On the uniform in bandwidth consistency of kernel–type estimators and conditional U–statistics

Julia Dony
Department of Mathematics, Free University of Brussels (VUB), Pleinlaan 2, B-1050 Brussels, Belgium.

Keywords: empirical processes, U–statistics, kernel estimation, regression estimation, Nadaraya–Watson, uniform in bandwidth consistency.
AMS: 60F15, 62G08

Abstract

Let $(X_i, Y_i), i \geq 1$ be i.i.d. random variables and denote the regression function by $m_\varphi(t) = \mathbb{E}[\varphi(Y)|X = t]$. A popular class of estimators consists of kernel–type estimators, which have been intensively studied for a long time. For estimating qualities such as $m_\varphi(\mathbf{t}) = \mathbb{E}[\varphi(Y_1, \dots, Y_m)|(X_1, \dots, X_m) = \mathbf{t}]$, one needs so–called conditional U–statistics, which were introduced by Stute. Among other things, he proved their pointwise consistency to $m_\varphi(\mathbf{t})$. We are now interested in uniform in bandwidth convergence of these estimators.

1. Kernel estimators

Let $X, X_1, X_2, \dots$ be i.i.d. random variables with values in $\mathbb{R}^d$, $d \geq 1$, and assume that the common distribution function of these variables has a Lebesgue density function, which we shall denote by f_X. A kernel $K : \mathbb{R}^d \to \mathbb{R}$ will be any measurable function that satisfies $\sup_{x\in\mathbb{R}^d} |K(x)| := \kappa < \infty$ and $\int_{\mathbb{R}^d} K(s)ds = 1$. Then the kernel density estimator of f_X based upon the sample $X_1, \dots, X_n$ and the bandwidth $0 < h < 1$ is defined as

$$f_{n,h}(x) := \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}^d. \qquad \text{[KDE]}$$

It is well known that if f_X is continuous and if one chooses a suitable bandwidth sequence $h_n \to 0$, one obtains a strongly consistent estimator $f_n \equiv f_{n,h_n}$ of f_X, i.e. one has almost surely $f_n(x) \to f_X(x)$ for all $x \in \mathbb{R}^d$. It is then natural to investigate other types of convergence, like for instance uniform convergence, or convergence with respect to weighted supremum norms (see Giné, Koltchinskii and Zinn [8], Dony and Einmahl [1]) and to ask what convergence rates are feasible.

For proving such results one usually writes the difference $f_n(x) - f_X(x)$ as the sum of a probabilistic term $f_n(x) - \mathbb{E}f_n(x)$ and a deterministic term

$\mathbb{E}f_n(x) - f_X(x)$, the so–called bias. The order of the bias depends on smoothness properties of f_X only, whereas the first (random) term can be studied via empirical process techniques. Giné and Guillou [7] have shown that if the density f_X is bounded and the kernel K is "regular", one has

$$\|f_n - \mathbb{E}f_n\|_\infty = O\big(\sqrt{|\log h_n|/nh_n}\big), \quad \text{a.s.}, \tag{1}$$

provided $h_n \to 0$ satisfies the following regularity conditions,

$$\frac{nh_n}{\log n} \to \infty, \quad \frac{h_n}{h_{2n}} = O(1), \quad \text{and} \quad \frac{\log(1/h_n)}{\log\log n} \to \infty,$$

and where $\|\cdot\|_\infty$ denotes the supremum norm on $\mathbb{R}^d$. Moreover, this rate cannot be improved. Interestingly, one does not need continuity of f_X for this result, but note that it is of course needed for controlling the bias.

Now let $Y, Y_1, Y_2, \ldots$ be a sequence of r–dimensional random variables, $(r \geq 1)$, so that the random vectors $(X, Y), (X_1, Y_1), \ldots$ are i.i.d. with common joint Lebesgue density function f. It is also of great interest to estimate the regression function $m_\varphi(x) = \mathbb{E}[\varphi(Y)|X = x]$, where $\varphi : \mathbb{R}^r \to \mathbb{R}$ is a suitable mapping. Therefore, a possible kernel–type estimator which has been extensively studied, is given by

$$\hat{m}_{n,h,\varphi}(x) = \frac{\sum_{i=1}^n \varphi(Y_i) K\big(\frac{x - X_i}{h}\big)}{\sum_{i=1}^n K\big(\frac{x - X_i}{h}\big)}, \tag{NW}$$

and is referred to as a Nadaraya–Watson type estimator. Exact convergence rates uniformly on compact subsets of $\mathbb{R}^d$ have been obtained in the case of deterministic bandwidth sequences by Einmahl and Mason [4].

Another popular way of estimating the regression function is by using local polynomial estimators. This method was introduced by Fan and Gijbels [6], and is based on the idea of fitting a polynomial of degree $p \geq 0$ to the observations by the method of weighted least squares. For simplicity, let $r = d = 1$ and $\varphi(y) = y$, and assume $m_\varphi(x)$ to be $p + 1$ times differentiable. Then the local polynomial regression function estimator of degree $p \geq 0$ is defined as

$$\hat{m}_{n,h}^{(p)}(x) = \mathbf{e}_1 \boldsymbol{\beta}^*(x), \tag{LP}$$

where $\mathbf{e}_1$ is the projection on the first component, and $\boldsymbol{\beta}^*(x) \in \mathbb{R}^{p+1}$ is defined as the solution of the following weighted least squares problem,

$$\boldsymbol{\beta}^*(x) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \ \frac{1}{nh} \sum_{i=1}^n [Y_i - \sum_{j=0}^p \beta_j (X_i - x)^j]^2 K\Big(\frac{x - X_i}{h}\Big).$$

2. Conditional U–statistics

A much wider class of estimators is the class of the so–called U–statistics. For a measurable space $(S, \mathcal{S})$ and S–valued random variables $Z_1, \ldots, Z_n$, one defines a U–statistic with kernel $G : S^k \to \mathbb{R}$ as

$$U_n^{(k)}(G) := \frac{(n-k)!}{n!} \sum_{\mathbf{i} \in I_n^k} G(Z_{i_1}, \ldots, Z_{i_k}), \quad 1 \leq k \leq n,$$

where $I_n^k = \{(i_1, \ldots, i_k) : 1 \leq i_j \leq n, i_j \neq i_l \text{ if } j \neq l\}$. To generalize the Nadaraya–Watson type estimators [NW] for the regression function, Stute [9] introduced *conditional* U–statistics, that he defined as

$$\hat{m}_{n,\varphi}(\mathbf{t}, h_n) = \frac{\sum_{\mathbf{i} \in I_n^m} \varphi(Y_{i_1}, \ldots, Y_{i_m}) K\left(\frac{t_1 - X_{i_1}}{h_n}\right) \cdots K\left(\frac{t_m - X_{i_m}}{h_n}\right)}{\sum_{\mathbf{i} \in I_n^m} K\left(\frac{t_1 - X_{i_1}}{h_n}\right) \cdots K\left(\frac{t_m - X_{i_m}}{h_n}\right)}, \qquad \text{[CUS]}$$

with K a kernel function and $0 < h_n < 1$ a bandwidth sequence. If h_n goes to zero at a certain rate, he established strong pointwise consistency to $m_\varphi(\mathbf{t}) := \mathbb{E}[\varphi(Y_1, \ldots, Y_m) | (X_1, \ldots, X_m) = \mathbf{t}]$, $\mathbf{t} \in \mathbb{R}^m$.

3. Motivation of the problem

It is well known that, although the choice of the kernel K is not really important, choosing a good bandwidth $h > 0$ is crucial for the consistency of the estimators. Unfortunately, the optimal bandwidth depends on the data, such that one needs also to consider estimators based upon *random* (or statistical) bandwidth sequences. For this reason, extensions of the consistency results that were based upon deterministic bandwidth sequences h_n are needed.

For the estimators [NW] and [KDE], optimal results have been obtained by Einmahl and Mason [5], while the consistency results for regression estimators such as [LP] and [CUS] based on random bandwidths, have been established in Dony, Einmahl and Mason [2], and Dony and Mason [3] respectively. The purpose of this paper is to present shortly these uniform in bandwidth results, without proofs. Local polynomial estimators are handled in the next section, and the conditional U–statistics are discussed in the last section.

4. Consistency of local polynomial regression estimators

Recall that for simplicity $r = d = 1$ and $\varphi(y) = y$. Assume now that the kernel K has support contained in $[-1/2, 1/2]$, and that f_X is continuous and strictly positive on $J = I^\eta$, the η–neighborhood of I, a compact interval that will be specified later on. Then, for $x \in \mathbb{R}$, define

$$\tilde{f}_{n,h,j}(x) := \frac{1}{nh}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K\left(\frac{x - X_i}{h}\right), \quad j = 0, \ldots, 2p,$$

$$\tilde{r}_{n,h,j}(x) := \frac{1}{nh}\sum_{i=1}^{n} Y_i \left(\frac{X_i - x}{h}\right)^j K\left(\frac{x - X_i}{h}\right), \quad j = 0, \ldots, p,$$

and set for $\mu_j := \int_{\mathbb{R}} (-u)^j K(u)du$,

$$f_j(x) := \mu_j f_X(x), \quad \text{and} \quad r_j(x) := \mu_j \int_{\mathbb{R}} y f(x,y)dy.$$

Then for some compact interval $I \subseteq \mathbb{R}$ and any sequences $a_n < b_n$ with $b_n \to 0$, one can show that uniformly in $h \in [a_n, b_n]$, $\|\mathbb{E}\tilde{f}_{n,h,j} - f_j\|_\infty \to 0$ and $\|\mathbb{E}\tilde{r}_{n,h,j} - r_j\|_I \to 0$. Moreover, the following result can be inferred from Theorems 1 and 2 of Dony, Einmahl and Mason [2].

Theorem 1 (Dony, Einmahl and Mason, 2006) *Under some usual assumptions, one has for all smooth functions* $\Phi : \mathbb{R}^{3p+2} \to \mathbb{R}$ *and suitable sequences* $0 < a_n < b_n \to 0$ *that,*

$$\sup_{a_n \le h \le b_n} \|\Phi(\tilde{\mathbf{f}}_{n,h}, \tilde{\mathbf{r}}_{n,h}) - \Phi(\tilde{\mathbf{f}}, \tilde{\mathbf{r}})\|_I \longrightarrow 0, \quad a.s.,$$

where $\tilde{\mathbf{f}}_{n,h} = (\tilde{f}_{n,h,0}, \ldots, \tilde{f}_{n,h,2p})$*,* $\tilde{\mathbf{r}}_{n,h} = (\tilde{r}_{n,h,0}, \ldots, \tilde{r}_{n,h,p})$*,* $\tilde{\mathbf{f}} = (f_0, \ldots, f_{2p})$ *and* $\tilde{\mathbf{r}} = (r_0, \ldots, r_p)$*.*

As a consequence of this result, the uniform in bandwidth consistency of the local polynomial estimators [LP] was established in Dony, Einmahl and Mason [2] by expressing $\hat{m}_{n,h}^{(p)}$ as functions of $\tilde{\mathbf{f}}_{n,h}$ and $\tilde{\mathbf{r}}_{n,h}$. As an example, consider the case $p = 0$, for which [LP] = [NW]. It is readily checked that

$$\hat{m}_{n,h}^{(0)}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} = \frac{\tilde{r}_{n,h,0}(x)}{\tilde{f}_{n,h,0}(x)},$$

such that Theorem 1 with $\Phi(x_1, x_2) = x_2/x_1$ implies that with probability one and uniformly in $x \in I$, $\sup_{a_n \le h \le b_n} |\hat{m}_{n,h}^{(0)}(x) - \mathbb{E}[Y|X = x]| \to 0$, proving the uniform in bandwidth consistency of the Nadaraya-Watson estimator.

5. Consistency of conditional U–statistics

Recall the definition of a conditional U–statistics in [CUS] and consider the function on $\mathbb{R}^m \times \mathbb{R}^m$ defined as

$$G_{\varphi,h,\mathbf{t}}(\mathbf{x}, \mathbf{y}) := \frac{1}{h^m}\varphi(\mathbf{y})\prod_{j=1}^{m} K\big(\frac{t_j - x_j}{h}\big) =: \varphi(\mathbf{y})\widetilde{K}_h(\mathbf{t} - \mathbf{x}), \quad \mathbf{x}, \mathbf{y}, \mathbf{t} \in \mathbb{R}^m,$$

where $0 < h < 1$ is a bandwidth and $\varphi : \mathbb{R}^m \to \mathbb{R}$ a measurable function satisfying $\mathbb{E}\varphi^2(Y_1, \ldots, Y_m) < \infty$. Then let $U_n(\varphi, h, \mathbf{t})$ denote the U–statistic with kernel $G_{\varphi,h,\mathbf{t}}(\mathbf{x}, \mathbf{y})$, such that $\hat{m}_{n,\varphi}(\mathbf{t}, h)$ can be rewritten as

$$\hat{m}_{n,\varphi}(\mathbf{t}, h) = \frac{\sum_{\mathbf{i}\in I_n^m} \varphi(\mathbf{Y_i})\widetilde{K}_h(\mathbf{t} - \mathbf{X_i})}{\sum_{\mathbf{i}\in I_n^m} \widetilde{K}_h(\mathbf{t} - \mathbf{X_i})} = \frac{U_n(\varphi, h, \mathbf{t})}{U_n(1, h, \mathbf{t})},$$

where $U_n(1, h, \mathbf{t})$ denotes the U–statistic $U_n(\varphi, h, \mathbf{t})$ with $\varphi \equiv 1$. By studying the asymptotic behavior of $U_n(\varphi, h, \mathbf{t})$, Dony and Mason [3] established the strong uniform consistency of conditional U–statistics to the general regression function $m_\varphi(\mathbf{t})$, uniformly in bandwidth. In order to obtain the convergence rates, another more appropriate centering factor than the expectation (which may not exist or be difficult to compute) has been considered, namely

$$\widehat{\mathbb{E}}\hat{m}_{n,\varphi}(\mathbf{t}, h) := \frac{\mathbb{E}U_n(\varphi, h, \mathbf{t})}{\mathbb{E}U_n(1, h, \mathbf{t})}.$$

In the next Theorem, the convergence rate of the process $\hat{m}_{n,\varphi}(\mathbf{t}, h) - \widehat{\mathbb{E}}\hat{m}_{n,\varphi}(\mathbf{t}, h)$ to zero, and the consistency of $\hat{m}_{n,\varphi}(\mathbf{t}, h)$, are stated, uniformly in bandwidth.

Theorem 2 (Dony and Mason, 2007) *Under some appropriate conditions on f_X, K and the class $\mathcal{F}$, it follows that for all $c > 0$ and all sequences $0 < \widetilde{a}_n \leq b_n \to 0$, there exists a constant $0 < M(c) < \infty$ such that,*

$$\limsup_{n\to\infty} \sup_{\widetilde{a}_n \leq h < b_n} \sup_{\varphi\in\mathcal{F}} \sup_{\mathbf{t}\in I^m} \frac{\sqrt{nh^m}|\hat{m}_{n,\varphi}(\mathbf{t}, h) - \widehat{\mathbb{E}}\hat{m}_{n,\varphi}(\mathbf{t}, h)|}{\sqrt{|\log h| \vee \log\log n}} \leq M(c), \quad a.s.,$$

where I is some compact interval in $\mathbb{R}$, and $\widetilde{a}_n = c(\log n/n)^{\gamma/m}$ with $\gamma = 1$ or $\gamma = 1 - 2/p$ depending on whether the class $\mathcal{F}$ is bounded or not. Moreover, for all sequences $0 < a_n \leq b_n < 1$ satisfying $b_n \to 0$ and $na_n/\log n \to \infty$,

$$\sup_{a_n \leq h < b_n} \sup_{\varphi\in\mathcal{F}} \sup_{\mathbf{t}\in I^m} |\hat{m}_{n,\varphi}(\mathbf{t}, h) - m_\varphi(\mathbf{t})| \longrightarrow 0, \quad a.s.$$

Acknowledgements: Research financed by a PhD grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

6. Bibliography

[1] Dony, J. and Einmahl, U. (2006). *Weighted uniform consistency of kernel density estimators with general bandwidth sequences.* Electr. J. Probab. **11,** paper 33, 844–859.

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

[3] Dony, J. and Mason, D.M. (2007). *Uniform in bandwidth consistency of conditional U–statistics.* Submitted.

[4] 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.

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

[6] Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. *Monographs on Statistics and Applied Probability*, 66. Chapman & Hall, London.

[7] Giné, E. and Guillou, A. (2002). *Rates of strong consistency for multivariate kernel density estimators.* Ann. Inst. H. Poincaré. **38,** 907–92.

[8] Giné, E., Koltchinskii, V. and Zinn, J. (2004). *Weighted uniform consistency of kernel density estimators.* Ann. Probab. **32,** 2570–2605.

[9] Stute, W. (1991). *Conditional U–statistics.* Ann. Probab. **19**, 812–825.