საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 12, №3, 2018
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 12, no.3, 2018

Mathematics

Nonparametric Estimates of Distribution Density Constructed by Dependent Observations and Approximation Accuracy

Zurab Kvatadze* **and Beqnu Pharjiani***

*Faculty of Informatics and Management System, Georgian Technical University, Tbilisi, Georgia

(Presented by Academy Member Elizbar Nadaraya)

ABSTRACT: On the probabilistic space (Ω ,F , P), a two-component stationary (in the narrow sense) sequence $\{\xi_i X_i\}_{i\geq 1}$ is given, where $\{\xi_i\}_{i\geq 1}$ ($\xi_i : \Omega \to \Xi$) is a controlling sequence, and the members of the sequence $\{X_i\}_{i\geq 1}$ $X_i : \Omega \to R$ are observations on some random variable X. The cases of conditional independence and chainwise dependence of these observations are considered. Using observations $\{X_i\}_{i\geq 1}$, kernel observations of Rosenblatt-Parzen type of an unknown density of the variable X are constructed. The upper bounds of the mathematical expectations are established for the integral of the standard deviation of the obtained estimates from $f(x)$. *© 2018 Bull. Georg. Natl. Acad. Sci.*

Key words: sequence with chainwise dependence, kernel estimate, Markov chain

In the previous works, nonparametric estimates were constructed by means of independent samples. In the present paper, we consider the estimates constructed by dependent observations.

On the probabilistic space (Ω,F,P) we consider a two-component stationary (in the narrow sense) sequence of random variables

$$\{\xi_i, X_i\}_{i\geq 1}, \tag{1}$$

where $\xi_i : \Omega \to \Xi$, and $X_i : \Omega \to R^m$ is some space.

Definition 1. The sequence $\{X_i\}_{i\geq 1}$ from (1) is called a conditionally independent sequence (see [1]) controlled by the sequence $\{\xi_i\}_{i\geq 1}$, if, when fixing the trajectory $\overline{\xi}_{1n} = (\xi_1, \xi_2, .., \xi_n)$, the variables

$X_1, X_2, \ldots, X_n$ become independent for any n and for natural numbers i , k , n, $j_1, j_2, \ldots, j_k$, $(2 \le k \le n;$ $i \le n$; $1 \le j_1 < j_2 < \cdots < j_k \le n)$ the equalities

$$\mathcal{P}_{\left(X_{j_1}, X_{j_2}, \ldots, X_{j_k}\right)\left|\xi_{1n}\right.} = \mathcal{P}_{X_{j_1}\left|\xi_{j_1}\right.} \times \mathcal{P}_{X_{j_2}\left|\xi_{j_2}\right.} \times \cdots \times \mathcal{P}_{X_{j_k}\left|\xi_{j_k}\right.},$$

$$\mathcal{P}_{X_i\left|\xi_{1n}\right.} = \mathcal{P}_{X_i\left|\xi_i\right.},$$

are fulfilled, where $\mathcal{P}_{X|Y}$ denotes the conditional distribution of the variable X for the condition Y .

Definition 2. The conditionally independent sequence $\{X_i\}_{i\ge 1}$ from (1) is called a sequence with a chainwise dependence if $\{\xi_i\}_{i\ge 1}$ is a finite Markov chain with a discrete time.

Some facts about the density nonparametric estimates constructed by independent observations.

Let the values X_i, , $x_i \in R$, $i = 1,2\ldots$, be independent observations on some random variable X with an unknown density $g(x)$. Different methods are available for obtaining estimates of $g(x)$. In the works of M. Rosenblatt and E. Parzen [2, 3], the density $g(x)$ is estimated by the class of estimates defined by the kernel $k(x)$

$$\hat{g}_n(x, a_n) = \frac{a_n}{n}\sum_{i=1}^{n} k\left(a_n\left(x - X_i\right)\right),$$

where $\hat{f}_{in}(x, a_n) = \dfrac{a_n}{\nu_n(i)}\sum\limits_{j=1}^{\nu_n(i)} k\left(a_n\left(x - X_{\tau_j(i)}\right)\right)$, $i = 1,2$ is a sequence of positive numbers such that

$$\lim_{n\to\infty} a_n = \infty, \quad a_n = o(n) \tag{2}$$

and the kernel $k(x)$ is some Lebesgue-integrable Borel function. The results of [2] are generalized in the works [4-6]. Along with estimates of Rosenblatt-Parzen type, estimates of projection type were also considered (see [6-8]). Deviations of estimates from $g(x)$ were studied using various characteristics [2-6, 8, 9].

In [6], E. Nadaraya obtained the sufficient conditions for the uniform convergence of $\hat{g}_n(x, a_n)$ to $g(x)$ with probability 1.

Let us introduce the notation. We will call a function $k(x)$ a function of the class H_s $(s \ge 2$, s is an even number$)$ if it satisfies the following conditions

$$k(-x) = k(x), \quad \int_{-\infty}^{\infty} k(x)\,dx = 1, \quad \sup\left|k(x)\right| \le A < \infty,$$

$$\int_{-\infty}^{\infty} x^i k(x)\,dx = 0, \ i = 1,2,\ldots,s-1; \quad \int_{-\infty}^{\infty} x^s k(x)\,dx \ne 0, \quad \int_{-\infty}^{\infty} x^s\left|k(x)\right|dx < \infty.$$

Denote by W_s the set of functions $\varphi(x)$ that have derivatives up to order s $(s \ge 2)$ inclusive and note that $\varphi^{(s)}(x)$ is a continuous bounded function belonging to the class $L_2(-\infty, \infty)$.

Lemma (see [6]). If $g(x) \in W_s$ and $k(x) \in H_s \cap L_2(-\infty, \infty)$, then the equalities

$$\int_{-\infty}^{\infty} D\hat{g}_n(x, a_n)\,dx = \frac{a_n}{n}\int_{-\infty}^{\infty} k^2(x)\,dx + O\left(\frac{a_n}{n}\right),$$

$$\int_{-\infty}^{\infty}\left[E\hat{g}_n(x, a_n) - g(x)\right]^2 dx = a_n^{2s}\frac{\alpha^2}{(s!)^2}\int_{-\infty}^{\infty}\left[g^{(s)}(x)\right]^2 dx + O\left(a_n^{-2s}\right)$$

hold as $n \to \infty$ and

$$\alpha = \int_{-\infty}^{\infty} x^s k(x)\,dx.$$

Main Results.

Let us consider the sequence (1) where ξ_i, $i=1,2,...$, are independent, equally distributed random variables. Let

$$\Xi=\{b_1,b_2\}\ ;\ P(\xi_1=b_i)=p_i\ ,\ i=\overline{1,2}\ ,\ p_1+p_2=1.$$

Let $\{X_i\}_{i\geq 1}$ be a conditionally independent sequence whose elements are observations on the value X. For the fixed trajectory $\overline{\xi}_{1n}$ the conditional distributions $\mathcal{P}_{X_1|\xi_1=b_i}$, $i=\overline{1,2}$, have the unknown densities $f_i(x)$, $i=\overline{1,2}$ respectively.

For the fixed trajectory $\overline{\xi}_{1n}=(\xi_1,\xi_2,..,\xi_n)$ of the sequence $\{\xi_i\}_{i\geq 1}$ we denote by $\nu_n(1)$ and $\nu_n(2)$ the time moment frequencies at which the first *n* member of this sequence takes the values b_1 and b_2, respectively.

Theorem 1. *Let us consider the sequence (1). The elements of the controlling sequence* $\{\xi_i\}_{i\geq 1}$ $(\xi_i:\Omega\to\{b_1,b_2\})$ *are independent, equally distributed random variables* $\xi_i=b_1I_{(\xi_1=b_1)}+b_2I_{(\xi_1=b_2)}$. *It is assumed that for each function* $\Psi:\Xi\to R^1$, *for which* $E\Psi(\xi_1)<\infty$ *as* $n\to\infty$, *there holds the convergence*

$$\frac{1}{n}\sum_{j=1}^{n}\Psi(\xi_i)\to E\Psi(\xi_1)\ \text{ a.p.} \tag{3}$$

The elements of the conditionally independent sequence $\{X_i\}_{i\geq 1}$ are observations on the value X. The conditional distributions $\mathcal{P}_{X_1|\xi_1=b_i}$, $i=1,2$, have the unknown densities $f_i(x)$, $i=1,2$, respectively from the class $f_i(x)\in W_s$, $k(x)\in H_s\cap L_2(-\infty,\infty)$. If the inequalities

$$D\left(\frac{\nu_n(i)}{n}\right)\leq\frac{c_i}{\sqrt{n}},\ i=1,2, \tag{4}$$

are fulfilled for the frequencies $\nu_n(i)$, $i=1,2$, then for any n the density $\overline{f}(x)=p_1f_1(x)+p_2f_2(x)$ is estimated by

$$\hat{f}_n(x,a_n)=\frac{a_n}{n}\sum_{j=1}^{n}k\left(a_n(x-X_j)\right),$$

$$E\int_{-\infty}^{\infty}[\hat{f}_n(x,a_n)-(p_1f_1(x)+p_2f_2(x))]^2dx\leq(M_1+M_2)^2+\frac{a_n}{n}\int_{-\infty}^{\infty}k^2(x)dx+(\frac{1}{\sqrt{n}}(C_1+C_2)+p_1^2+p_2^2)\cdot O(\frac{a_n}{n}),$$

and the following equality

$$u(a_n)=(M_1+M_2)^2+\frac{a_n}{n}\int_{-\infty}^{\infty}k^2(x)dx+((C_1+C_2)n^{-1/2}+p_1^2+p_2^2)\cdot O(\frac{a_n}{n})$$

is valid, where

$$M_i=T_i^{1/2}+\left(C_in^{-1/2}\int_{-\infty}^{\infty}f_i^2(x)dx\right)^{1/2}$$

$$T_i=(a_n^{-2s}\frac{\alpha^2}{(s!)^2}\int_{-\infty}^{\infty}[f_i^{(s)}(x)]^2dx+O(a_n^{-2s}))(C_in^{-1/2}+p_i^2)\ \ i=1,2$$

and

$$\alpha = \int_{-\infty}^{\infty} x^s k(x)dx .$$

Proof. First we prove the finiteness of $E\hat{f}_n(x,a_n)$ and $D\hat{f}_n(x,a_n)$. The proof of the theorem is based on the decomposition of $\hat{f}_n(x,a_n)$ (for the fixed trajectory $\overline{\xi}_{1n}$) into two uncorrelated sums of independent summands. For the fixed trajectory $\overline{\xi}_{1n}$, we enumerate time moments, each separately, at which the members of the sequence $\{\xi_i\}_{i\geq 1}$ take the value b_i, $i=1,2$,

$$\tau_0(i)=0, \quad \tau_m(i)=\min\{j \mid \tau_{m-1}<j\leq n; \xi_i=b_i\}; \quad m=\overline{1,\nu_n(i)}, \quad i=1,2.$$

For the fixed trajectory $\overline{\xi}_{1n}$, the sum $\hat{f}_n(x,a_n)$ can be decomposed as follows

$$\hat{f}_n(x,a_n)=\frac{\nu_n(1)}{n}\hat{f}_{1n}(x,a_n)+\frac{\nu_n(2)}{n}\hat{f}_{2n}(x,a_n),$$

where

$$\hat{f}_{in}(x,a_n)=\frac{a_n}{\nu_n(i)}\sum_{m=1}^{\nu_n(i)}k\left(a_n\left(x-X_{\tau_{m(i)}}\right)\right), \; i=1,2.$$

Naturally, if $\nu_n(i)=0$, where $i=1,2$, then the summand $\hat{f}_{in}(x,a_n)$ does not exist.

Further we apply the lemma and the fact that the functions $\nu_n(1)$ and $\nu_n(2)$ are measurable with respect to the σ-algebra generated by the partitioning of Ω because of fixing of $\overline{\xi}_{1n}$.

Let us now consider the case when the controlling sequence $\{\xi_i\}_{i\geq 1}$ is a finite regular stationary Markov chain.

Theorem 2. Let in the sequence (1) the controlling sequence $\{\xi_i\}_{i\geq 1}$ be a finite homogeneous regular Markov chain, where $\Xi=\{b_1,b_2\}$, $\pi=(\pi_1,\pi_2)$ is the initial distribution of the chain, $\pi_i=P(\xi_1=b_i)$, $i=1,2$ and $P=(p_{ij})_{ij=1,2}$ is the matrix of transient probabilities, $\{X_i\}_{i\geq 1}$ is the sequence with a chainwise dependence whose elements are observations on the random value X. Assume that the conditional distributions $\mathcal{P}_{X_i|\xi_i=b_i}$, $i=1,2$, have the unknown densities $f_i(x)$, $i=1,2$, respectively, $f_i(x)\in W_s$ and $k(x)\in H_s\cap L_2(-\infty,\infty)$. Then for any n the estimate of the density $\overline{f}(x)=\pi_1 f_1(x)+\pi_2 f_2(x)$ is $\hat{f}_n(x,a_n)$ and the estimate

$$E\int_{-\infty}^{\infty}[\hat{f}_n(x,a_n)-(\pi_1 f_1(x)+\pi_2 f_2(x))]^2dx\leq$$

$$\leq (M_1^*+M_2^*)^2+\frac{a_n}{n}\int_{-\infty}^{\infty}k^2(x)dx+\left(\frac{1}{n}(C_1(\pi p)+C_2(\pi p))+\pi_1^2+\pi_2^2\right)O\left(\frac{a_n}{n}\right)$$

is valid, where

$$M_1^*=T_i^{*1/2}+\left(\frac{c_i(\pi,p)}{n}\int_{-\infty}^{\infty}f_i^2(x)dx\right)^{1/2}$$

and

$$T_i^{*1/2}=\left(a_n^{-2s}\frac{\alpha^2}{(s!)^2}\int_{-\infty}^{\infty}[f_i^{(s)}(x)]^2dx+O(a_n^{-2s})\right)\left(\frac{c_i(\pi,p)}{n}+\pi_i^2\right), \; i=1,2.$$

Proof. The proof is carried out in a manner analogous to that of Theorem 1 but without the ergodicity condition by virtue of which the convergence (3) is fulfilled for any function of the chain and the relations conditions (3) and (4). As is known (see [10]), for regular Markov chains we need to use the so-called

$$E\frac{\nu_n(i)}{n}=\pi_i,\quad D\left(\frac{\nu_n(i)}{n}\right)\le\frac{c_i(\pi,p)}{n}$$

are fulfilled, where $c_i(\pi,p)$ are the constants depending on the chain parameters.

It should be noted that the obtained upper bound of estimate approximation is much more accurate for observations with chainwise dependence than for conditionally independent observations.

The paper is dedicated to the blessed memory of our scientific leader Professor Tengiz Shervashidze.

მათემატიკა

დამოკიდებული დაკვირვებებით აგებული განაწილების სიმკვრივის არაპარამეტრული შეფასებები და მათი მიახლოების სიზუსტე

ზ. ქვათაძე* და ბ. ფარჯიანი*

**საქართველოს ტექნიკური უნივერსიტეტი, ინფორმატიკისა და მართვის სისტემების ფაკულტეტი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის ე. ნადარაიას მიერ)

(Ω,F,P) ალბათურ სივრცეზე მოცემულია ორკომპონენტიანი ვიწრო აზრით სტაციონარული მიმდევრობა $\{\xi_i X_i\}_{i\ge1}$, სადაც $\{\xi_i\}_{i\ge1}$, $(\xi_i:\Omega\to\Xi)$ მმართველი მიმდევრობაა, ხოლო $\{X_i\}_{i\ge1}$, $(X_i:\Omega\to\mathrm{R})$ მიმდევრობის წევრები წარმოადგენენ რაიმე X შემთხვევით სიდიდეზე დაკვირვებებს. განხილულია პირობითად დამოუკიდებელი და ჯაჭვურად დამოკიდებული დაკვირვებების შემთხვევები. $\{X_i\}_{i\ge1}$, დაკვირვებების საშუალებით აგებულია X სიდიდის უცნობი $f(x)$ სიმკვრივის როზენბლატ-პარზენის ტიპის გულოვანი შეფასებები. დადგენილია ამ შეფასებათა $f(x)$ სიმკვრივიდან გადახრის კვადრატის ინტეგრალის მათემატიკური ლოდინის ზედა საზღვრები.

REFERENCES

1. Bokuchava I. V., Kvatadze Z. A. and Shervashidze T. L. (1985) On limit theorems for random vectors controlled by a Markov chain. *Probability Theory and Mathematical Statistics,* **I**: 231-250. Vilnius (Lithuania).
2. Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.,* 27: 832-837, Chicago, USA.
3. Parzen E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.,* 33: 1065-1076. Stanford, USA.
4. Watson G. S., Leadbetter M. R. (1963) On the estimation of the probability density. I. *Ann. Math. Statist.,* 34: 480-491, Toronto, Canada.
5. Mania G. M. (1974) Statisticheskoe otsenivanie raspredeleniia veroiatnostei. Tbilisi (in Russian).
6. Nadaraya E. A. (1983) Neparametricheskoe otsenivanie plotnosti veroiatnostei i krivoi regressii. Tbilisi (in Russian).
7. Chencov. N. N. (1962) Otsenka neizvestnoi plotnosti raspredeleniia po nabliudeniiam. *Dokl. AN SSSR,* **147**, I: 643-648 M. (in Russian).
8. Devroi L., Diorfi L. (1988) Neparametricheskoe otsenivanie plotnosti L_1 podkhod 408s. M. (in Russian).
9. Mnacakanov R. M., Khmaladze E. B. (1981) Ob L_1 skhodimosti statisticheskikh iadernykh otsenok plotnosti raspredelenii. *Dokl. AN SSSR* **258**, 5: 1052-1055. M. (in Russian).
10. 10. Kemeni Dj., Snell Dj. (1970) Konechnye tsepi Markova. M. (in Russian).

Received May, 2018