*The Annals of Probability*
1991, Vol. 19, No. 2, 812–825

# CONDITIONAL *U*-STATISTICS

BY WINFRIED STUTE

*University of Giessen*

We introduce a class of so-called conditional $U$-statistics, which generalize the Nadaraya–Watson estimate of a regression function in the same way as Hoeffding's classical $U$-statistic is a generalization of the sample mean. Asymptotic normality and weak and strong consistency are proved.

**1. Introduction.** In this paper we introduce a class of so-called conditional $U$-statistics, which may be viewed as a generalization of the Nadaraya–Watson estimate of a regression function. This extension is similar to Hoeffding's (1948) generalization of sample means to what we now call $U$-statistics.

To be precise, assume that $(X_i, Y_i)$, $1 \le i \le n$, are i.i.d. random vectors in some Euclidean space. For notational convenience, we shall restrict ourselves to real $X$'s, though our results may be generalized to the multivariate case without difficulties. Let $h$ be any function of $k$ variates (the $U$ kernel), $k \le n$, such that $h(Y_1, \ldots, Y_k)$ is integrable. We are interested in the estimation of

$$m(x_1, \ldots, x_k) = \mathbb{E}\big[h(Y_1, \ldots, Y_k) | X_1 = x_1, \ldots, X_k = x_k\big].$$

When $k = 1$ and $h = Id$, then clearly

$$m(x_1) = \mathbb{E}\big[Y_1 | X_1 = x_1\big],$$

the regression of $Y_1$ given $X_1 = x_1$. Examples for $k \ge 2$ will be discussed in greater detail in Section 4.

For estimation of $m(x_1)$, Nadaraya (1964) and Watson (1964) independently proposed

$$m_n(x_1) = \frac{\sum_{i=1}^n Y_i K\big[(x_1 - X_i)/a_n\big]}{\sum_{i=1}^n K\big[(x_1 - X_i)/a_n\big]}.$$

Here $K$ is a so-called smoothing kernel and $(a_n)_n$ is a sequence of bandwidths tending to zero at appropriate rates. For $K$, at least integrability w.r.t. Lebesgue measure with nonvanishing integral $\int K(u)\,du \ne 0$ is assumed. Since, unlike the case of density estimation, $m_n$ has ratio structure, $\int K(u)\,du = 1$ is not required, but is usually assumed for the sake of convenience. Often $K$ is chosen so as to satisfy further smoothness and tail conditions, which, for example, guarantee smoothness of $m_n$. For an arbitrary $k$, we shall consider

Received April 1989; revised October 1989.
*AMS* 1980 *subject classifications*. 60F05, 60F15, 62J99.
*Key words and phrases*. Conditional $U$-statistics, smoothing, asymptotic normality, strong convergence.

statistics of the form

$$u_n(\mathbf{x}) \equiv u_n(x_1, \ldots, x_k) = \frac{\Sigma_\beta h(Y_{\beta_1}, \ldots, Y_{\beta_k}) \Pi_{j=1}^k K\left[(x_j - X_{\beta_j})/a_n\right]}{\Sigma_\beta \Pi_{j=1}^k K\left[(x_j - X_{\beta_j})/a_n\right]}.$$

Here summation extends over all permutations $\beta = (\beta_1, \ldots, \beta_k)$ of length $k$, that is, over all pairwise distinct $\beta_1, \ldots, \beta_k$ taken from $1, \ldots, n$. In Section 2, we shall derive the limit distribution of $u_n(\mathbf{x})$ when properly standardized (Theorem 1). In Section 3, weak and strong consistency are proved (Theorems 2 and 3). Examples are discussed in Section 4.

There are two remarks in order about the choice of $h$ and $K$. In classical $U$-statistics theory, one may assume w.l.o.g. that $h$ is symmetric. Since the nominator of $u_n(\mathbf{x})$ is a $U$-statistic with kernel $h \prod K$, such a symmetrization would also involve the factor $\prod K$. Since the role of $h$ differs from that of $\prod K$ it turns out, after a moment of thought, that little if nothing would be gained from a symmetrization. So, in this paper, no symmetry of $h$ will be imposed. In several places, we shall make use of the variance formula for not necessarily symmetric $U$-statistics. For the ease of reference, this is included in the Appendix.

Together with the variance formula, we shall frequently apply a version of the differentiation theorem. For the rectangular kernel, Theorem 10.49 in Wheeden and Zygmund (1977) is appropriate, if one is interested in limit results which hold almost everywhere without any continuity assumption on the function to be smoothed. Greblicki, Krzyźak and Pawlak (1984) present an extension to a slightly larger class of (nonproduct) kernels. Theorem 2 on page 63 of Stein (1970) is widely applicable if the underlying measure is Lebesgue-continuous.

In Section 2 of this paper we preferred to state our results under the assumption that $\mathbf{x}$ is a point of continuity for the function of interest. This allows for consideration of arbitrary measures $\mu$ as well as for a large class of kernels $K$. It will be easy to restate the main results so that one of the above mentioned differentiation results may be applied. This is exemplified for the consistency results proved in Section 3. Note, however, that for Theorem 1, it will be essential to have a product kernel.

Needless to say, that for $d$-dimensional $X$'s, $K$ will be a kernel on $\mathbb{R}^d$.

**2. Asymptotic normality.** Let $\mathbf{x} = (x_1, \ldots, x_k)$ be fixed throughout. In this section, $h$ will be assumed square-integrable. Set

$$U_n(h, \mathbf{x}) \equiv U_n$$
$$= \frac{(n-k)!}{n!} \sum_\beta h(Y_{\beta_1}, \ldots, Y_{\beta_k}) \prod_{j=1}^k K\left[\frac{x_j - X_{\beta_j}}{a_n}\right] \Big/ \prod_{j=1}^k \mathbb{E} K\left[\frac{x_j - X_1}{a_n}\right].$$

Hence

$$u_n(\mathbf{x}) = U_n(h, \mathbf{x}) / U_n(1, \mathbf{x}).$$

Note that $U_n(h, \mathbf{x})$, for each $n \geq k$, is a classical $U$-statistic with a kernel depending on $n$. The expectation of $U_n$ equals

$$\theta_n = \int m(z_1, \ldots, z_k) \prod_{j=1}^{k} K\left[\frac{x_j - z_j}{a_n}\right] \mu(dz_1) \cdots \mu(dz_k) \Big/ \prod_{j=1}^{k} \mathbb{E}K\left[\frac{x_j - X_1}{a_n}\right],$$

with $\mu$ denoting the distribution of $X_1$. The Hájek projection $\hat{U}_n$ of $U_n$ satisfies

$$\hat{U}_n - \theta_n = n^{-1} \sum_{i=1}^{n} \bar{h}_n(X_i, Y_i),$$

where

$$\bar{h}_n(x, y) = \sum_{j=1}^{k} \left[h_{nj}(x, y) - \theta_n\right]$$

and $h_{nj}$ is defined by

$$h_{nj}(x, y) = N^{-1} \int h(Y_1, \ldots, Y_{j-1}, y, Y_{j+1}, \ldots, Y_k) \times \prod_{\substack{r=1 \\ r \neq j}}^{k} K\left[\frac{x_r - X_r}{a_n}\right] K\left[\frac{x_j - x}{a_n}\right] d\mathbb{P}.$$

For brevity's sake we have set

$$N = \prod_{j=1}^{k} \mathbb{E}K\left[\frac{x_j - X_1}{a_n}\right]$$

and $\mathbb{P} \equiv$ underlying probability measure. By independence,

$$n\mathbb{E}\left(\hat{U}_n - \theta_n\right)^2 = \mathbb{E}\bar{h}_n^2(X, Y) = \sum_{j=1}^{k} \sum_{l=1}^{k} \mathbb{E}\left[h_{nj}(X, Y) - \theta_n\right]\left[h_{nl}(X, Y) - \theta_n\right].$$

In the following $(X, Y)$, $(X_i, Y_i)_{1 \leq i \leq 2k}$ are assumed to be i.i.d. Then

$$\mathbb{E}h_{nj}(X, Y)h_{nl}(X, Y) = N^{-2} \int h(Y_1, \ldots, Y_{j-1}, Y, Y_{j+1}, \ldots, Y_k) \times h(Y_{k+1}, \ldots, Y_{k+l-1}, Y, Y_{k+l+1}, \ldots, Y_{2k}) \times \prod_{\substack{r=1 \\ r \neq j}}^{k} K\left[\frac{x_r - X_r}{a_n}\right] \prod_{\substack{s=1 \\ s \neq l}}^{k} K\left[\frac{x_s - X_{k+s}}{a_n}\right] K\left[\frac{x_l - X}{a_n}\right] K\left[\frac{x_l - X}{a_n}\right] d\mathbb{P}.$$

Set, when $x_j = x_l$,

$$m_{jl}(\mathbf{x}) = \mathbb{E}\big[h(Y_1, \ldots, Y_{j-1}, Y, Y_{j+1}, \ldots, Y_k) \times h(Y_{k+1}, \ldots, Y_{k+l-1}, Y, Y_{k+l+1}, \ldots, Y_{2k}) | X_r = x_r \text{ for } r \neq j,\ X_{k+s} = x_s \text{ for } s \neq l \text{ and } X = x_j = x_l\big],$$

and zero when $x_j \neq x_l$. Now, if $K$ has compact support,

$$K\left[\frac{x_j - X}{a_n}\right] K\left[\frac{x_l - X}{a_n}\right] = 0,$$

whenever $x_j \neq x_l$ and $n$ is sufficiently large. Thus

$$\mathbb{E}h_{nj}(X, Y)h_{nl}(X, Y) = 0 \quad \text{when } x_j \neq x_l.$$

When $x_j = x_l$ and $\mathbf{x}$ is a point of continuity for $m_{jl}$, then

$$\frac{\mathbb{E}^2 K\big[(x_j - X_1)/a_n\big]}{\mathbb{E}K^2\big[(x_j - X_1)/a_n\big]} \mathbb{E}\big[h_{nj}(X, Y)h_{nl}(X, Y)\big] \to m_{jl}(\mathbf{x}), \tag{2.1}$$

by an obvious differentiation argument. Now, when $X_1$ admits a density $f$ with $f(x_j) = f(x_l) > 0$, then (provided $f$ is continuous at $x_j = x_l$)

$$\frac{\mathbb{E}^2 K\big[(x_j - X_1)/a_n\big]}{\mathbb{E}K^2\big[(x_j - X_1)/a_n\big]} \sim a_n \frac{f(x_j)}{\int K^2(u)\, du},$$

that is,

$$a_n \mathbb{E}\big[h_{nj}(X, Y)h_{nl}(X, Y)\big] \to m_{jl}(\mathbf{x}) \int K^2(u)\, du / f(x_j).$$

Furthermore, when $m$ is bounded in a neighborhood of $\mathbf{x}$, then $\theta_n$, $n \geq 1$, is bounded so that $a_n \theta_n^2 \to 0$. We may conclude that

$$\lim_{n \to \infty} n a_n \mathbb{E}\big[\hat{U}_n - \theta_n\big]^2 = \sigma^2 = \sigma^2(h), \tag{2.2}$$

where

$$\sigma^2(h) = \sum_{j=1}^{k} \sum_{l=1}^{k} 1_{\{x_j = x_l\}} m_{jl}(\mathbf{x}) \int K^2(u)\, du / f(x_j). \tag{2.3}$$

In the following lemma we shall state some conditions under which $\hat{U}_n$ is asymptotically normal. For this we need also consider the functions

$$m_{jlm}(z_1, \ldots, z_{j-1}, z, \ldots, z_k; z_{k+1}, \ldots, z_{k+l-1}, z, \ldots, z_{2k}; z_{2k+1}, \ldots, z_{2k+m-1}, z, \ldots, z_{3k})$$
$$:= \mathbb{E}\big[|h(Y_1, \ldots, Y_{j-1}, Y, \ldots, Y_k)h(Y_{k+1}, \ldots, Y_{k+l-1}, Y, \ldots, Y_{2k}) \times h(Y_{2k+1}, \ldots, Y_{2k+m-1}, Y, \ldots, Y_{3k})| \,|X_i = z_i \quad \text{for } 1 \leq i \leq 3k,\ i \neq j, k+1, 2k+m,\ X = z\big].$$

LEMMA 2.1. *Assume that*

(i) $a_n \to 0$ *and* $na_n \to \infty$,
(ii) $K$ *is bounded and has compact support,*
(iii) $\mathbf{x}$ *is a point of continuity for each* $m_{jl}$,
(iv) $f$ *is continuous at each* $x_j$, $1 \le j \le k$, *with* $f(x_j) > 0$,
(v) $m$ *is bounded in a neighborhood of* $\mathbf{x}$,
(vi) $m_{jlm}(\cdot;\cdot;\cdot)$ *is bounded in a neighborhood of* $(\mathbf{x}, \mathbf{x}, \mathbf{x})$ *for all* $1 \le j, l, m \le k$.

*Then*

$$(na_n)^{1/2}\left[\hat{U}_n - \theta_n\right] \to \mathcal{N}(0, \sigma^2) \quad \textit{in distribution},$$

*where* $\sigma^2$ *is given by* (2.3).

PROOF. We verify Ljapunov's condition for third moments. It is easy to see that this amounts to showing

$$n^{-1/2}a_n^{3/2}\mathbb{E}|\overline{h}_n(X, Y)|^3 \to 0. \tag{2.4}$$

Since

$$|a - b|^3 \le 3(|a|^3 + |a|^2|b| + |a||b|^2 + |b|^3),$$

it turns out that an upper bound for the absolute third moment of $\overline{h}_n(X, Y)$ is dominated by sums of the form

$$\mathbb{E}|h_{nj}(X, Y)h_{nl}(X, Y)h_{nm}(X, Y)|.$$

Similar to before, we may restrict ourselves to triples $(j, l, m)$ such that $x_j = x_l = x_m$. Under (vi), the last integral is easily seen to be of the order $a_n^{-2}$. Since $na_n \to \infty$, this proves (2.4). □

To show that $U_n$ has the same asymptotic distribution as $\hat{U}_n$, we want to bound the variance of $U_n - \hat{U}_n$. In this context, we shall have to deal with integrals [cf. (A1)] of the form

$$I(\Delta_1, \Delta_2) = \mathbb{E}\left[g(Z_{i_1}, \ldots, Z_{i_k})g(Z_{j_1}, \ldots, Z_{j_k})\right],$$

where $\Delta_1$ and $\Delta_2$ are positions of some length $1 \le r \le k$, and the $i$'s in position $\Delta_1$ coincide with the $j$'s in position $\Delta_2$, and are pairwise distinct otherwise. Rather than $I(\Delta_1, \Delta_2)$, to bound the variances, we consider

$$m_{\Delta_1, \Delta_2}(\mathbf{x}_1, \mathbf{x}_2) := \mathbb{E}\Big[\big|h(Y_{i_1}, \ldots, Y_{i_k})h(Y_{j_1}, \ldots, Y_{j_k})\big|(X_{i_1}, \ldots, X_{i_k}) = \mathbf{x}_1, (X_{j_1}, \ldots, X_{j_k}) = \mathbf{x}_2\Big].$$

In addition to the assumptions of Lemma 2.1, we shall assume that

(vii) $m_{\Delta_1, \Delta_2}$ is bounded in a neighborhood of $(\mathbf{x}, \mathbf{x})$.

LEMMA 2.2. *Under the assumptions of Lemma* 2.1 *and* (vii),

$$(na_n)^{1/2}[U_n - \theta_n] \to \mathcal{N}(0, \sigma^2) \quad \textit{in distribution}.$$

PROOF. In view of 2.1, it suffices to show

$$(na_n)^{1/2}\left[U_n - \hat{U}_n\right] \to 0 \quad \text{in } L^2.$$

For this, we need the variance formula for a (centered) $U$-statistic

$$V_n = \frac{(n-k)!}{n!} \sum_{\beta} \frac{\tilde{h}(Z_{\beta_1}, \ldots, Z_{\beta_k})}{N}$$

based on a not necessarily symmetric $U$ kernel $\tilde{h}$ [cf. (A1)]:

$$\operatorname{Var}(V_n) = \left[\frac{(n-k)!}{n!}\right]^2 \sum_{r=1}^{k} \frac{(n-r)!}{(n-2k+r)!} \sum^{(r)} \frac{I(\Delta_1, \Delta_2)}{N^2}.$$

Here

$$I(\Delta_1, \Delta_2) = \int \tilde{h}(z_1, \ldots, z_k)\tilde{h}(y_1, \ldots, y_k)F(dz_1) \cdots F(dz_{2k-r})$$

and the $y$'s in position $\Delta_2$ coincide with the $z$'s in position $\Delta_1$ and are taken from $z_{k+1}, \ldots, z_{2k-r}$ otherwise. Furthermore, $\Sigma^{(r)}$ denotes summation over all positions $\Delta_1, \Delta_2$ with cardinality $r$ and $F$ is the common d.f. of the $Z$'s. When applied to $V_n = U_n - \hat{U}_n$, upon recalling $\tilde{h}$ from Serfling [(1980), page 188] (in the symmetric case), we get

$$\Sigma^{(1)} I(\Delta_1, \Delta_2) = 0.$$

Furthermore, by (vii),

$$N^{-2} I(\Delta_1, \Delta_2) = O(a_n^{-r}) \quad \text{for each } 2 \le r \le k.$$

In summary,

$$\begin{aligned} na_n \operatorname{Var}\left(U_n - \hat{U}_n\right) &= O\left[na_n \sum_{r=2}^{k} \binom{n}{k}^{-1} \binom{k}{r} \binom{n-k}{k-r} a_n^{-r}\right] \\ &= O\left[\sum_{r=2}^{k} (na_n)^{1-r}\right] = O\left[(na_n)^{-1}\right] = o(1). \end{aligned}$$

This completes the proof. □

In the following, we shall investigate the asymptotic behavior of the two-dimensional vector

$$(U_n(h_1, \mathbf{x}) - \theta_n(h_1), U_n(h_2, \mathbf{x}) - \theta_n(h_2)),$$

where $h_1$ and $h_2$ are two $U$ kernels satisfying the smoothness assumptions of 2.2. We would like to apply the Cramér–Wold device. So, let $c_1, c_2$ denote any

two real numbers. Clearly,

$$c_1U_n(h_1,\mathbf{x}) + c_2U_n(h_2,\mathbf{x}) = U_n(c_1h_1 + c_2h_2,\mathbf{x}) \equiv U_n(h,\mathbf{x}),$$

so that Lemma 2.2 applies. Specification of $\sigma^2(h)$ immediately leads to:

LEMMA 2.3. *Under the stated assumptions,*

$$(na_n)^{1/2}[U_n(h_1,\mathbf{x}) - \theta_n(h_1), U_n(h_2,\mathbf{x}) - \theta_n(h_2)] \to \mathcal{N}(\mathbf{0},\Sigma)$$

*in distribution, with*

$$\Sigma = \begin{bmatrix} \sigma^2(h_1,h_1) & \sigma^2(h_1,h_2) \\ \sigma^2(h_1,h_2) & \sigma^2(h_2,h_2) \end{bmatrix}$$

*and where for two functions g and h,*

$$\sigma^2(g,h) = \sum_{j=1}^{k}\sum_{l=1}^{k} 1_{\{x_j=x_l\}} m_{jl}^{gh}(\mathbf{x}) \int K^2(u)\,du/f(x_j)$$

*and*

$$m_{jl}^{gh}(\mathbf{x}) = \mathbb{E}[g(Y_1,\ldots,Y,\ldots,Y_k)h(Y_{k+1},\ldots,Y,\ldots,Y_{2k})|\cdots],$$

*with Y entering in the jth and lth positions.*

From the preceding lemma it is now easy to deduce the limit distribution of $u_n(\mathbf{x})$.

THEOREM 1. *Under the assumptions of Lemma* 2.2, *with*

(v′) *m is continuous at* $\mathbf{x}$

*rather than* (v), *we have*

$$(na_n)^{1/2}[u_n(\mathbf{x}) - \mathbb{E}U_n(h,\mathbf{x})] \to \mathcal{N}(0,\rho^2) \quad \textit{in distribution},$$

*where*

$$\rho^2 = \sum_{j=1}^{k}\sum_{l=1}^{k} 1_{\{x_j=x_l\}}[m_{jl}^{hh}(\mathbf{x}) - m^2(\mathbf{x})]\int K^2(u)\,du/f(x_j). \tag{2.5}$$

It is instructive to compare the values of $\rho^2$ for the two cases $x_1 = x_2$ and $x_1 \neq x_2$, when $k = 2$, say. Irrespective of different values of $m_{jl}^{hh}$ and $f$, the case $x_1 \neq x_2$ only gives rise to two summands, while the other two vanish. The tendency toward a larger variance in the case $x_1 = x_2$ is due to the fact that only data from a (single) neighborhood (of $x_1$) are used, while for $x_1 \neq x_2$ data from two eventually disjoint sets are incorporated.

PROOF. We have

$$u_n(\mathbf{x}) = U_n(h,\mathbf{x})/U_n(1,\mathbf{x}).$$

Define

$$g(x_1, x_2) = x_1/x_2 \quad \text{for } x_2 \neq 0.$$

Then

$$D = \left[\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}\right] = \left[x_2^{-1}, -x_1 x_2^{-2}\right].$$

Since, by continuity of $m$ at $\mathbf{x}$,

$$\mathbb{E}U_n(h, \mathbf{x}) \to m(\mathbf{x})$$

and

$$\mathbb{E}U_n(1, \mathbf{x}) = 1,$$

we may infer from Lemma 2.3

$$(na_n)^{1/2}[u_n(\mathbf{x}) - \mathbb{E}U_n(h, \mathbf{x})] \to \mathcal{N}(0, \rho^2),$$

where

$$\rho^2 = (1, -m(\mathbf{x}))\Sigma\begin{bmatrix} 1 \\ -m(\mathbf{x}) \end{bmatrix}.$$

It is easy to see that $\rho^2$ is of the form (2.5). □

Under appropriate smoothness assumptions on the marginal density $f$ and the function $m$, Theorem 1 immediately yields asymptotic normality of $u_n(\mathbf{x}) - m(\mathbf{x})$. Let $m$ admit an expansion

$$m(\mathbf{y} + \boldsymbol{\Delta}) = m(\mathbf{y}) + \{m'(\mathbf{y})\}^t\boldsymbol{\Delta} + \tfrac{1}{2}\boldsymbol{\Delta}^t\{m''(\mathbf{y})\}\boldsymbol{\Delta} + o(\boldsymbol{\Delta}^t\boldsymbol{\Delta}) \tag{2.6}$$

as $\boldsymbol{\Delta} \to \mathbf{0}$, for all $\mathbf{y}$ in a neighborhood of $\mathbf{x}$.

Also assume that

$$f \text{ is twice differentiable in neighborhoods of } x_j,\ 1 \leq j \leq k, \tag{2.7}$$

and

$$K \text{ is symmetric at zero.} \tag{2.8}$$

Then

$$a_n^{-2}[\mathbb{E}U_n(h, \mathbf{x}) - m(\mathbf{x})] = \tfrac{1}{2}\left[\int \prod_{i=1}^{k} K(y_i)\mathbf{y}^t\{r''(\mathbf{x})\}\mathbf{y}\, d\mathbf{y}/\tilde{f}(\mathbf{x}) - \int \prod_{i=1}^{k} K(y_i)\mathbf{y}^t\{\tilde{f}''(\mathbf{x})\}\mathbf{y}\, d\mathbf{y}\, m(\mathbf{x})/\tilde{f}(\mathbf{x})\right] + o(1),$$

where

$$r(\mathbf{x}) = m(\mathbf{x})\tilde{f}(\mathbf{x}), \qquad \tilde{f}(\mathbf{x}) = \prod_{i=1}^{k} f(x_i).$$

COROLLARY 2.4. *If in addition to Theorem* 1, (2.6)–(2.8) *hold, then*

$$(na_n)^{1/2}[u_n(\mathbf{x}) - m(\mathbf{x})] \to \mathcal{N}(0, \rho^2) \quad \textit{in distribution}$$

*provided that* $na_n^5 \to 0$.

Under squared loss the optimal $a_n$ satisfies $na_n^5 \to c$, some finite $c$ depending on $m$ and $f$. In this case the conclusion of Corollary 2.4 also holds, but with $\mathcal{N}(0, \rho^2)$ replaced by $\mathcal{N}(a, \rho^2)$ for some $a \neq 0$.

**3. Consistency.** Our first result states that $u_n(\mathbf{x})$ is a weakly consistent estimator of $m(\mathbf{x})$. I decided to formulate it for kernels satisfying the assumptions in Greblicki, Krzyźak and Pawlak (1984).

THEOREM 2. *Assume that*

(i) $a_n \to 0$ *and* $na_n \to \infty$,

(ii) $K(x) \geq c1_{\{|x| \leq r\}}$ *for some* $c, r > 0$,

(iii) $c_1 H(|x|) \leq K(x) \leq c_2 H(|x|)$ *for some positive constants* $c_1, c_2$ *and some decreasing function* $H$ (*defined on the positive half-line*) *satisfying* $tH(t) \to 0$ *as* $t \to \infty$. *Then, for* $\mu \otimes \cdots \otimes \mu$ *almost all* $\mathbf{x}$,

$$u_n(\mathbf{x}) \to m(\mathbf{x}) \quad \textit{in probability}.$$

PROOF. The method of proof is similar to the regression case, with some modifications due to the $U$-statistic structure. Set

$$A_n(\mathbf{x}) = \frac{\int m(z_1, \ldots, z_k) \prod_{j=1}^k K\left[(x_j - z_j)/a_n\right] \mu(dz_1) \cdots \mu(dz_k)}{\int \prod_{j=1}^k K\left[(x_j - z_j)/a_n\right] \mu(dz_1) \cdots \mu(dz_k)},$$

which for a given $\mathbf{x}$, equals $\theta_n$ from the last section. From Lemma 1 in Greblicki, Krzyźak and Pawlak (1984), almost surely

$$A_n(\mathbf{x}) \to m(\mathbf{x}). \tag{3.1}$$

For each $\beta = (\beta_1, \ldots, \beta_k)$, write

$$V_{n\beta}^* = h(Y_{\beta_1}, \ldots, Y_{\beta_k}) \prod_{j=1}^k K\left[\frac{x_j - X_{\beta_j}}{a_n}\right] \Big/ \prod_{j=1}^k \mathbb{E}K\left[\frac{x_j - X_1}{a_n}\right],$$

$$W_{n\beta}^* = \prod_{j=1}^k K\left[\frac{x_j - X_{\beta_j}}{a_n}\right] \Big/ \prod_{j=1}^k \mathbb{E}K\left[\frac{x_j - X_1}{a_n}\right]$$

and

$$V_{n\beta} = V_{n\beta}^* - \mathbb{E}V_{n\beta}^*, \qquad W_{n\beta} = W_{n\beta}^* - \mathbb{E}W_{n\beta}^*.$$

Finally, set

$$B_{n1} = \frac{(n-k)!}{n!} \sum_{\beta} V_{n\beta}$$

and

$$B_{n2} = \frac{(n-k)!}{n!} \sum_{\beta} W_{n\beta}.$$

It follows that

$$u_n(\mathbf{x}) = \frac{A_n + B_{n1}}{1 + B_{n2}}.$$

So, in view of (3.1), it remains to show

$$B_{n1} \to 0 \quad \text{and} \quad B_{n2} \to 0 \quad \text{in probability.}$$

Since $B_{n2}$ is a special case of $B_{n1}$ (put $h \equiv 1$), we have only to deal with $B_{n1}$. Decompose $h$ into

$$h = h'_M + h''_M,$$

where

$$h'_M = h1_{\{|h| \le M\}} \quad \text{and} \quad h''_M = h1_{\{|h| > M\}}$$

and $M$ is a prescribed constant which typically will be chosen large. Let $B'_{n1}$ and $B''_{n1}$ be defined as $B_{n1}$, but with $h$ replaced by $h'_M$, respectively, $h''_M$. It suffices to prove that

$$B'_{n1} \to 0 \quad \text{in probability,}$$

and that $B''_{n1}$ can be made small with large probability by choosing $M$ large. To show this, apply Chebyshev's inequality and (A1) to get

$$\mathbb{P}(|B'_{n1}| > \varepsilon)$$

$$\le \varepsilon^{-2}\left[\frac{(n-k)!}{n!}\right]^2 \sum_{r=1}^{k} \frac{(n-r)!}{(n-2k+r)!} \sum{}^{(r)} I(\Delta_1, \Delta_2) \Big/ \prod_{j=1}^{k} \mathbb{E}^2 K\left[\frac{x_j - X_1}{a_n}\right],$$

where of course $I(\Delta_1, \Delta_2)$ is computed for the $U$ kernel $h'_M$. Since $h'_M$ is bounded, so are the functions $m_{\Delta_1, \Delta_2}$ from Lemma 2.2. Hence, as in the previous section, we obtain [use (A3)]

$$\mathbb{P}(|B'_{n1}| \ge \varepsilon) = O\left[\varepsilon^{-2} \sum_{r=1}^{k} (na_n)^{-r}\right] = o(1).$$

Finally, again by Lemma 1 of Greblicki, Krzyźak and Pawlak (1984),

$$\mathbb{P}(|B''_{n1}| \ge \varepsilon) \le 2\varepsilon^{-1}\mathbb{E}\left|h''_M(Y_1, \ldots, Y_k) \prod_{j=1}^{k} K\left[\frac{x_j - X_j}{a_n}\right]\right| \Big/ \prod_{j=1}^{k} \mathbb{E}K\left[\frac{x_j - X_1}{a_n}\right]$$

$$\to 2\varepsilon^{-1}\mathbb{E}\left[|h''_M(Y_1, \ldots, Y_k)| \, |X_1 = x_1, \ldots, X_k = x_k\right]$$

almost surely as $n \to \infty$. The last term may be made arbitrarily small by letting $M \uparrow \infty$. The proof is complete. □

Theorem 3 presents strong convergence of $u_n(\mathbf{x})$, under a finite third moment assumption on $h$.

THEOREM 3. *In addition to* (i)–(iii) *from Theorem* 2, *assume that*

(iv) $$\mathbb{E}|h(Y_1, \ldots, Y_k)|^3 < \infty,$$

(v) $$\sum_n n^{-3/2} a_n^{-2} < \infty.$$

*Then, for almost all* $\mathbf{x}$,

$$u_n(\mathbf{x}) \to m(\mathbf{x}) \quad \textit{with probability } 1.$$

REMARK 3.1. (v) is satisfied for the optimal choice $a_n = cn^{-1/5}$ (cf. Corollary 2.4) and for all slightly suboptimal $a_n$ satisfying $na_n^5 \to 0$.

PROOF OF THEOREM 3. It remains to show that

$$B_{n1} \to 0 \quad \text{with probability } 1.$$

Recall from the proof of Lemma 2.2 that

$$\mathbb{E}\big(U_n - \hat{U}_n\big)^2 = O\big((na_n)^{-2}\big).$$

In fact, in Lemma 2.2, condition (vii) was needed to yield the above bound for a given $\mathbf{x}$. If the kernel satisfies the assumptions of Greblicki, Krzyźak and Pawlak (1984), their Lemma 1 may be used alternatively to bound the integrals $I(\Delta_1, \Delta_2)$ almost surely.

Since $B_{n1}$ equals $U_n$ after centering we may conclude from (v) and Borel–Cantelli that it suffices to show

(3.2) $$\hat{U}_n - \theta_n \to 0 \quad \text{with probability } 1.$$

For (3.2) we need to prove, for each $1 \le j \le k$,

$$S_n \equiv n^{-1} \sum_{i=1}^{n} \big[h_{nj}(X_i, Y_i) - \theta_n\big] \to 0 \quad \text{with probability } 1.$$

By application of the Marcinkiewicz–Zygmund inequality [cf., e.g, Chow and Teicher (1978), page 356] we have for some universal constant $C$,

$$\mathbb{E}|S_n|^3 \le Cn^{-3/2}\mathbb{E}|h_{nj}(X, Y)|^3.$$

As in the proof of Lemma 2.1, but using Lemma 1 from Greblicki, Krzyźak and Pawlak (1984) rather than boundedness of the functions $m_{jlm}$, the last expectation is shown to be $O(a_n^{-2})$, because of $K(x) \ge c1_{\{|x| \le r\}}$ and (A3), at least for $\mu \otimes \cdots \otimes \mu$ almost all $\mathbf{x}$. Apply (v) and Borel–Cantelli to complete the proof. □

**4. Examples.** Generally speaking, we may take for $h$ any function which has been found interesting in the unconditional setup; cf. Serfling (1980).

As mentioned before, the case $k = 1$ leads to the Nadaraya–Watson estimator if we set $h = id$; $h = 1_{(-\infty, t]}$ yields the conditional d.f. evaluated at $t$; cf. Stute (1986). We now discuss several examples for $k = 2$.

EXAMPLE 4.1. Put $h(y_1, y_2) = y_1 y_2$; then

$$\begin{aligned} m(x_1, x_2) &= \mathbb{E}[Y_1 Y_2 | X_1 = x_1, X_2 = x_2] \\ &= \mathbb{E}[Y_1 | X_1 = x_1]\mathbb{E}[Y_2 | X_2 = x_2] \qquad \text{(by independence)} \\ &= \overline{m}(x_1)\overline{m}(x_2), \end{aligned}$$

with $\overline{m}$ denoting the regression of $Y$ on $X$. When $x_1 = x_2$, the variance formula (2.5) yields

$$\rho^2 = 4\,\mathrm{Var}(Y_1 | X_1 = x_1)\overline{m}^2(x_1) \int K^2(u)\, du / f(x_1),$$

while for $x_1 \neq x_2$, we get

$$\begin{aligned} \rho^2 = \big[&\mathrm{Var}(Y_1 | X_1 = x_1)\overline{m}^2(x_2)/f(x_1) \\ &+ \mathrm{Var}(Y_1 | X_1 = x_2)\overline{m}^2(x_1)/f(x_2)\big] \int K^2(u)\, du. \end{aligned}$$

The above $h$ is a simple example of a $U$-statistic where one is interested in functions of $\overline{m}$. More generally, by way of forming linear combinations and incorporating higher order $U$-statistics, the prescribed method yields estimates for polynomial functions of $\overline{m}$.

EXAMPLE 4.2. For

$$h(y_1, y_2) = \tfrac{1}{2}(y_1 - y_2)^2$$

we obtain

$$m(x_1, x_1) = \mathrm{Var}(Y_1 | X_1 = x_1).$$

In this case,

$$\begin{aligned} \rho^2 &= \Big\{\mathbb{E}\big[(Y - Y_2)^2 (Y - Y_3)^2 | X = X_2 = X_3 = x_1\big] - 4m^2(x_1, x_1)\Big\} \\ &\quad \times \int K^2(u)\, du / f(x_1) \\ &= \Big\{\mathbb{E}\big[(Y - \overline{m}(x_1))^4 | X = x_1\big] - \mathrm{Var}^2(Y | X = x_1)\Big\} \int K^2(u)\, du / f(x_1). \end{aligned}$$

Compare $\rho^2$ with $\zeta_1$ in Serfling (1980), page 182.

EXAMPLE 4.3. For $h(y_1, y_2) = 1_{\{y_1 + y_2 > 0\}}$, we obtain a conditional $U$-statistic which may be viewed as a conditional version of the Wilcoxon one-sample statistic. It may be used for testing the hypothesis that the conditional

distribution at $x_1$ is symmetric at zero. Obviously,

$$\rho^2 = 4\{\mathbb{P}[Y + Y_1 > 0, Y + Y_2 > 0 | X = X_1 = X_2 = x_1] - m^2(x_1)\} \times \int K^2(u)\, du / f(x_1),$$

with

$$m(x_1) = \mathbb{P}(Y_1 + Y_2 > 0 | X_1 = x_1 = X_2).$$

EXAMPLE 4.4. For $h(y_1, y_2) = 1_{\{y_1 \le y_2\}}$,

$$m(x_1, x_2) = \mathbb{P}(Y_1 \le Y_2 | X_1 = x_1, X_2 = x_2), \qquad x_1 \ne x_2$$

equals the probability that the output pertaining to $x_1$ is less than or equal to the one pertaining to $x_2$.

We close this section with one example where the $Y$'s are bivariate.

EXAMPLE 4.5. Assume $Y_i = (Y_{i1}, Y_{i2})^t$, and define $h$ by

$$h\left[\begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix}, \begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix}\right] = \tfrac{1}{2}(y_{11}y_{12} + y_{21}y_{22} - y_{11}y_{22} - y_{12}y_{21}),$$

that is, $k = 2$, and

$$m(x_1, x_2) = \tfrac{1}{2}[\mathbb{E}(Y_{11}Y_{12} | X_1 = x_1) + \mathbb{E}(Y_{21}Y_{22} | X_2 = x_2) - \mathbb{E}(Y_{11}Y_{22} | X_1 = x_1, X_2 = x_2) - \mathbb{E}(Y_{12}Y_{21} | X_1 = x_1, X_2 = x_2)].$$

In particular,

$$m(x_1, x_1) = \mathbb{E}(Y_{11}Y_{12} | X_1 = x_1) - \mathbb{E}(Y_{11} | X_1 = x_1)\mathbb{E}(Y_{12} | X_1 = x_1),$$

the conditional covariance of $Y_1$ given $X_1 = x_1$.

## APPENDIX

We quote two things which are used in several places. First, let

$$V_n = \frac{(n-k)!}{n!} \sum_{\beta} g(Z_{\beta_1}, \ldots, Z_{\beta_k})$$

be any zero mean $U$-statistic of degree $k$, with square-integrable $g$, the $Z$'s being i.i.d. Then

$$\text{(A1)} \quad \operatorname{Var}(V_n) = \left[\frac{(n-k)!}{n!}\right]^2 \sum_{r=1}^{k} \frac{(n-r)!}{(n-2k+r)!} \sum_{|\Delta_1| = r = |\Delta_2|}^{(r)} I(\Delta_1, \Delta_2),$$

where $\sum^{(r)}$ denotes summation over all positions $\Delta_1, \Delta_2$ of length $r$,

$$I(\Delta_1, \Delta_2) = \mathbb{E}[g(Z_{i_1}, \ldots, Z_{i_k}) g(Z_{j_1}, \ldots, Z_{j_k})],$$

and the $i$'s in position $\Delta_1$ coincide with the $j$'s in position $\Delta_2$. All $i$'s and $j$'s are pairwise distinct otherwise.

(A1) reduces to (*) in Lemma A, page 183, in Serfling (1980) when $g$ is symmetric.

Another result, which is especially used in Section 3, is Corollary (10.50) from Wheeden and Zygmund (1977). It states that, if $Q_x(h)$ denotes a cube with center $x$ and length $h$ and edges parallel to the coordinate axes and if $\nu_1$ and $\nu$ are two Borel measures on $\mathbb{R}^k$, then

$$\text{(A2)} \qquad \lim_{h\to 0} \frac{\nu_1(Q_x(h))}{\nu(Q_x(h))} = f(x), \qquad \nu \text{ almost everywhere,}$$

where $f$ is the Radon–Nikodym derivative of the $\nu$ continuous part of $\nu_1$. When $\nu_1$ is Lebesgue measure, (A2) may be restated to give

$$\text{(A3)} \qquad \lim_{h\to 0} h^{-k}\nu(Q_x(h)) \quad \text{exists } \nu \text{ almost everywhere,}$$

and is positive, possibly infinite. Together with

$$K(x) \geq c1_{\{|x|\leq r\}},$$

(A3) allows for bounding the integral

$$N = \int \prod_{i=1}^{k} K\left[\frac{x_i - z_i}{a_n}\right]\mu(dz_1)\cdots\mu(dz_k),$$

from below, for $\nu = \mu \otimes \cdots \otimes \mu$ almost all $\mathbf{x} = (x_1, \ldots, x_k)$.

**Acknowledgment.** I am very grateful to a referee for his careful and critical reading of the original manuscript.

## REFERENCES

CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales*. Springer, New York.

GREBLICKI, W., KRZYŹAK, A. and PAWLAK, M. (1984). Distribution-free pointwise consistency of kernel regression estimate. *Ann. Statist.* **12** 1570–1575.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Princeton Univ. Press.

STUTE, W. (1986). On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.* **14** 891–901.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral*. Dekker, New York.

MATHEMATICAL INSTITUTE
JUSTUS-LIEBIG UNIVERSITY
ARNDTSTRASSE 2
D-6300 GIESSEN
GERMANY