Electronic Journal of Statistics
Vol. 9 (2015) 1518–1539
ISSN: 1935-7524
DOI: 10.1214/15-EJS1049

Uniform in bandwidth consistency of kernel estimators of the density of mixed data

David M. Mason*

Department of Applied Economics and Statistics, University of Delaware, 206 Townsend Hall, Newark, DE 19716, USA
and
Department of Statistics, North-West University, Potchefstroom, South Africa
e-mail: davidm@udel.edu

and

Jan W. H. Swanepoel†

Department of Statistics, North-West University, Potchefstroom, South Africa
e-mail: Jan.Swanepoel@nwu.ac.za

Abstract: We establish a general uniform in bandwidth consistency result for kernel estimators of the unconditional and conditional joint density of a distribution, which is defined by a mixed discrete and continuous random variable.

MSC 2010 subject classifications: Primary 60F15, 62G07; secondary 62G08.
Keywords and phrases: Kernel estimators, uniform in bandwidth, empirical process methods, mixed data.

Received November 2014.

1. Introduction

Kernel nonparametric function estimation methods have long attracted a great deal of attention. Although they are popular, they present only one of many approaches to the construction of good function estimators. These include, for example, nearest-neighbor, spline, neural network, and wavelet methods. These methods have been applied to a wide variety of data. In this article, we shall restrict attention to the construction of consistent kernel-type estimators of joint (unconditional and conditional) densities based on mixed data, that is data with both discrete and continuous components.

When faced such data, researchers have traditionally resorted to a "frequency" approach. This involves breaking the continuous data into subsets according to the realizations of the discrete data ("cells"), in order to produce consistent estimators. However, as the number of subsets increases, the amount

*The author is Extraordinary Professor at North-West University, Potchefstroom, South Africa.

†Research partially supported by National Research Foundation of South Africa.

of data in each cell tends to decrease, leading to a "sparse data" problem. In such cases, there may be insufficient data in each subset to deliver sensible density estimators (they will be highly variable). Aitchison and Aitken [1] proposed a novel extension of the kernel density estimation method to a discrete data setting in a multivariate binary discrimination context.

The approach we consider below uses "generalized product kernels". For the continuous component of a variable we use standard kernels (Epanechnikov, etc.) and for a general multivariate unordered discrete component we apply the kernels suggested by Aitchison and Aitken [1]. In case of ordered categorical data, alternative approaches can be used by essentially applying near-neighbor weights (see, e.g., Wang and van Ryzin [20]; Burman [3] and Hall and Titterington [10]). Smoothing methods for ordered categorical data have been surveyed by Simonoff [18, Sec. 6]. For illustration purposes, we show how this can be done using a kernel estimator proposed by Wang and van Ryzin [20].

Mason and Swanepoel [13] introduced a general method based on empirical process techniques to prove uniform in bandwidth consistency of a wide variety of kernel-type estimators. It is a distillation of results of Einmahl and Mason [8] and Dony et al. [5], whose work was motivated by the original groundwork of Nolan and Marron [14]. The goal of the present paper is to provide a general uniform in bandwidth consistency result for kernel estimators of the joint density of a distribution, which is defined by a mixed discrete and continuous random variable. We shall use the setup of Li and Racine [11] and show that the general Theorem of Mason and Swanepoel [13] applies to it. Our results will imply uniform in bandwidth consistency of the kernel density estimators for mixed discrete and continuous data of Li and Racine [11] and the kernel estimator of the conditional density for such data of Hall, Racine and Li [9].

In Section 2 we introduce and describe our basic setup, and some needed notation, constructions and assumptions. We prove our main technical result in Section 3 and in Section 4 we use it to prove a uniform in bandwidth consistency theorem for kernel density estimators of mixed data. Applications are given in Section 5. Section 6 contains the material from Mason and Swanepoel [13] that we use to prove our results. We conclude in Section 7 with an appendix on pointwise measurability.

2. Some basic notation, a probability construction and assumptions

In order to state and prove our results we shall need the following basic setup, notation, probability constructions and assumptions. First, we focus on the case when we have a mix of continuous and general multivariate unordered (nominal) variables. The case when the discrete variables are ordered (ordinal) will be dealt with at the end of Section 4.

2.1. The Li and Racine setup

We shall take our basic setup from Li and Racine [11], using the notation (with some modifications) of Hall, Racine and Li [9]. Let for $p \geq 1$, $q \geq 1$,

$$\mathbf{X} = \left(\mathbf{X}^c, \mathbf{X}^d\right) = \left(\left(X_1^c, \ldots, X_p^c\right), \left(X_1^d, \ldots, X_q^d\right)\right) \in \mathbb{R}^p \times \mathbb{R}^q,$$

be a random vector. Assume that $\mathbf{X}^d$ takes on a finite number of values $\mathbf{x}^d = (x_1^d, \ldots, x_q^d)$ in an arbitrary finite subset $\mathcal{D}$ of $\mathbb{R}^q$ for which

$$P\left\{\mathbf{X}^d = \left(x_1^d, \ldots, x_q^d\right)\right\} =: p\left(\left(x_1^d, \ldots, x_q^d\right)\right) = p\left(\mathbf{x}^d\right) > 0.$$

Also, given $\mathbf{X}^d = (x_1^d, \ldots, x_q^d) = \mathbf{x}^d \in \mathcal{D}$, assume that $\mathbf{X}^c = (X_1^c, \ldots, X_p^c)$ has conditional density on $\mathbb{R}^p$,

$$f\left(\left(x_1^c, \ldots, x_p^c\right) \mid \left(x_1^d, \ldots, x_q^d\right)\right) = f\left(\mathbf{x}^c | \mathbf{x}^d\right),$$

for $\mathbf{x}^c = (x_1^c, \ldots, x_p^c) \in \mathbb{R}^p$. This says that $\mathbf{X} = (\mathbf{X}^c, \mathbf{X}^d)$ has joint density

$$f\left(\mathbf{x}^c, \mathbf{x}^d\right) = f\left(\mathbf{x}^c | \mathbf{x}^d\right) p\left(\mathbf{x}^d\right),$$

for $(\mathbf{x}^c, \mathbf{x}^d) \in \mathbb{R}^p \times \mathcal{D}$.

For each $\mathbf{x}^c \in \mathbb{R}^p$ and $\mathbf{h} = (h_1, \ldots, h_p) \in (0,1]^p$ introduce the kernel function of $\mathbf{z}^c = (z_1^c, \ldots, z_p^c) \in \mathbb{R}^p$

$$K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{z}^c\right) := \Pi_{j=1}^p h_j^{-1/p} K\left(\frac{x_j^c - z_j^c}{h_j^{1/p}}\right),$$

where K is a measurable real-valued function on $\mathbb{R}$ satisfying conditions (K.i)–(K.iv) stated in Subsection 2.4.1 below.

From now on we assume for convenience of labeling that for each $1 \le k \le q$, X_k^d takes on values $0, 1, \ldots, r_k - 1$, where $r_k \ge 2$, and thus

$$\mathcal{D} \subset \{0, 1, \ldots, r_1 - 1\} \times \cdots \times \{0, 1, \ldots, r_q - 1\}. \tag{2.1}$$

For any

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q) \in [0, (r_1 - 1)/r_1] \times \cdots \times [0, (r_q - 1)/r_q] =: \boldsymbol{\Gamma}, \tag{2.2}$$

set for $\mathbf{z}^d = (z_1^d, \ldots, z_q^d) \in \mathbb{R}^q$

$$K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{z}^d\right) := \Pi_{k=1}^q \left(\frac{\lambda_k}{r_k - 1}\right)^{I\left(z_k^d \ne x_k^d\right)} (1 - \lambda_k)^{I\left(z_k^d = x_k^d\right)}.$$

In particular, we have

$$K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{X}^c\right) = \Pi_{j=1}^p h_j^{-1/p} K\left(\frac{x_j^c - X_j^c}{h_j^{1/p}}\right)$$

and

$$K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}^d\right) = \Pi_{k=1}^q \left(\frac{\lambda_k}{r_k - 1}\right)^{I\left(X_k^d \ne x_k^d\right)} (1 - \lambda_k)^{I\left(X_k^d = x_k^d\right)}.$$

Whenever $\mathbf{X}_1 = (\mathbf{X}_1^c, \mathbf{X}_1^d)$, $\mathbf{X}_2 = (\mathbf{X}_2^c, \mathbf{X}_2^d)$, ..., is an i.i.d. $\mathbf{X} = (\mathbf{X}^c, \mathbf{X}^d)$ sequence, for each $i \geq 1$ we define $K_{\mathbf{h}}^c(\mathbf{x}^c, \mathbf{X}_i^c)$ and $K_{\boldsymbol{\lambda}}^d(\mathbf{x}^d, \mathbf{X}_i^d)$ as above with $(\mathbf{X}_i^c, \mathbf{X}_i^d)$ replacing $(\mathbf{X}^c, \mathbf{X}^d)$, $X_{i,j}^c$ replacing X_j^c, for $j = 1, \ldots, p$, and $X_{i,k}^d$ replacing X_k^d, for $k = 1, \ldots, q$.

For any vector $\mathbf{z}$ let $\max \mathbf{z}$ denote the maximum of its components. In particular,

$$\max \boldsymbol{\lambda} = \max \{\lambda_1, \ldots, \lambda_q\}.$$

Notice that for each $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}$

$$(1 - \max \boldsymbol{\lambda})^q I\left\{\mathbf{X}^d = \mathbf{x}^d\right\} \leq K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}^d\right) \leq \max \boldsymbol{\lambda} I\left\{\mathbf{X}^d \neq \mathbf{x}^d\right\} + I\left\{\mathbf{X}^d = \mathbf{x}^d\right\}.$$

For any $0 < \delta < 1$ let

$$\boldsymbol{\Gamma}(\delta) = \{\boldsymbol{\lambda} \in \boldsymbol{\Gamma} : \max \boldsymbol{\lambda} \leq \delta\}.$$

We see that uniformly in $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}(\delta)$

$$n^{-1} N_n\left(\mathbf{x}^d\right)(1 - \delta)^q \leq n^{-1} \sum_{i=1}^n K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right) \leq \delta + n^{-1} N_n\left(\mathbf{x}^d\right), \tag{2.3}$$

where

$$N_n\left(\mathbf{x}^d\right) = \sum_{i=1}^n I\left\{\mathbf{X}_i^d = \mathbf{x}^d\right\}. \tag{2.4}$$

Consider the Aitchison and Aitken [1] kernel estimator of $p(\mathbf{x}^d)$,

$$\widehat{p}_n\left(\mathbf{x}^d, \boldsymbol{\lambda}\right) := n^{-1} \sum_{i=1}^n K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right).$$

Remark 1. Although $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ was initially proposed by Aitchison and Aitken [1] as a smooth estimator of $p(\mathbf{x}^d)$ in a multivariate binary data discrimination context, it has since then often been applied to the analysis of general multivariate unordered discrete variables. Note that when $\boldsymbol{\lambda} = 0$, the estimator $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ reduces to the conventional frequency estimator $\tilde{p}_n(\mathbf{x}^d) = n^{-1} N_n(\mathbf{x}^d)$. Therefore, the smoothed estimator $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ includes the frequency estimator as a special case.

From a statistical perspective it is known (see, e.g., Brown and Rundell [2], and Ouyang et al. [16]) that the smooth estimator $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ may introduce some finite sample bias; however, it may also reduce the variance substantially, leading (using a bandwidth $\boldsymbol{\lambda}$ which balances bias and variance) to a reduction in the mean squared error of $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ relative to the frequency estimator $\tilde{p}_n(\mathbf{x}^d)$. Ouyang et al. [16] provide an informative discussion on some further interesting properties of $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$. It is, among others, pointed out that $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$ can be viewed as a Bayes-type estimator because it is a weighted average of a uniform probability and a frequency estimator. Their simulation studies also show that $\widehat{p}_n(\mathbf{x}^d, \boldsymbol{\lambda})$, particularly when used in conjunction with a data-driven method

of bandwidth selection such as least-squares cross-validation, performs much better than the commonly used frequency estimator $\tilde{p}_n(\mathbf{x}^d)$, especially in the case when some of the discrete variables are uniformly distributed (a specific definition of "uniformly distributed variables" is provided in their Section 2).

Lemma 1. *With probability 1,*

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta)} \sup_{\mathbf{x}^d\in\mathcal{D}} \left|\widehat{p}_n\left(\mathbf{x}^d,\boldsymbol{\lambda}\right) - p\left(\mathbf{x}^d\right)\right| \to 0, \ \text{as } \delta \searrow 0. \tag{2.5}$$

Proof. Since, with probability 1, $n^{-1}N_n(\mathbf{x}^d) \to p(\mathbf{x}^d)$, we readily conclude from inequality (2.3) that (2.5) holds with probability 1. □

Our aim is firstly to study the uniform in bandwidth consistency of estimators of the joint density $f(\mathbf{x}^c,\mathbf{x}^d)$ of $\mathbf{X} = (\mathbf{X}^c,\mathbf{X}^d)$ of the form

$$\widehat{f}_n\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{h}}^c\left(\mathbf{x}^c,\mathbf{X}_i^c\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{X}_i^d\right).$$

Our objective is to establish the result stated in Theorem 2, which is given in Section 4. In order to do this we must first build some needed framework and machinery.

2.2. Some useful classes of functions

In order to apply the Mason and Swanepoel [13] general uniform in bandwidth consistency theorem we must introduce the following classes of functions.

$$\mathcal{T} = \left\{\mathbf{t} = (t_1\ldots,t_p) \in (0,1]^p : \text{at least one } t_j = 1\right\}.$$

Notice there is a one to one correspondence between

$$\mathcal{T}\times(0,1] \ \text{ and } \ (0,1]^p$$

given by

$$\mathbf{h} = (h_1,\ldots,h_p) \in (0,1]^p \leftrightarrows (\mathbf{t},h), \ \text{where } h = \max \mathbf{h} \text{ and } t_j = h_j/h. \tag{2.6}$$

Also note that for any $\mathbf{t} = (t_1\ldots,t_p) \in \mathcal{T}$ and $h \in (0,1]$, we have $h = \max \mathbf{h}$, where $h_j = t_j h$ for $1 \le j \le p$.

Choose $\mathbf{t} \in \mathcal{T}$ and $\mathbf{x}^c = (x_1^c,\ldots,x_p^c) \in \mathbb{R}^p$. Define the function

$$g_{\mathbf{t},\mathbf{x}^c} : \mathbb{R}^p \times (0,1] \to \mathbb{R},$$

by

$$(\mathbf{z},h) \longmapsto g_{t,\mathbf{x}^c}(\mathbf{z},h) = \Pi_{j=1}^p K\left(\frac{x_j^c - z_j}{t_j^{1/p} h^{1/p}}\right)$$

for $\mathbf{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$ and $h \in (0, 1]$. Choose a measurable subset $\mathbb{A}$ of $\mathbb{R}^p$. Denote the class of measurable functions of $(\mathbf{z}, h) \in \mathbb{R}^p \times (0, 1]$ indexed by $(\mathbf{x}^c, \mathbf{t}) \in \mathbb{A} \times \mathcal{T}$,

$$\mathcal{G}_K = \{g_{\mathbf{t},\mathbf{x}^c} \ : (\mathbf{x}^c, \mathbf{t}) \in \mathbb{A} \times \mathcal{T}\}. \tag{2.7}$$

From this class we form the class $\mathcal{G}_{K,0}$ of measurable real valued functions of $\mathbf{z} \in \mathbb{R}^p$ defined as

$$\mathcal{G}_{K,0} = \{\mathbf{z} \mapsto g_{\mathbf{t},\mathbf{x}^c}(\mathbf{z},h) : g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K,\ 0 < h \leq 1\}. \tag{2.8}$$

Using this notation we see that

$$\widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) = \frac{1}{n\left(\Pi_{j=1}^p t_j\right)^{1/p} h} \sum_{i=1}^n g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{X}_i^c, h\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right).$$

where we use the one to one correspondence given in (2.6).

Remark 2. The class of functions given in this subsection can be used to apply the Theorem in Mason and Swanepoel [13] to obtain uniform in bandwidth consistency results for multivariate kernel estimators based on a vector of smoothing parameters, where the components may be different.

2.3. A useful probability construction

We shall see that the following probability construction will come in very handy. Let $\mathbf{X}_1 = (\mathbf{X}_1^c, \mathbf{X}_1^d), \mathbf{X}_2 = (\mathbf{X}_2^c, \mathbf{X}_2^d), \ldots$, be a sequence of i.i.d. $\mathbf{X} = (\mathbf{X}^c, \mathbf{X}^d)$ random vectors. Also for each $\mathbf{x}^d \in \mathcal{D}$, let $\mathbf{Z}(\mathbf{x}^d)$ be a random vector with density $f(\mathbf{x}^c|\mathbf{x}^d)$ on $\mathbb{R}^p$, and $\mathbf{Z}_1(\mathbf{x}^d)$, $\mathbf{Z}_2(\mathbf{x}^d), \ldots$, be a sequence of i.i.d $\mathbf{Z}(\mathbf{x}^d)$ random vectors. Further we assume that the sequences $\{\mathbf{X}_i\}_{i\geq 1}, \{\mathbf{Z}_i(\mathbf{x}^d)\}_{i\geq 1}$, $\mathbf{x}^d \in \mathcal{D}$, are independent of each other. For each $\mathbf{x}^d$ and $n \geq 1$, recall the definition of $N_n(\mathbf{x}^d)$ given (2.4). We find that for any class $\mathcal{F}$ of measurable real valued functions φ defined on $\mathbb{R}^p \times \mathcal{D} \times (0, 1]$,

$$\left\{\sum_{i=1}^n \varphi\left(\mathbf{X}_i, h\right) : \varphi \in \mathcal{F},\ h \in (0, 1]\right\}_{n\geq 1}$$

$$\stackrel{\mathrm{D}}{=} \left\{\sum_{\mathbf{x}^d \in \mathcal{D}} \sum_{i \leq N_n(\mathbf{x}^d)} \varphi\left(\mathbf{Z}_i\left(\mathbf{x}^d\right), \mathbf{x}^d, h\right) : \varphi \in \mathcal{F},\ h \in (0, 1]\right\}_{n\geq 1}.$$

To see the kind of argument that establishes this distributional identity consult the proof of Proposition 3.1 of Einmahl and Mason [6].

2.4. Assumptions

Here are our basic assumptions on the kernel and the joint and marginal densities.

2.4.1. Assumptions on the kernel K

The kernel K satisfies the following conditions:

(K.i) $K = K_1 - K_2$, where K_1 and K_2 are bounded, nondecreasing, right continuous functions on $\mathbb{R}$,
(K.ii) $|K| \leq \kappa < \infty$, for some $\kappa > 0$,
(K.iii) $\int K(u)du = 1$,
(K.iv) K has support contained in $[-B, B]$, for some $B > 0$.

Note that (K.ii) and (K.iv) imply that for any $h > 0$

$$\frac{1}{h}\int |K|\,(u/h)du = \frac{1}{h}\int_{-Bh}^{Bh} |K|\,(u/h)du = \int_{-B}^{B} |K|\,(v)dv \leq 2B\kappa. \tag{2.9}$$

2.4.2. Assumptions on the joint and marginal densities

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ set $|\mathbf{x} - \mathbf{y}| = \max\{|x_i - y_i| : i = 1, \ldots, p\}$ and for a measurable subset $\mathbb{A} \subset \mathbb{R}^p$ and $\varepsilon > 0$ we define

$$\mathbb{A}^\varepsilon = \{\mathbf{x} \in \mathbb{R}^p : |\mathbf{x} - \mathbf{y}| \ \leq \varepsilon \text{ for some } \mathbf{y} \in \mathbb{A}\}. \tag{2.10}$$

(f.i) For some $\varepsilon > 0$ and $M > 0$

$$\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{x}^c \in \mathbb{A}^\varepsilon} f\left(\mathbf{x}^c | \mathbf{x}^d\right) \leq M.$$

(f.ii) For some $\varepsilon > 0$ and $\delta > 0$

$$\min_{\mathbf{x}^d \in \mathcal{D}} \inf_{\mathbf{x}^c \in \mathbb{A}^\varepsilon} f\left(\mathbf{x}^c\right) \geq \delta.$$

3. Technical result

In this section we establish a technical result that will be used in the next section to prove our uniform in bandwidth theorem for kernel density estimators for mixed discrete and continuous data.

For any $i \geq 1$ and $\mathbf{x}^d \in \mathcal{D}$, set

$$\mathbf{Z}_i\left(\mathbf{x}^d\right) = \left(Z_{i,1}\left(\mathbf{x}^d\right), \ldots, Z_{i,p}\left(\mathbf{x}^d\right)\right),$$

where $\{\mathbf{Z}_i(\mathbf{x}^d)\}_{i \geq 1}$ are i.i.d. $\mathbf{Z}(\mathbf{x}^d)$.

In the following proposition, for $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$,

$$\begin{aligned} s_n(g_{t,\mathbf{x}^c}, \mathbf{x}^d, h) &:= \sum_{i=1}^{n} g_{t,\mathbf{x}^c}\left(\mathbf{Z}_i\left(\mathbf{x}^d\right), h\right) \\ &= \sum_{i=1}^{n} \Pi_{j=1}^{p} K\left(\frac{x_j^c - Z_{i,j}\left(\mathbf{x}^d\right)}{t_j^{1/p} h^{1/p}}\right). \end{aligned}$$

(Here and elsewhere in these notes $\log x$ denotes the natural logarithm of the maximum of x and e.)

Proposition 1. *Let K satisfy (K.i)–(K.iv) and the marginal densities fulfill (f.i). Then for any $\mathbf{x}^d \in \mathcal{D}$, choice of $c > 0$ and $0 < b_0 < 1$ we have, with probability 1,*

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{c_n \le h \le b_0} \sup_{g_{t,\mathbf{x}^c} \in \mathcal{G}_K} \frac{|s_n(g_{t,\mathbf{x}^c}, \mathbf{x}^d, h) - \mathbb{E}s_n(g_{t,\mathbf{x}^c}, \mathbf{x}^d, h)|}{\sqrt{nh\left(|\log h| \vee \log\log n\right)}} \\ = \;& A(c, \mathbf{x}^d), \end{aligned} \tag{3.11}$$

where $c_n = \frac{c \log n}{n}$, $A(c, \mathbf{x}^d)$ is a finite constant depending on c, $\mathbf{x}^d$, and the stated assumptions on the kernel K and the marginal densities.

Proof. Throughout the proof keep in mind that $\mathbb{A}$ is the set used in assumption (f.i) and to define the class $\mathcal{G}_K$ in (2.7). Choose any $\mathbf{x}^d \in \mathcal{D}$. Notice that for any $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$

$$s_n(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{x}^d, h) = \sum_{i=1}^{n} g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}_i\left(\mathbf{x}^d\right), h\right) = nh\hat{\vartheta}_{n,h}(g_{\mathbf{t},\mathbf{x}^c}).$$

(See the notation (6.33) below.) The assumptions of Proposition 1 allow us to apply the general Theorem of Mason and Swanepoel [13] (see below) with $\mathcal{G} = \mathcal{G}_K$ to conclude (3.11). In particular we see that (K.ii) implies that (G.i) holds (assumptions (G.i)–(G.iv) are stated in Subsection 6.2). Also it is readily shown using (f.i) and (K.ii) that (G.ii) is fulfilled, that is, for some constant $C > 0$ for all $\mathbf{t} \in \mathcal{T}$, $h \in (0, 1]$, $\mathbf{x}^c \in \mathbb{A}$ and $\mathbf{x}^d \in \mathcal{D}$

$$\mathbb{E}\left(g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{x}^d\right), h\right)\right)^2 \le C\left(\Pi_{j=1}^{p} t_j\right)^{1/p} h \le Ch. \tag{3.12}$$

To see this, observe that $g_{\mathbf{t},\mathbf{x}^c}(\cdot, h)$ is zero off the set

$$B_{\mathbf{t},h}\left(\mathbf{x}^c\right) = \mathbf{x}^c + \left[-Bt_1^{1/p}h^{1/p}, Bt_1^{1/p}h^{1/p}\right] \times \cdots \times \left[-Bt_p^{1/p}h^{1/p}, t_p^{1/p}Bh^{1/p}\right]$$

and for all h small enough uniformly in $\mathbf{x}^c \in \mathbb{A}$ and $\mathbf{t} \in \mathcal{T}$, $B_{\mathbf{t},h}(\mathbf{x}^c) \subset \mathbb{A}^\varepsilon$ so that (f.i) holds. From these observations (3.12) follows.

The results in the Appendix prove that (K.i) implies that the pointwise measurable assumption (G.iii) holds for the class $\mathcal{G}_{K,0}$. (Note that in assumption (F.ii) of Mason and Swanepoel [13] $\mathcal{G}$ should be $\mathcal{G}_\gamma$.) For any $1 \le j \le p$, define the class of functions

$$\mathcal{K}_j = \left\{ z_j \mapsto K\left(\frac{x_j^c - z_j}{h_j^{1/p}}\right) : \left(x_j^c, h_j\right) \in \mathbb{R} \times (0, 1] \right\}.$$

Using assumption (K.i), an application of Lemma 22 of Nolan and Pollard [15] shows that each $\mathcal{K}_j$ satisfies (G.iv). Further since by assumption (K.ii), $|K|$ is assumed to be bounded by some $\kappa > 0$, we can apply Lemma A.1 of Einmahl and Mason [7] to infer that $\mathcal{G}_{K,0}$ satisfies (G.iv). □

3.1. Main technical result

Here is our main technical result. In the following, for any $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}$, $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$ and $\mathbf{x}^d \in \mathcal{D}$

$$\begin{aligned}\widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d) &:= \sum_{i=1}^n g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{X}_i^c,h\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{X}_i^d\right) \\ &= \sum_{i=1}^n \Pi_{j=1}^p K\left(\frac{x_j^c - X_{i,j}^c}{t_j^{1/p}h^{1/p}}\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{X}_i^d\right).\end{aligned}$$

Theorem 1. *Let K satisfy (K.i)–(K.iv) and the marginal densities fulfill (f.i). Then for any choice of $c > 0$ and $0 < b_0 < 1$ we have, with probability 1,*

$$\begin{aligned}&\limsup_{n\to\infty} \max_{\mathbf{x}^d\in\mathcal{D}} \sup_{c_n\le h\le b_0} \sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}} \sup_{g_{t,\mathbf{x}^c}\in\mathcal{G}_K} \frac{|\widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d) - \mathbb{E}\widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d)|}{\sqrt{nh\left(|\log h| \vee \log\log n\right)}} \\ =\ & B(c), \qquad (3.13)\end{aligned}$$

where $c_n = \frac{c\log n}{n}$, $B(c)$ is a finite constant depending on c, and the stated assumptions on the kernel K and the marginal densities.

In order to prove the theorem we require the following lemma.

Lemma 2. *Let K satisfy (K.i)–(K.iv) and the marginal densities fulfill (f.i). Then for any $\mathbf{z}^d \in \mathcal{D}$, choice of $c > 0$ and $0 < b_0 < 1$ we have, with probability 1,*

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{c_n\le h\le b_0} \sup_{g_{\mathbf{t},\mathbf{x}^c}\in\mathcal{G}_K} \frac{|s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) - \mathbb{E}s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h)|}{\sqrt{nh\left(|\log h| \vee \log\log n\right)}} \\ =\ & C(c,\mathbf{z}^d), \qquad (3.14)\end{aligned}$$

where $c_n = \frac{c\log n}{n}$, $C(c,\mathbf{z}^d)$ is a finite constant depending on c, $\mathbf{z}^d$ and the stated assumptions on the kernel K and the marginal densities.

Proof. Choose any $\mathbf{z}^d \in \mathcal{D}$. Notice that by Wald's identity

$$\mathbb{E}s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) = np\left(\mathbf{z}^d\right)\mathbb{E}g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right),h\right).$$

Thus

$$\begin{aligned}& s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) - \mathbb{E}s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) \\ =\ & s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) - np\left(\mathbf{z}^d\right)\mathbb{E}g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right),h\right) \\ =\ & s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h) - N_n\left(\mathbf{z}^d\right)\mathbb{E}g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right),h\right) \\ & + \left(N_n\left(\mathbf{z}^d\right) - np\left(\mathbf{z}^d\right)\right)\mathbb{E}g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right),h\right).\end{aligned}$$

Since the assumptions of Proposition 1 hold, the sequence of random variables $\{N_n(\mathbf{z}^d)\}_{n\ge1}$ is independent of $\{\mathbf{Z}_n(\mathbf{z}^d)\}_{n\ge1}$, and $N_n(\mathbf{z}^d) \to \infty$, with probability 1, we see that, for every $d_0 > 0$, with probability 1, for some finite constant

$A(d_0, \mathbf{z}^d)$ depending on d_0 and $\mathbf{z}^d$, we have

$$\begin{aligned} & \limsup_{n\to\infty} \sup_{d_{N_n(\mathbf{z}^d)} \le h \le b_0} \sup_{g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K} \frac{|A_n\left(h, \mathbf{z}^d, g_{\mathbf{t},\mathbf{x}^c}\right)|}{\sqrt{N_n\left(\mathbf{z}^d\right) h \left(|\log h| \vee \log\log N_n\left(\mathbf{z}^d\right)\right)}} \\ = \; & A(d_0, \mathbf{z}^d), \end{aligned} \tag{3.15}$$

where

$$A_n\left(h, \mathbf{z}^d, g_{\mathbf{t},\mathbf{x}^c}\right) = s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{z}^d, h) - N_n\left(\mathbf{z}^d\right) \mathbb{E} g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right), h\right),$$

and $d_{N_n(\mathbf{z}^d)} = \frac{d_0 \log N_n(\mathbf{z}^d)}{N_n(\mathbf{z}^d)}$. Now since, with probability 1, $N_n(\mathbf{z}^d)/n \to p(\mathbf{z}^d) > 0$, and thus $d_{N_n(\mathbf{z}^d)} \le \frac{2d_0 \log n}{np(\mathbf{z}^d)}$ for all large enough n and $\frac{2d_0 \log n}{np(\mathbf{z}^d)} \le c_n$ for small enough $d_0 > 0$, we see from (3.15) that

$$\begin{aligned} & \limsup_{n\to\infty} \sup_{c_n \le h \le b_0} \sup_{g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K} \frac{|s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{z}^d, h) - N_n\left(\mathbf{z}^d\right) \mathbb{E} g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right), h\right)|}{\sqrt{nh\left(|\log h| \vee \log\log n\right)}} \\ = \; & \sqrt{p\left(\mathbf{z}^d\right)} A(d_0, \mathbf{z}^d) < \infty. \end{aligned} \tag{3.16}$$

Next, for each $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$, we get using the assumptions on K, (f.i) and (2.9) that for all $h > 0$ small enough

$$\left|\mathbb{E} g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{x}\right), h\right)\right| \le h\left(2B\kappa\right)^p M.$$

Thus, by the law of the iterated logarithm, with probability 1, for some $C_0 > 0$,

$$\begin{aligned} & \limsup_{n\to\infty} \sup_{c_n \le h \le b_0} \sup_{g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K} \frac{|\left(N_n\left(\mathbf{z}^d\right) - np\left(\mathbf{z}^d\right)\right) \mathbb{E} g_{\mathbf{t},\mathbf{x}^c}\left(\mathbf{Z}\left(\mathbf{z}^d\right), h\right)|}{\sqrt{nh\left(|\log h| \vee \log\log n\right)}} \\ \le \; & \limsup_{n\to\infty} \frac{\left|N_n\left(\mathbf{z}^d\right) - np\left(\mathbf{z}^d\right)\right| C_0}{\sqrt{n \log\log n}} = \sqrt{2p\left(\mathbf{z}^d\right)\left(1 - p\left(\mathbf{z}^d\right)\right)} C_0. \end{aligned} \tag{3.17}$$

The proof of (3.14) now follows from (3.16) and (3.17) and the Kolmogorov zero one law. □

Proof of Theorem 1. Notice that as a process in $(\mathbf{X}_i^c, \mathbf{X}_i^d)_{i\ge 1}$, $h \in (0,1]$, $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}$, $\mathbf{x}^d \in \mathcal{D}$ and $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$,

$$\begin{aligned} \widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{x}^d) &= \sum_{i=1}^n \Pi_{j=1}^p K\left(\frac{x_j^c - X_{i,j}^c}{t_j^{1/p} h^{1/p}}\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right) \\ &\overset{\mathrm{D}}{=} \sum_{\mathbf{z}^d \in \mathcal{D}} \sum_{i \le N_n(\mathbf{z}^d)} \Pi_{j=1}^p K\left(\frac{x_j^c - Z_{i,j}\left(\mathbf{z}^d\right)}{t_j^{1/p} h^{1/p}}\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{z}^d\right) \\ &= \sum_{\mathbf{z}^d \in \mathcal{D}} s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{z}^d, h) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{z}^d\right). \end{aligned} \tag{3.18}$$

(Recall the probability construction in Subsection 2.3.) From this we see that

$$\begin{aligned}&\widetilde{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d)-\mathbb{E}\widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d)\\&=\sum_{\mathbf{z}^d\in\mathcal{D}}\left(s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h)-\mathbb{E}s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h)\right)K^d_{\boldsymbol{\lambda}}\left(\mathbf{x}^d,\mathbf{z}^d\right).\end{aligned}\tag{3.19}$$

Noting that each $|K^d_{\boldsymbol{\lambda}}(\mathbf{x}^d,\mathbf{z}^d)|\le 1$, we see then using (3.19), with $|\mathcal{D}|$ denoting the cardinality of $\mathcal{D}$, that by Lemma 2, with probability 1,

$$\begin{aligned}&\limsup_{n\to\infty}\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{c_n\le h\le b_0}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}}\sup_{g_{\mathbf{t},\mathbf{x}^c}\in\mathcal{G}_K}\frac{|\widetilde{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d)-\mathbb{E}\widehat{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{x}^d)|}{\sqrt{nh\left(|\log h|\vee\log\log n\right)}}\\&\le\sum_{\mathbf{z}^d\in\mathcal{D}}\limsup_{n\to\infty}\sup_{c_n\le h\le b_0}\sup_{g_{\mathbf{t},\mathbf{x}^c}\in\mathcal{G}_K}\frac{|s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h)-\mathbb{E}s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c},\mathbf{z}^d,h)|}{\sqrt{nh\left(|\log h|\vee\log\log n\right)}}\\&\le\max_{\mathbf{z}^d\in\mathcal{D}}C\left(c,\mathbf{z}^d\right)|\mathcal{D}|.\end{aligned}$$

The Kolmogorov zero one law now completes the proof. □

4. Uniform in bandwidth consistency theorem

For any $\delta>0$ let

$$\boldsymbol{\Gamma}(\delta)=\{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}:\max\boldsymbol{\lambda}\le\delta\},$$

where $\boldsymbol{\Gamma}$ is as in (2.2). Given sequences $0<a_n<b_n<1$, set

$$\mathcal{H}_n=\left\{\mathbf{h}\in(0,1]^p:a_n\le\frac{\left(\Pi_{j=1}^p h_j\right)^{2/p}}{\max\mathbf{h}}\le\max\mathbf{h}\le b_n\right\}.$$

Note that if $h_1=\cdots=h_p=h$, then $\mathcal{H}_n$ becomes

$$\mathcal{H}_n=\{h\in(0,1]:a_n\le h\le b_n\}.$$

Theorem 2. *Let K satisfy (K.i)–(K.iv) and the marginal densities fulfill (f.i). For any sequences $0<a_n<b_n<1$, $0<\delta_n<1$ satisfying $b_n\to 0$, $\delta_n\to 0$, and $na_n/\log n\to\infty$, and density f on $\mathbb{R}^p\times\mathcal{D}$ such that for each $\mathbf{z}^d\in\mathcal{D}$, $f(\cdot|\mathbf{z}^d)$ is uniformly continuous on the subset $\mathbb{A}^\varepsilon$ of $\mathbb{R}^p$ for some $\varepsilon>0$, we have, with probability 1,*

$$\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{f}_n\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right)-f\left(\mathbf{x}^c,\mathbf{x}^d\right)\right|\to 0.\tag{4.20}$$

In order to prove the theorem we require the following lemma. Let $\{\varepsilon_n\}_{n\ge1}$ be a sequence of positive constants such that $\varepsilon_n\to 0$ as $n\to\infty$ and set

$$\mathcal{H}(\varepsilon_n)=\{\mathbf{h}\in(0,1]^p:\max\mathbf{h}\le\varepsilon_n\}.$$

Lemma 3. *Let K satisfy (K.i)–(K.iv) and the marginal densities fulfill (f.i). Whenever for a given $\mathbf{z}^d \in \mathcal{D}$, $f(\cdot|\mathbf{z}^d)$ is uniformly continuous on $\mathbb{A}^\varepsilon$ for some $\varepsilon > 0$, we have with $\{\varepsilon_n\}_{n\geq 1}$ as above*

$$\sup_{\mathbf{h}\in\mathcal{H}(\varepsilon_n)} \sup_{\mathbf{x}^c\in\mathbb{A}} \left|\mathbb{E}K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{Z}\left(\mathbf{z}^d\right)\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| \to 0. \tag{4.21}$$

Proof. Fix $\mathbf{z}^d \in \mathcal{D}$ and $\varepsilon > 0$. Choose $\mathbf{h} \in \mathcal{H}(\varepsilon_n)$, $\mathbf{x}^c \in \mathbb{A}$ and set

$$B_{\mathbf{h}}\left(\mathbf{x}^c\right) = \mathbf{x}^c + \left[-Bh_1^{1/p}, Bh_1^{1/p}\right] \times \cdots \times \left[-Bh_p^{1/p}, Bh_p^{1/p}\right].$$

Notice that when (K.i)–(K.iv) are satisfied, we get by using (2.9) that

$$\begin{aligned}
&\left|\mathbb{E}K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{Z}\left(\mathbf{z}^d\right)\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| \\
= &\left|\int_{B_{\mathbf{h}}(\mathbf{x}^c)} \Pi_{j=1}^p h_j^{-1/p} K\left(\frac{x_j^c - y_j}{h_j^{1/p}}\right)\left(f\left(\mathbf{y}|\mathbf{z}^d\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right) dy_1 \ldots dy_p\right| \\
\leq &\sup_{\mathbf{y}\in B_{\mathbf{h}}(\mathbf{x}^c)} \left|f\left(\mathbf{y}|\mathbf{z}^d\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| \int_{B_{\mathbf{h}}(\mathbf{x}^c)} \Pi_{j=1}^p h_j^{-1/p} \left|K\right|\left(\frac{x_j^c - y_j}{h_j^{1/p}}\right) dy_j \\
\leq &\sup_{\mathbf{y}\in B_{\mathbf{h}}(\mathbf{x}^c)} \left|f\left(\mathbf{y}|\mathbf{z}^d\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| (2B\kappa)^p.
\end{aligned}$$

Hence, with $\boldsymbol{\varepsilon}_n(p) = (\varepsilon_n^{1/p}, \ldots, \varepsilon_n^{1/p})$, we deduce that

$$\begin{aligned}
&\sup_{\mathbf{h}\in\mathcal{H}(\varepsilon_n)} \sup_{\mathbf{x}^c\in\mathbb{A}} \left|\mathbb{E}K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{Z}\left(\mathbf{z}^d\right)\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| \\
\leq &\sup_{\mathbf{x}^c\in\mathbb{A}} \sup_{\mathbf{y}\in B_{\boldsymbol{\varepsilon}_n(p)}(\mathbf{x}^c)} \left|f\left(\mathbf{y}|\mathbf{z}^d\right) - f\left(\mathbf{x}^c|\mathbf{z}^d\right)\right| (2B\kappa)^p,
\end{aligned}$$

and using the assumption that $f(\cdot|\mathbf{z}^d)$ is uniformly continuous on $\mathbb{A}^\varepsilon$, we get (4.21), keeping in mind that $\varepsilon_n \to 0$ as $n \to \infty$. □

Proof of Theorem 2. Notice that by the one to one correspondence given in (2.6), for any $\mathbf{x}^d \in \mathcal{D}$,

$$\begin{aligned}
\widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) &= \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}}^c\left(\mathbf{x}^c, \mathbf{X}_i^c\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right) \\
&= \frac{1}{n\left(\Pi_{j=1}^p h_j\right)^{1/p}} \sum_{i=1}^n \Pi_{j=1}^p K\left(\frac{x_j^c - X_{i,j}^c}{t_j^{1/p} h^{1/p}}\right) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{X}_i^d\right),
\end{aligned}$$

where $h = \max \mathbf{h}$. Since by the probability construction in Subsection 2.3, as a process in $(\mathbf{X}_i^c, \mathbf{X}_i^d)_{i\geq 1}$, $h \in (0, 1]$, $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}$, $\mathbf{x}^d \in \mathcal{D}$ and $g_{\mathbf{t},\mathbf{x}^c} \in \mathcal{G}_K$, recalling that $h_j = t_j h$,

$$\left\{\widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right)\right\}_{n\geq 1} \overset{\mathrm{D}}{=} \left\{\frac{\sum_{\mathbf{z}^d\in\mathcal{D}} s_{N_n(\mathbf{z}^d)}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{z}^d, h) K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d, \mathbf{z}^d\right)}{n\left(\Pi_{j=1}^p t_j\right)^{1/p} h}\right\}_{n\geq 1}$$

$$= \left\{ \frac{\widetilde{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{x}^d)}{n \left(\Pi_{j=1}^p t_j\right)^{1/p} h} \right\}_{n \geq 1}, \tag{4.22}$$

we can assume for the purpose of proving limit results that we have equality in (4.22). We see then, keeping in mind the one to one correspondence given in (2.6), that

$$\begin{aligned} &\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{h} \in \mathcal{H}_n} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Gamma}(\delta_n)} \sup_{\mathbf{x}^c \in \mathbb{A}} \left| \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) - \mathbb{E} \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) \right| \\ = &\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{h} \in \mathcal{H}_n} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Gamma}(\delta_n)} \sup_{g_{t,\mathbf{x}^c} \in \mathcal{G}_K} \left\{ \frac{\left| \widetilde{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{x}^d) - \mathbb{E} \widetilde{\Upsilon}_{n,h,\boldsymbol{\lambda}}(g_{\mathbf{t},\mathbf{x}^c}, \mathbf{x}^d) \right|}{n \left(\Pi_{j=1}^p t_j\right)^{1/p} h} \right\}, \end{aligned}$$

which by (3.13) is almost surely for some constant $C > 0$

$$\begin{aligned} \leq &\sup_{\mathbf{h} \in \mathcal{H}_n} \frac{C}{\left(\Pi_{j=1}^p t_j\right)^{1/p}} \sqrt{\frac{\log n}{nh}} = \sup_{\mathbf{h} \in \mathcal{H}_n} \frac{C}{\left(\Pi_{j=1}^p h_j\right)^{1/p}} \sqrt{\frac{h \log n}{n}} \\ = &C \sup_{\mathbf{h} \in \mathcal{H}_n} \frac{\sqrt{\max \mathbf{h}}}{\left(\Pi_{j=1}^p h_j\right)^{1/p}} \sqrt{\frac{\log n}{n}} = C \sup_{\mathbf{h} \in \mathcal{H}_n} \sqrt{\frac{\max \mathbf{h}}{\left(\Pi_{j=1}^p h_j\right)^{2/p}} \frac{\log n}{n}}. \end{aligned}$$

Now, since for each $\mathbf{h} \in \mathcal{H}_n$,

$$a_n \leq \frac{\left(\Pi_{j=1}^p h_j\right)^{2/p}}{\max \mathbf{h}} \text{ and } n a_n / \log n \to \infty,$$

we get, with probability 1,

$$\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{h} \in \mathcal{H}_n} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Gamma}(\delta_n)} \sup_{\mathbf{x}^c \in \mathbb{A}} \left| \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) - \mathbb{E} \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) \right| \to 0. \tag{4.23}$$

Now

$$\begin{aligned} &\mathbb{E} \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) = \mathbb{E} \left(K_{\mathbf{h}}^c \left(\mathbf{x}^c, \mathbf{X}^c\right) K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{X}^d\right) \right) \\ = &\sum_{\mathbf{z}^d \in \mathcal{D}} \mathbb{E} K_{\mathbf{h}}^c \left(\mathbf{x}^c, \mathbf{Z} \left(\mathbf{z}^d\right)\right) K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{z}^d\right) p \left(\mathbf{z}^d\right). \end{aligned}$$

Let $\max \boldsymbol{\lambda} = \max\{\lambda_1, \ldots, \lambda_q\}$. Notice that for each $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}$

$$(1 - \max \boldsymbol{\lambda})^q I\left\{\mathbf{z}^d = \mathbf{x}^d\right\} \leq K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{z}^d\right) \leq \max \boldsymbol{\lambda} I\left\{\mathbf{z}^d \neq \mathbf{x}^d\right\} + I\left\{\mathbf{z}^d = \mathbf{x}^d\right\}.$$

Thus, uniformly in $\mathbf{x}^d, \mathbf{z}^d \in \mathcal{D}$,

$$\max_{\mathbf{x}^d, \mathbf{z}^d \in \mathcal{D}} \left| K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{z}^d\right) - I\left\{\mathbf{z}^d = \mathbf{x}^d\right\} \right| \to 0, \text{ as } \max \boldsymbol{\lambda} \searrow 0. \tag{4.24}$$

Next, Lemma 3 implies that

$$\max_{\mathbf{z}^d \in \mathcal{D}} \sup_{\mathbf{h} \in \mathcal{H}_n} \sup_{\mathbf{x}^c \in \mathbb{A}} \left| \mathbb{E} K_{\mathbf{h}}^c \left(\mathbf{x}^c, \mathbf{Z} \left(\mathbf{z}^d\right)\right) - f\left(\mathbf{x}^c | \mathbf{z}^d\right) \right| \to 0. \tag{4.25}$$

In turn, (4.24) and (4.25) imply that

$$\max_{\mathbf{x}^d,\mathbf{z}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\mathbb{E}K_{\mathbf{h}}^c\left(\mathbf{x}^c,\mathbf{Z}\left(\mathbf{z}^d\right)\right)K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{z}^d\right)\right.$$
$$\left.-f\left(\mathbf{x}^c|\mathbf{z}^d\right)I\left\{\mathbf{z}^d=\mathbf{x}^d\right\}\right|\to 0.$$

This implies that

$$\begin{aligned}&\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\sum_{\mathbf{z}^d\in\mathcal{D}}\mathbb{E}K_{\mathbf{h}}^c\left(\mathbf{x}^c,\mathbf{Z}\left(\mathbf{z}^d\right)\right)K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{z}^d\right)p\left(\mathbf{z}^d\right)\right.\\&\qquad\left.-f\left(\mathbf{x}^c|\mathbf{x}^d\right)p\left(\mathbf{x}^d\right)\right|\\&=\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\mathbb{E}\left(K_{\mathbf{h}}^c\left(\mathbf{x}^c,\mathbf{X}^c\right)K_{\boldsymbol{\lambda}}^d\left(\mathbf{x}^d,\mathbf{X}^d\right)\right)-f\left(\mathbf{x}^c,\mathbf{x}^d\right)\right|\\&=\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\mathbb{E}\widehat{f}_n\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right)-f\left(\mathbf{x}^c,\mathbf{x}^d\right)\right|\to 0.\end{aligned}\tag{4.26}$$

Finally, (4.23) and (4.26) imply that, with probability 1,

$$\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{f}_n\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right)-f\left(\mathbf{x}^c,\mathbf{x}^d\right)\right|\to 0.$$ □

Remark 3. When the components of $\mathbf{X}^d$ have a natural ordering, for example in the case $x_k^d, z_k^d \in \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, for $k = 1, \ldots, q$, Wang and van Ryzin [20] suggested the following kernel

$$\begin{aligned}&K_{\boldsymbol{\lambda}}^{d,o}\left(\mathbf{x}^d,\mathbf{z}^d\right)\\:=\;&\Pi_{k=1}^q\left\{\left(\frac{1-\lambda_k}{2}\right)\lambda_k^{\left|x_k^d-z_k^d\right|}I\left(\left|x_k^d-z_k^d\right|\geq 1\right)+\left(1-\lambda_k\right)I\left(x_k^d=z_k^d\right)\right\},\end{aligned}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q) \in [0, 1]^q =: \boldsymbol{\Gamma}^o$. Here we take $\mathcal{D} = \mathbb{Z}^q$. The corresponding smooth estimator is

$$p_n^o\left(\mathbf{x}^d,\boldsymbol{\lambda}\right) := n^{-1}\sum_{i=1}^n K_{\boldsymbol{\lambda}}^{d,o}\left(\mathbf{x}^d,\mathbf{X}_i^d\right).$$

Mean squared error comparisons with the maximum likelihood estimator (frequency estimator) $\tilde{p}_n(\mathbf{x}^d) = n^{-1}N_n(\mathbf{x}^d)$ based on large-sample theory and small-sample simulations were obtained by the authors. Typically, $p_n^o(\mathbf{x}^d, \boldsymbol{\lambda})$ yielded significantly smaller mean squared error in these comparisons.

Notice that for each $\boldsymbol{\lambda} \in \boldsymbol{\Gamma}^o$ we have

$$(1-\max\boldsymbol{\lambda})^q I\left\{\mathbf{X}^d=\mathbf{x}^d\right\}\leq K_{\boldsymbol{\lambda}}^{d,o}\left(\mathbf{x}^d,\mathbf{X}^d\right)\leq\max\boldsymbol{\lambda}+I\left\{\mathbf{X}^d=\mathbf{x}^d\right\},$$

so that (2.3) again holds with $\boldsymbol{\Gamma}$ and $K_{\boldsymbol{\lambda}}^d$ replaced by $\boldsymbol{\Gamma}^o$ and $K_{\boldsymbol{\lambda}}^{d,o}$ respectively. Now, consider the estimator

$$\widehat{f}_n^o\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right) := \frac{1}{n}\sum_{i=1}^n K_{\mathbf{h}}^c\left(\mathbf{x}^c,\mathbf{X}_i^c\right)K_{\boldsymbol{\lambda}}^{d,o}\left(\mathbf{x}^d,\mathbf{X}_i^d\right),$$

for $\mathbf{x}^c \in \mathbb{R}^p$ and $\mathbf{x}^d \in \mathcal{D}$. Theorems 1 and 2 then again hold with $\mathbf{\Gamma}$, $K^d_{\boldsymbol{\lambda}}$ and $\mathcal{D}$ replaced by $\mathbf{\Gamma}^o$, $K^{d,o}_{\boldsymbol{\lambda}}$ and $\mathcal{D}^o$ respectively, where $\mathcal{D}^o$ is a finite subset of $\mathcal{D}$. This follows from the inequality above and an exact repetition of the steps in the proofs above.

In practice, it is likely that some of the discrete variables will have natural orderings while the others will be unordered. Following Section 2.5 of Racine [17], let $\tilde{\mathbf{X}}^d$ denote a $q_1 \times 1$ vector (say the first q_1 components of $\mathbf{X}^d$) of discrete variables that do not have a natural ordering ($1 \leq q_1 \leq q$), and let $\bar{\mathbf{X}}^d$ denote the remaining discrete variables that do have a natural ordering. In this case, we can construct a product kernel of the form

$$K^c_{\mathbf{h}}\left(\mathbf{x}^c, \mathbf{X}^c\right) K^d_{\boldsymbol{\lambda}}\left(\tilde{\mathbf{x}}^d, \tilde{\mathbf{X}}^d\right) K^{d,o}_{\boldsymbol{\lambda}}\left(\bar{\mathbf{x}}^d, \bar{\mathbf{X}}^d\right),$$

where $\mathbf{x}^c = (x^c_1, \ldots, x^c_p)$, $\tilde{\mathbf{x}}^d = (x^d_1, \ldots, x^d_{q_1})$ and $\bar{\mathbf{x}}^d = (x^d_{q_1+1}, \ldots, x^d_q)$. Then the conclusions of Theorems 1 and 2 remain unchanged using this kernel. The proofs of this claim are identical to those above.

5. Applications

5.1. Application to Li and Racine estimator

In this Subsection we shall apply Theorem 3.1 of Li and Racine [11] to obtain a uniform in bandwidth consistency result for their estimator. They treat the density estimator of $f(\mathbf{x}^c, \mathbf{x}^d)$ in the case $h_i = h$ for $i = 1, \ldots, p$ and $\lambda_j = \lambda$ for $j = 1, \ldots, q$. Also their h_i is our $h_i^{1/p}$. So in our notation

$$\widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) = \frac{1}{n}\sum_{i=1}^n K^c_{\mathbf{h}}\left(\mathbf{x}^c, \mathbf{X}^c_i\right) K^d_{\boldsymbol{\lambda}}\left(\mathbf{x}^d, \mathbf{X}^d_i\right),$$

where for $\mathbf{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$

$$K^c_{\mathbf{h}}\left(\mathbf{x}^c, \mathbf{z}^c\right) = \frac{1}{h}\Pi_{j=1}^p K\left(\frac{x^c_j - z^c_j}{h^{1/p}}\right)$$

and for $\mathbf{z}^d = (z^d_1, \ldots, z^d_q) \in \mathbb{R}^q$

$$K^d_{\boldsymbol{\lambda}}\left(\mathbf{x}^d, \mathbf{z}^d\right) = \Pi_{k=1}^q \left(\frac{\lambda}{r_k - 1}\right)^{I\left(z^d_k \neq x^d_k\right)} (1-\lambda)^{I\left(z^d_k = x^d_k\right)}.$$

Their version of $K^d_{\boldsymbol{\lambda}}(\mathbf{x}^d, \mathbf{X}^d_i)$ is bit different than ours. However, this does not affect the conclusion of their Theorem 3.1. See their comment on the general multivariate discrete case following the statement of Theorem 3.1. Keeping in mind that their h_i is our $h_i^{1/p}$, if one assumes in addition to the conditions of our Theorem 2, those of their Theorem 3.1 one gets for their cross-validation estimators $\widehat{h}$ and $\widehat{\lambda}$ of the smoothing parameters h and λ that

$$\left(\left(\widehat{h}\right)^{1/p} - (h_0)^{1/p}\right) / (h_0)^{1/p} = O_p\left(n^{-\alpha/(4+p)}\right) \text{ and } \widehat{\lambda} - \lambda_0 = O_p\left(n^{-\beta/(4+p)}\right),$$

where for appropriate $c_1 > 0$ and $c_2 > 0$

$$(h_0)^{1/p} = c_1 n^{-1/(4+p)} \text{ and } \lambda_0 = c_2 n^{-2/(4+p)},$$

and $\alpha = \min\{2, p/2\}$ and $\beta = \min\{1/2, 4/(4+p)\}$. This implies that $\widehat{\lambda} = o_p(1)$ and for appropriate $0 < a < b < \infty$, with probability converging to 1,

$$\widehat{h} \in \left[a n^{-p/(4+p)}, b n^{-p/(4+p)} \right].$$

Thus, we can apply Theorem 2 to conclude that

$$P\left\{ \max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{x}^c \in \mathbb{A}} \left| \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, \widehat{\mathbf{h}}, \widehat{\boldsymbol{\lambda}} \right) - f\left(\mathbf{x}^c, \mathbf{x}^d \right) \right| \to 0 \right\} \to 1,$$

where$(\widehat{\mathbf{h}}, \widehat{\boldsymbol{\lambda}}) \in \mathbb{R}^p \times \mathbb{R}^q$ is defined as

$$\widehat{\mathbf{h}} = \left(\widehat{h}, \ldots, \widehat{h} \right) \text{ and } \widehat{\boldsymbol{\lambda}} = \left(\widehat{\lambda}, \ldots, \widehat{\lambda} \right).$$

5.2. Application to Hall, Racine and Li estimator

The Hall, Racine and Li [9] setup is a follows. Assume that for $p \geq 1$, $q \geq 1$,

$$\mathbf{X} = \left(\mathbf{X}^c, \mathbf{X}^d \right) = \left(\left(X_1^c, \ldots, X_p^c \right), \left(X_1^d, \ldots, X_q^d \right) \right) \in \mathbb{R}^p \times \mathbb{R}^q,$$

is as in the Li and Racine [11] setup. Introduce an additional continuous real valued random variable Y and assume that $(\mathbf{X}, Y) = (\mathbf{X}^c, \mathbf{X}^d, Y)$ has joint density $f(\mathbf{x}, y) = f(\mathbf{x}^c, \mathbf{x}^d, y)$ with marginal density $m(\mathbf{x}) = f(\mathbf{x})$. They study the kernel estimator of the conditional density of Y given $\mathbf{X} = \mathbf{x}$, i.e.,

$$g\left(y | \mathbf{x} \right) = f\left(\mathbf{x}, y \right) / m\left(\mathbf{x} \right),$$

defined by

$$\widehat{g}_n \left(y | \mathbf{x}, \mathbf{h}, \boldsymbol{\lambda} \right) = \widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, y, \mathbf{h}, \boldsymbol{\lambda} \right) / \widehat{m}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda} \right),$$

where

$$\widehat{f}_n \left(\mathbf{x}^c, \mathbf{x}^d, y, \mathbf{h}, \lambda \right) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{h}}^c \left(\mathbf{x}^c, \mathbf{X}_i^c \right) K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{X}_i^d \right) L_{h_0} \left(y, Y_j \right),$$

and

$$\widehat{m}_n \left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda} \right) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{h}}^c \left(\mathbf{x}^c, \mathbf{X}_i^c \right) K_{\boldsymbol{\lambda}}^d \left(\mathbf{x}^d, \mathbf{X}_i^d \right).$$

In order to apply our Theorem 2 we assume that

$$\mathbf{h} = (h_0, h_1, \ldots, h_p) \in (0, 1]^{p+1},$$

for $\mathbf{x}^c$and $\mathbf{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$,

$$K_{\mathbf{h}}^c(\mathbf{x}^c, \mathbf{z}) = \Pi_{j=1}^p h_j^{-1/(p+1)} K\left(\frac{x_j^c - z_j}{h_j^{1/(p+1)}}\right)$$

and for y and $z_0 \in \mathbb{R}$,

$$L_{h_0}(y, z_0) = h_0^{-1/(p+1)} L\left(\frac{y - z_0}{h_0^{1/(p+1)}}\right),$$

with L being a kernel with the same properties as K. Notice that the Hall, Racine and Li [9] h_j are $h_j^{1/(p+1)}$ in our notation. If one assumes in addition to the conditions of our Theorem 2, those of their Theorem 2 one gets for their cross-validation estimators $\widehat{\mathbf{h}}$ and $\widehat{\boldsymbol{\lambda}}$ of the smoothing vector $\mathbf{h}$ and $\boldsymbol{\lambda}$ that

$$P\left\{n^{1/(p+5)}\left(\widehat{h}_i\right)^{1/(p+1)} \to a_i\right\} = 1 \text{ and } P\left\{n^{2/(p+5)}\widehat{\lambda}_j \to b_j\right\} = 1,$$

for appropriate $a_i > 0$, $i = 0, \ldots, p$, and $b_j > 0$, $j = 1, \ldots, q$, whenever all of the variables $(\mathbf{X}^c, \mathbf{X}^d)$ are *relevant* in the sense of Hall, Racine and Li [9]. Therefore we can apply Theorem 2 to get that

$$P\left\{\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{(\mathbf{x}^c, y) \in \mathbb{A}} \left|\widehat{g}_n\left(y|\mathbf{x}^c, \mathbf{x}^d, \widehat{\mathbf{h}}, \widehat{\boldsymbol{\lambda}}\right) - g\left(y|\mathbf{x}^c, \mathbf{x}^d\right)\right| \to 0\right\} \to 1, \tag{5.27}$$

where it is assumed that $m(\mathbf{x}) = f(\mathbf{x})$ satisfies (f.ii) for the $\mathbb{A}$ in (5.27).

5.3. Further applications to estimating conditional densities

An obvious estimator of

$$f\left(\mathbf{x}^c|\mathbf{x}^d\right) = f\left(\mathbf{x}^c, \mathbf{x}^d\right)/p\left(\mathbf{x}^d\right)$$

is

$$\widehat{f}_n\left(\mathbf{x}^c|\mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) := \widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right)/\widehat{p}_n\left(\mathbf{x}^d, \boldsymbol{\lambda}\right),$$

which under the assumptions of Theorem 2 is readily shown to satisfy, with probability 1,

$$\max_{\mathbf{x}^d \in \mathcal{D}} \sup_{\mathbf{h} \in \mathcal{H}_n} \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Gamma}(\delta_n)} \sup_{\mathbf{x}^c \in \mathbb{A}} \left|\widehat{f}_n\left(\mathbf{x}^c|\mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right) - f\left(\mathbf{x}^c|\mathbf{x}^d\right)\right| \to 0. \tag{5.28}$$

Observe that we can estimate the density function

$$f\left(\mathbf{x}^c\right) = f\left(x_1^c, \ldots, x_p^c\right)$$

of $\mathbf{X}^c = (X_1^c, \ldots, X_p^c)$ using the estimator

$$\widehat{f}_n\left(\mathbf{x}^c, \mathbf{h}, \boldsymbol{\lambda}\right) := \sum_{\mathbf{x}^d \in \mathcal{D}} \widehat{f}_n\left(\mathbf{x}^c, \mathbf{x}^d, \mathbf{h}, \boldsymbol{\lambda}\right).$$

Clearly, under the assumptions of Theorem 2, we conclude that, with probability 1,

$$\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{f}_n\left(\mathbf{x}^c,\mathbf{h},\boldsymbol{\lambda}\right)-f\left(\mathbf{x}^c\right)\right|\to 0. \tag{5.29}$$

Further, we can estimate

$$p\left(\mathbf{x}^d|\mathbf{x}^c\right)=f\left(\mathbf{x}^d,\mathbf{x}^c\right)/f\left(\mathbf{x}^c\right),$$

the conditional probability that $\mathbf{X}^d=\mathbf{x}^d$ given $\mathbf{X}^c=\mathbf{x}^c$, by

$$\widehat{p}_n\left(\mathbf{x}^d|\mathbf{x}^c,\mathbf{h},\boldsymbol{\lambda}\right):=\widehat{f}_n\left(\mathbf{x}^c,\mathbf{x}^d,\mathbf{h},\boldsymbol{\lambda}\right)/\widehat{f}_n\left(\mathbf{x}^c,\mathbf{h},\boldsymbol{\lambda}\right).$$

If we also assume (f.ii) we get, with probability 1, that

$$\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{h}\in\mathcal{H}_n}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Gamma}(\delta_n)}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{p}_n\left(\mathbf{x}^d|\mathbf{x}^c,\mathbf{h},\boldsymbol{\lambda}\right)-p\left(\mathbf{x}^d|\mathbf{x}^c\right)\right|\to 0. \tag{5.30}$$

Moreover, using the Li and Racine [11] cross-validation estimators $(\widehat{\mathbf{h}},\widehat{\boldsymbol{\lambda}})$ of $(\mathbf{h},\boldsymbol{\lambda})$ mentioned in Subsection 5.2, we get under appropriate regularity conditions

$$P\left\{\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{p}_n\left(\mathbf{x}^d|\mathbf{x}^c,\widehat{\mathbf{h}},\widehat{\boldsymbol{\lambda}}\right)-p\left(\mathbf{x}^d|\mathbf{x}^c\right)\right|\to 0\right\}\to 1$$

and

$$P\left\{\max_{\mathbf{x}^d\in\mathcal{D}}\sup_{\mathbf{x}^c\in\mathbb{A}}\left|\widehat{f}_n\left(\mathbf{x}^c|\mathbf{x}^d,\widehat{\mathbf{h}},\widehat{\boldsymbol{\lambda}}\right)-f\left(\mathbf{x}^c|\mathbf{x}^d\right)\right|\to 0\right\}\to 1.$$

Remark 4. The applications in Subsections 5.1–5.3 can also be extended to cover the case of ordered discrete variables by applying, for example, the kernel $K_{\boldsymbol{\lambda}}^{d,o}(\mathbf{x}^d,\mathbf{z}^d)$. The proofs are slightly more involved and are therefore omitted.

Kernel regression function estimation versions of the results above, using Einmahl and Mason [8] and Mason [12] as a guide, follow in a routine manner from our methods.

6. Material from Mason and Swanepoel (2011) paper

6.1. The general setup

Mason and Swanepoel [13] introduced the following general setup for studying kernel-type estimators. Let $X,X_1,X_2,\ldots$ be i.i.d. random variables on a probability space $(\Omega,\mathcal{A},P)$ with values in a measure space $(S,\mathcal{S})$. (Typically S will be a Fréchet space.) Let $\mathcal{G}$ denote a class of measurable real valued functions of $(x,h)\in S\times(0,1]$

$$g:(x,h)\mapsto g(x,h). \tag{6.31}$$

From this class we form the class of measurable real valued functions $\mathcal{G}_0$ of $x\in S$ defined as

$$\mathcal{G}_0=\{x\mapsto g(x,h):g\in\mathcal{G},\ 0<h\leq 1\}. \tag{6.32}$$

It will be necessary in our presentation to distinguish between $\mathcal{G}$ and $\mathcal{G}_0$. Always keep in mind that functions $g \in \mathcal{G}$ are defined on $S \times (0,1]$ and functions $g_0 \in \mathcal{G}_0$ are defined on S. Introduce the class of estimators

$$\hat{\vartheta}_{n,h}(g) := \frac{1}{nh} \sum_{i=1}^{n} g(X_i, h), \quad g \in \mathcal{G} \text{ and } 0 < h < 1. \tag{6.33}$$

6.2. The underlying assumptions and basic definitions

Let X be a random variable from a probability space $(\Omega, \mathcal{A}, P)$ to a measure space $(S, \mathcal{S})$. In the sequel, $||\cdot||_\infty$ denotes the supremum norm on the space of bounded real valued measurable functions on S. To formulate our basic theoretical results we shall need the following class of functions. Let $\mathcal{G}$ denote the class of measurable real valued functions g of $(u,h) \in S \times (0,1]$ introduced in our general setup (6.31) and recall the class of functions $\mathcal{G}_0$ on S defined in (6.32). We shall assume the following conditions on $\mathcal{G}$ and $\mathcal{G}_0$:

(G.i) $\sup_{g\in\mathcal{G}} \sup_{0<h\le 1} \|g(\cdot,h)\|_\infty =: \eta < \infty$,
(G.ii) $\sup_{g\in\mathcal{G}} \mathbb{E}g^2(X,h) \le Dh$, for some $D > 0$ and all $0 < h \le 1$,
(G.iii) $\mathcal{G}_0$ is a pointwise measurable class,
(G.iv) $\mathcal{N}(\epsilon, \mathcal{G}_0) \le C\epsilon^{-\nu}, 0 < \epsilon < 1$, for some $C > 0$ and $\nu \ge 1$.

Note that (G.iii) is a measurability condition that we assume in order to avoid using outer probability measures in all of our statements. A *pointwise measurable class* $\mathcal{G}_0$ has a countable subclass $\mathcal{G}_c$ such that we can find for any function $g \in \mathcal{G}_0$ a sequence of functions $\{g_m, m \ge 1\}$ in $\mathcal{G}_c$ for which $\lim_{m\to\infty} g_m(x) = g(x)$ for all $x \in S$. See Example 2.3.4 in [19].

Condition (G.iv) is a so–called *uniform entropy condition.* As is usual, we define the covering numbers

$$\mathcal{N}(\epsilon, \mathcal{G}_0) = \sup_Q \mathcal{N}\left(\epsilon\sqrt{Q(G^2)}, \mathcal{G}_0, d_Q\right), \tag{6.34}$$

where G is an envelope function for $\mathcal{G}_0$, and where the supremum is taken over all probability measures Q on $(S, \mathcal{S})$ with $Q(G^2) < \infty$. We shall now define the notation in (6.34). By an *envelope function* G for $\mathcal{G}_0$ we mean a measurable function $G : S \to [0, \infty]$, such that

$$G(u) \ge \sup_{g_0\in\mathcal{G}_0} |g_0(u)|, \quad u \in S.$$

Note that by the definition of the class $\mathcal{G}_0$,

$$\sup_{g_0\in\mathcal{G}_0} |g_0(u)| = \sup\{|g(u,h)| : g \in \mathcal{G}, 0 < h \le 1\}.$$

The d_Q in (6.34) is the $L_2(Q)$–metric and for any $\gamma > 0$, $\mathcal{N}(\gamma, \mathcal{G}_0, d_Q)$ is the minimal number of d_Q–balls with radius γ which is needed to cover the entire function class $\mathcal{G}_0$.

We use η as our (constant) envelope function, when condition (G.i) holds. (In this case $\mathbb{E}G^2(X) < \infty$ is trivially satisfied.)

For future reference, recall that we say that a class $\mathcal{F}$ is of *VC–type* for the envelope function F, if $\mathcal{N}(\epsilon, \mathcal{F}) \leq C\epsilon^{-\nu}$, $0 < \epsilon < 1$, for some constants $C > 0, \nu \geq 1$. (Here $\mathcal{N}(\epsilon, \mathcal{F})$ is defined as in (6.34) with $\mathcal{F}$ and F replacing $\mathcal{G}_0$ and G, respectively.) This condition is automatically fulfilled if the class is a *VC subgraph class* (see Theorem 2.6.7 on page 141 of [19], where we refer the reader for a definition of a *VC subgraph class*).

6.3. A uniform in bandwidth result

We shall need the following special case of the Theorem in Mason and Swanepoel [13]. Note that when we apply this result, we should keep in mind that in condition (F.ii) given there, $\mathcal{G}$ should be $\mathcal{G}_\gamma$.

Theorem 3 (General Theorem (Mason and Swanepoel [13]))**.** *Suppose that $\mathcal{G}$ is a class of functions that satisfies all of the conditions in (G.i)–(G.iv). Then we have for any choice of $c > 0$ and $0 < b_0 < 1$ that, with probability 1,*

$$\limsup_{n\to\infty} \sup_{c_n \leq h \leq b_0} \sup_{g\in\mathcal{G}} \frac{\sqrt{nh}|\hat{\vartheta}_{n,h}(g) - \mathbb{E}\hat{\vartheta}_{n,h}(g)|}{\sqrt{|\log h| \vee \log\log n}} = A(c), \tag{6.35}$$

where $c_n = \frac{c\log n}{n}$, $A(c)$ is a finite constant depending on c and the constants in (G.i), (G.ii) and (G.iv).

For an even more general uniform in bandwidth result see Theorem 4.1 of Mason [12].

7. Appendix: Pointwise measurability

We say that a class $\mathcal{G}_0$ of measurable functions $g : S \to \mathbb{R}$ is pointwise measurable if there exists a countable subclass $\mathcal{G}_c \subseteq \mathcal{G}_0$, so that for any function g in $\mathcal{G}_0$, we can find a sequence of functions $g_n \in \mathcal{G}_c, m \geq 1$ for which $g_m(x) \to g(x), x \in S$.

Example. Consider a real valued right–continuous function $K : \mathbb{R} \to \mathbb{R}$, and define the class of functions

$$\mathcal{F}^K := \{x \mapsto K(\gamma x + \rho) : \gamma > 0, \rho \in \mathbb{R}\}. \tag{7.36}$$

Then this class is always pointwise measurable. Let $\mathbb{Q}$ denote the rationals. The subclass that will do the job here is

$$\mathcal{F}^K_c := \{x \mapsto K(\gamma x + \rho) : \gamma > 0, \quad \gamma, \rho \in \mathbb{Q}\}.$$

Proof. We claim that $\mathcal{F}^K$ is a pointwise measurable class. To see this choose any $g(u) = K(\gamma u + \rho) \in \mathcal{F}^K$, $u \in \mathbb{R}$ and set for $m \geq 1$, $g_m(u) = K(\gamma_m u + \rho_m)$,

$u \in \mathbb{R}$, where $\gamma_m = \frac{1}{m^2}\lfloor m^2\gamma \rfloor + \frac{1}{m^2}$ and $\rho_m = \frac{1}{m}\lfloor m\rho \rfloor + \frac{2}{m}$, with $\lfloor x \rfloor$ denoting the integer part of x. With $\varepsilon_m = \gamma_m - \gamma$ and $\delta_m = \rho_m - \rho$, we can write

$$\Delta_m := \gamma_m u + \rho_m - (\gamma u + \rho) = \varepsilon_m u + \delta_m.$$

Now since $\frac{2}{m^2} \geq \varepsilon_m > 0$ and $\frac{3}{m} \geq \delta_m > \frac{1}{m}$, we get for all large enough m that

$$\Delta_m = \delta_m \left(1 + o(1)\right) > 0.$$

Thus since $\gamma_m u + \rho_m \to \gamma u + \rho$ and K is right continuous at $\gamma u + \rho$, we see that $g_m(u) \to g(u)$ as $m \to \infty$. □

This proof is taken from that of Lemma A.1 of Deheuvels and Mason [4] with a couple of misprints fixed, and for the benefit of the reader is repeated here.

Trivially we get that if $K_1, \ldots, K_p$ are right continuous functions on $\mathbb{R}$ and φ is a fixed measurable real-valued function on $\mathbb{R}$, then the class of functions

$$\left\{(x_1, \ldots, x_p, y) \longmapsto \Pi_{j=1}^{p} K_j\left(\gamma_j x_j + \rho_j\right) :, \gamma_j > 0, \rho_j \in \mathbb{R},\, 1 \leq j \leq p\right\},$$

is pointwise measurable.

Acknowledgements

The authors thank the editor, associate editor and referee for their valuable remarks and suggestions.

References

[1] Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420. MR0443222

[2] Brown, P. J. and Rundell, P. W. K. (1985). Kernel estimates for categorical data. *Technometrics* **27** 293–299. MR0797568

[3] Burman, P. (1987). Smoothing sparse contingency tables. *Sankhyā, Ser. A* **49** 24–36. MR0917903

[4] Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.* **7** 225–277. MR2111291

[5] Dony, J., Einmahl, U. and Mason, D. M. (2006). Uniform in bandwidth consistency of local polynomial regression function estimators. *Aust. J. Stat.* **35** 105–120.

[6] Einmahl, U. and Mason, D. M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Theory Rel.* **107** 283–311. MR1440134

[7] Einmahl, U. and Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theor. Probab.* **13** 1–37. MR1744994

[8] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Stat.* **33** 1380–1403. MR2195639

[9] Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Am. Stat. Assoc.* **99** 1015–1026. MR2109491

[10] Hall, P. and Titterington, D. M. (1987). On smoothing sparse multinomial data. *Aust. J. Stat.* **29** 19–37. MR0899373

[11] Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *J. Multivariate Anal.* **86** 266–292. MR1997765

[12] Mason, D. M. (2012). Proving consistency of non-standard kernel estimators. *Stat. Inference Stoch. Process.* **15** 151–176. MR2928244

[13] Mason, D. M. and Swanepoel, J. W. H. (2011). A general result on the uniform in bandwidth consistency of kernel–type function estimators. *Test* **20** 72–94. MR2806311

[14] Nolan, D. and Marron, J. S. (1989). Uniform consistency of automatic and location-adaptive delta-sequence estimators. *Probab. Theory Rel.* **80** 619–632. MR0980690

[15] Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Stat.* **15** 780–799. MR0888439

[16] Ouyang, D., Li, Q. and Racine, J. (2006). Cross-validation and the estimation of probability distributions with categorical data. *J. Nonparametric Stat.* **18** 69–100. MR2214066

[17] Racine, J. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics* **3** 1–88.

[18] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York. MR1391963

[19] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics. Springer Series in Statistics*. Springer-Verlag, New York. MR1385671

[20] Wang, M. C. and van Ryzin, J. A. (1981). A class of smooth estimators for discrete distributions. *Biometrika* **68** 301–309. MR0614967