Ann. Inst. Statist. Math.
Vol. 56, No. 4, 819–832 (2004)
©2004 The Institute of Statistical Mathematics

UNIVERSALLY CONSISTENT CONDITIONAL U-STATISTICS FOR ABSOLUTELY REGULAR PROCESSES AND ITS APPLICATIONS FOR HIDDEN MARKOV MODELS*

MICHEL HAREL[1] AND MADAN L. PURI[2]

[1] *UMRC 55830 C.N.R.S. and IUFM du Limousin, 209 Bd de Vanteaux, F87036 Limoges Cedex, France,* e-mail: harel@unilim.fr
[2] *Department of Mathematics, Indiana University, Bloomington, IN 47405, U.S.A.*

(Received November 18, 2002; revised November 11, 2003)

Abstract. A general class of conditional U-statistics was introduced by W. Stute as a generalization of the Nadaraya-Watson estimates of a regression function. It was shown that such statistics are universally consistent. Also, universal consistencies of the window and k_n-nearest neighbor estimators (as two special cases of the conditional U-statistics) were proved. In this paper, we extend these results from the independent case to dependent case. The result is applied to verify the Bayes risk consistency of the corresponding discrimination rules.

Key words and phrases: Universally consistent conditional U-statistics, absolute regularity, Bayes risk, Hidden Markov Models.

1. Introduction

In this paper we work with the so-called conditional U-statistics introduced by Stute (1991). These statistics may be viewed as generalizations of the Nadaraya-Watson estimates of a regression function.

To be precise, let $\{(X_i, Y_i), i \geq 1\}$ be a sequence of random vectors in some Euclidean space $\mathbb{R}^d \times \mathbb{R}^s$, defined on some probability space $(\Omega, \mathcal{A}, P)$. We assume that $\{(X_i, Y_i), i \geq 1\}$ is absolutely regular with rates

$$(1.1) \qquad \sum_{m \geq 1} m\beta^{1-1/r}(m) < +\infty,$$

where $0 < \beta(m) < 1$ and r is a positive integer. Also assume that the random vectors (r.v.'s) $\{(Y_i \mid X_i), i \geq 1\}$ are independent.

Recall that a sequence of random vectors $\{X_i, i \geq 1\}$ is absolutely regular if

$$\max_{j \geq 1} E\left\{\sup_{A \in \sigma(X_i, i \geq j+m)} |P(A \mid \sigma(X_i, 1 \leq i \leq j)) - P(A)|\right\} = \beta(m) \downarrow 0.$$

Here $\sigma(X_i, 1 \leq i \leq j)$ and $\sigma(X_i, i \geq j+m)$ are the σ-fields generated by $(X_1, \dots, X_j)$ and $(X_{j+m}, X_{j+m+1}, \dots, X_n)$, respectively. Also recall that $\{X_i\}$ satisfies the strong mixing condition if $\max_{j \geq 1}\{\sup |P(A \cap B) - P(A)P(B)|;\ A \in \sigma(X_i, 1 \leq i \leq j),\ B \in$

*Research supported by the Office of Naval Research Contract N00014-91-J-1020.

$\sigma(X_i, i \geq j+m)\} = \alpha(m) \downarrow 0$. Since $\alpha(m) \leq \beta(m)$, it follows that if $\{X_i\}$ is absolutely regular, then it is also strong mixing.

Let h be a function of k-variates (the U kernel) such that for some $r \geq 1$, $h \in \mathcal{L}_r^*$, which means that $E\{\sup_\beta |h(Y_\beta)|^r\} \leq +\infty$ (where sup 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 $\mathbb{N}^*$) which implies that for all integers $i_1, i_2, \ldots, i_k$ $(i_1 < i_2 < \cdots < i_k)$ $h(Y_{i_1}, \ldots, Y_{i_k}) \in \mathcal{L}_r$ the space of all random variables Z for which $|Z|^r$ is integrable. In order to measure the impact of a few X's, say $(X_1, \ldots, X_k)$, on a function $h(Y_1, \ldots, Y_k)$ of the pertaining Y's, set

$$m(\boldsymbol{x}) \equiv m(x_1, \ldots, x_k) := E[h(Y_1, \ldots, Y_k) \mid X_1 = x_1, \ldots, X_k = x_k] \tag{1.2}$$

where m is defined on $\mathbb{R}^{dk}$.

For estimation of $m(\boldsymbol{x})$, Stute (1991) proposed a statistic of the form

$$u_n(\boldsymbol{x}) = u_n(x_1, \ldots, x_k) = \frac{\sum_\beta h(Y_{\beta_1}, \ldots, Y_{\beta_k}) \prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]}{\sum_\beta \prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]} \tag{1.3}$$

where u_n is defined on $\mathbb{R}^{dk}$, K is the so-called smoothing kernel satisfying $\int K(u)du = 1$ and $\{h_n, n \geq 1\}$ is a sequence of bandwidth tending to zero at appropriate rates. 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 $1, \ldots, n$. Stute (1991) proved the asymptotic normality and weak and strong consistency of $u_n(\boldsymbol{x})$ when the random variables $\{(X_i, Y_i), i \geq 1\}$ are independent and identically distributed. Harel and Puri (1996) extended the results of Stute (1991) from independent case to the case when the underlying random variables are absolutely regular. Stute (1994*b*) also derived the $\mathcal{L}_r$ convergence of the conditional U-statistics under the i.i.d. set up.

If a number of the X_i's in the random sample are exactly equal to x which can happen if X is a discrete random variable, $P^Y(\cdot \mid X = x)$ can be estimated by the empirical distribution of the Y_i's corresponding to X_i's equal to x. If few or none of the X_i's are exactly equal to x, it is necessary to use Y_i's corresponding to X_i's near x. This leads to estimators $\hat{P}_n^Y(\cdot \mid X = x)$ of the form

$$\hat{P}_n^Y(\cdot \mid X = x) = \sum_{i=1}^n W_{ni}(x) I_{[Y_i \in \cdot]}$$

where $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ $(1 \leq i \leq n)$ weights those values of i for which X_i is close to x more heavily than these values of i for which X_i is far from x and I_A denotes the indicator function of A.

Let g be a Borel function on $\mathbb{R}^s$ such that $g(Y) \in \mathcal{L}_r$. Corresponding to W_n is the estimator $\ell_n(x)$ of $\ell(x) = E(g(Y) \mid X = x)$ defined by

$$\ell_n(x) = \sum_{i=1}^n W_{ni}(x) g(Y_i).$$

More generally if we now consider the estimates of $m(\boldsymbol{x})$ defined in (1.2), this leads to weighting those values of β for which $\boldsymbol{X}_\beta = (X_{\beta_1}, \ldots, X_{\beta_k})$ is close to $\boldsymbol{x}$ more heavily than the values of β for which $\boldsymbol{X}_\beta$ is far from $\boldsymbol{x}$.

This is why, as in Stute (1994a), we study a fairly general class of conditional U-statistics of the form

$$m_n(\boldsymbol{x}) = \sum_{\beta} W_{\beta,n}(\boldsymbol{x}) h(\boldsymbol{Y}_\beta) \tag{1.4}$$

designed to estimate $m(\boldsymbol{x})$, where $W_{\beta,n}(\boldsymbol{x})$ is defined from a function $W_n(\boldsymbol{x}, \boldsymbol{y})$ by $W_{\beta,n}(\boldsymbol{x}) = W_n(\boldsymbol{x}, \boldsymbol{X}_\beta)$, $\boldsymbol{Y}_\beta = (Y_{\beta_1}, \ldots, Y_{\beta_k})$, and the summation in (1.4) takes place over all permutations $\beta = (\beta_1, \ldots, \beta_k)$ of length k such that $1 \leq \beta_i \leq n$, $i = 1, \ldots, k$.

Remark 1.1. The estimator defined in (1.3) is a special case of the estimator defined in (1.4), see (2.6).

In order to make $m_n(\boldsymbol{x})$ a local average, $W_{\beta,n}(\boldsymbol{x})$ has to give larger weights to those $h(\boldsymbol{Y_\beta})$ for which $\boldsymbol{X_\beta}$ is close to $\boldsymbol{x}$. For this general class of conditional U-statistics (defined in (1.4)) and for i.i.d. random variables, Stute (1994a) derived the universal consistency. We extend his results for the case of absolutely regular r.v.'s which allow broader applications that include, among others, hidden Markov models described in detail in Section 3.

We shall call $\{W_{\beta,n}\}$ **universally consistent** if and only if

$$m_n(\boldsymbol{X}) \to m(\boldsymbol{X}) \quad \text{in } \mathcal{L}_r$$

under no conditions on h (up to integrability) or the distribution of $\{(X_i, Y_i), i \geq 1\}$. Here $\boldsymbol{X} = (X_1^0, \ldots, X_k^0)$ is a vector of X's with the same distribution as $(X_1, \ldots, X_k)$ and independent of $\{(X_i, Y_i), i \geq 1\}$.

For the ease of convenience, we shall write W_β for $W_{\beta,n}$.

Assumptions and main results are gathered in Section 2. In Section 3, we will show how our results are useful for the problem of discrimination that is considering an unobservable random vector Y which is correlated to an observable vector X. To estimate the value of Y from the value of X by using the minimal conditional risk, we need to know the distribution of (X, Y) which is unknown. That is why we use a sequence of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent of (X, Y) and often called a training sequence in pattern recognition to estimate the unknown conditional probabilities. The most adapted estimates to this situation are those which have the form given in (1.4) because we need to use the (X_i, Y_i) where X_i is near x. At last, we will see an application to the hidden Markov model. Then we give the proofs in Section 4. The main idea is to show that the estimator m_n is the ratio of two U-statistics.

2. Assumptions and main results

Consider the following set of assumptions.

(i) There exist functions $V_n(\boldsymbol{x}, \boldsymbol{y})$ on $I\!R_+^{2dk}$ such that for each $\ell \in \mathcal{L}_r^*$, $\boldsymbol{z}^{(n)} = (z_1, \ldots, z_n) \in I\!R^{dn}$ and $\boldsymbol{y}^{(n)} = (y_1, \ldots, y_n) \in I\!R^{sn}$

$$\sum_{\beta} W_n(\boldsymbol{x}, \boldsymbol{z}_\beta) \ell(\boldsymbol{y}_\beta) = \frac{\sum_\beta V_n(\boldsymbol{x}, \boldsymbol{z}_\beta) \ell(\boldsymbol{y}_\beta)}{\sum_\beta V_n(\boldsymbol{x}, \boldsymbol{z}_\beta)}$$

where $\boldsymbol{z}_\beta = (z_{\beta_1}, \ldots, z_{\beta_k})$ and $\boldsymbol{y}_\beta = (y_{\beta_1}, \ldots, y_{\beta_k})$.

(ii) There exists a function $V(\boldsymbol{x})$ on $\mathbb{R}^{dk}$ satisfying

$$\int |V(\boldsymbol{x})| d\boldsymbol{x} < \infty$$

such that for each scalar function q on $\mathbb{R}^{dk}$ verifying

$$\sup_{\boldsymbol{x} \in \mathbb{R}^{d \times k}} |q(\boldsymbol{x})| < \infty$$

we have

$$\lim_{n \to \infty} \int q(\boldsymbol{z}) V_n(\boldsymbol{x}, \boldsymbol{z}) \prod_{j=1}^{k} F(dz_j) = q(\boldsymbol{x}) \tilde{f}(\boldsymbol{x}) \int V(\boldsymbol{z}) d\boldsymbol{z}$$

where F is the d.f. of X_1 and $\tilde{f}(\boldsymbol{x}) = \prod_{j=1}^{k} f(x_j)$ where f is the density function of F.

Remark 2.1. Our conditions (i) and (ii) are completely different from conditions (ii) to (v) in Stute (1994*a*). Our conditions are more general and more easy to verify. More, the condition (i) in Stute (1994*a*) is not necessary.

The following theorems generalize Theorems 1.1, 1.2 and 3.1 in Stute (1994*a*) from the independent case to the absolute regularity case.

THEOREM 2.1. *Assume that $h \in \mathcal{L}_r^*$. Then under* (i), (ii), *and* (1.1),

$$m_n(\boldsymbol{X}) \to m(\boldsymbol{X}) \quad \textit{in } \mathcal{L}_r,$$

that is

$$E\left[\int |m_n(\boldsymbol{x}) - m(\boldsymbol{x})|^r \mu(d\boldsymbol{x})\right] \longrightarrow 0 \tag{2.1}$$

where μ denotes the distribution of $(X_1, X_2, \ldots, X_k)$.

COROLLARY 2.1. *Assume that h is a bounded function, and*

$$\sum_{n \geq 1} n^\rho \exp(-n^{1-\rho} d_n) < \infty \tag{2.2}$$

for some ρ $(0 < \rho < 1)$ and

$$d_n = \sup_{\substack{\boldsymbol{x} \in \mathbb{R}^{dk} \\ \boldsymbol{z} \in \mathbb{R}^{dk} \\ i \in \{1,\ldots,k\}}} \int V_n(\boldsymbol{x}, \boldsymbol{z}) \prod_{j \neq i} F(dz_j) \Big/ \int V_n(\boldsymbol{x}, \boldsymbol{z}) \prod_{j=1}^{k} F(dz_j)$$

then, under the conditions of Theorem 2.1, *$m_n(\boldsymbol{x}) \to m(\boldsymbol{x})$ with probability one for μ-almost all $\boldsymbol{x}$.*

Theorems 2.2 and 2.3 deal with two special cases: window weights and NN-weights. Consistency of window estimates for the regression function has been obtained by

Devroye and Wagner (1980) and Spiegelman and Sacks (1980). NN-weights for the regression function have been studied in Stone ((1977), Theorem 2).

In what follows, $|\cdot|$ denotes the maximum norm on $\mathbb{R}^d$. We also write

$$\|\boldsymbol{X}_\beta - \boldsymbol{x}\| := \max_{1\le i\le k} |X_{\beta_i} - x_i|.$$

To define window weights, put (see Stute (1994a))

$$W_\beta(\boldsymbol{x}) = \begin{cases} 1_{[\|\boldsymbol{X}_\beta - \boldsymbol{x}\|\le h_n]} / \sum_\beta 1_{[\|\boldsymbol{X}_\beta - \boldsymbol{x}\|\le h_n]} & \text{if well defined,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Here $h_n > 0$ is a given window size to be chosen by the statistician. Then we have the following result:

THEOREM 2.2. *Assume $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$. Then, under* (i), (ii) *and* (1.1), *we have*

$$m_n(\boldsymbol{X}) \to m(\boldsymbol{X}) \quad \text{in } \mathcal{L}_r,$$

where $W_\beta(\boldsymbol{x})$ in (1.4) *is given by* (2.3).

For the NN-weights, recall that X_j is among the k_n-NN of $x \in \mathbb{R}^d$ iff $d_j(x) := \|X_j - x\|$ is among the k_n-smallest ordered values $d_{1:n}(x) \le \cdots \le d_{n:n}(x)$ of the d's. Ties may be broken by randomization.

For a given $1 \le k_n \le n$, set

$$W_\beta(\boldsymbol{x}) = \begin{cases} k_n^{-d} & \text{if } X_{\beta_i} \text{ is among the } k_n\text{-}NN \text{ of } x_i \text{ for } 1 \le i \le k \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

THEOREM 2.3. *Assume that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then under* (i), (ii) *and* (1.1)

$$m_n(\boldsymbol{X}) \to m(\boldsymbol{X}) \quad \text{in } \mathcal{L}_r,$$

where $W_\beta(\boldsymbol{x})$ in (1.4) *is given by* (2.4).

We now consider as estimator of $m(\boldsymbol{x})$, the statistics of the form

$$m_n(\boldsymbol{x}) = u_n(\boldsymbol{x}) \tag{2.5}$$

where $u_n(\boldsymbol{x})$ is defined in (1.3). Then, in view of (1.4), we have

$$W_{\beta,n}(\boldsymbol{x}) = \frac{\prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]}{\sum_\beta \prod_{j=1}^k K[(x_j - X_{\beta_j})/h_n]}, \tag{2.6}$$

where $K(\boldsymbol{x})$ is a so-called smoothing kernel satisfying $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$ and $\lim_{\boldsymbol{u}\to\infty} |\boldsymbol{u}|K(\boldsymbol{u}) = 0$ and $\{h_n, n \ge 1\}$ is a sequence of bandwidths tending to zero. This special case was studied by Stute (1991) for i.i.d. random variables, and further investigated by Harel and Puri (1996) for dependent random variables. The following theorem establishes that the universal consistency still holds for conditional U-statistics involving kernel K and a sequence of bandwidth h_n.

THEOREM 2.4. *Assume that* $h_n \to 0$ *and* $nh_n^d \to \infty$ *as* $n \to \infty$. *Then under conditions* (i), (ii) *and* (1.1), *we have*

$$m_n(\boldsymbol{X}) \to m(\boldsymbol{X}) \quad \text{in } \mathcal{L}_r,$$

where $m(\boldsymbol{x})$ *is given by* (1.2).

3. Application to the Bayes risk consistency in discrimination

Now, we apply the results of Section 2 to the problem of discrimination described in Section 3 of Stute (1994*a*). Then we apply it to the Hidden Markov Model (HMM) which satisfies condition (1.1). At last, we give an example such as a multivariate mixing process defined in (3.4) below. We give a generalization of Theorem 3.1 of Stute (1994*a*).

Let h be any function taking at most finitely many values, say $1, \ldots, M$. The sets

$$A_j = \{(y_1, \ldots, y_k); h(y_1, \ldots, y_k) = j\}, \qquad 1 \leq j \leq M$$

then yield a partition of the feature space. Predicting the value of $h(Y_1, \ldots, Y_k)$ is tantamount to predicting the set in the partition to which $(Y_1, \ldots, Y_k)$ belongs. For any discrimination rule g, we have

$$P(g(\boldsymbol{X}) = h(\boldsymbol{Y})) \leq \sum_{j=1}^{M} \int_{\{\boldsymbol{x}: g(\boldsymbol{x})=j\}} \max m^j(\boldsymbol{x}) \mu(dx_1) \cdots \mu(dx_k)$$

where

$$m^j(\boldsymbol{x}) = P(h(\boldsymbol{Y}) = j \mid \boldsymbol{X} = \boldsymbol{x}), \qquad \boldsymbol{x} \in I\!\!R^p. \tag{3.1}$$

The above inequality becomes an equality if

$$g_0(\boldsymbol{x}) = \arg \max_{1 \leq j \leq M} m^j(\boldsymbol{x}) \tag{3.2}$$

g_0 is called the Bayes rule, and the pertaining probability of error

$$L^* = 1 - P_0(g_0(\boldsymbol{X}) = h(\boldsymbol{Y})) = 1 - E\left[\max_{1 \leq j \leq M} m^j(\boldsymbol{x})\right] \tag{3.3}$$

is called the Bayes risk. Each of the above unknown function m^j's can be consistently estimated by one of the methods discussed in Section 2. Let

$$m_n^j(\boldsymbol{x}) = \sum_{\beta} W_\beta(\boldsymbol{x}) 1_{[h(\boldsymbol{Y}_\beta)=j]}, \qquad 1 \leq j \leq M$$

and set

$$g_{n0}(\boldsymbol{x}) = \arg \max_{1 \leq j \leq M} m_n^j(\boldsymbol{x}).$$

Write

$$L_n := P(g_{n0}(\boldsymbol{X}) \neq h(\boldsymbol{Y})).$$

Then, the following theorem shows that the discrimination rule g_{n_0} is asymptotically Bayes' risk consistent (i.e. $L_n \to L^*$).

THEOREM 3.1. *Assume that the weights $\{W_{\beta n}\}$ are universally consistent. Then for almost all x*

$$L_n \to L^* \quad \textit{as} \quad n \to \infty.$$

PROOF. Follows from the obvious relation

$$|L_n - L^*| \le 2E\left[\max_{1\le j\le M} |m_n^j(\boldsymbol{X}) - m^j(\boldsymbol{X})|\right]. \qquad \square$$

Remark 3.1. From Theorem 2.1, as h is bounded, we can apply Theorem 3.1 when the conditions (i), (ii) and (1.1) are satisfied.

Now, we consider models for which the Bayes risk consistency in discrimination is available. One very useful should be the Hidden Markov Model (HMM) introduced by Baum and Petrie (1966). First we explain the elements and the mechanism of the type of HMM's.

There are a finite number, say M, of states in the model; we shall not rigorously define what a state is but simply say that within a state the signal possesses some measurable, distinctive properties. At each time i, a new state is entered based upon a transition probability which depends on the previous state (Markovian property). After each transition is made, an observation is produced according to a probability distribution which depends on the current state. This probability distribution is held fixed for the state regardless of when and how the state is entered.

For example, let us consider an "urn and ball" model (Rabiner and Juang (1986)). There are M urns, each filled with a large number of colored balls. There are m possible colors for each ball. The observations sequence is generated by initially choosing one of the M urns (according to an initial probability distribution), selecting a ball from the initial urn, recording its color, replacing the ball, and then choosing a new ball according to a transition probability distribution associated with the current urn.

Define now a Hidden Markov Model. Let $(Y_i)_{i\ge1}$ be a Markov chain with state space $\mathcal{X} \subset I\!R^s$ and let $(X_i)_{i\ge1}$ be a stochastic process with state space $\mathcal{X} \subset I\!R^d$.

We call $(X_i, Y_i)_{i\ge1}$ a hidden Markov Model (HMM) if the (Y_i) are conditionally independent given $(X_i)_{i\ge1}$ such that for a family $(Q_x)_{x\in\mathcal{X}}$ of probability measures on $\mathcal{Y}$.

$$P\left((Y_i)_{i\ge1} \in \prod_{i\ge1} A_i \,\middle|\, (X_i)_{i\ge1} = (x_i)_{i\ge1}\right) = \prod_{i\ge1} Q_{x_i}(A_i)$$

for any measurable $A_i \in \mathcal{Y}$ where

$$Q_x(A) = P(Y_i \in A \mid X_i = x)$$

is the conditional distribution independent of i.

If such a process satisfies the condition (1.1), we can apply the discrimination rule and the Bayes risk consistency is verifed if the weights $\{W_{\beta,n}\}$ satisfy the conditions (i), (ii) and particularly the window weights, the kernel weights, and the NN-weights.

We give an example of HMM process $(X_i, Y_i)_{i\ge1}$ which satisfies condition (1.1). It implies that we can apply Theorem 3.1.

Consider the model

$$X_i = \Psi(Y_i) + \epsilon_i, \qquad i \geq 1 \tag{3.4}$$

where X_i denotes a $I\!R^d$-vector of observed values, Ψ is a measurable known function, ϵ_i is a multivariate white noise corresponding to the measurement errors (that is, $\{\epsilon_i, i \in I\!N\}$ is a sequence of i.i.d. random $I\!R^d$-vectors with strictly positive density) and Y_i is an $I\!R^s$ predictor vector. If the sequence $\{Y_i, i \geq 1\}$ of the random vectors is absolutely regular with a geometric rate, the process (X_i, Y_i) satisfies condition (1.1).

It is well known that any Markov process which is Harris recurrent, aperiodic and geometrically ergodic is absolutely regular with a geometric rate.

For example, consider the sequence of random vectors $(Y_i)_{i\geq 1}$ defined by:

$$Y_i + \sum_{j=1}^{p_1} A_j Y_{i-j} = e_i + \sum_{l=1}^{p_2} B_l e_{i-l}, \qquad i \in Z\!\!Z \tag{3.5}$$

where $A_1, \ldots, A_{p_1}$ and $B_1, \ldots, B_{p_2}$ are $p \times p$ real matrices, A_{p_1} and B_{p_2} are invertible and $e_i = (e_{i_1}, \ldots, e_{i_p})$ is a multivariate white noise where each e_{i_j}, $i \geq 1$, $1 \leq j \leq p$ admits the same density g such that $\int |x|^\delta g(x)dx < \infty$ and $\int |g(x) - g(x - \theta)|dx = 0(|\theta|^\gamma)$ for some $\delta > 0$ and $\gamma > 0$. From Pham and Tran (1985), Y_i admits a Markovian representation

$$Y_i = HZ_i, \qquad Z_i = FZ_{i-1} + Ge_i$$

where $\{Z_i, i \geq 1\}$ is a sequence of random vectors, and H, F, G are appropriate matrices. If the eigenvalues of the matrices H have a modulus less than 1, then Y_i is absolutely regular with a geometrical rate and the process $(X_i, Y_i)_{i\geq 1}$ satisfies condition (1.1).

If $p = 1$, $q = 1$ and $k = 2$, we can write the following particular case of (3.4) and (3.5) as

$$X_i = aY_i + \epsilon_i, \qquad a \in I\!R \tag{3.6}$$

where Y_i is an AR(1) process defined by

$$Y_i = bY_{i-1} + e_i \quad \text{where} \quad |b| < 1. \tag{3.7}$$

4. Proof of theorems and Corollary 2.1

First we show that m_n is the ratio of two U-statistics.

Let $\boldsymbol{x} = (x_1, \ldots, x_k)$ be fixed throughout. Let

$$\begin{aligned} U_n(h, \boldsymbol{x}) &\equiv U_n(\boldsymbol{x}) \equiv U_n \\ &= \frac{(n-k)!}{n!} \sum_\beta h(\boldsymbol{Y}_\beta) V_n(\boldsymbol{x}, \boldsymbol{X}_\beta) \Big/ \int V_n(\boldsymbol{x}, \boldsymbol{u}) \prod_{j=1}^k F(du_j). \end{aligned} \tag{4.1}$$

Hence $m_n(\boldsymbol{x}) = U_n(h, \boldsymbol{x})/U_n(1, \boldsymbol{x})$ and $U_n(h, \boldsymbol{x})$, for each $n \geq k$ is a classical U-statistic with a kernel depending on n.

Consider a sequence of functionals

$$\theta_n(h, \boldsymbol{x}) \equiv \theta_n = \int m(z_1, \ldots, z_k) V_n(\boldsymbol{x}, \boldsymbol{z}) \prod_{j=1}^k F(dz_j) \Big/ \int V_n(\boldsymbol{x}, \boldsymbol{u}) \prod_{j=1}^k F(du_j).$$

Note that $\theta_n = E(U_n)$. For every c, $(0 \le c \le k)$ put

$$\begin{aligned} g_{c,n}(\boldsymbol{z}^{(c)}, \boldsymbol{y}^{(c)}) &\equiv g_c(\boldsymbol{z}^{(c)}, \boldsymbol{y}^{(c)}) \equiv g_c \\ &= \int h(\boldsymbol{y}) V_n^0(\boldsymbol{x}, \boldsymbol{z}) \prod_{j>c} \tilde{G}(z_j; dy_j) F(dz_j) \Big/ \int V_n(\boldsymbol{x}, \boldsymbol{u}) \prod_{j=1}^{k} F(du_j), \end{aligned}$$

where $\tilde{G}(x; \cdot)$ is the conditional density function of $(Y_1 \mid X_1 = x)$,

$$V_n^0(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{k!} \sum_{\alpha(k)} V_n(\boldsymbol{x}, \boldsymbol{z}_{\alpha(k)})$$

and where the summation is taken over all permutations $(\alpha^{(1)}(k), \dots, \alpha^{(k)}(k))$ of $\{1, \dots, k\}$. We have $g_{0,n} = \theta_n$ and

$$g_k(\boldsymbol{z}, \boldsymbol{y}) = h(\boldsymbol{y}) V_n(\boldsymbol{x}, \boldsymbol{z}) \Big/ \int V_n(\boldsymbol{x}, \boldsymbol{u}) \prod_{j=1}^{k} F(du_j).$$

Let $n^{-[r]} = \{n(n-1)\cdots(n-r+1)\}^{-1}$. Set

$$U_n^{(c)} = n^{-[c]} \sum_{\beta^{(c)}} \int g_c(\boldsymbol{z}^{(c)}, \boldsymbol{y}^{(c)}) \prod_{j=1}^{c} d(I_{[(X_{\beta_j}, Y_{\beta_j}) \le (z_j, y_j)]} - H(z_j, y_j))$$

where $\beta^{(c)}$ is the summation over all permutations $\beta^{(c)} = (\beta_1, \dots, \beta_c)$ of $\{1, \dots, n\}$ of length c. Then

$$U_n = \theta_n + \sum_{c=1}^{k} \binom{k}{c} U_n^{(c)} \tag{4.2}$$

from the Hoeffding decomposition.

To prove Theorem 2.1, the following lemmas are needed.

LEMMA 4.1. *Under the conditions of Theorem* 2.1

$$(U_n^{(c)})^2 = O(n^{-2}), \qquad 2 \le c \le k. \tag{4.3}$$

PROOF. We shall consider the case $c = 2$. The proofs in the cases $c = 3, \dots, k$ are analogous and so they are omitted. We first note that

$$\begin{aligned} U_n^{(2)} = n^{[-2]} \sum_{1 \le i_1 < i_2 \le n} \{ & g_2((X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2})) \\ & - g_1(X_{i_1}, Y_{i_1}) - g_1(X_{i_2}, Y_{i_2}) + \theta_n \}. \end{aligned} \tag{4.4}$$

So we have

$$(n^{-[2]})^{-2} E(U_n^{(2)})^2 = \sum_{1 \le i_1 < i_2 \le n} \sum_{1 \le j_1 < j_2 \le n} J((i_1, i_2), (j_1, j_2)) \tag{4.5}$$

where

$$\begin{aligned} &J((i_1,i_2,),(j_1,j_2)) \\ &\qquad = E\{(g_2(X_{i_1},Y_{i_1}),(X_{i_2},Y_{i_2})) - g_1(X_{i_1},Y_{i_1}) - g_1(X_{i_2},Y_{i_2}) + \theta_n\} \\ &\qquad\quad \{g_2((X_{j_1},Y_{j_1}),(X_{j_2},Y_{j_2})) - g_1(X_{j_1},Y_{j_1}) - g_1(X_{j_2},Y_{j_2}) + \theta_n\}. \end{aligned}$$

Since

$$\int \{g_2((z_1,y_1),(z_2,y_2)) - g_1(z_1,y_1) - g_1(z_2,y_2) + \theta_n\} H(dz_1,dy_1) = 0,$$

we have from Lemma 2.1 in Yoshihara (1976) the following inequalities:
if $1 \leq i_1 < i_2 \leq j_1 < j_2 \leq n$ and $j_2 - j_1 \geq i_2 - i_1$ then

$$J((i_1,i_2),(j_1,j_2)) \leq 4DM^{1/r}(r,h)\beta^{1-1/r}(j_2-j_1) \tag{4.6}$$

where $M(r,h) = E\{\sup_\beta |h(\boldsymbol{Y}_\beta)|^r\}$, and similarly, if $1 \leq i_1 < i_2 \leq j_1 < j_2 \leq n$ and $i_2 - i_1 \geq j_2 - j_1$, then

$$J((i_1,i_2),(j_1,j_2)) \leq 4DM^{1/r}(r,h)\beta^{1-1/r}(i_2-i_1). \tag{4.7}$$

Thus, from (4.5), (4.6) and the Assumption (1.1)

$$\begin{aligned} &\left| \sum_{1 \leq i_1 < i_2 \leq j_1 < j_2 \leq n} J((i_1,i_2),(j_1,j_2)) \right| \\ &\qquad \leq 4DM^{1/r}(r,h)n^2 \sum_{p=1}^{n} (p+1)\beta^{1-1/r}(p) = O(n^2). \end{aligned} \tag{4.8}$$

Similarly, we have

$$\begin{aligned} &\left| \sum_{1 \leq i_1 < j_1 \leq i_2 < j_2 \leq n} J((i_1,i_2),(j_1,j_2)) \right| \\ &\qquad \leq 4DM^{1/r}(r,h)n^2 \sum_{p=1}^{n} (p+1)\beta^{1-1/r}(p) = O(n^2), \end{aligned} \tag{4.9}$$

$$\left| \sum_{1 \leq i_1 < j_1 < j_2 < i_2 \leq n} J((i_1,i_2),(j_1,j_2)) \right| = O(n^2), \tag{4.10}$$

and

$$\begin{aligned} &\left| \sum_{1 \leq i_1, j_1 \leq n} \sum_{i_2=1}^{n} J((i_1,i_2),(j_1,j_2)) \right| \\ &\qquad \leq 4DM^{1/r}(r,h)n^2 \left(1 + \sum_{p=1}^{n} \beta^{1-1/r}(p)\right) = O(n^2). \end{aligned} \tag{4.11}$$

From (4.8)–(4.11) and (4.5), we obtain (4.3) for $c = 2$. The proof follows. □

LEMMA 4.2. *Under the conditions of Theorem* 2.1, *for* μ*-almost all* $\boldsymbol{x}$

$$\theta_n(h,\boldsymbol{x}) \longrightarrow m(\boldsymbol{x}) \quad as \quad n \to \infty. \tag{4.12}$$

PROOF. From condition (ii) in Section 2, we have

$$\lim_{n\to\infty} \int m(\boldsymbol{x})V_n(\boldsymbol{x},\boldsymbol{z})\prod_{j=1}^{k} F(dz_j) = m(\boldsymbol{x})\tilde{f}(\boldsymbol{x})\int V(\boldsymbol{z})d\boldsymbol{z}$$

and so

$$\lim_{n\to\infty} \int V_n(\boldsymbol{x},\boldsymbol{z})\prod_{j=1}^{k} F(dz_i) = \tilde{f}(\boldsymbol{x})\int V(\boldsymbol{z})d\boldsymbol{z}.$$

Thus

$$\lim_{n\to\infty} \theta_n(h,\boldsymbol{x}) = \frac{m(\boldsymbol{x})\tilde{f}(\boldsymbol{x})\int V(\boldsymbol{z})d\boldsymbol{z}}{\tilde{f}(\boldsymbol{x})\int V(\boldsymbol{z})d\boldsymbol{z}} = m(\boldsymbol{x}).$$

To prove Theorem 2.1, from Lemmas 4.1 and 4.2 and the fact that $h \in \mathcal{L}_r^*$, we now have to show that μ-almost all $\boldsymbol{x}$

$$U_n^{(1)}(\boldsymbol{x}) \to 0 \quad \text{in probability.}$$

Since

$$U_n^{(1)}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n}(g_1(X_i,Y_i) - \theta_n),$$

we have

$$\begin{aligned} E(U_n^{(1)})^2 &= n^{-2}E\left(\sum_{i=1}^{n}(g_1(X_i,Y_i)-\theta_n)\right)^2 \\ &= n^{-2}\sum_{i=1}^{n}E(g_1(X_i,Y_i)-\theta_n)^2 \\ &\quad + 2n^{-2}\sum_{1\le i<j\le n} E\{(g_1(X_i,Y_i)-\theta_n)(g_1(X_j,Y_j)-\theta_n)\} \end{aligned}$$

(from Lemma 2.1 of Yoshihara (1976))

$$\begin{aligned} &\le 2n^{-2}nM(2,h) + 4n^{-2}M^{1/r}(r,h)\sum_{p=1}^{n}(p+1)\beta^{1-1/r}(p) \\ &= O(n^{-1}) \end{aligned}$$

which implies

$$E(U_n^{(1)})^2 = O(n^{-1}). \tag{4.13}$$

From Lemmas 4.1 and 4.2 and from (4.13) we have $U_n(h,\boldsymbol{x}) \to m(\boldsymbol{x})$ and $U_n(1,\boldsymbol{x}) \to 1$ in probability, as $n \to \infty$ for μ almost all $\boldsymbol{x}$. It remains to prove the uniform integrability. By Jensen's inequality, we have

$$\begin{aligned}
&\sup_{n\in I\!N^*} E\left\{\left[\sum_\beta V_n(\boldsymbol{X},\boldsymbol{X}_\beta)|h(\boldsymbol{Y}_\beta)|/\sum_\beta V_n(\boldsymbol{X},\boldsymbol{X}_\beta)\right]^r\right\}\\
&\quad\le \sup_{n\in I\!N^*} E\left\{\sum_\beta V_n(\boldsymbol{X},\boldsymbol{X}_\beta)|h(\boldsymbol{Y}_\beta)|^r/\sum_\beta V_n(\boldsymbol{X},\boldsymbol{X}_\beta)\right\}\\
&\quad\le E\left\{\sup_\beta |h(\boldsymbol{Y}_\beta)|^r\right\} < +\infty
\end{aligned}$$

from (i), and Theorem 2.1 is proved. □

Proof of Corollary 2.1. From Lemma 4.1, we have

$$E(U_n - kU_n^{(1)})^2 = O(n^{-2}). \tag{4.14}$$

Then, from the Borel-Cantelli lemma, it suffices to show that

$$U_n^{(1)} \to 0 \quad \text{with probability 1.} \tag{4.15}$$

Clearly

$$U_n^{(1)} = n^{-1}\sum_{i=1}^n \{S_{i,n} - E(S_{i,n})\}$$

where

$$S_{i,n} = g_1(X_i, Y_i) - \theta_n.$$

As h is bounded, there exists two positive constants b and c such that

$$|S_{i,n}| \le b/d_n \quad \text{and} \quad E(S_{i,n}^2) \le c/d_n. \tag{4.16}$$

If $U_1, U_2, \ldots, U_n$ are independent random variables with $|U_i| \le m$, $E(U_i^2) \le \sigma_i^2$, then an inequality due to Bennett ((1962), p. 39) states that

$$P\left[\left|n^{-1}\sum U_n\right| \ge \varepsilon\right] \le 2\exp\{-n\varepsilon^2/2(\sigma^2 + m\varepsilon)\}$$

where $\sigma^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. Put $q = q_n = [n^\rho] + 1$ for some $0 \le \rho \le 1$, and write

$$U_n^{(1)} = \sum_{j=1}^q S_{j,n}^*$$

where

$$S_{j,n}^* = \sum_{p=0}^{\ell_j} \{S_{j+pq,n} - E(S_{j+pq,n})\}$$

and ℓ_j is the largest integer such that $j + \ell_j q \leq n$. Then, proceeding as in Harel and Puri (1996), we get

$$P[U_n^{(1)} \geq \varepsilon] \leq 2n^\rho[\exp\{-\alpha \ell_j d_n\} + n^{1-p}\beta([n^\rho] + 1)]$$

where $\alpha = \varepsilon^2/(2c + 4b\varepsilon)$.

From the Borel-Cantelli lemma and conditions (2.1) and (2.3), we deduce (4.15) and Corollary 2.1 is proved. □

Proofs of Theorems 2.2 to 2.4. We have only to show that conditions (i) and (ii) are satisfied.

For Theorem 2.2, we can write for every $\ell \in \mathcal{L}_r^*$

$$\sum_\beta W_\beta(\boldsymbol{x}_\beta)\ell(\boldsymbol{y}_\beta) = \frac{\sum_\beta h_n^{-k} 1_{[\|\boldsymbol{x}_\beta - \boldsymbol{x}\| \leq h_n]}\ell(\boldsymbol{y}_\beta) / \int 1_{[\|\boldsymbol{u}-\boldsymbol{x}\| \leq h_n]} \prod_{j=1}^k F(du_j)}{\sum_\beta h_n^{-k} 1_{[\|\boldsymbol{x}_\beta - \boldsymbol{x}\| \leq h_n]} / \int 1_{[\|\boldsymbol{u}-\boldsymbol{x}\| \leq h_n]} \prod_{j=1}^k F(du_j)}$$

and (i) is proved.

Now we have

$$\begin{aligned}
&\lim_{n\to\infty} h_n^{-k} \int 1_{[\|\boldsymbol{x}-\boldsymbol{z}\| \leq h_n]} q(\boldsymbol{z}) \prod_{j=1}^k F(dz_j) \\
&= \lim_{n\to\infty} h_n^{-k} \int 1_{[\|\boldsymbol{u}\| \leq 1]} q(\boldsymbol{x} + \boldsymbol{u}h_n) f(\boldsymbol{x} + \boldsymbol{u}h_n) h_n^k \prod_{j=1}^k du_j \\
&= q(\boldsymbol{x})\tilde{f}(\boldsymbol{x}) \int 1_{[\|\boldsymbol{u}\| \leq 1]} \prod_{j=1}^k du_j = 2^k q(\boldsymbol{x})\tilde{f}(\boldsymbol{x})
\end{aligned}$$

and (ii) is also proved for Theorem 2.2 where $V_n(\boldsymbol{x}, \boldsymbol{z}) = 1_{[\|\boldsymbol{x}-\boldsymbol{z}\} \leq h_n]}$ and $V(\boldsymbol{x}) = 1_{[\|\boldsymbol{x}\| \leq 1]}$.

For Theorem 2.3, from the fact that $\sum_\beta W_\beta(\boldsymbol{x}) = 1$, we can put $V_n(\boldsymbol{x}, \boldsymbol{y}) = W_n(\boldsymbol{x}, \boldsymbol{y})$ and we have

$$\sum_\beta W_n(\boldsymbol{x}, \boldsymbol{z}_\beta)\ell(\boldsymbol{y}_\beta) = \frac{\sum_\beta W_n(\boldsymbol{x}, \boldsymbol{z}_\beta)\ell(\boldsymbol{y}_\beta)}{\sum_\beta W_n(\boldsymbol{x}, \boldsymbol{z}_\beta)}$$

and (i) is proved. Now we get (ii) if we put $V(\boldsymbol{x}) \equiv 1$.

Theorem 2.4 follows analogously, if we put

$$V_n(\boldsymbol{x}, \boldsymbol{z}) = h_n^{-k} \prod_{j=1}^k K\left(\frac{\boldsymbol{x} - \boldsymbol{z}}{h_n}\right)$$

then (i) is immediate. From Cacoullos (1966), we deduce

$$\begin{aligned}
\lim_{n\to\infty} \int q(\boldsymbol{z}) h_n^{-k} \prod_{j=1}^k K\left(\frac{\boldsymbol{x} - \boldsymbol{z}}{h_n}\right) \prod_{j=1}^k F(dz_j) &= q(\boldsymbol{x})\tilde{f}(\boldsymbol{x}) \int K(\boldsymbol{z}) d\boldsymbol{z} \\
&= q(\boldsymbol{x})\tilde{f}(\boldsymbol{x})
\end{aligned}$$

and (ii) is proved for Theorem 2.4 if we put $V(\boldsymbol{x}) = K(\boldsymbol{x})$. □

References

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics*, **37**, 1554–1563.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables, *Journal of the American Statistical Association*, **19**, 33–45.

Cacoullos, T. (1966). Estimation of a multivariate density, *Annals of the Institute of Statistical Mathematics*, **18**, 179–189.

Devroye, L. P. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Annals of Statistics*, **8**, 231–239.

Harel, M. and Puri, M. L. (1996). Conditional U-statistics for dependent random variables, *Journal of Multivariate Analysis*, **57**, 84–100.

Pham, T. D. and Tran, L. T. (1985). Some mixing properties of time series models, *Stochastic Processes and Their Applications*, **19**, 297–303.

Rabiner, L. R. and Juang, B.H. (1986). An introduction to hidden Markov models, *I.E.E.E. A.S.S.P.*, **24**, 4–16.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression, *Annals of Statistics*, **8**, 240–246.

Stone, C. (1977). Consistent nonparametric regression, *Annals of Statistics*, **5**, 595–620.

Stute, W. (1991). Conditional U-statistics, *Annals of Probability*, **19**, 812–825.

Stute, W. (1994*a*). Universally consistent conditional U-statistics, *Annals of Statistics*, **22**, 460–473.

Stute, W. (1994*b*). L^p convergence of conditional U-statistics, *Journal of Multivariate Analysis*, **51**, 71–82.

Yoshihara, K. I. (1976). Limiting behavior of U-statistics for stationary absolutely regular processes, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **35**, 237–252.