G. Morvai, S. Yakowitz, and L. Györfi:

Nonparametric inference for ergodic, stationary time series.

Ann. Statist. 24 (1996), no. 1, 370–379.

Abstract

The setting is a stationary, ergodic time series. The challenge is to construct a sequence of functions, each based on only finite segments of the past, which together provide a strongly consistent estimator for the conditional probability of the next observation, given the infinite past. Ornstein gave such a construction for the case that the values are from a finite set, and recently Algoet extended the scheme to time series with coordinates in a Polish space.

The present study relates a different solution to the challenge. The algorithm is simple and its verification is fairly transparent. Some extensions to regression, pattern recognition, and on-line forecasting are mentioned.

1 Introduction

In this section, we give brief overview of the situation with respect to nonparametric inference under the most lenient mixing conditions. Impetus for this line of study follows Roussas (1969) and Rosenblatt (1970) who extended ideas in the nonparametric regression literature for i.i.d. variables to give a theory adequate for showing, for example, that for $\{X_i\}$ a real Markov sequence, under Doeblin-like assumptions, the obvious kernel forecaster is an asymptotically normal estimator of the conditional expectation $E(X_0|X_{-1} = x)$. In the 1980's, there was an explosion of works which showed consistency in various senses for nonparametric auto-regression and density estimators under more and more general mixing assumptions (e.g., Castellana and Leadbetter (1986), Collomb (1985), Györfi (1981), and Masry (1986)). The monograph by Györfi *et al.* (1989) gives supplemental information about nonparametric estimation for dependent series.

Such striving for generality stems from the inconvenience of mixing conditions; satisfactory statistical tests are not available. Some recent developments have succeeded in disposing of these conditions altogether. In the Markov case, aside from some smoothness assumptions, it is enough that an invariant law exist to get the usual pointwise asymptotic normality of kernel regression (Yakowitz (1989)). In case of Harris recurrence but no invariant law, one can still attain a.s. pointwise convergence of a nearest-neighbor regression algorithm in which the neighborhood is chosen in advance and observations continue until a prescribed number of points fall into that neighborhood (Yakowitz (1993)).

Pushing beyond the Markov hypothesis, by a histogram estimate (Györfi *et al.* (1989)) or a recursive-type estimator (Györfi and Masry (1990)), one can infer the marginal density of an ergodic stationary time series provided only that there exist an absolutely continuous transition density. Here the limit may have been attained; it is now known (Györfi *et al.* (1989) and Györfi and Lugosi (1992), respectively) that without the conditional density

assumption, the histogram estimator and the kernel and recursive kernel estimates for the marginal density are not generally consistent.

The situation with respect to (auto-) regression is more inclusive for ergodic, stationary sequences. In a landmark paper, following developments by Ornstein (1978) for the case that the time series values are from a finite set, for time series with values in a Polish space, Algoet (1992, §5) has provided a data-driven distribution function construction $F_n(x|X_{-1}, X_{-2}, \ldots)$ which a.s. converges in distribution to

$$P(X_0 \leq x|X_{-1}, X_{-2}, \ldots) = P(X_0 \leq x|\mathbf{X}^-),$$

where $\mathbf{X}^- = (X_{-1}, X_{-2}, \ldots)$.

The goal of the present study is to relate a simpler rule the consistency of which is easy to establish. In concluding sections, it is noted that as a result of these developments, one has a consistent regression estimate in the bounded time-series case, and implications to problems of pattern recognition and online forecasting are mentioned. It is to be conceded that our algorithm, as well as those of Algoet's and Ornstein's, can be expected to require very large data segments for acceptable precision.

As a final general comment, we note that the assumption of ergodicity may be relaxed somewhat. Thus in view of Sections 7.4 and 8.5 of Gray (1988), one sees that a nonergodic stationary process has an ergodic decomposition. With probability one, a realization of the time series falls into an invariant event on which the process is ergodic and stationary. Then one may apply the developments of this study to that event as though it were the process universe. Thus the analysis here also remains valid for stationary nonergodic processes. Our analysis is restricted to the case that the coordinates of the time series are real, but it is evident that the proofs extend directly to the vector-valued case. In view of Theorem 2.2 of Billingsley (1968, p. 14) it will be clear that the formulas and derivations to follow also hold if the $X_i's$ are in a Polish space.

2 Estimation of conditional distributions

Let $\mathbf{X} = \{X_n\}$ denote a real-valued doubly infinite stationary ergodic time series. Let

$$X_{-j}^{-1} = (X_{-j}, X_{-j+1}, \ldots, X_{-1})$$

be notation for a data segment into the j-past, where j may be infinite. For a Borel set C one wishes to infer the conditional probability

$$P(C|\mathbf{X}^-) = P(X_0 \in C|X_{-\infty}^{-1}).$$

The algorithm to be promoted here is iterative on an index $k = 1, 2, \ldots$ For each k, the data-driven estimate of $P(C|\mathbf{X}^-)$ requires only a segment of finite (but random) length of $\mathbf{X}^-$. One may proceed by simply repeating the estimation process for k=1,2,..., until a given finite data record no longer suffices for the demands of the algorithm. The goal of the study will be to show that a.s. convergence can be attained. That is, our estimation is strongly consistent in the topology of weak convergence.

The estimation algorithm is now revealed in the simple context of binary sequences, and afterwards, we show alterations necessary for more general processes.

Define the sequences λ_{k-1} and τ_k recursively ($k = 1, 2, \ldots$). Put $\lambda_0 = 1$ and let τ_k be the time between the occurrence of the pattern

$$B(k) = (X_{-\lambda_{k-1}}, \ldots, X_{-1}) = X_{-\lambda_{k-1}}^{-1}$$

at time -1 and the last occurrence of the same pattern prior to time -1. More precisely, let

$$\tau_k = \min\{t > 0 : X_{-\lambda_{k-1}-t}^{-1-t} = X_{-\lambda_{k-1}}^{-1}\}.$$

Put

$$\lambda_k = \tau_k + \lambda_{k-1}.$$

The observed vector $B(k)$ a.s. takes a value having positive probability; thus by ergodicity, with probability 1 the string $B(k)$ must appear infinitely often in the sequence $X_{-\infty}^{-2}$. One denotes the kth estimate of $P(C|\mathbf{X}^-)$ by $P_k(C)$, and defines it to be

$$P_k(C) = \frac{1}{k} \sum_{1 \le j \le k} 1_C(X_{-\tau_j}). \tag{1}$$

Here 1_C is the indicator function for C.

For the general case, we use a sub-sigma-field structure motivated by Algoet (1992, Section 5.2), which is more general. Let $\mathcal{P}_k = \{A_{k,i},\, i = 1, 2, \ldots, m_k\}$ be a sequence of finite partitions of the real line by (finite or infinite) right semi-closed intervals such that $\sigma(\mathcal{P}_k)$ is an increasing sequence of finite σ-algebras that asymptotically generate the Borel σ-field. Let G_k denote the corresponding quantizer:

$$G_k(x) = A_{k,i} \,\text{if}\, x \in A_{k,i}.$$

The role of the feature vector in (1) is now played by the discrete quantity,

$$B(k) = (G_k(X_{-\lambda_{k-1}}), \ldots, G_k(X_{-1})) = G_k(X_{-\lambda_{k-1}}^{-1}).$$

Now

$$\tau_k = \min\{t > 0 : G_k(X_{-\lambda_{k-1}-t}^{-1-t}) = G_k(X_{-\lambda_{k-1}}^{-1})\}.$$

Again, ergodicity implies that $B(k)$ is almost surely to be found in the sequence $G_k(X_{-\infty}^{-2})$, and with this generalization of notation, the kth estimate of $P(C|\mathbf{X}^-)$ is still provided by formula (1).

As in Algoet's construct, the estimate P_k is calculated from observations of random size. Here the random sample size is λ_k. To obtain a fixed sample size $t > 0$ version, let κ_t be the maximum of integers k for which $\lambda_k \le t$. Put

$$\hat{P}_{-t}(C) = P_{\kappa_t}(C). \tag{2}$$

Theorem 1 *Under the stationary ergodic assumption regarding* $\{X_n\}$ *and under the estimator constructs (1) and (2) described above,*

$$\lim_{k\to\infty} P_k(\cdot) = P(\cdot|\mathbf{X}^-) \quad \text{a.s.}, \tag{3}$$

and

$$\lim_{t\to\infty} \hat{P}_{-t}(\cdot) = P(\cdot|\mathbf{X}^-) \quad \text{a.s.}, \tag{4}$$

in the weak topology of distributions.

Proof. To begin with, assume that for some m, $C \in \sigma(\mathcal{P}_m)$. The first chore is to show that a.s.,

$$P_k(C) \to P(C|\mathbf{X}^-).$$

For $k > m$ we have that

$$\begin{aligned}
&P_k(C) - P(C|\mathbf{X}^-) \\
&= \frac{1}{k}\sum_{1\le j\le m}[1_C(X_{-\tau_j}) - P(X_{-\tau_j}\in C|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))] \\
&+ \frac{(k-m)}{k}\frac{1}{(k-m)}\sum_{m<j\le k}[1_C(X_{-\tau_j}) - P(X_{-\tau_j}\in C|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))] \\
&+ \frac{1}{k}\sum_{1\le j\le k} P(X_{-\tau_j}\in C|G_{j-1}(X^{-1}_{-\lambda_{j-1}})) - P(C|\mathbf{X}^-) \\
&= P1_k + \frac{(k-m)}{k}P2_k + P3_k.
\end{aligned}$$

Obviously,

$$P1_k \to 0 \quad \text{a.s.}$$

Toward mastering $P2_k$, one observes that $P2_k$ is an average of bounded martingale differences. To see this note that $\sigma(G_j(X^{-1}_{-\lambda_j}))$ $j = 0, 1, \ldots$ is monotone increasing, and that $1_C(X_{-\tau_j})$ is measurable on $\sigma(G_j(X^{-1}_{-\lambda_j}))$ for $j \ge m$. The convergence of $P2_k$ can be established by Lévy's classical result, namely, the Cesàro means of a bounded sequence of martingale differences

converge to zero almost surely. For a version suited to our needs, see, for example, Theorem 3.3.1 in Stout (1974). One may even obtain rates for $P2_k$ through the use of Azuma's (1967) exponential bound for martingale differences. We have to prove that

$$P3_k \to 0 \quad \text{a.s.}$$

By Lemma 1 in the appendix,

$$P(X_{-\tau_j} \in C | G_{j-1}(X^{-1}_{-\lambda_{j-1}})) = P(X_0 \in C | G_{j-1}(X^{-1}_{-\lambda_{j-1}})).$$

Using this we get

$$P3_k = \frac{1}{k} \sum_{1 \le j \le k} P(X_{-\tau_j} \in C | G_{j-1}(X^{-1}_{-\lambda_{j-1}})) - P(C|\mathbf{X}^-)$$

$$= \frac{1}{k} \sum_{1 \le j \le k} P(X_0 \in C | G_{j-1}(X^{-1}_{-\lambda_{j-1}})) - P(C|\mathbf{X}^-).$$

By assumption,

$$\sigma(B(j)) \uparrow \sigma(\mathbf{X}^-),$$

which implies that

$$\sigma(G_j(X^{-1}_{-\lambda_j})) \uparrow \sigma(\mathbf{X}^-).$$

Consequently by the a.s. martingale convergence theorem we have that

$$P(X_0 \in C | G_j(X^{-1}_{-\lambda_j})) \to P(C|\mathbf{X}^-) \quad \text{a.s.},$$

and thus by the Toeplitz lemma (cf. Ash (1972))

$$P3_k \to 0 \quad \text{a.s.}$$

Let D denote the countably infinite set of x's for which $(-\infty, x] \in \sigma(\mathcal{P}_k)$ for sufficiently large k. By assumption, D is dense in $\mathbb{R}$. Define

$$F_k(x) = P_k((-\infty, x]).$$

Also, set

$$F(x) = P((-\infty, x]|\mathbf{X}^-).$$

By the preceding development we have the almost sure event H such that on H for all $x \in D$

$$F_k(x) \to F(x). \tag{5}$$

Since D is dense in $\mathbb{R}$, we have (5) on H and for all continuity points of $F(\cdot)$, and (3) is proved. The convergence (4) is an obvious consequence of (3).

3 Estimation of auto-regression functions

The next result uses estimators

$$R_k = \frac{1}{k} \sum_{1 \le j \le k} X_{-\tau_j} \tag{6}$$

and

$$\hat{R}_{-t} = \frac{1}{\kappa_t} \sum_{1 \le j \le \kappa_t} X_{-\tau_j}. \tag{7}$$

Corollary 1 *Assume that for some number D, a.s., $|X_0| \le D < \infty$. Under the stationary ergodic assumption regarding $\{X_n\}$ and under the estimator constructs (6) and (7) described above,*

$$\lim_{k \to \infty} R_k = E(X_0|\mathbf{X}^-) \quad \text{a.s.}, \tag{8}$$

and

$$\lim_{t \to \infty} \hat{R}_{-t} = E(X_0|\mathbf{X}^-) \quad \text{a.s.} \tag{9}$$

Proof. Define the function

$$\phi(x) = \begin{cases} D, & \text{if } x > D \\ x, & \text{if } -D \le x \le D \\ -D, & \text{if } x < -D \end{cases}$$

Then

$$R_k = \int xP_k(dx) = \int \phi(x)P_k(dx)$$

$$\rightarrow \int \phi(x)P(dx|\mathbf{X}^-) = \int xP(dx|\mathbf{X}^-) = E(X_0|\mathbf{X}^-).$$

because of Theorem 1 and the fact that convergence in distribution implies the convergence of integrals of the bounded continuous function ϕ with respect to the actual distributions (Billingsley (1968)). Thus the proof of (8) is complete. The proof of (9) follows in the same way; just put $\hat{P}_{-t}$ in place of P_k.

The estimates $\hat{R}_{-t}$ converge almost surely to $E(X_0|\mathbf{X}^-)$ and are uniformly bounded so $|\hat{R}_{-t} - E(X_0|X_{-t}^{-1})| \rightarrow 0$ also in mean. Motivated by Bailey (1976), consider the estimator $\hat{R}_t(\omega) = \hat{R}_{-t}(T^t\omega)$ which is defined in terms of $(X_0, \ldots, X_{t-1})$ in the same way as $\hat{R}_{-t}(\omega)$ was defined in terms of $(X_{-t}, \ldots, X_{-1})$. (T denotes the left shift operator.) The estimator $\hat{R}_t$ may be viewed as an on-line predictor of X_t. This predictor has special significance not only because of potential applications, but additionally because Bailey (1976) proved that it is impossible to construct estimators $\hat{R}_t$ such that always $\hat{R}_t - E(X_t|X_0^{t-1}) \rightarrow 0$ almost surely. An immediate consequence of Corollary 1 is that convergence in probability is verified. That is, the shift transformation T is measure preserving hence convergence $\hat{R}_{-t} - E(X_0|X_{-t}^{-1}) \rightarrow 0$ in L^1 implies convergence $\hat{R}_t - E(X_t|X_0^{t-1}) \rightarrow 0$ in L^1 and in probability.

4 Pattern recognition

Consider the 2-class pattern recognition problem with d-dimensional feature vector X_0 and binary valued label Y_0. Let $\mathcal{D}^- = (X_{-\infty}^{-1}, Y_{-\infty}^{-1})$ be the data. In conventional pattern recognition problems (X_0, Y_0) and $\mathcal{D}^-$ are independent, so the best possible decision based on X_0 and based on $(X_0, \mathcal{D}^-)$ are the same. Here assume that $\{(X_i, Y_i)\}$ is a doubly infinite stationary and

ergodic sequence. The classification problem is to decide on Y_0 for given data $(X_0, \mathcal{D}^-)$ in order to minimize the probability of misclassification. The Bayes decision g^* is the best possible one. Let $\eta(X_0, \mathcal{D}^-)$ be the *a posteriori* probability of $Y_0 = 1$ (regression function):

$$\eta(X_0, \mathcal{D}^-) = P(Y_0 = 1 | X_0, \mathcal{D}^-) = E(Y_0 | X_0, \mathcal{D}^-).$$

Then $g^*(X_0, \mathcal{D}^-) = 1$ if $\eta(X_0, \mathcal{D}^-) \geq 1/2$ and 0 otherwise. For an arbitrary approximation $\eta_k = \eta_k(X_0, \mathcal{D}^-)$ put $g_k = g_k(X_0, \mathcal{D}^-) = 1$ if $\eta_k \geq 1/2$ and 0 otherwise. Then it is easy to see (cf. Devroye and Györfi (1985), Chapter 10) that

$$\begin{aligned} 0 &\leq P(g_k \neq Y_0 | X_0, \mathcal{D}^-) - P(g^*(X_0, \mathcal{D}^-) \neq Y_0 | X_0, \mathcal{D}^-) \\ &\leq 2|\eta_k - \eta(X_0, \mathcal{D}^-)|. \end{aligned} \tag{10}$$

The estimation is a slight modification of (1). Define the sequences λ_{k-1} and τ_k recursively ($k = 1, 2, \ldots$). Put $\lambda_0 = 1$ and τ_k be the time between the occurrence of the pattern

$$B(k) = (G_k(X_{-\lambda_{k-1}}), Y_{-\lambda_{k-1}}, \ldots, G_k(X_{-1}), Y_{-1}, G_k(X_0))$$

at time 0 and the last occurrence of the same pattern in $\mathcal{D}^-$. More precisely,

$$\tau_k = \min\{t > 0 : G_k(X^{-t}_{-\lambda_{k-1}-t}) = G_k(X^0_{-\lambda_{k-1}}), Y^{-1-t}_{-\lambda_{k-1}-t} = Y^{-1}_{-\lambda_{k-1}})\}.$$

Put

$$\lambda_k = \tau_k + \lambda_{k-1}.$$

The observed vector $B(k)$ a.s. takes a value of positive probability; thus by ergodicity $B(k)$ has occurred with probability 1. One denotes the kth estimate of $\eta(X_0, \mathcal{D}^-)$ by η_k, and defines it to be

$$\eta_k = \frac{1}{k} \sum_{1 \leq j \leq k} Y_{-\tau_j}. \tag{11}$$

Corollary 2 *Under the stationary ergodic assumption regarding the process* $\{(X_n, Y_n)\}$ *and under the estimator construct (11) described above,*

$$P(g_k \neq Y_0 | X_0, \mathcal{D}^-) \to P(g^*(X_0, \mathcal{D}^-) \neq Y_0 | X_0, \mathcal{D}^-) \quad \text{a.s.} \tag{12}$$

Proof. Because of (10), we get (12) from

$$\eta_k \to \eta(X_0, \mathcal{D}^-) \quad \text{a.s.},$$

the proof of which is similar to the proof of Theorem 1.
Remark. It is also possible to construct a version of this estimate with fixed sample size $t > 0$ in the same way as in (2) and (7).

5 Appendix

In the sequel, we use the notation of Section 2.

Lemma 1 *Under the stationary ergodic assumption regarding* $\{X_n\}$*, for* $j = 1, 2, \ldots,$

$$P(X_{-\tau_j} \in C | G_{j-1}(X_{-\lambda_{j-1}}^{-1})) = P(X_0 \in C | G_{j-1}(X_{-\lambda_{j-1}}^{-1})).$$

Proof. First of all, note that by definition,

$$\begin{aligned} &\sigma(G_{j-1}(X_{-\lambda_{j-1}}^{-1})) = \mathcal{F}_{j-1} \\ &= \quad \sigma(\{G_{j-1}(X_{-m}^{-1}) = b_{-m}^{-1}, \lambda_{j-1} = m\}; b_{-m}^{-1}, m = 1, 2, \ldots), \end{aligned}$$

where b_{-m}^{-1} is an m-vector of sets from the finite partition $\mathcal{P}_{j-1}$.
Note also that

$$B = \{G_{j-1}(X_{-m}^{-1}) = b_{-m}^{-1}, \lambda_{j-1} = m\}$$

are the (countable many) generating atoms of $\mathcal{F}_{j-1}$, so we have to show that for any atom B the following equality holds:

$$P(B \cap \{X_{-\tau_j} \in C\}) = P(B \cap \{X_0 \in C\}).$$

λ_{j-1} is a stopping time, B is an m-dimensional cylinder set, which means that b_{-m}^{-1} determines whether $\lambda_{j-1} \neq m$ (in which case $B = \emptyset$ and the statement is trivial) or $\lambda_{j-1} = m$ and then

$$B = \{G_{j-1}(X_{-m}^{-1}) = b_{-m}^{-1}\}.$$

For $j = 1, 2, \ldots$ let

$$\tilde{\tau}_j = \min\{0 < t : G_j(X_{-\lambda_{j-1}+t}^{-1+t}) = G_j(X_{-\lambda_{j-1}}^{-1})\}.$$

Now

$$\begin{aligned}
&T^{-l}[B \cap \{\tau_j = l, X_{-l} \in C\}] \\
&= T^{-l}[\{G_{j-1}(X_{-m}^{-1}) = b_{-m}^{-1}, G_j(X_{-m-l}^{-1-l}) = G_j(X_{-m}^{-1}), \\
&\quad G_j(X_{-m-t}^{-1-t}) \neq G_j(X_{-m}^{-1}), 0 < t < l, X_{-l} \in C\}] \\
&= \{G_{j-1}(X_{-m+l}^{-1+l}) = b_{-m}^{-1}, G_j(X_{-m}^{-1}) = G_j(X_{-m+l}^{-1+l}), \\
&\quad G_j(X_{-m-t+l}^{-1-t+l}) \neq G_j(X_{-m+l}^{-1+l}), 0 < t < l, X_0 \in C\} \\
&= \{G_{j-1}(X_{-m+l}^{-1+l}) = b_{-m}^{-1}, G_j(X_{-m}^{-1}) = G_j(X_{-m+l}^{-1+l}), \\
&\quad G_j(X_{-m+t}^{-1+t}) \neq G_j(X_{-m+l}^{-1+l}), 0 < t < l, X_0 \in C\} \\
&= \{G_{j-1}(X_{-m}^{-1}) = b_{-m}^{-1}, G_j(X_{-m}^{-1}) = G_j(X_{-m+l}^{-1+l}), \\
&\quad G_j(X_{-m+t}^{-1+t}) \neq G_j(X_{-m}^{-1}), 0 < t < l, X_0 \in C\} \\
&= B \cap \{\tilde{\tau}_j = l, X_0 \in C\},
\end{aligned}$$

where T denotes the left shift operator.
By stationarity, it follows that

$$\begin{aligned}
&P(B \cap \{X_{-\tau_j} \in C\}) \\
&= \sum_{l=1}^{\infty} P(B \cap \{\tau_j = l, X_{-l} \in C\}) \\
&= \sum_{l=1}^{\infty} P(T^{-l}[B \cap \{\tau_j = l, X_{-l} \in C\}]) \\
&= \sum_{l=1}^{\infty} P(B \cap \{\tilde{\tau}_j = l, X_0 \in C\}) \\
&= P(B \cap \{X_0 \in C\}),
\end{aligned}$$

and the proof of Lemma 1 is complete.

Acknowledgements

The authors thank P. Algoet for his comments, suggestions, and encouragement. Suggestions by the referees have been helpful. The second author's work has been supported, in part, by NIH grant No. R01 A129426.

References

[1] Ash, R. (1972) *Real Analysis and Probability*. Academic Press.

[2] Algoet, P. (1992). Universal schemes for prediction, gambling and portfolio selection. *Annals of Probability*, **20**, pp. 901-941.

[3] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, **37**, pp. 357-367.

[4] Bailey D. (1976). Sequential Schemes for Classifying and Predicting Ergodic Processes. *Ph.D. thesis, Stanford University.*

[5] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[6] Castellana, J. V., and Leadbetter, M. R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Processes and their Application*, **21**, pp. 179-193.

[7] Collomb, G. (1985). Nonparametric time series analysis and prediction: uniform almost sure convergence. *Statistics*, 2, pp. 197-307.

[8] Cover T. (1975). Open Problems in Information Theory. *1975 IEEE-USSR Joint Workshop on Information Theory* pp. 35-36.

[9] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1-View*. Wiley, New York.

[10] Gray, R. (1988) *Probability, Random Processes, and Ergodic Properties.* Springer-Verlag, New York.

[11] Györfi, L. (1981). Strong consistent density estimation from ergodic sample. *J. Multivariate Analysis*, **11**, pp. 81-84.

[12] Györfi, L., Haerdle, W., Sarda, P., and Vieu, Ph. (1989) *Nonparametric Curve Estimation from Time Series,* Springer Verlag, Berlin.

[13] Györfi, L. and Lugosi, G. (1992). Kernel density estimation from ergodic sample is not universally consistent. *Computational Statistics and Data Analysis*, **14**, pp. 437-442.

[14] Györfi, L. and Masry, E. (1990). The L_1 and L_2 strong consistency of recursive kernel density estimation from time series. *IEEE Trans. on Information Theory*, **36**, pp. 531-539.

[15] Masry, E. (1986). Recursive probability density estimation for weakly dependent stationary processes. *IEEE Trans. Information Theory*, **IT-32**, pp 249-254.

[16] Ornstein, D. (1978). Guessing the next output of a stationary process. *Israel J. of Math.*, **30**, pp. 292-296.

[17] Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, M. Puri, Ed. London: Cambridge University, pp. 199-210.

[18] Roussas, G. (1969). Non-parametric estimation of the transition distribution of a Markov processes. *Annals of Inst. Statist. Math.* bf 21, pp.73-87.

[19] Stout, W. F. (1974). *Almost Sure Convergence,* Academic Press, New York.

[20] Yakowitz S. (1989). Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Analysis*, **30**, pp. 124-136.

[21] Yakowitz S. (1993). Nearest neighbor regression estimation for null-recurrent Markov time series. *Stochastic Processes and their Applications*, **37**, pp. 311-318.