{p_i^2 \Lambda_j \Lambda_k} + \frac{\lambda^{d_{ij}+d_{kj}}}{p_j^2 \Lambda_i \Lambda_k} + \frac{\lambda^{d_{ik}+d_{jk}}}{p_k^2 \Lambda_i \Lambda_j} \right] \\ &\equiv \frac{(n-1)(n-2)}{n^2} U_3, \end{aligned} \tag{25}$$

where

$$U_3 = \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{k>j}^n \frac{1}{3} \left[\frac{\lambda^{d_{ji}+d_{ki}}}{p_i^2 \Lambda_j \Lambda_k} + \frac{\lambda^{d_{ij}+d_{kj}}}{p_j^2 \Lambda_i \Lambda_k} + \frac{\lambda^{d_{ik}+d_{jk}}}{p_k^2 \Lambda_i \Lambda_j} \right].$$

Let

$$\begin{aligned}
H_{ijk} &= \frac{1}{3} \left[\frac{\lambda^{d_{ji}+d_{ki}}}{p_i^2 \Lambda_j \Lambda_k} + \frac{\lambda^{d_{ij}+d_{kj}}}{p_j^2 \Lambda_i \Lambda_k} + \frac{\lambda^{d_{ik}+d_{jk}}}{p_k^2 \Lambda_i \Lambda_j} \right] \\
&\equiv \frac{1}{3} [H_{ijk,1} + H_{ijk,2} + H_{ijk,3}],
\end{aligned}$$

where $H_{ijk,1} = (1/3)\lambda^{d_{ji}+d_{ki}}/(p_i^2 \Lambda_j \Lambda_k)$, $H_{ijk,2} = (1/3)\lambda^{d_{ij}+d_{kj}}/(p_j^2 \Lambda_i \Lambda_k)$, and $H_{ijk,3} = (1/3)\lambda^{d_{ik}+d_{jk}}/(p_k^2 \Lambda_i \Lambda_j)$.

Then

$$U_3 = \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{k>j}^n H_{ijk}.$$

Note that U_3 is a third order U-statistic that is symmetric in i, j, k .

Define $H_i = E(H_{ijk}|X_i)$, $H_{i,l} = E(H_{ijk,l}|X_i)$, $H_{ij} = E(H_{ijk}|X_i, X_j)$, and $H_{ij,l} = E(H_{ijk,l}|X_i, X_j)$.

It is straightforward to see that $H_i = H_{i,1} + H_{i,2} + H_{i,3}$, and $H_{ij} = H_{ij,1} + H_{ij,2} + H_{ij,3}$.

By U-statistic H-decomposition,

$$\begin{aligned}
U_3 &= E(H_{ijk}) + \frac{3}{n} \sum_{i=1}^n [H_i - E(H_{ijk})] + \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [H_{ijk} - H_i - H_j + E(H_{ijk})] \\
&\quad + \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{k>j}^n [H_{ijk} - H_{ij} - H_{ik} - H_{jk} + H_i + H_j + H_k - E(H_{ijk})] \\
&\equiv E(H_{ijk}) + U_{3,1} + U_{3,2} + U_{3,3},
\end{aligned} \tag{26}$$

where $U_{3,1} = \frac{3}{n} \sum_{i=1}^n [H_i - E(H_{ijk})]$, $U_{3,2} = \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [H_{ij} - H_i - H_j + E(H_{ijk})]$, and $U_{3,3} = \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{k>j}^n [H_{ijk} - H_{ij} - H_{ik} - H_{jk} + H_i + H_j + H_k - E(H_{ijk})]$.

We first analyze the term $H_{ijk,1}$. The other two terms $H_{ijk,2}$ and $H_{ijk,3}$ follow similar arguments.

$$\begin{aligned}
H_{ijk,1} &= \frac{1}{3} \frac{\lambda^{d_{ji}+d_{ki}}}{p_i^2 \Lambda_j \Lambda_k} \\
&= \frac{1}{3} \frac{\mathbf{1}_{0,i,j} \mathbf{1}_{0,i,k} + \lambda [\mathbf{1}_{0,i,j} \mathbf{1}_{1,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,i,k}] + \lambda^2 [\mathbf{1}_{0,i,j} \mathbf{1}_{2,i,k} + \mathbf{1}_{2,i,j} \mathbf{1}_{0,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{1,i,k}] + O_p(\lambda^3)}{p_i^2 \left\{ 1 + \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,j,z} \lambda + \mathbf{1}_{2,j,z} \lambda^2] \right\} \left\{ 1 + \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,k,z} \lambda + \mathbf{1}_{2,k,z} \lambda^2] \right\} + O_p(\lambda^3)} \\
&= \frac{1}{3p_i^2} \left\{ \mathbf{1}_{0,i,j} \mathbf{1}_{0,i,k} + \lambda [\mathbf{1}_{0,i,j} \mathbf{1}_{1,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,i,k}] + \lambda^2 [\mathbf{1}_{0,i,j} \mathbf{1}_{2,i,k} + \mathbf{1}_{2,i,j} \mathbf{1}_{0,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{1,i,k}] \right\} \\
&\times \left\{ 1 - \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,j,z} \lambda + \mathbf{1}_{2,j,z} \lambda^2] \right\} \left\{ 1 - \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,k,z} \lambda + \mathbf{1}_{2,k,z} \lambda^2] \right\} + O_p(\lambda^3) \\
&= \frac{1}{3p_i^2} \mathbf{1}_{0,i,j} \mathbf{1}_{0,i,k} + \frac{\lambda}{3p_i^2} \left\{ \mathbf{1}_{0,i,j} \mathbf{1}_{1,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,i,k} - \mathbf{1}_{0,i,j} \mathbf{1}_{0,i,k} \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right] \right\} \\
&+ \frac{\lambda^2}{3p_i^2} \left\{ \mathbf{1}_{0,i,j} \mathbf{1}_{0,i,k} \left[\left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} \right) \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right) - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,j,z} - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,k,z} \right] \right. \\
&- \left. \left[\mathbf{1}_{0,i,j} \mathbf{1}_{1,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,i,k} \right] \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right] + \left[\mathbf{1}_{0,i,j} \mathbf{1}_{2,i,k} + \mathbf{1}_{2,i,j} \mathbf{1}_{0,i,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{1,i,k} \right] \right\} \\
&+ O_p(\lambda^3). \tag{27}
\end{aligned}$$

$$\begin{aligned}
H_{ij,1} &= E(H_{ijk,1} | X_i, X_j) \\
&= \frac{1}{3p_i} \mathbf{1}_{0,i,j} + \frac{\lambda}{3p_i} \sum_{x \in \mathcal{D}} \left\{ \frac{1}{p_i} p(x) \mathbf{1}_{0,i,j} \mathbf{1}_{1,i,x} + \mathbf{1}_{1,x,j} - \mathbf{1}_{0,i,j} \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,j,z} + \mathbf{1}_{1,x,z}] \right\} + O_p(\lambda^2) \\
&\equiv D_{ij,1,1} + \lambda D_{ij,1,2} + O_p(\lambda^2), \tag{28}
\end{aligned}$$

where the definitions of $D_{ij,1,1}$ and $D_{ij,1,2}$ should be apparent.

$$\begin{aligned}
H_{i,1} &= E(H_{ijk,1}|X_i) \\
&= \frac{1}{3} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \frac{1}{p_i^2} p(x)p(y) \mathbf{1}_{0,i,x} \mathbf{1}_{0,i,y} \\
&\quad + \frac{\lambda}{3p_i^2} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} p(x)p(y) \left\{ \mathbf{1}_{0,i,x} \mathbf{1}_{1,i,y} + \mathbf{1}_{1,i,x} \mathbf{1}_{0,i,y} - \mathbf{1}_{0,i,x} \mathbf{1}_{0,i,y} \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,x,z} + \mathbf{1}_{1,y,z}] \right\} \\
&\quad + \frac{\lambda^2}{3p_i^2} \left\{ \mathbf{1}_{0,i,x} \mathbf{1}_{0,i,y} \left[\left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right) + \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,y,z} \right) - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,x,z} - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,y,z} \right] \right. \\
&\quad \left. - \left[\mathbf{1}_{0,i,x} \mathbf{1}_{1,i,y} + \mathbf{1}_{1,i,x} \mathbf{1}_{0,i,y} \right] \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,x,z} + \mathbf{1}_{1,y,z}] + \left[\mathbf{1}_{0,i,x} \mathbf{1}_{2,i,y} + \mathbf{1}_{2,i,x} \mathbf{1}_{0,i,y} + \mathbf{1}_{1,i,x} \mathbf{1}_{1,i,y} \right] \right\} \\
&= \frac{1}{3} + \frac{2\lambda}{3} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] + \frac{\lambda^2}{3} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - 2\mathbf{1}_{2,i,x} \right. \\
&\quad \left. - \frac{2}{p_i} p(y) \mathbf{1}_{1,i,y} \left[\mathbf{1}_{1,i,x} + \mathbf{1}_{1,y,x} \right] + \frac{2}{p_i} p(y) \mathbf{1}_{2,i,y} + \frac{1}{p_i^2} p(x)p(y) \mathbf{1}_{1,i,x} \mathbf{1}_{1,i,y} \right\} + O_p(\lambda^3), \tag{29}
\end{aligned}$$

where in the last equality we use the same argument as in Equation (19).

Similarly, we have that

$$\begin{aligned}
H_{ijk,2} &= \frac{1}{3} \frac{\lambda^{d_{ij}+d_{kj}}}{p_j^2 \Lambda_i \Lambda_k} \\
&= \frac{1}{3p_j^2} \mathbf{1}_{0,i,j} \mathbf{1}_{0,j,k} + \frac{\lambda}{3p_j^2} \left\{ \mathbf{1}_{0,i,j} \mathbf{1}_{1,j,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,j,k} - \mathbf{1}_{0,i,j} \mathbf{1}_{0,j,k} \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right] \right\} \\
&\quad + \frac{\lambda^2}{3p_j^2} \left\{ \mathbf{1}_{0,i,j} \mathbf{1}_{0,j,k} \left[\left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} \right) \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right) - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,i,z} - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,k,z} \right] \right. \\
&\quad \left. - \left[\mathbf{1}_{0,i,j} \mathbf{1}_{1,j,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{0,j,k} \right] \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,k,z} \right] + \left[\mathbf{1}_{0,i,j} \mathbf{1}_{2,j,k} + \mathbf{1}_{2,i,j} \mathbf{1}_{0,j,k} + \mathbf{1}_{1,i,j} \mathbf{1}_{1,j,k} \right] \right\} \\
&\quad + O_p(\lambda^3). \tag{30}
\end{aligned}$$

$$\begin{aligned}
H_{ij,2} &= E(H_{ijk,2}|X_i, X_j) \\
&= \frac{1}{3p_j} \mathbf{1}_{0,i,j} + \frac{\lambda}{3p_j} \sum_{x \in \mathcal{D}} \left\{ \frac{1}{p_j} p(x) \mathbf{1}_{0,i,j} \mathbf{1}_{1,j,x} + \mathbf{1}_{1,x,i} - \mathbf{1}_{0,i,j} \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,i,z} + \mathbf{1}_{1,x,z}] \right\} + O_p(\lambda^2) \\
&\equiv D_{ij,2,1} + \lambda D_{ij,2,2} + O_p(\lambda^2), \tag{31}
\end{aligned}$$

where the definitions of $D_{ij,2,1}$ and $D_{ij,2,2}$ should be apparent.

$$\begin{aligned}
H_{ijk,3} &= \frac{1}{3} \frac{\lambda^{d_{ik}+d_{kj}}}{p_k^2 \Lambda_i \Lambda_j} \\
&= \frac{1}{3p_k^2} \mathbf{1}_{0,i,k} \mathbf{1}_{0,j,k} + \frac{\lambda}{3p_k^2} \left\{ \mathbf{1}_{0,i,k} \mathbf{1}_{1,j,k} + \mathbf{1}_{1,i,k} \mathbf{1}_{0,j,k} - \mathbf{1}_{0,i,k} \mathbf{1}_{0,j,k} \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} \right] \right\} \\
&\quad + \frac{\lambda^2}{3p_k^2} \left\{ \mathbf{1}_{0,i,k} \mathbf{1}_{0,j,k} \left[\left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} \right) \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} \right) - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,i,z} - \sum_{z \in \mathcal{D}} \mathbf{1}_{2,j,z} \right] \right. \\
&\quad \left. - \left[\mathbf{1}_{0,i,k} \mathbf{1}_{1,j,k} + \mathbf{1}_{1,i,k} \mathbf{1}_{0,j,k} \right] \left[\sum_{z \in \mathcal{D}} \mathbf{1}_{1,i,z} + \sum_{z \in \mathcal{D}} \mathbf{1}_{1,j,z} \right] + \left[\mathbf{1}_{0,i,k} \mathbf{1}_{2,j,k} + \mathbf{1}_{2,i,k} \mathbf{1}_{0,j,k} + \mathbf{1}_{1,i,k} \mathbf{1}_{1,j,k} \right] \right\} \\
&\quad + O_p(\lambda^3). \tag{32}
\end{aligned}$$

$$\begin{aligned}
H_{ij,3} &= E(H_{ijk,3} | X_i, X_j) \\
&= \frac{1}{3p_i} \mathbf{1}_{0,i,j} + \frac{\lambda}{3p_i} \sum_{x \in \mathcal{D}} \left\{ 2\mathbf{1}_{1,i,j} - \mathbf{1}_{0,i,j} \sum_{z \in \mathcal{D}} [\mathbf{1}_{1,i,z} + \mathbf{1}_{1,j,z}] \right\} + O_p(\lambda^2) \\
&\equiv D_{ij,3,1} + \lambda D_{ij,3,2} + O_p(\lambda^2), \tag{33}
\end{aligned}$$

where the definitions of $D_{ij,3,1}$ and $D_{ij,3,2}$ should be apparent.

$$\begin{aligned}
H_{i,2} &= \frac{1}{3} + \frac{\lambda}{3} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] + \frac{\lambda^2}{3} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - \mathbf{1}_{2,i,x} \right. \\
&\quad \left. - 2\mathbf{1}_{1,i,x} [\mathbf{1}_{1,i,y} + \mathbf{1}_{1,x,y}] + \frac{1}{p_i} p(y) \mathbf{1}_{1,i,x} \mathbf{1}_{1,x,y} \right\} + O_p(\lambda^3), \tag{34}
\end{aligned}$$

and

$$\begin{aligned}
H_{i,3} &= \frac{1}{3} + \frac{\lambda}{3} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] + \frac{\lambda^2}{3} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - \mathbf{1}_{2,i,x} \right. \\
&\quad \left. - 2\mathbf{1}_{1,i,x} [\mathbf{1}_{1,i,y} + \mathbf{1}_{1,x,y}] + \frac{1}{p_i} p(y) \mathbf{1}_{1,i,x} \mathbf{1}_{1,x,y} \right\} + O_p(\lambda^3). \tag{35}
\end{aligned}$$

It is straightforward to calculate that

$$\begin{aligned}
E(H_{ijk,1}) &= \frac{1}{3} + \frac{\lambda^2}{3} \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - p(y) \mathbf{1}_{1,w,x} \left[\mathbf{1}_{1,w,y} + \mathbf{1}_{1,x,y} \right] \right. \\
&\quad \left. + \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} \right\} + O(\lambda^3).
\end{aligned}$$

Therefore,

$$\begin{aligned}
E(H_{ijk}) &= 3E(H_{ijk,1}) \\
&= 1 + \lambda^2 \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - p(y) \mathbf{1}_{1,w,x} \left[\mathbf{1}_{1,w,y} + \mathbf{1}_{1,x,y} \right] + \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} \right\} + O(\lambda^3) \\
&\equiv 1 + \lambda^2 \overline{H}_2 + O(\lambda^3),
\end{aligned} \tag{36}$$

where $\overline{H}_2$ is defined in Lemma 9.2.

Combining equations (29), (34), and (35), we have

$$\begin{aligned}
H_i &= H_{i,1} + H_{i,2} + H_{i,3} \\
&= 1 + \frac{4\lambda}{3} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] + O_p(\lambda^2).
\end{aligned} \tag{37}$$

Hence, combining equations (36) and (37), we have that

$$\begin{aligned}
U_{3,1} &= \frac{1}{n} \sum_{i=1}^n [H_i - E(H_{ijk})] \\
&= \frac{4\lambda}{n} \sum_{i=1}^n \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right] + O_p(n^{-1/2} \lambda^2) \\
&= \frac{8\lambda}{n} \sum_{i=1}^n B_{1,i} + O_p(n^{-1/2} \lambda^2),
\end{aligned} \tag{38}$$

where $B_{1,i} = \frac{1}{2} \sum_{x \in \mathcal{D}} \mathbf{1}_{1,i,x} p(x) \left[\frac{1}{p_i} - \frac{1}{p(x)} \right]$ as defined in Equation (19).

Adding up equations (28), (31), and (33), we have that

$$\begin{aligned}
H_{ij} &= H_{ij,1} + H_{ij,2} + H_{ij,3} \\
&= [D_{ij,1,1} + D_{ij,2,1} + D_{ij,3,1}] + \lambda [D_{ij,1,2} + D_{ij,2,2} + D_{ij,3,2}] + O_p(\lambda^2).
\end{aligned}$$

$$\begin{aligned}
U_{3,2} &= \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [H_{ij} - H_i - H_j + E(H_{ijk})] \\
&= \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [D_{ij,1,1} + D_{ij,2,1} + D_{ij,3,1} - 1] \\
&\quad + \frac{6\lambda}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [D_{ij,1,2} + D_{ij,2,2} + D_{ij,3,2}] + O_p(n^{-1}\lambda^2).
\end{aligned} \tag{39}$$

It is straightforward to see that in $U_{3,3}$ the term proportional to λ has an order $O_p(n^{-3/2}\lambda)$.

Combining equations (23), (24), (25), (26), (36), (38), and (39), we have that

$$\begin{aligned}
A_2 &= 1 + \frac{\lambda}{n} \bar{H}_1 + \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [D_{ij,1,1} + D_{ij,2,1} + D_{ij,3,1} - 1] + \lambda^2 \bar{H}_2 \\
&\quad + \frac{8\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{6\lambda}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^n [D_{ij,1,2} + D_{ij,2,2} + D_{ij,3,2}] + o_p(\lambda^2 + \lambda/n) \\
&= \lambda^2 \bar{H}_2 + \frac{8\lambda}{n} \sum_{i=1}^n B_{1,i} + \frac{\lambda}{n} C_{n2} + o_p(\lambda^2 + \lambda/n) + \text{terms unrelated to } \lambda
\end{aligned}$$

where $C_{n2} = \bar{H}_1 + \frac{6}{n-1} \sum_{i=1}^{n-1} \sum_{j>i}^n [D_{ij,1,2} + D_{ij,2,2} + D_{ij,3,2}]$.

It is straightforward to see that $C_{n2} = O_p(1)$. □

Lemma 9.3. *Let $\bar{B}_2$ and $\bar{H}_2$ be defined as in lemmas 9.1 and 9.2. Then both $\bar{B}_2$ and $\bar{H}_2$ are positive constants.*

Proof. Note that

$$\bar{B}_2 = \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{z \in \mathcal{D}} p(x) \mathbf{1}_{1,x,y} \mathbf{1}_{1,x,z} > 0.$$

We only need to show $\overline{H}_2 = \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - p(y) \mathbf{1}_{1,w,x} \left[\mathbf{1}_{1,w,y} + \mathbf{1}_{1,x,y} \right] + \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} \right\} > 0$.

Note that for any $x, y \in \mathcal{D}$, since $\frac{1}{p(x)} > 1$, $p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} < \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y}$. Hence,

$$\sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \frac{1}{p(x)} p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} - p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,w,y} \right\} > 0. \quad (40)$$

Next, note that

$$\begin{aligned} & \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) \left\{ \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,w,y} \right\} \\ &= \sum_{x \in \mathcal{D}} \left(\sum_{z \in \mathcal{D}} \mathbf{1}_{1,x,z} \right)^2 - \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(w) p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,w,y} \\ &= \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} \mathbf{1}_{1,x,y} \mathbf{1}_{1,x,w} - \sum_{x \in \mathcal{D}} \sum_{y \in \mathcal{D}} \sum_{w \in \mathcal{D}} p(x) p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,x,y} > 0, \end{aligned} \quad (41)$$

where we use the fact that for any $x, y \in \mathcal{D}$, $\mathbf{1}_{1,w,x} \mathbf{1}_{1,w,y} > p(w) p(y) \mathbf{1}_{1,w,x} \mathbf{1}_{1,w,y}$, since $p(x) p(y) < 1$.

Equations (40) and (41) imply that $\overline{H}_2 > 0$. Therefore,

$$2\overline{B}_2 + \frac{1}{2}\overline{H}_2 > 0.$$

□

References

- Ahmad, I. A. and 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. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.
- Burman, P. (1987). Smoothing sparse contingency tables. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 49(1):24–36.

- Chen, L., Guo, G., Wang, S., and Kong, X. (2014). Kernel learning method for distance-based classification of categorical data. *2014 14th UK Workshop on Computational Intelligence (UKCI)*, pages 1–7.
- Chu, C.-Y., Henderson, D. J., and Parmeter, C. F. (2017). On discrete Epanechnikov kernel functions. *Computational Statistics and Data Analysis*, 116(Supplement C):79 – 105.
- Dong, J. and Simonoff, J. S. (1994). The construction and properties of boundary kernels for smoothing sparse multinomials. *Journal of Computational and Graphical Statistics*, 3(1):57–66.
- Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, 68(343):683–691.
- Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4):1491–1519.
- Hall, P., Li, Q., and Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics*, 89:784–789.
- Hall, P., Racine, J. S., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026.
- Hall, P. and Titterton, D. M. (1987). On smoothing sparse multinomial data. *Australian Journal of Statistics*, 29(1):19–37.
- Harfouche, L., Adjabi, S., Zougab, N., and Funke, B. (2017). Multiplicative bias correction for discrete kernels. *Statistical Methods & Applications*.
- Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application of the patents-R&D relationship. *Econometrica*, 52(4):909–938.
- Ouyang, D., Li, Q., and Racine, J. S. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, 18(1):69–100.
- Rajagopalan, B. and Lall, U. (1995). A kernel estimator for discrete distributions. *Journal of Nonparametric Statistics*, 4(4):409–426.
- Titterton, D. M. and Bowman, A. W. (1985). A comparative study of smoothing procedures for ordered categorical data. *Journal of Statistical Computation and Simulation*, 21(3-4):291–312.
- van Ryzin, J. and Wang, H. (1978). The methods for smooth estimation of discrete distributions. *Communications in Statistics - Theory and Methods*, 7(3):211–228.
- Wang, H. and van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, 68:301–309.
- White, H. L. (1984). *Asymptotic Theory for Econometricians*. Academic Press, Orlando, FL.