Far East Journal of Theoretical Statistics

Published Online: July 2014
Available online at http://pphmj.com/journals/fjts.htm
Volume 47, Number 1, 2014, Pages 41-86

UNIFORMITY INSPIRED BANDWIDTH SELECTORS FOR THE KERNEL DISTRIBUTION FUNCTION ESTIMATOR

É. Youndjé

Laboratoire Raphaël Salem
Université de Rouen
UMR 6085 CNRS, France
e-mail: elie.youndje@univ-rouen.fr

Abstract

It is a well-known result in probability that if X is a random variable with a continuous distribution F, then $F(X)$ is uniformly distributed over the interval $]0, 1[$. This idea helps us to build new score functions to choose the regularization parameter of kernel distribution estimator. Two classes of selectors are considered: the indirect L^2 class and the L^1 class. Although all the methods work well, those in the L^1 class seem more appealing because they appear to be more robust in the simulations. A small scale comparison study (by simulations) between the popular method of Bowman et al. [3] and the indirect L^2 methods is carried out in the paper. An optimality result of the indirect L^2 methods is given in the article.

1. Introduction

Since the earlier 1960s, people have been interested in smooth estimators

Received: March 27, 2014; Accepted: June 1, 2014
2010 Mathematics Subject Classification: 62G05.
Keywords and phrases: kernel estimation, bandwidth selection, distribution function, order statistics.

of a distribution function. The estimator mostly considered is the kernel estimator. Let X be a random variable with a density function f and a distribution function F. Let $X_1, ..., X_n$ be identically and independently distributed from the distribution function F. The kernel estimator of F was introduced by Nadaraya [10] and is defined by

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right),$$

where H is the distribution function of a nonnegative kernel function K, i.e.,

$$H(x) = \int_{-\infty}^{x} K(t)dt,$$

and $h = h_n > 0,$ is the smoothing parameter. Convergence properties of the estimator F_h can be found in Nadaraya [10], Reiss [11] and Lejeune and Sarda [6]. It appears from these theoretical results that the most influential parameter is the bandwidth. In the literature, there are two cross-validation methods in Sarda [12] and Bowman et al. [3], and the plug-in method proposed by Altman and Léger [1] for choosing the bandwidth h.

The methods introduced in this paper are based on the fact that $F(X_1), ..., F(X_n)$ is an i.i.d. sample of the uniform distribution over $]0, 1[$. The fundamental ideas and results of our methods are presented in Section 2. In Section 3, the ingredients used to build the bandwidths are displayed. The general methods are presented in Section 4. The particular approaches based on the indirect L^2 distance are outlined in Section 5. The approaches derived from the L^1 distance are studied by simulations in Section 6. The remainder of the paper consists of proofs.

2. Basic Idea and Fundamental Result of Our Methods

In what follows the conditions on the kernel function, the distribution function and the bandwidth are:

(A.1) K is nonnegative and $\int K(u)du = 1$;

(A.2) K is compactly supported on $[-1, 1]$;

(A.3) K is symmetric and $0 < \int u^2 K(u)du < +\infty$;

(A.4) F is twice continuously differentiable, and $F' = f$, $F'' = f'$ are bounded;

(A.5) $\lim_{n\to+\infty} h_n = 0$.

The main idea behind our methods is to choose $h = h_n$ such that

$$F_h(X_1), ..., F_h(X_n)$$

behaves like an i.i.d. sample from the uniform distribution function defined by

$$V(x) = x\mathbb{1}_{]0,1[}(x) + \mathbb{1}_{[1,+\infty[}(x), \tag{1}$$

where $\mathbb{1}_A$ is the indicator function of the set A. "Theoretical justification" of the approaches might be based on the fact that

$$\sup_{1\le i\le n} |F_h(X_i) - F(X_i)| \xrightarrow{a.e.} 0$$

which in turn is based on the uniform convergence of F_h namely:

Proposition 1 (Yamato [14] and Chacón and Rodríguez-Casal [4])**.** *Under the assumptions* (A.1), (A.4) *and* (A.5), *we have*:

$$\sup_{x\in\mathbb{R}} |F_h(x) - F(x)| \xrightarrow{a.s.} 0.$$

As in the cross-validation approach, it is helpful to introduce the following leave-one-out estimators

$$F_h^{-i}(x) = \frac{1}{n-1}\sum_{j\neq i} H\left(\frac{x - X_j}{h}\right),$$

$$F_h^{-(i)}(x) = \frac{1}{n-1}\sum_{j\neq i} H\left(\frac{x - X_{(j)}}{h}\right)$$

with $X_{(1)}, ..., X_{(n)}$ being the order statistics associated with $X_1, ..., X_n$. For computational purposes, it is important to note that

$$F_h^{-(1)}(X_{(1)}), ..., F_h^{-(n)}(X_{(n)})$$

are the order statistics of

$$F_h^{-1}(X_1), ..., F_h^{-n}(X_n),$$

which means that

$$F_h^{-(1)}(X_{(1)}) \le F_h^{-(2)}(X_{(2)}) \le \cdots \le F_h^{-(n)}(X_{(n)}).$$

The validity of this remark is based on the following argument; for $k \in \{1, ..., n-1\}$, we have

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_{(i)}}{h}\right),$$

and since F_h is nondecreasing, it is true that

$$F_h(X_{(k)}) \le F_h(X_{(k+1)})$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{X_{(k)} - X_{(i)}}{h}\right) \le \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{X_{(k+1)} - X_{(i)}}{h}\right)$$

$$\Leftrightarrow \sum_{\substack{i=1 \\ i\neq k}}^{n} H\left(\frac{X_{(k)} - X_{(i)}}{h}\right) + H(0) \le \sum_{\substack{i=1 \\ i\neq k+1}}^{n} H\left(\frac{X_{(k+1)} - X_{(i)}}{h}\right) + H(0)$$

$$\Rightarrow F_h^{-(k)}(X_{(k)}) \le F_h^{-(k+1)}(X_{(k+1)}).$$

Let

$$V_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{F_h^{-i}(X_i)\le x\}} \tag{2}$$

be the "noisy" empirical distribution based on

$$F_h^{-1}(X_1), ..., F_h^{-n}(X_n).$$

The cornerstone of this paper is that V_h is a consistent estimator of V (see equation (1)), this fact is stated in the following theorem whose proof is the Appendix.

Theorem 1. *Under the assumptions* (A.1), (A.4) *and* (A.5), *we have*:

$$\sup_{x\in\mathbb{R}} |V_h(x) - V(x)| \overset{a.s.}{\to} 0.$$

3. The Ingredients Used in Our Selection Criteria

3.1. The uniform samples and their role

The uniform samples are the most important tools of our approaches. Their role stems from the arguments contain in the remainder of this paragraph. Let

$$\hat{V}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F(X_i)\le x\}} \tag{3}$$

be the empirical distribution based on $F(X_1), ..., F(X_n)$. Let d be any reasonable *distance* between distribution functions. In a practical situation (finite distance), it would be desirable to use the value of h minimizing $d(V_h, \hat{V}_n)$ as bandwidth for the estimator F_h. Unhappily, the sample $F(X_1), ..., F(X_n)$ is not available. At the same time, it is well-known that the distribution of $F(X)$ is uniform over $]0, 1[$. Let $S = \{U_1, ..., U_n\}$ with $U_1, ..., U_n$ an i.i.d. sample from the uniform distribution. Set

$$V_n^S(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i\le x\}}, \tag{4}$$

since $\hat{V}_n(x)$ and $V_n^S(x)$ have the same theoretical properties, the basic idea

of the methods developed in this paper consists (loosely speaking) in using the minimizer of $d(V_h, V_n^S)$ (with respect to h for a uniform sample S) as an approximation to the minimizer of $d(V_h, \hat{V}_n)$. The difficulty of this approach will be how to choose the sample S. We are not going to build S element by element, we will go the other way around. Roughly speaking, we will

– start with N independent samples $S_1, ..., S_N$, of size n of i.i.d. uniform observations;

– use the lower and upper bounds on the bandwidth (presented in the next section) to select the ones, which will help to build the final bandwidth for each method.

3.2. Lower and upper bounds on the smoothing parameter

Let d be a distance between distribution functions, S be an i.i.d. sample of size n drawn from the uniform distribution and V_n^S the empirical distribution built with the aid of S. As mentioned earlier, the core of our methodology consists in using the minimizer of $d(V_h, V_n^S)$ as bandwidth. Experimentally, we have found that the bandwidths obtained from *many* uniform samples are very small, and a *few* uniform samples lead to very large bandwidths. The aim of this section is to design a lower and an upper bound for the smoothing parameter of the estimator F_h. These bounds will help us to select the uniform samples in the process of building the bandwidth of F_h. In this process, we will retain only the samples whose bandwidth are within these bounds.

It is a well-known fact in nonparametric estimation that small bandwidths lead to highly variable estimates, this is not desirable, because our goal is to obtain smooth estimates. In fact, it is known (see, for instance, Bowman et al. [3]) that

$$\lim_{h \to 0} F_h(x) = F_n(x)$$

with

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}},$$

i.e., the empirical distribution of F, which of course is discontinuous. To push this argument further if we set

$$\delta_{i-1} = \min(X_{(i+1)} - X_{(i)},\ X_{(i)} - X_{(i-1)}), \quad i = 2, ..., n-1,$$

then (because K is compactly supported on $[-1, 1]$) we have the implication:

$$h < \delta_{i-1} \Rightarrow F_h^{-(i)}(X_{(i)}) = \frac{i-1}{n-1}, \quad i = 2, ..., n-1. \tag{5}$$

Let

$$\delta_0 = \min_{1 \le i \le n-2} \delta_i.$$

It can be shown that (using the implication (5))

$$h < \delta_0 \Rightarrow F_h^{-(i)}(X_{(i)}) = \frac{i-1}{n-1}, \quad i = 1, ..., n.$$

Hence, when $h < \delta_0$, F_h is not very different from the empirical distribution. The lower bound for the bandwidth of our estimator (F_h) is built with the aid of the sequence $(\delta_i,\ i = 1, ..., n-2)$, and is given by

$$\delta^* = \min(\text{mean}\{\delta_i,\ i = 1, ..., n-2\},\quad \text{median}\{\delta_i,\ i = 1, ..., n-2\}). \tag{6}$$

The lower bound δ^* will allow us to avoid the empirical distribution (under smooth estimates). A few remarks in regard of the quantity δ^* are in order here. First, observe that

$$\delta^* \le \min\left\{\frac{X_{(n)} - X_{(2)}}{n-2},\ \frac{X_{(n-1)} - X_{(1)}}{n-2}\right\},$$

so, for example, in the case where X is bounded, δ^* is at most of order $1/n$.

In fact, this quantity appears to be very small in the simulations. Let us set

$$A_h = \left\{ i : F_h^{-(i)}(X_{(i)}) = \frac{i-1}{n-1} \right\};$$

it can easily be shown that

$$h < \delta^* \Rightarrow \#A_h \geq \left[\frac{n}{2} \right],$$

where for any set A, $\#A$ denotes its cardinality and $[x]$ is the integer part of x. This means that, when $h < \delta^*$, the behavior of F_h is not very far from that of F_n. Another argument in favor of this "censoring" process can be obtained by looking at the kernel density estimator f_h of f. More specifically, let f_h^{-i} be the leave-one-out estimator of f given by

$$f_h^{-i}(x) = \frac{1}{h(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x - X_j}{h} \right)$$

and set

$$B_h = \{ i : f_h^{-i}(X_i) = 0 \};$$

it can be shown that

$$h < \delta^* \Rightarrow \#B_h \geq \left[\frac{n}{2} \right].$$

It follows that the estimator f_h will behave very wildly when $h < \delta^*$. Since it is our belief that a good estimator F_h would lead to a "reasonable" estimator f_h, we are not inclined to use such values of h to estimate F.

The intuition behind the upper bound lies in the statistic (empirical measure of the interval $[1/4, 3/4]$)

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\left[\frac{1}{4}, \frac{3}{4}\right]}(F(X_i)) \tag{7}$$

and the following equality:

$$\lim_{h \to +\infty} F_h(x) = \frac{1}{2}. \tag{8}$$

Let us define the "noisy" empirical measure of the interval $\left[\frac{1}{4}, \frac{3}{4}\right]$ by

$$\mu_h = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\left[\frac{1}{4}, \frac{3}{4}\right]}(F_h^{-i}(X_i)). \tag{9}$$

Because of (8), we have

$$\lim_{h \to +\infty} \mu_h = 1. \tag{10}$$

By the central limit theorem in setting

$$\nu_n = 2\sqrt{n}\left(\mu_n - \frac{1}{2}\right),$$

we have

$$\nu_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Because of Theorem 1, the "noisy" counterpart of ν_n defined by

$$\nu_h = 2\sqrt{n}\left(\mu_h - \frac{1}{2}\right)$$

for "good" values of h will behave like a standard normal random variable. Let t_α be the quantile of order α of the standard normal distribution and set

$$\Delta^* = \inf\{h > 0, \nu_h \geq t_\alpha\}, \tag{11}$$

in this paper, the upper bound for the smoothing parameter is Δ^* (with $\alpha = 0.95$). This bound requires ν_h to fail the normal test at level α. The existence of Δ^* is ensured by (10).

4. The Three Main Methods

The three methods introduced are relative to the distance used, in order words, each distance d between V_h and V_n^S gives birth to three different methods. So, the first thing consists in choosing a reasonable distance d between distribution functions.

Let S_k be a sample of n i.i.d. uniform observations. In order to simplify writings, let $V_n^{(k)}$ (instead of $V_n^{S_k}$) denote the empirical distribution function built with S_k and h_k be the minimizer of $d(V_h, V_n^{(k)})$. We will retain the sample S_k in the process of building a bandwidth for F_h if $\delta^* < h_k < \Delta^*$ (see equations (6) and (11) for the definitions of these quantities). The next step of our algorithm to build a bandwidth is as follows: let

$$S_1, ..., S_N$$

be N samples of size n of i.i.d. uniform observations, the procedure presented immediately above permits us to obtain N_1 samples

$$S_1^*, ..., S_{N1}^* \ (N1 \le N),$$

therefore the bandwidth h_j obtained with the sample S_j^* satisfies $\delta^* < h_j < \Delta^*$. The other steps of the algorithm will consist in using $S_1^*, ..., S_{N1}^*$ to construct the final bandwidth. In the sequel, we will say that a sample is "*censored*" if the bandwidth derived from it is not in the interval $]\delta^*, \Delta^*[$.

We are now ready to describe the three methods.

The average or median bandwidth method. Let h_k be the minimizer of $d(V_h, V_n^{(k)})$ with $V_n^{(k)}$ being the empirical uniform distribution built with the sample S_k^* $(1 \le k \le N1)$. This method consists of using

$$\hat{h} = \text{mean}\{h_i, \ i = 1, ..., \ N1\} \quad \text{or} \quad \hat{h} = \text{median}\{h_i, \ i = 1, ..., \ N1\}$$

as our smoothing parameter. The main idea behind this method is that each h_k is a reasonable bandwidth, thus the sequence can be summarized by its mean or median.

The one step pilot bandwidth method. In this approach, we assume that we possess a "trusted" bandwidth h_0 suitable for our estimator F_h. This trusted bandwidth can be obtained by any existing selection method (Bowman et al. [3], Altman and Léger [1] or Sarda [12]) in the literature. Indeed, if h_0 is a "good" smoothing parameter, then $F_{h_0}^{-1}(X_1)$, ..., $F_{h_0}^{-n}(X_n)$ would be closed to $F(X_1)$, ..., $F(X_n)$ in the sense that $d(V_{h_0}, \hat{V}_n)$ is small. This method consist in using h_0 to choose a "proper" sample from

$$S_1^*, ..., S_{N1}^* \ (N1 \leq N)$$

and use it to build the bandwidth. More precisely, let $V_n^{(i)}$ be the empirical distribution function built with the sample S_i^*. Let i_1 be such that

$$d(V_{h_0}, V_n^{(i_1)}) = \min_{1 \leq i \leq N1} d(V_{h_0}, V_n^{(i)}).$$

Let us denote

$$S_1^{**} = S_{i_1}^*, \quad V_n^{(*1)} = V_n^{(i_1)}$$

and

$$\hat{h}_1 = \arg\min_h d(V_h, V_n^{(*1)}),$$

the final bandwidth for this approach is $\hat{h}_1$.

The fixed point pilot bandwidth approach. This approach is an extension of the previous one. We start with a trusted bandwidth h_0 and then build $\hat{h}_1$ using the above method. Now we can use $\hat{h}_1$ as pilot to build $\hat{h}_2$ and so on. Let i_2 be such that

$$d(V_{\hat{h}_1}, V_n^{(i_2)}) = \min_{1\le i\le N1} d(V_{\hat{h}_1}, V_n^{(i)}),$$

we can then set

$$S_2^{**} = S_{i_2}^*, \quad V_n^{(*2)} = V_n^{(i_2)}$$

and

$$\hat{h}_2 = \arg\min_h d(V_h, V_n^{(*2)}).$$

It is important to note that

$$\begin{aligned} d(V_{\hat{h}_2}, V_n^{(*2)}) &= \inf_h d(V_h, V_n^{(*2)}) \\ &\le d(V_{\hat{h}_1}, V_n^{(*2)}) \\ &= \min_{1\le i\le N1} d(V_{\hat{h}_1}, V_n^{(i)}) \\ &\le d(V_{\hat{h}_1}, V_n^{(*1)}), \end{aligned}$$

therefore we have

$$d(V_{\hat{h}_2}, V_n^{(*2)}) \le d(V_{\hat{h}_1}, V_n^{(*1)}).$$

Using the same procedure, we then build a sequence $(\hat{h}_i)$ such that

$$d(V_{\hat{h}_{i+1}}, V_n^{(*(i+1))}) \le d(V_{\hat{h}_i}, V_n^{(*i)}).$$

It follows from this last inequality that the sequence (u_i) defined by

$$u_i = d(V_{\hat{h}_i}, V_n^{(*i)})$$

is convergent, but since there is a finite number of samples $S_1^*, ..., S_{N1}^*$ $(N1 \le N)$, there exists a smallest natural number j_0 such that

$$d(V_{\hat{h}_{j_0+1}}, V_n^{(*(j_0+1))}) = d(V_{\hat{h}_{j_0}}, V_n^{(*j_0)}).$$

The smoothing parameter selected by this approach is $\hat{h}_{j_0}$.

5. The Indirect L^2 Approaches

The distance considered here is the Mallows metric. Let G_1, G_2 be two distribution functions. The left continuous inverse of G_1 is defined by

$$G_1^{-1}(u) = \inf\{x : G_1(x) \geq u\}.$$

The Mallows distance (d_M) between G_1 and G_2 is defined by

$$d_M^2(G_1, G_2) = \int_0^1 | G_1^{-1}(u) - G_2^{-1}(u) |^2 du.$$

The interesting properties of this distance can be found in Levina and Bickel [7] or Munk and Czado [9]. Let us define $\hat{V}_h$ by

$$\hat{V}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F_h(X_{(i)}) \leq x\}}$$

(note the difference with V_h, equation (2)); we have that

$$d_M^2(\hat{V}_h, \hat{V}_n) = \frac{1}{n}\sum_{i=1}^{n} (F_h(X_{(i)}) - F(X_{(i)}))^2$$

this equality is due to the fact that the order statistics of $F_h(X_1)$, ..., $F_h(X_n)$ are $F_h(X_{(1)})$, ..., $F_h(X_{(n)})$ and those of $F(X_1)$, ..., $F(X_n)$ are $F(X_{(1)})$, ..., $F(X_{(n)})$. To state and proof the optimality result, we will use a modified version of the distance d_M. Let W be a nonnegative and bounded weight function, using the weight function the new measure of accuracy $\bar{d}_M$ is defined by

$$\bar{d}_M^2(\hat{V}_h, \hat{V}_n) = \frac{1}{n}\sum_{i=1}^{n}(F_h(X_{(i)}) - F(X_{(i)}))^2 W(X_{(i)}).$$

In setting as usual

$$ASE(h) = \frac{1}{n}\sum_{i=1}^{n}(F_h(X_i) - F(X_i))^2 W(X_i),$$

it is obvious that

$$\bar{d}_M^2(\hat{V}_h, \hat{V}_n) = ASE(h). \tag{12}$$

By the equivalence of measures of accuracy Sarda [12] or Marron and Härdle [8], under the assumptions (A.1)-(A.5), if in addition W the weight function is compactly supported, we have:

$$ASE(h) = \frac{C_1}{n} - C_2\frac{h}{n} + C_3 h^4 + o\left(\frac{h}{n} + h^4\right) \text{ a.s.} \tag{13}$$

with C_1, C_2 and C_3 being positive constants. To continue, let us introduce the following quantities:

$$\bar{d}_M^2(V_h, \hat{V}_n) = \frac{1}{n}\sum_{i=1}^{n}(F_h^{-(i)}(X_{(i)}) - F(X_{(i)}))^2 W(X_{(i)}),$$

$$\overline{ASE}(h) = \frac{1}{n}\sum_{i=1}^{n}(F_h^{-i}(X_i) - F(X_i))^2 W(X_i).$$

As above, we have:

$$\bar{d}_M^2(V_h, \hat{V}_n) = \overline{ASE}(h). \tag{14}$$

Since it is true that

$$E[(F_h^{-i}(X_i) - F(X_i))^2 W(X_i) \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n]$$
$$= \int (F_h^{-i}(x) - F(x))^2 W(x) f(x) dx,$$

we can deduce that

$$E[\overline{ASE}(h)] = \frac{1}{h}\sum_{i=1}^{n} E[(F_h^{-i}(X_i) - F(X_i))^2 W(X_i)]$$

$$= E\int (F_h^{-n}(x) - F(x))^2 W(x)\, f(x)\, dx.$$

It follows from the MISE representation given in Reiss [11] and Sarda [12] that

$$E[\overline{ASE}(h)] = \frac{C_1}{n} - C_2\,\frac{h}{n} + C_3 h^4 + o\left(\frac{h}{n} + h^4\right) \tag{15}$$

under the conditions (A.1)-(A.5), the constants C_1-C_3 being the same as in equation (13).

Let $S = \{U_1, ..., U_n\}$ be a fixed set composed of an i.i.d. sample of uniform observations, for sake of simplicity, from *here on* we set

$$V_n(x) = \sum_{i=1}^{n} \mathbb{1}_{\{U_i \le x\}}, \tag{16}$$

note that to stress the dependence on the set *S*, this quantity was written as V_n^S in equation (4).

To continue, let us observe that

$$d_M^2(V_h, V_n) = \frac{1}{n}\sum_{i=1}^{n} (F_h^{-(i)}(X_{(i)}) - U_{(i)})^2,$$

where $U_{(1)}, ..., U_{(n)}$ are the order statistics of $U_1, ..., U_n$. So, correspondingly, the distance $\bar{d}_M^2(V_h, V_n)$ is given by

$$\bar{d}_M^2(V_h, V_n) = \frac{1}{n}\sum_{i=1}^{n} (F_h^{-(i)}(X_{(i)}) - U_{(i)})^2 W(X_{(i)}). \tag{17}$$

Note that, unlike in equation (12), it is not true, in general, that

$$\bar{d}_M^2(V_h, V_n) = \frac{1}{n}\sum_{i=1}^{n}(F_h^{-i}(X_i) - U_i)^2 W(X_i),$$

the role of the order statistics appears in equation (17), is crucial in the formula of $\bar{d}_M^2(V_h, V_n)$ and the methods developed in this paper. In the theorem below, the bandwidth will be chosen in the set

$$H_n = [An^{-\frac{1}{3}}, Bn^{-\frac{1}{3}}], \quad 0 < A < B < \infty.$$

It is clear from the right hand side of equation (15) that the asymptotic minimizer of $E\bar{d}_M^2(V_h, \hat{V}_n) = E[\overline{ASE}(h)]$ is of the form $Cn^{-\frac{1}{3}}$, this form in turn justifies the form of H_n. We have the following optimality result for the indirect L^2 methods. It shows that minimizing $\bar{d}_M^2(V_h, \hat{V}_n)$ over H_n is equivalent of minimizing $\bar{d}_M^2(V_h, V_n)$.

Theorem 2. *Let $h \in H_n$, assume that the weight function W is bounded, nonnegative, compactly supported with support $[a_0, b_0]$ such that $F(a_0) > 0$ and $1 - F(b_0) > 0$. Then under the conditions* (A.1)-(A.4)*, there exists a random variable T independent of h such that*

$$\left| \frac{E(\bar{d}_M^2(V_h, V_n) - \bar{d}_M^2(V_h, \hat{V}_n) - T)}{E(\bar{d}_M^2(V_h, \hat{V}_n))} \right| \to 0.$$

The proof of this result is the Appendix.

We will end this section by a simulation study. The kernel used in this section and Section 6 is the Epanechnikov kernel defined by

$$K(x) = \frac{3}{4}\mathbb{1}_{[-1,1]}(x)(1 - x^2). \tag{18}$$

Samples sizes, $n = 200$, $n = 300$ and $n = 500$ are considered throughout the paper. The distribution functions studied in this section are the standard normal and the Pareto distribution functions. We have found that the indirect L^2 methods work well with or without the weight function being compactly supported. In other words, it appears that the compact support assumption is a technical condition needed in the proof of Theorem 2. Therefore, our simulations are done with $W = 1$. Let us specify the notations used in the simulation below (see the tables):

h_{m0} is the minimizer of $d_M^2(\hat{V}_h, \hat{V}_n) = ASE(h)(W = 1)$;

h_{m1} is the bandwidth obtained by the average bandwidth method with $d = d_M$;

h_{m2} is the bandwidth obtained by one step pilot bandwidth method with $d = d_M$ and the pilot bandwidth being h_{m1};

h_{m3} is the bandwidth obtained by fixed point pilot bandwidth method with $d = d_M$ and the pilot bandwidth being h_{m1}.

The reader is referred to Section 3 for the definition of each method.

Case of the standard normal distribution $\mathcal{N}(0, 1)$: The one step pilot bandwidth method is illustrated in Figure 1. Specifically, Figure 1 plots the "selection criterion" $d_M^2(V_h, V_n^{(*1)})$ for the one step pilot bandwidth method with the true error $d_M^2(\hat{V}_h, \hat{V}_n)$. Both curves have the same intervals of variations (the intervals of increase and decrease are roughly the same): this is a validation of Theorem 2.

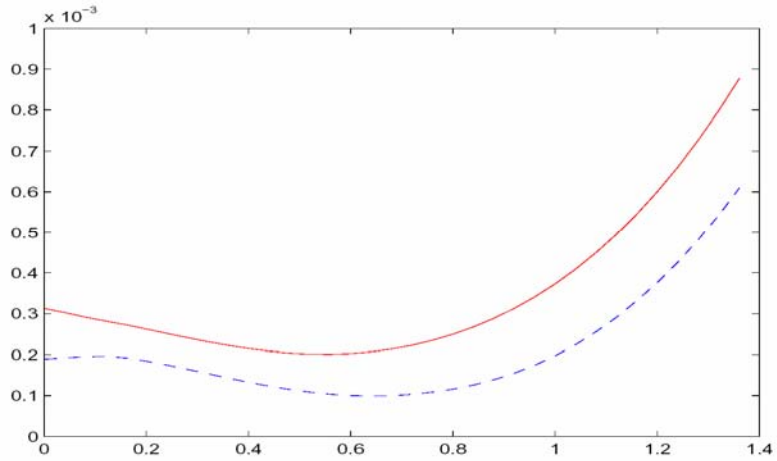

Figure 1. $d_M^2(\hat{V}_h, \hat{V}_n)$ solid curve and $d_M^2(V_h, V_n^{(*1)})$ dashed curve as functions of *h* for one sample, $n = 300$.

In this section, the "CV method" is short for the cross-validation method developed in Bowman et al. [3] and h_b is the bandwidth built by that method. We have chosen to compare the methods of this section to the CV method because they (the methods) are all related to the L^2 distance, and also because as shown in the simulations done in Bowman et al. [3], in general, better results are obtained with the CV method compared to the others usually considered methods. In Table 1, a simulation study regarding the fixed point pilot bandwidth and the CV method is summarized. Note that similar computations were done for the mean and one step pilot bandwidth methods, and the fixed point pilot bandwidth method appeared to be only marginally superior to mean and one step pilot bandwidth methods. To save space the results in Table 1 are written in the format $d * E$, where *d* is a number and *E* a constant equal to 10^{-6}. Table 1 shows that for the standard normal distribution, the CV method is only slightly superior to the fixed point pilot bandwidth method in terms of mean and standard deviation when $n = 200$ or $n = 300$.

Table 1. Average square error for h_{m3}, h_b and h_{m0}; mean and (std) standard deviation over 100 replications

Fixed point and CV methods						
	$ASE(h_{m3})$		$ASE(h_b)$		$ASE(h_{m0})$	
	mean	std	mean	std	mean	std
$n = 200$	664E	599E	647E	585E	565E	582E
$n = 300$	437E	399E	426E	386E	358E	370E
$n = 500$	317E	308E	321E	311E	266E	286E

Case of a Pareto distribution: Here we consider the Pareto distribution function with parameter β defined by

$$F(x) = \mathbb{1}_{[1,+\infty]}(x)\left(1 - \frac{1}{x^{\beta}}\right).$$

The Pareto distribution is chosen to *make fail* the CV method, however, we would like to remind the reader that, the CV method works well for most of the usually considered distribution functions as evidenced by the results of Table 1 or the simulations done in Bowman et al. [3]. The reason for the failure is that, if $\int |u| K(u)du < +\infty$ (which is implied by (A.3)) and F is a Pareto distribution whose parameter satisfies $0 < \beta \leq \frac{1}{2}$, the quantity $ISE(h)$ as defined on page 803 of Bowman et al. [3] is infinite, i.e.,

$$ISE(h) = +\infty.$$

When the kernel function is compactly supported, the proof of this result is obvious since for x sufficiently large, we have

$$F_h(x) - F(x) = \frac{1}{x^{\beta}}.$$

Table 2. Average square error for h_{m1}, h_b and h_{m0}; mean and (std) standard deviation over 100 replications, $E = 10^{-6}$

Mean and CV methods							
	$ASE(h_{m1})$		$ASE(h_b)$			$ASE(h_{m0})$	
	mean	std	mean	std	MNoM	mean	std
$n = 200$	793E	585E	781E	589E	10.60	736E	576E

Table 2 contains some statistics on $ASE(h_{m1})$, $ASE(h_b)$ and $ASE(h_{m0})$, over 100 replications of size $n = 200$ of samples drawn from the Pareto distribution with parameter $\beta = 0.2$. It can be seen from this table that the mean of $ASE(h_b)$ is good bet of that of $ASE(h_{m0})$, it appears to be a bit better approximation than the mean of $ASE(h_{m1})$. So, what is going on here? Things seem to go against our assumption. Now, look at the (odd) column headed MNoM! The number under that column represents the **M**ean **N**umber **o**f **M**imina of the function $h \mapsto CV(h)$ (see the CV function in Bowman et al. [3]). Let us recall that h_b the bandwidth of the CV method is the argument of the minimum of the $h \mapsto CV(h)$. A natural question raises by our situation is how to compute h_b in case of multiple minima? To compute the results of Table 2, we have:

– selected grid of points referred to as G1 contained in the interval $[10^{-6}, 5]$;

– chosen h_b to be the point of the grid (G1) closest to the mean of the arguments of the minima.

It appears that in case of multiple minima, the mean of the minimizers tends to fall around the middle of the interval containing G1, and this point is actually a good bandwidth for the example under consideration. It becomes suspicious that the interval containing G1 plays a role in the good result obtained by the CV method in Table 2. So, what happen if we extend the

interval? Unfortunately, it is nearly an "impossibility" to do the computations for all the 100 samples for larger grid of bandwidths. The domain of integration (where the quantities to be integrated are different from 0) of the integrals involved in the CV criterion depends on the range of the random sample, and this range can be extremely large for a sample drawn from a Pareto distribution with parameter $\beta = 0.2$. This makes the computations of the CV criterion for this example very difficult and time consuming. To be concrete, it took us 4 days and a $1/2$ to compute the results (regarding the CV method) contained in Table 2 where the grid G1 is contained in $[10^{-6}, 5]$. We will study the effect of extending the grid of bandwidth on samples already having multiple minima on G1. Out of the 100 samples considered in Table 2, 11 present the maximum number of minima which is 31. The grid of bandwidths G1 is extended to G2 contained in $[10^{-6}, 8.5]$. The results of computations on this new grid are presented in Table 3. The notations are:

ID is the identifier of the sample under study;

$$YCV = \{CV(h),\ h \in G2\};$$

mean represents the mean of *YCV*;

std is the standard deviation of *YCV* ;

NoM represents the **N**umber **o**f **M**inima of *YCV*;

$$\Omega = 10^{16}, \quad \kappa = 10^{5}.$$

Table 3 reveals what was expected when this example was introduced:

– The number of minimum initially 31 for each sample has risen to 32 for one sample and, 49 for the others.

– Even for a sample size of $n = 200,$ there are samples for which the CV method cannot produce a bandwidth, the function $h \mapsto CV(h)$ is a constant for those samples. Seeing the size (hugeness) of the means, we believe that the small variations symbolized by non-zero standard deviation are only due to round-off errors in the computations of the CV criterion. In fact, we have

plotted the graphs of $h \mapsto CV(h)$ for all the 11 samples and, they are totally flat.

Table 3. Some statistics on YCV for 11 samples of size $n = 200$

YCV					YCV			
ID	mean	std	NoM	‖	ID	mean	Std	NoM
a	0.013Ω	0.069	32	‖	*g*	0.029Ω	0.189	49
b	2.752Ω	24.24	49	‖	*h*	0.130Ω	0.758	49
c	1.076Ω	8.082	49	‖	*i*	0.192Ω	1.263	49
d	117.6Ω	0	49	‖	*j*	6.665Ω	40.41	49
e	0.187Ω	0.252	49	‖	*k*	0.045Ω	0.505	49
f	94147Ω	2.65κ	49	‖				

Table 4. Average square error for h_{m3} and h_{m0}; mean, median and standard deviation over 100 replications

Fixed point pilot bandwidth method						
	$ASE(h_{m3})$			$ASE(h_{m0})$		
	mean	median	standard deviation	mean	median	standard deviation
$n = 200$	0.00077	0.00063	0.00059	0.00073	0.00060	0.00058
$n = 300$	0.00052	0.00040	0.00039	0.00051	0.00039	0.00038
$n = 500$	0.00034	0.00024	0.00033	0.00033	0.00024	0.00032

In Table 4, the results of computations for the fixed point pilot bandwidth method on the Pareto distribution with parameter $\beta = 0.2$ are displayed. We have done similar computations for the mean and one step

pilot methods. The order of performance is fixed point pilot (best), one step pilot and the mean bandwidth method, but the difference between the methods is slim. Table 4 shows that the indirect L^2 methods do a pretty good job on this example.

6. The L^1 Approaches

The distance considered in this section is the Wasserstein metric. Let G_1, G_2 be two distribution functions. The Wasserstein distance (d_W) between G_1 and G_2 is defined by

$$d_W(G_1, G_2) = \int_0^1 | G_1^{-1}(u) - G_2^{-1}(u) | du.$$

An interesting property of this distance is the following equality (see, for instance, Shorack and Wellner [13])

$$d_W(G_1, G_2) = \int | G_1(x) - G_2(x) | dx,$$

this is why we consider the methods based on this metric to be both indirect and direct L^1 approaches. We have that

$$d_W(\hat{V}_h, \hat{V}_n) = \frac{1}{n} \sum_{i=1}^{n} | F_h(X_{(i)}) - F(X_{(i)}) |,$$

this equality is due to the fact that the order statistics of $F_h(X_1)$, ..., $F_h(X_n)$ are $F_h(X_{(1)})$, ..., $F_h(X_{(n)})$ and those of $F(X_1)$, ..., $F(X_n)$ are $F(X_{(1)})$, ..., $F(X_{(n)})$. As in Section 4, we introduce a modified version of the distance d_W. Using the weight function *W*, this new measure of accuracy d_W is given by

$$\bar{d}_W(\hat{V}_h, \hat{V}_n) = \frac{1}{n} \sum_{i=1}^{n} | F_h(X_{(i)}) - F(X_{(i)}) | W(X_{(i)}).$$

In setting as usual

$$AAE(h) = \frac{1}{n}\sum_{i=1}^{n} | F_h(X_i) - F(X_i) | W(X_i),$$

it is obvious that

$$\bar{d}_W(\hat{V}_h, \hat{V}_n) = AAE(h). \tag{19}$$

To continue, let us introduce the following quantities

$$\bar{d}_W(V_h, \hat{V}_n) = \frac{1}{n}\sum_{i=1}^{n} | F_h^{-(i)}(X_{(i)}) - F(X_{(i)}) | W(X_{(i)}),$$

$$\overline{AAE}(h) = \frac{1}{n}\sum_{i=1}^{n} | F_h^{-i}(X_i) - F(X_i) | W(X_i)$$

as above, we have:

$$\bar{d}_W(V_h, \hat{V}_n) = \overline{AAE}(h). \tag{20}$$

To continue, let us observe that (see the definition of V_n in equation (16))

$$d_W(V_h, V_n) = \frac{1}{n}\sum_{i=1}^{n} | F_h^{-(i)}(X_{(i)}) - U_{(i)} |,$$

where $U_{(1)}, ..., U_{(n)}$ are the order statistics of $U_1, ..., U_n$. So, the associated distance $\bar{d}_W(V_h, V_n)$ is given by

$$\bar{d}_W(V_h, V_n) = \frac{1}{n}\sum_{i=1}^{n} | F_h^{-(i)}(X_{(i)}) - U_{(i)} | W(X_{(i)}).$$

We will assess the methods based on the L^1 distance by simulations. The example considered is the standard normal distribution $\mathcal{N}(0, 1)$. The kernel used is the Epanechnikov given in equation (18). As in Section 5, the L^1

methods work well whether the weight function is compactly supported or not; in the simulation below $W = 1$. Let us specify the notations used to present our results:

h_{w0} is the minimizer of $d_W(\hat{V}_h, \hat{V}_n) = AAE(h)$ (with $W = 1$);

h_{w1} is the bandwidth obtained by the median bandwidth method with $d = d_W$;

h_{w2} is the bandwidth obtained by one step pilot bandwidth method with $d = d_W$ and the pilot bandwidth being h_{w1};

h_{w3} is the bandwidth obtained by fixed point pilot bandwidth method with $d = d_W$ and the pilot bandwidth being h_{w1}.

Figure 2 plots the "selection criterion" $d_W(V_h, V_n^{(*1)})$ for the one step pilot bandwidth method with the true error $d_W(\hat{V}_h, \hat{V}_n)$.

In Table 5 below, the results of the simulation study for the median bandwidth approach are summarized. It is clear from this table that the bandwidth obtained by this method is a good guess of h_{w0} (the minimizer of $AAE(h)$), because the values of $AAE(h_{w0})$ and $AAE(h_{w1})$ are of comparable magnitude in terms of mean, median and standard deviation. Similar computations were done for the other two (L^1) methods. Surprisingly, the best method in terms of mean was the median bandwidth method followed by the one step bandwidth method, but there is no significant difference between the results obtained by the three methods. So, the performances over replications of the (L^1) methods on the example are equivalent. However, the L^1 approaches are more robust because the standard deviations are small compared to the means. This is not true for L^2 approaches.

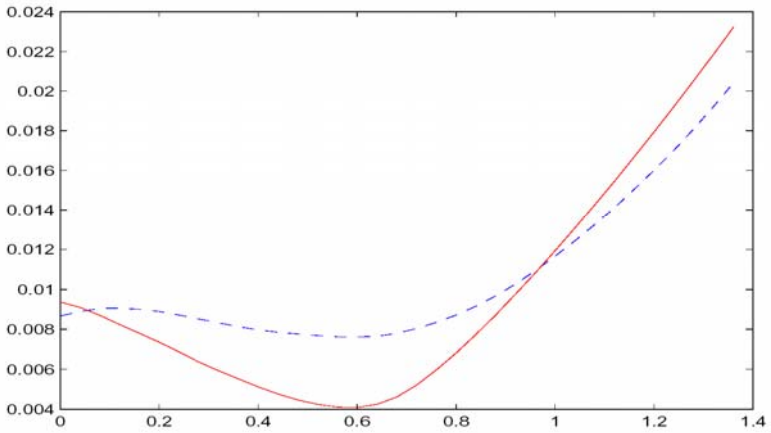

Figure 2. $\bar{d}_W(\hat{V}_h, \hat{V}_n)$ solid curve and $\bar{d}_W(V_h, V_n^{(*1)})$ dashed curve as functions of h for one sample, $n = 300$.

Table 5. Average absolute error for h_{w1} and h_{w0}; mean, median and standard deviation over 100 replications

Median bandwidth method						
	$AAE(h_{w1})$			$AAE(h_{w0})$		
	mean	median	standard deviation	mean	median	standard deviation
$n = 200$	0.01967	0.01832	0.00953	0.01805	0.01670	0.01008
$n = 300$	0.01620	0.01466	0.00741	0.01470	0.01365	0.00754
$n = 500$	0.01346	0.01226	0.00677	0.01230	0.01142	0.00697

A. Proofs

A.1. Proof of Theorem 1

Let us recall that the definition of $\hat{V}_h$ is

$$\hat{V}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F_h(X_{(i)}) \le x\}}.$$

First, we will establish the following result:

$$\sup_{x\in\mathbb{R}} |\hat{V}_h(x) - V(x)| \xrightarrow{a.s.} 0 \tag{21}$$

because it will be used in the proof of the theorem. We have

$$\sup_{x\in\mathbb{R}} |\hat{V}_h(x) - V(x)| = \sup_{x\in[0,1]} |\hat{V}_h(x) - x|.$$

We know that (Kolmogorov-Smirnov statistic) that

$$\sup_{x\in[0,1]} |\hat{V}_h(x) - x| = \max\left\{ \max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i}{n} \right| \right\}, \max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i-1}{n} \right| \right\} \right\}.$$

So to prove (21), it is enough to show that

$$\max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i}{n} \right| \right\} \xrightarrow{a.s.} 0 \tag{22}$$

and

$$\max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i-1}{n} \right| \right\} \xrightarrow{a.s.} 0. \tag{23}$$

On the other hand, we have

$$\left| F_h(X_{(i)}) - \frac{i}{n} \right| \le |F_h(X_{(i)}) - F(X_{(i)})| + \left| F(X_{(i)}) - \frac{i}{n} \right|.$$

It follows that

$$\max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i}{n} \right| \right\} \le \sup_{x\in\mathbb{R}} |F_h(x) - F(x)| + \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|.$$

The proof of (22) is completed using Proposition 1 and Glivenko-Cantelli theorem.

Using the same arguments as immediately above, we have

$$\max_{i=1,\ldots,n} \left\{ \left| F_h(X_{(i)}) - \frac{i-1}{n} \right| \right\} \le \sup_{x\in\mathbb{R}} |F_h(x) - F(x)| + \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|;$$

this inequality allows us to obtain (23).

To continue, let us set

$$\hat{F}_h^{-(i)}(x) = \frac{1}{n}\sum_{j\neq i} H\left(\frac{x - X_{(j)}}{h}\right);$$

we have

$$\hat{F}_h^{-(i)}(x) - F_h^{-(i)}(x) = \left(\frac{1}{n-1} - \frac{1}{n}\right)\sum_{\substack{j=1\\ j\neq i}}^{n} H\left(\frac{x - X_{(j)}}{h}\right)$$

$$= \left(\frac{1}{n(n-1)}\right)\sum_{\substack{j=1\\ j\neq i}}^{n} H\left(\frac{x - X_{(j)}}{h}\right).$$

It follows that

$$| \hat{F}_h^{-(i)}(x) - F_h^{-(i)}(x) | \leq \frac{1}{n}.$$

We also have

$$F_h(x) - \hat{F}_h^{-(i)}(x) = \frac{1}{n}\left(\sum_{j=1}^{n} H\left(\frac{x - X_{(j)}}{h}\right) - \sum_{\substack{j=1\\ j\neq i}}^{n} H\left(\frac{x - X_{(j)}}{h}\right)\right)$$

$$= \frac{1}{n} H\left(\frac{x - X_{(i)}}{h}\right),$$

therefore we get

$$| F_h(x) - \hat{F}_h^{-(i)}(x) | \leq \frac{1}{n}.$$

We then obtain

$$| F_h(x) - F_h^{-(i)}(x) | \leq | F_h(x) - \hat{F}_h^{-(i)}(x) | + | \hat{F}_h^{-(i)}(x) - F_h^{-(i)}(x) | \leq \frac{2}{n}.$$

To prove our Theorem 1, observe that

$$\sup_{x\in\mathbb{R}} |V_h(x) - V(x)| = \sup_{x\in[0,1]} |V_h(x) - x|.$$

We have (Kolmogorov-Smirnov statistic)

$$\sup_{x\in[0,1]} |V_h(x) - x| = \max\left\{ \max_{i=1,\ldots,n} \left\{ \left| F_h^{-(i)}(X_{(i)}) - \frac{i}{n} \right| \right\}, \right.$$

$$\left. \max_{i=1,\ldots,n} \left\{ \left| F_h^{-(i)}(X_{(i)}) - \frac{i-1}{n} \right| \right\} \right\}.$$

It is also obvious that

$$\left| F_h^{-(i)}(X_{(i)}) - \frac{i}{n} \right| \le | F_h^{-(i)}(X_{(i)}) - F_h(X_{(i)}) | + \left| F_h(X_{(i)}) - \frac{i}{n} \right|$$

$$\le \frac{2}{n} + \left| F_h(X_{(i)}) - \frac{i}{n} \right|$$

and

$$\left| F_h^{-(i)}(X_{(i)}) - \frac{i-1}{n} \right| \le \frac{2}{n} + \left| F_h(X_{(i)}) - \frac{i-1}{n} \right|.$$

Putting everything together and using Kolmogorov-Smirnov statistic, we get

$$\sup_{x\in\mathbb{R}} |V_h(x) - V(x)| \le \frac{2}{n} + \sup_{x\in\mathbb{R}} |\hat{V}_h(x) - V(x)|$$

and the proof of Theorem 1 is completed using (21).

A.2. Proof of Theorem 2

We have:

$$(F_h^{-(i)}(X_{(i)}) - U_{(i)})^2 = (F_h^{-(i)}(X_{(i)}) - F(X_{(i)}))^2$$

$$+ 2(F_h^{-(i)}(X_{(i)}) - F(X_{(i)}))(F(X_{(i)}) - U_{(i)})$$

$$+ (F(X_{(i)}) - U_{(i)})^2$$

$$
\begin{aligned}
&= (F_h^{-(i)}(X_{(i)}) - F(X_{(i)}))^2 \\
&\quad + 2(F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}]) \\
&\quad \cdot (F(X_{(i)}) - U_{(i)}) \\
&\quad + 2(E[F_h^{-(i)}(X_{(i)}) | X_{(i)}] - F(X_{(i)}))(F(X_{(i)}) - U_{(i)}) \\
&\quad + (F(X_{(i)}) - U_{(i)})^2.
\end{aligned}
$$

It follows that

$$\bar{d}_M^2(V_h, V_n) = \bar{d}_M^2(V_h, \hat{V}_n) + 2CT^1(h) + 2CT^2(h) + T_1$$

with

$$CT^1(h) = \frac{1}{n}\sum_{i=1}^{n}(F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}])(F(X_{(i)}) - U_{(i)})W(X_{(i)}),$$

$$CT^2(h) = \frac{1}{n}\sum_{i=1}^{n}(E[F_h^{-(i)}(X_{(i)}) | X_{(i)}] - F(X_{(i)}))(F(X_{(i)}) - U_{(i)})W(X_{(i)}),$$

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}(F(X_{(i)}) - U_{(i)})^2 W(X_{(i)}).$$

Since T_1 is independent of h, Theorem 2 is a straightforward consequence of the following lemmas:

Lemma 1. *Under the conditions of Theorem* 2, *we have*:

$$E(CT^1(h)) = 0.$$

Lemma 2. *Under the conditions of Theorem* 2, *there exists a random variable* T_2 *independent of h such that*

$$\left| \frac{E(CT^2(h) - T_2)}{E(\bar{d}_M^2(V_h, \hat{V}_n))} \right| \to 0.$$

A.2.1. Proof of Lemma 1

It is enough to prove that $E[\Gamma_i] = 0,$ where

$$\Gamma_i = (F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}])(F(X_{(i)}) - U_{(i)})W(X_{(i)}).$$

We have:

$$\Gamma_i = \Gamma_i^+ - \Gamma_i^-$$

with

$$\Gamma_i^+ = (F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}])F(X_{(i)})W(X_{(i)}),$$

$$\Gamma_i^- = (F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}])U_{(i)}W(X_{(i)}).$$

We have

$$E[\Gamma_i^+ | X_{(i)}] = 0,$$

and using independence between $(U_i)_{i=1,\ldots,n}$ and $(X_i)_{i=1,\ldots,n},$ we get

$$E[\Gamma_i^-] = E[(F_h^{-(i)}(X_{(i)}) - E[F_h^{-(i)}(X_{(i)}) | X_{(i)}])W(X_{(i)})]E(U_{(i)}) = 0.$$

It follows that

$$E[\Gamma_i] = E[\Gamma_i^+] - E[\Gamma_i^-] = 0.$$

This ends the proof of Lemma 1.

A.2.2. Proof of Lemma 2

Let us set

$$\gamma_i = E[F_h^{-(i)}(X_{(i)}) | X_{(i)}] - F(X_{(i)}),$$

$$\Delta_i = F(X_{(i)}) - U_{(i)}.$$

First, we are going to split γ_i into "simpler" components. Let $f_{jk}(u, v)$ be the joint density of $(X_{(j)}, X_{(k)})$ $(j < k),$ we have (see, for instance,

Arnold et al. [2])

$$f_{jk}(u, v) = \frac{n!F^{j-1}(u)[F(v) - F(u)]^{k-j-1}[1 - F(v)]^{n-k}}{(j-1)!(k-j-1)!(n-k)!} f(u) f(v)$$

for $-\infty < u \le v < +\infty$. The density f_j of $X_{(j)}$ is given by

$$f_j(x) = \frac{n!F^{j-1}(x)[1 - F(x)]^{n-j}}{(j-1)!(n-j)!} f(x).$$

Therefore, for $j_0 < k$, the conditional density φ_{j_0k} of $X_{(k)}$ knowing $X_{(j_0)}$ is given by

$$\varphi_{j_0k}(y \mid X_{(j_0)})$$

$$= \frac{f_{j_0k}(X_{(j_0)}, y)}{f_{j_0}(X_{(j_0)})} \qquad (X_{(j_0)} \le y)$$

$$= \frac{n!F^{j_0-1}(X_{(j_0)})[F(y) - F(X_{(j_0)})]^{k-j_0-1}[1 - F(y)]^{n-k}}{(j_0-1)!(k-j_0-1)!(n-k)!} f(X_{(j_0)}) f(y)$$

$$\times \frac{(j_0-1)!(n-j_0)!}{n!F^{j_0-1}(X_{(j_0)})[1 - F(X_{(j_0)})]^{n-j_0} f(X_{(j_0)})}$$

$$= \frac{(n-j_0)!}{(n-k)!(k-j_0-1)!} \frac{[F(y) - F(X_{(j_0)})]^{k-j_0-1}[1 - F(y)]^{n-k}}{[1 - F(X_{(j_0)})]^{n-j_0}} f(y).$$

For $k < j_0$, the conditional density φ_{kj_0} of $X_{(k)}$ given $X_{(j_0)}$ is given by

$$\varphi_{kj_0}(y \mid X_{(j_0)})$$

$$= \frac{f_{kj_0}(y, X_{(j_0)})}{f_{j_0}(X_{(j_0)})} \qquad (y \le X_{(j_0)})$$

$$= \frac{n!F^{k-1}(y)[F(X_{(j_0)}) - F(y)]^{j_0-k-1}[1 - F(X_{(j_0)})]^{n-j_0}}{(k-1)!(j_0-k-1)!(n-j_0)!} f(y) f(X_{(j_0)})$$

$$\times \frac{(j_0-1)!(n-j_0)!}{n!F^{j_0-1}(X_{(j_0)})[1-F(X_{(j_0)})]^{n-j_0} f(X_{(j_0)})}$$

$$= \frac{(j_0-1)!}{(j_0-k-1)!(k-1)!} \frac{F^{k-1}(y)\,[F(X_{(j_0)})-F(y)]^{j_0-k-1}}{F^{j_0-1}(X_{(j_0)})} f(y).$$

In the sequel, we will need the values of the quantities

$$\psi^+(y \mid X_{(j_0)}) = \sum_{k=j_0+1}^{n} \varphi_{j_0 k}(y \mid X_{(j_0)}),$$

$$\psi^-(y \mid X_{(j_0)}) = \sum_{k=1}^{j_0-1} \varphi_{k j_0}(y \mid X_{(j_0)}).$$

To compute these quantities, we are going to use the following obvious binomial identity. Let $p, q \in \mathbb{R}$. Then we have:

$$np(p+q)^{n-1} = \sum_{k=1}^{n} k\binom{n}{k} p^k q^{n-k}. \tag{24}$$

We have:

$$\psi^+(y \mid X_{(j_0)}) = \frac{f(y)}{[1-F(X_{(j_0)})]^{n-j_0}} \sum_{k=j_0+1}^{n} \frac{(n-j_0)!}{(n-k)!(k-j_0-1)!}$$

$$\times [F(y)-F(X_{(j_0)})]^{k-j_0-1}[1-F(y)]^{n-k}$$

$$= \frac{f(y)}{[1-F(X_{(j_0)})]^{n-j_0}[F(y)-F(X_{(j_0)})]}$$

$$\times \sum_{k=j_0+1}^{n} \frac{(k-j_0)(n-j_0)!}{(n-j_0-(k-j_0))!(k-j_0)!}$$

$$\times [F(y)-F(X_{(j_0)})]^{k-j_0}[1-F(y)]^{n-j_0-(k-j_0)}$$

$$= \frac{f(y)}{[1-F(X_{(j_0)})]^{n-j_0}[F(y)-F(X_{(j_0)})]} \sum_{\ell=1}^{n-j_0} \frac{\ell(n-j_0)!}{(n-j_0-\ell)!\,\ell!}$$

$$\times [F(y)-F(X_{(j_0)})]^{\ell}[1-F(y)]^{n-j_0-\ell}$$

$$= \frac{f(y)}{[1-F(X_{(j_0)})]^{n-j_0}[F(y)-F(X_{(j_0)})]} \sum_{\ell=1}^{n-j_0} \ell \binom{n-j_0}{\ell}$$

$$\times [F(y)-F(X_{(j_0)})]^{\ell}[1-F(y)]^{n-j_0-\ell}$$

$$= \frac{f(y)(n-j_0)}{[1-F(X_{(j_0)})]}.$$

The binomial identity (equation (24)) is used to obtain the last equality above. In the same spirit, we have:

$$\psi^{-}(y \mid X_{(j_0)}) = \frac{f(y)}{F^{j_0-1}(X_{(j_0)})} \sum_{k=1}^{j_0-1} \frac{(j_0-1)!}{(j_0-k-1)!(k-1)!}$$

$$\times F^{k-1}(y)[F(X_{(j_0)})-F(y)]^{j_0-1-k}$$

$$= \frac{f(y)}{F^{j_0-1}(X_{(j_0)})F(y)} \sum_{k=1}^{j_0-1} \frac{k(j_0-1)!}{(j_0-1-k)!\,k!}$$

$$\times F^{k}(y)[F(X_{(j_0)})-F(y)]^{j_0-1-k}$$

$$= \frac{f(y)}{F^{j_0-1}(X_{(j_0)})F(y)} \sum_{k=1}^{j_0-1} k \binom{j_0-1}{k}$$

$$\times F^{k}(y)[F(X_{(j_0)})-F(y)]^{j_0-1-k}$$

$$= \frac{f(y)(j_0-1)}{F(X_{(j_0)})}.$$

Observe that

$$E[F_h^{-(j_0)}(X_{(j_0)})|\, X_{(j_0)}] = \mu_{j_0}^- + \mu_{j_0}^+,$$

where

$$\mu_{j_0}^- = \frac{1}{n-1}\sum_{k=1}^{j_0-1} E\left[H\left(\frac{X_{(j_0)} - X_{(k)}}{h}\right)|\, X_{(j_0)}\right],$$

$$\mu_{j_0}^+ = \frac{1}{n-1}\sum_{k=j_0+1}^{n} E\left[H\left(\frac{X_{(j_0)} - X_{(k)}}{h}\right)|\, X_{(j_0)}\right].$$

We have that

$$\mu_{j_0}^+ = \frac{1}{n-1}\sum_{k=j_0+1}^{n}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right)\varphi_{j_0k}(y|\, X_{(j_0)})dy$$

$$= \frac{1}{n-1}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right)\sum_{k=j_0+1}^{n}\varphi_{j_0k}(y|\, X_{(j_0)})dy$$

$$= \frac{1}{n-1}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right)\psi^+(y|\, X_{(j_0)})dy$$

$$= \frac{(n-j_0)}{(n-1)[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right) f(y)dy.$$

We can then split $\mu_{j_0}^+$ as follows:

$$\mu_{j_0}^+ = \left[\frac{n-j_0}{n-1} - \frac{n-j_0}{n}\right]\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right) f(y)dy$$

$$+ \frac{n-j_0}{n}\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)} - y}{h}\right) f(y)dy$$

$$= \frac{n-j_0}{(n-1)n}\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy$$

$$+ \frac{n-j_0}{n}\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy.$$

This quantity can further be split as

$$\mu_{j_0}^{+} = \frac{n-j_0}{(n-1)n}\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy$$

$$+\left(\frac{1-\hat{V}_n(F(X_{(j_0)}))}{[1-F(X_{(j_0)})]}-1\right)\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy$$

$$+\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy.$$

We get the following decomposition of $\mu_{j_0}^{+}$:

$$\mu_{j_0}^{+} = \alpha_{j_0}^{+} + \beta_{j_0}^{+} + \theta_{j_0}^{+}$$

with

$$\alpha_{j_0}^{+} = \frac{n-j_0}{(n-1)n}\frac{1}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy,$$

$$\beta_{j_0}^{+} = \frac{F(X_{(j_0)})-\hat{V}_n(F(X_{(j_0)}))}{[1-F(X_{(j_0)})]}\int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy,$$

$$\theta_{j_0}^{+} = \int_{X_{(j_0)}}^{+\infty} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy.$$

In the same spirit, we have:

$$\mu_{j_0}^{-} = \frac{1}{n-1}\sum_{k=1}^{j_0-1}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)\varphi_{kj_0}(y \mid X_{(j_0)})dy$$

$$= \frac{1}{n-1}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)\sum_{k=1}^{j_0-1}\varphi_{kj_0}(y\,|\,X_{(j_0)})dy$$

$$= \frac{1}{n-1}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)\psi^{-}(y\,|\,X_{(j_0)})dy$$

$$= \frac{j_0-1}{(n-1)F(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy.$$

We can also split $\mu_{j_0}^{-}$ as follows:

$$\mu_{j_0}^{-} = \left[\frac{j_0-1}{n-1}-\frac{j_0}{n}\right]\frac{1}{F(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy$$

$$+\frac{j_0}{nF(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy$$

$$= \frac{-(n-j_0)}{(n-1)nF(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy$$

$$+\frac{j_0}{nF(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy.$$

This quantity can be rewritten as

$$\mu_{j_0}^{-} = \frac{-(n-j_0)}{(n-1)nF(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy$$

$$+\frac{\hat{V}_n(F(X_{(j_0)}))-F(X_{(j_0)})}{F(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy$$

$$+\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right)f(y)dy.$$

From this last equality, we see that $\mu_{j_0}^{-}$ has the following decomposition:

$$\mu_{j_0}^{-} = \alpha_{j_0}^{-} + \beta_{j_0}^{-} + \theta_{j_0}^{-} + \tau_{j_0},$$

where

$$\alpha_{j_0}^{-} = \frac{-(n-j_0)}{(n-1)nF(X_{(j_0)})}\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy,$$

$$\beta_{j_0}^{-} = \frac{\hat{V}_n(F(X_{(j_0)}))-F(X_{(j_0)})}{F(X_{(j_0)})}\left[\int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy - F(X_{(j_0)})\right],$$

$$\theta_{j_0}^{-} = \int_{-\infty}^{X_{(j_0)}} H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy,$$

$$\tau_{j_0} = \hat{V}_n(F(X_{(j_0)})) - F(X_{(j_0)}).$$

Note that τ_{j_0} is independent of *h*. To sum up, at this stage, we have:

$$E[F_h^{-(j_0)}(X_{(j_0)})|\, X_{(j_0)}] = \alpha_{j_0}^{+} + \beta_{j_0}^{+} + \theta_{j_0}^{+} + \alpha_{j_0}^{-} + \beta_{j_0}^{-} + \theta_{j_0}^{-} + \tau_{j_0}$$

and since

$$\gamma_{j_0} = E[F_h^{-(j_0)}(X_{(j_0)})|\, X_{(j_0)}] - F(X_{(j_0)}),$$

we see that

$$\gamma_{j_0} = \alpha_{j_0}^{+} + \alpha_{j_0}^{-} + \beta_{j_0}^{+} + \beta_{j_0}^{-} + \theta_{j_0} + \tau_{j_0}$$

with

$$\theta_{j_0} = \theta_{j_0}^{+} + \theta_{j_0}^{-} - F(X_{(j_0)}) = \int H\left(\frac{X_{(j_0)}-y}{h}\right) f(y)dy - F(X_{(j_0)}).$$

Because

$$CT^2(h) = \frac{1}{n}\sum_{i=1}^{n} \gamma_i \Delta_i W(X_{(i)}),$$

we have

$$CT^2(h) = \sum_{k=1}^{5} CT^{2,k} + T_2$$

with

$$CT^{2,1}(h) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i^{+} \Delta_i W(X_{(i)}),$$

$$CT^{2,2}(h) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i^{-} \Delta_i W(X_{(i)}),$$

$$CT^{2,3}(h) = \frac{1}{n}\sum_{i=1}^{n} \beta_i^{+} \Delta_i W(X_{(i)}),$$

$$CT^{2,4}(h) = \frac{1}{n}\sum_{i=1}^{n} \beta_i^{-} \Delta_i W(X_{(i)}),$$

$$CT^{2,5}(h) = \frac{1}{n}\sum_{i=1}^{n} \theta_i \Delta_i W(X_{(i)}),$$

$$T_2 = \frac{1}{n}\sum_{i=1}^{n} (\hat{V}_n(F(X_{(j_0)})) - F(X_{(j_0)}))\Delta_i W(X_{(i)})$$

and note that as stated in Lemma 2, T_2 is independent of h. So, to prove Lemma 2, it is enough to prove the following Lemma 3.

Lemma 3. *Under the conditions of Theorem* 2, *we have*

$$\left| \frac{E(CT^{2,k}(h))}{E(\bar{d}_M^2(V_h, \hat{V}_n))} \right| \to 0 \ \textit{for} \ k = 1, ..., 5.$$

A.2.3. Proof of Lemma 3

(i) Case $k = 1$: We have:

$$| \alpha_i^{+} \Delta_i W(X_{(i)}) | = \left| \frac{n-i}{(n-1)n} \frac{W(X_{(i)})\Delta_i}{[1 - F(X_{(i)})]} \int_{X_{(i)}}^{+\infty} H\left(\frac{X_{(i)} - y}{h} \right) f(y) dy \right|$$

$$\leq \frac{C}{n(1-F(b_0))} \mid F(X_{(i)}) - U_{(i)} \mid$$

$$\leq \frac{C}{n} \mid F(X_{(i)}) - U_{(i)} \mid,$$

where from hereafter C denotes a generic constant. Therefore, we have:

$$\mid CT^{2,1}(h) \mid \leq \frac{C}{n} \sum_{i=1}^{n} \frac{1}{n} \mid F(X_{(i)}) - U_{(i)} \mid.$$

It follows that

$$\mid CT^{2,1}(h) \mid^2 \leq \frac{C}{n^2} \sum_{i=1}^{n} \frac{1}{n} \mid F(X_{(i)}) - U_{(i)} \mid^2$$

hence

$$E \mid CT^{2,1}(h) \mid^2 \leq \frac{C}{n^2} \sum_{i=1}^{n} \frac{1}{n} E \mid F(X_{(i)}) - U_{(i)} \mid^2$$

and using Jensen's inequality, we get

$$E \mid CT^{2,1}(h) \mid^2 \leq \frac{C}{n} \left(\sum_{i=1}^{n} \frac{1}{n} E \mid F(X_{(i)}) - U_{(i)} \mid^2 \right)^{\frac{1}{2}}.$$

The proof of this case is completed using the following Lemma 4, whose proof is given below.

Lemma 4. *We have*:

$$\sum_{i=1}^{n} \frac{1}{n} E \mid F(X_{(i)}) - U_{(i)} \mid^2 = \frac{2}{6(n+1)}.$$

(ii) Case $k = 2$: We have:

$$\mid \alpha_i^{-} \Delta_i W(X_{(i)}) \mid = \left| \frac{n-i}{(n-1)n} \frac{W(X_{(i)}) \Delta_i}{F(X_{(i)})} \int_{-\infty}^{X_{(i)}} H\left(\frac{X_{(i)} - y}{h} \right) f(y) dy \right|$$

$$\leq \frac{C}{nF(a_0)} \mid F(X_{(i)}) - U_{(i)} \mid$$

$$\leq \frac{C}{n} \mid F(X_{(i)}) - U_{(i)} \mid$$

and the proof is completed using exactly the same steps as in the case $k = 1$.

(iii) Case $k = 3$: To prove this case, we will need Proposition 2 which is a consequence of Dvoretzky et al. [5] theorem.

Proposition 2. *There exists a constant* κ_0 *such that*

$$E(\sup_{x\in[0,1]} \mid \hat{V}_n(x) - V(x) \mid)^2 \leq \frac{\kappa_0}{n}.$$

Observe that

$$\int_{X_{(i)}}^{+\infty} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy = \left[H\left(\frac{X_{(i)} - y}{h}\right)F(y)\right]_{X_{(i)}}^{+\infty}$$

$$+ \int_{X_{(i)}}^{+\infty} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)F(y)dy$$

$$= \int_{X_{(i)}}^{+\infty} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)F(y)dy - \frac{F(X_{(i)})}{2}.$$

This quantity can be rewritten as

$$\int_{X_{(i)}}^{+\infty} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy = \int_{X_{(i)}}^{+\infty} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)[F(y) - F(X_{(i)})]dy$$

$$= \int_{-\infty}^{0} K(u)[F(X_{(i)} - uh) - F(X_{(i)})]du$$

$$= h\int_{-\infty}^{0} -uK(u)\, f(X_{(i)} - \nu uh)du.$$

And, since f is assumed bounded, we have

$$\left| \int_{X_{(i)}}^{+\infty} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy \right| \leq Ch.$$

We have:

$$\begin{aligned} |\beta_i^+ \Delta_i W(X_{(i)})| &= \left| \frac{(F(X_{(i)}) - \hat{V}_n F(X_{(i)}))}{[1 - F(X_{(i)})]} \Delta_i W(X_{(i)}) \right. \\ &\quad \left. \times \int_{X_{(i)}}^{+\infty} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy \right| \\ &\leq \frac{Ch|\ F(X_{(i)}) - \hat{V}_n(F(X_{(i)}))\ |}{[1 - F(b_0)]} |\ \Delta_i\ | W(X_{(i)}) \\ &\leq Ch|\ F(X_{(i)}) - \hat{V}_n F(X_{(i)})\ ||\ \Delta_i\ | W(X_{(i)}). \end{aligned}$$

Therefore,

$$|\ CT^{2,3}(h)\ | \leq ChD_n \frac{1}{n} \sum_{i=1}^{n} |\ \Delta_i\ |$$

with

$$D_n = \sup_{x \in [0,1]} |\ \hat{V}_n(x) - V(x)\ |$$

hence

$$E|\ CT^{2,3}(h)\ | \leq Ch(ED_n^2)^{\frac{1}{2}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} E(F(X_{(i)} - U_{(i)})^2}.$$

It then follows from Proposition 2 and Lemma 4 that

$$E|\ CT^{2,3}(h)\ | = O\left(\frac{h}{n}\right)$$

and this equation completes the proof of this case.

(iv) Case $k = 4$: Recall that

$$\beta_i^- = \frac{\hat{V}_n(F(X_{(i)})) - F(X_{(i)})}{F(X_{(i)})}\left[\int_{-\infty}^{X_{(i)}} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy - F(X_{(i)})\right].$$

Also, we have:

$$\begin{aligned}\int_{-\infty}^{X_{(i)}} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy &= \left[H\left(\frac{X_{(i)} - y}{h}\right)F(y)\right]_{-\infty}^{X_{(i)}} \\ &\quad + \int_{-\infty}^{X_{(i)}} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)F(y)dy \\ &= \int_{-\infty}^{X_{(i)}} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)F(y)dy + \frac{F(X_{(i)})}{2}.\end{aligned}$$

It follows that

$$\begin{aligned}&\int_{-\infty}^{X_{(i)}} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy - F(X_{(i)}) \\ &= \int_{-\infty}^{X_{(i)}} \frac{1}{h} K\left(\frac{X_{(i)} - y}{h}\right)[F(y) - F(X_{(i)})]dy \\ &= \int_{0}^{-\infty} K(u)[F(X_{(i)} - uh) - F(X_{(i)})]du \\ &= h\int_{0}^{-\infty} -uK(u)\, f(X_{(i)} - \nu uh)du.\end{aligned}$$

We thus have:

$$\left|\int_{-\infty}^{X_{(i)}} H\left(\frac{X_{(i)} - y}{h}\right) f(y)dy - F(X_{(i)})\right| \leq Ch.$$

From this inequality, we get:

$$| \beta_i^- \Delta_i W(X_{(i)}) | = \left| Ch \frac{F(X_{(i)}) - \hat{V}_n(F(X_{(i)}))}{F(X_{(i)})} \Delta_i W(X_{(i)}) \right|$$

$$\leq \frac{Ch| F(X_{(i)}) - \hat{V}_n(F(X_{(i)})) |}{F(a_0)} | \Delta_i | W(X_{(i)})$$

$$\leq Ch| F(X_{(i)}) - \hat{V}_n F(X_{(i)}) | | \Delta_i | W(X_{(i)}).$$

So, the remainder of the proof for this case follows exactly the same steps as that of the case $k = 3$.

(v) Case $k = 5$: Let

$$\omega(h) = \sup_{x \in [a_0, b_0]} \left| \int H\left(\frac{x - y}{h}\right) f(y) dy - F(x) \right|,$$

to prove this case, we will need the following Proposition 3 which is a consequence of Taylor expansion of order 2 and the fact that f' is assumed bounded.

Proposition 3. *Under the conditions of Theorem* 2, *there exists a constant* κ_1 *such that*

$$\omega(h) \leq \kappa_1 h^2.$$

Since

$$| \theta_i W(X_{(i)}) | \leq C\omega(h),$$

we have:

$$| CT^{2,5}(h) | \leq Ch\omega(h) \frac{1}{n} \sum_{i=1}^{n} | \Delta_i |.$$

It follows that

$$E| CT^{2,5}(h) | \leq C\omega(h) \sqrt{\frac{1}{n} \sum_{i=1}^{n} E(F(X_{(i)}) - U_{(i)})^2}$$

$$= \frac{C\omega(h)}{\sqrt{6(n+1)}}$$

$$= O\left(\frac{h^2}{\sqrt{n}}\right)$$

using Lemma 4 and Proposition 3. This last equation completes the proof of this case.

A.2.4. Proof of Lemma 4

Because of independence and equidistribution, we have:

$$E(F(X_{(i)}) - U_{(i)})^2 = EF(X_{(i)})^2 - 2EF(X_{(i)})EU_{(i)} + EU_{(i)}^2$$

$$= 2(EU_{(i)}^2 - (EU_{(i)})^2)$$

$$= 2\mathrm{Var}(U_{(i)}).$$

It is known that $U_{(i)}$ is a $B(i,\ n-i+1)$ (Beta distribution with parameters i and $n-i+1,$ see, for instance, Arnold et al. [2]) random variable. Therefore,

$$\mathrm{Var}(U_{(i)}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}.$$

It follows that

$$\sum_{i=1}^{n} \mathrm{Var}(U_{(i)}) = \frac{n}{6(n+1)}$$

and Lemma 4 is a consequence of this equality.

References

[1] N. Altman and C. Léger, Bandwidth selection for kernel distribution function estimation, J. Statist. Plann. Inference 46(2) (1995), 195-214.

[2] B. C. Arnold, N. Balakrishnan and H. N. Nagaraja, A first course in order statistics, Volume 54 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, Unabridged Republication of the 1992 Original, 2008.

[3] A. Bowman, P. Hall and T. Prvan, Bandwidth selection for the smoothing of distribution functions, Biometrika 85(4) (1998), 799-808.

[4] J. E. Chacón and A. Rodríguez-Casal, A note on the universal consistency of the kernel distribution function estimator, Statist. Probab. Lett. 80(17-18) (2010), 1414-1419.

[5] A. Dvoretzky, J. Kiefer and J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, Ann. Math. Statist. 27 (1956), 642-669.

[6] M. Lejeune and P. Sarda, Smooth estimators of distribution and density functions, Comput. Statist. Data Anal. 14(4) (1992), 457-471.

[7] E. Levina and P. Bickel, The earthmover's distance is the Mallows distance: some insights from statistics, 2002. http://www.stat.lsa.umich.edu/ elevia/EDM.pdf.

[8] J. S. Marron and W. Härdle, Random approximations to some measures of accuracy in nonparametric curve estimation, J. Multivariate Anal. 20(1) (1986), 91-113.

[9] A. Munk and C. Czado, Nonparametric validation of similar distributions and assessment of goodness of fit, J. R. Stat. Soc. Ser. B Stat. Methodol. 60(1) (1998), 223-241.

[10] E. A. Nadaraya, Some new estimates for distribution functions, Theory Probab. Appl. 9 (1964), 497-500.

[11] R.-D. Reiss, Nonparametric estimation of smooth distribution functions, Scand. J. Statist. 8(2) (1981), 116-119.

[12] P. Sarda, Smoothing parameter selection for smooth distribution functions, J. Statist. Plann. Inference 35(1) (1993), 65-75.

[13] G. R. Shorack and J. A. Wellner, Empirical processes with applications to statistics, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986.

[14] H. Yamato, Some statistical properties of estimators of density and distribution functions, Bull. Math. Statist. 15(1-2) (1973), 113-131.