საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 14, №1, 2020
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 14, no. 1, 2020

Mathematics

Kernel Estimations of the Density Distribution Constructed by Dependent Observations and the Accuracy of their Approximation by L_1 Metric

Zurab Kvatadze* **and Beqnu Pharjiani***

* *Department of Mathematics, Georgian Technical University, Tbilisi, Georgia*

(Presented by Academy Member Elizbar Nadaraya)

Kernel estimations of the Rosenblatt-Parzen type of unknown density distribution by conditionally independent and chain-dependent observations are constructed. The upper boundaries for the approximations of these densities constructed by estimates for L_1 metric are determined. The obtained results are specified for the case of Bartlett kernel and smoothing coefficient $a_n = \sqrt{n}$.

Markov chain, kernel estimate, conditionally independent sequence, sequence with chain dependence

The construction of nonparametric estimations of the density distribution is one of the urgent issues of mathematical statistics. Until recently estimations were constructed on independent samples. M. Rosenblatt [1] and E. Parzen [2] considered kernel-type estimations. G. Mania constructed a multidimensional kernel estimation [3]. E. Khmaladze and R. Mnatsakanov [4] studied a class of estimations with a kernel of a more general type. E. Nadaraya obtained sufficient conditions for the uniform convergence of the obtained estimation to density with probability 1 [5]. L. Devroye [6] constructed an estimate of density with a finite number of discontinuity points. Various characteristics in terms of the metric L_1 [4, 6], L_2 [1, 5] were considered as a measure of deviation of estimates from density. Building estimations with dependent observations began. Sidney Yakowitz [7], concerning density estimation by observations bound with Markov chain, in which Markov chain is considered by precision of general phase states, should be noted. In [8, 9] the upper boundaries of the approximation of the density with estimations constructed by dependent observations for L_2 metric are determined. In the present paper density estimations by conditionally independent observations and chain-dependent observations are constructed. In addition, accuracy in the approximation of density constructed by estimations on L_1 metric is determined.

Definition 1. Denote by F the set of functions $f(x)$ which satisfy the conditions: $f(x)$ is absolutely continuous function and has nearly everywhere the derivative f', f' is absolutely continuous function and has nearly everywhere the derivative f'', f'' is continuous and bounded.

Definition 2. Denote by Φ the set of functions $\varphi(x)$ which satisfy the conditions: φ density with compact medium which has four continuous forth order derivatives (inclusive), $\varphi \in F$, $\varphi'' \in F$ and $\varphi_a(x) = (1/a)\varphi(x/a)$.

Definition 3. Denote by $K^{\bullet}$ the class of bounded densities on R with compact medium for which $k(-x) = k(x)$.

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

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

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

Definition 4. The sequence $\{X_i\}_{i\geq 1}$ from (1) is called a conditionally independent sequence (see [10]) controlled by the sequence $\{\xi_i\}_{i\geq 1}$ if for any natural n and the fixed trajectory $\overline{\xi}_{1n} = (\xi_1, \xi_2, .., \xi_n)$, the values $X_1, X_2, \ldots, X_n$ become independent and for all natural numbers, $i, k, n, j_1, j_2, \ldots, j_k$, $(2 \leq k \leq n;\ i \leq n\ ;\ 1 \leq j_1 < j_2 <, \cdots, < j_k \leq n)$ the equalities

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

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

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

Definition 5. The conditionally independent sequence $\{X_i\}_{i\geq 1}$ in (1) is called a sequence with chain dependence [11], if $\{\xi_i\}_{i\geq 1}$ is a finite Markov chain with discrete time.

Consider the sequence (1). Let ξ_i, $i = 1, 2, ..$, be independent, identically distributed random variables and let

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

On the fixed trajectory $\overline{\xi}_{1n} = (\xi_1, \xi_2, .., \xi_n)$ of the sequence $\{\xi_i\}_{i\geq 1}$, we denote by $\nu_n(1)$, $\nu_n(2)$,…, $\nu_n(r)$ the frequencies with which the first n members of the sequence adopt the values b_1, b_2,… b_r.

Theorem 1. [5] Let us consider the sequence (3). $\Xi = \{b_1, b_2, ..., b_r\}$. The elements of the controlling sequence $\{\xi_i\}_{i\geq 1}$, $(\xi_i : \Omega \to \{b_1, b_2, ..., b_r\})$ are independent, identically distributed discrete random values $\xi_i = \sum_{j=1}^{r} b_j I_{(\xi_i = b_j)}$, $i = 1, 2, \ldots$. Let for every function $\Psi : \Xi \to R^1$, for which $E\Psi(\xi_i) < \infty$, the convergence

$$\frac{1}{n}\sum_{j=1}^{n} \Psi(\xi_j) \to E\Psi(\xi_1) \text{ a. s.}, \tag{2}$$

hold as $n \to \infty$.

The elements of the conditionally independent sequence $\{X_i\}_{i\geq 1}$ are the observations of the value *X*. The conditional distributions $\mathcal{P}_{X_i|\xi_i=b_j}$ $j=\overline{1,r}$ have respectively the unknown densities $f_j(x)$ (with compact support), $j=\overline{1,r}$, if the equalities

$$D\left(\frac{\nu_n(i)}{n}\right)\leq\frac{c_i}{\sqrt{n}},\quad i=\overline{1,r}$$

are fulfilled for the frequencies $\nu_n(i)$, $i=\overline{1,r}$, then for any natural *n* the estimate of $\overline{f}(x)=\sum_{i=1}^{r}p_i f_i(x)$ is $f_n(x,a_n)=\frac{a_n}{n}\sum_{j=1}^{n}k\left(a_n\left(x-X_j\right)\right)$, where $k(x)\in K^{\bullet}$ is a density limited by certain finite constant and for value $J(a_n)=\int_{-\infty}^{\infty}\left|f_n(x,a_n)-\overline{f}(x)\right|dx$ is valid the estimate

$$EJ(a_n)\leq\sqrt{\frac{a_n}{n}}\alpha\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx+$$

$$+\frac{\beta}{2a_n^2}\sum_{i=1}^{r}p_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\sqrt{\frac{a_n}{n}}\right)\tag{3}$$

If also $f_i(x)\in F$, $i=\overline{1,r}$, then

$$EJ(a_n)\leq\sqrt{\frac{a_n}{n}}\alpha\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx+$$

$$+\frac{\beta}{2a_n^2}\sum_{i=1}^{r}p_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\sqrt{\frac{a_n}{n}}\right).\tag{4}$$

where

$$\alpha=\sqrt{\int_{-\infty}^{\infty}k^2(x)dx},\ \beta=\int_{-\infty}^{\infty}x^2k(x)dx.$$

Corollary 1. If in the conditions of Theorem 1 $k(x)$ represents the Bartlett kernel

$$k(x)=\overline{k}(x)=\frac{3}{4}(1-x^2)\mathrm{I}_{[|x|\leq 1]}$$

then for arbitrary natural *n*, the estimate of density $\overline{f}(x)=\sum_{i=1}^{r}p_i f_i(x)$ will be the sum

$$\overline{f}_n(x,a_n)=\frac{3a_n}{4n}\sum_{i=1}^{n}(1-[a_n(x-X_i)]^2)\mathrm{I}_{[|x-X_i|\leq\frac{1}{a_n}]}\tag{5}$$

and for the value $\overline{J}(a_n)=\int_{-\infty}^{\infty}\left|\overline{f}_n(x,a_n)-\overline{f}(x)\right|dx$ the following estimate is valid

$$E\overline{J}(a_n)\leq\sqrt{\frac{3a_n}{5n}}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx+$$

$$+\frac{0.1}{a_n^2}\sum_{i=1}^{r}p_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\sqrt{\frac{a_n}{n}}\right).\tag{6}$$

Also, if $f_i(x) \in F$, $i = \overline{1,r}$, the following estimate will be valid:

$$E\bar{J}(a_n) \le \sqrt{\frac{3a_n}{5n}}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx +$$

$$+\frac{0.1}{a_n^2}\sum_{i=1}^{r}p_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\sqrt{\frac{a_n}{n}}\right). \tag{7}$$

Corollary 2. If in the conditions of Corollary 1 if we select the $\hat{f}_{in}(x,a_n)=\frac{a_n}{\nu_n(i)}\sum_{j=1}^{\nu_n(i)}k\left(a_n\left(x-X_{\tau_j(i)}\right)\right)$, $i=1,2$ sequence by the $a_n=\sqrt{n}$ formula in the conditions of Corollary 1, the estimate of density $\bar{f}(x)=\sum_{i=1}^{r}p_i f_i(x)$ for any natural *n* will be the following sum:

$$\bar{f}_n\left(x,\sqrt{n}\right)=\frac{3}{4\sqrt{n}}\sum_{i=1}^{n}(1-n(x-X_i)^2)\mathrm{I}_{[|x-X_i|\le\frac{1}{\sqrt{n}}]} \tag{8}$$

and for the $\bar{J}(\sqrt{n})=\int_{-\infty}^{\infty}\left|\bar{f}_n(x,\sqrt{n})-\bar{f}(x)\right|dx$ value, the following estimate is valid:

$$E\bar{J}(\sqrt{n}) \le \frac{1}{\sqrt[4]{n}}\sqrt{\frac{6}{5\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx +$$

$$+\frac{0.1}{n}\sum_{i=1}^{r}p_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\frac{1}{\sqrt[4]{n}}\right). \tag{9}$$

Also, if $f_i(x) \in F$, $i=\overline{1,r}$, the following estimate will be valid:

$$E\bar{J}(\sqrt{n}) \le \frac{1}{\sqrt[4]{n}}\sqrt{\frac{6}{5\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{p_i f_i(x)}dx +$$

$$+\frac{0.1}{n}\sum_{i=1}^{r}p_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i}+o\left(\frac{1}{\sqrt[4]{n}}\right). \tag{10}$$

Theorem 2. Let in the sequence (3) $\{X_i\}_{i\ge1}$ be the sequence with chain dependence, the elements of which represents the observations on value *X*. The conditional distributions, $\mathcal{P}_{X_i|\xi_i=b_j}$, $j=\overline{1,r}$, has the unknown densities $f_j(x)$, having the compact supports, $j=\overline{1,r}$. Then for arbitrary $\boldsymbol{n}$ estimate of density $\tilde{f}(x)=\sum_{i=1}^{r}\pi_i f_i(x)$ represents $f_n(x,a_n)=\frac{a_n}{n}\sum_{i=1}^{n}k\left(a_n(x-X_i)\right)$, where $k(x)\in K^{\bullet}$ and for value $J_1(a_n)=\int_{-\infty}^{\infty}\left|f_n(x,a_n)-\tilde{f}(x)\right|dx$ is valid the estimate

$$EJ_1(a_n) \le \sqrt{\frac{a_n}{n}}\alpha\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{\beta}{2a_n^2}\sum_{i=1}^{r}\pi_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt[4]{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\sqrt{\frac{a_n}{n}}\right). \tag{11}$$

If also $f_i(x)\in F$, $i=\overline{1,r}$, then

$$E J_1(a_n) \le \sqrt{\frac{a_n}{n}}\alpha\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{\beta}{2a_n^2}\sum_{i=1}^{r}\pi_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\sqrt{\frac{a_n}{n}}\right). \quad (12)$$

Corollary 3. If in the conditions of Theorem 2 $k(x)=\bar{k}(x)=\frac{3}{4}\left(1-x^2\right)\mathrm{I}_{[|x|\le 1]}$ then for arbitrary natural n, the estimate of density $\tilde{f}(x)=\sum_{i=1}^{r}\pi_i f_i(x)$ will be the sum

$$\bar{f}_n(x,a_n)=\frac{3a_n}{4n}\sum_{i=1}^{n}(1-[a_n(x-X_i)]^2)\mathrm{I}_{[|x-X_i|\le\frac{1}{a_n}]} \quad (13)$$

and for the value $\bar{J}_1(a_n)=\int_{-\infty}^{\infty}\left|\bar{f}_n(x,a_n)-\tilde{f}(x)\right|dx$ the following estimate is valid

$$E\bar{J}_1(a_n)\le\sqrt{\frac{3a_n}{5n}}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{0.1}{a_n^2}\sum_{i=1}^{r}\pi_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\sqrt{\frac{a_n}{n}}\right). \quad (14)$$

Also, if $f_i(x)\in F$, $i=\overline{1,r}$, the following estimate will be valid:

$$E\bar{J}_1(a_n)\le\sqrt{\frac{3a_n}{5n}}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{0.1}{a_n^2}\sum_{i=1}^{r}\pi_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\sqrt{\frac{a_n}{n}}\right). \quad (15)$$

Corollary 4. If in the conditions of Corollary 3 we select the $\hat{f}_{in}(x,a_n)=\frac{a_n}{\nu_n(i)}\sum_{j=1}^{\nu_n(i)}k\left(a_n\left(x-X_{\tau_j(i)}\right)\right)$, $i=1,2$ sequence by the $a_n=\sqrt{n}$ formula in the conditions of Corollary 1, the estimate of density $\tilde{f}(x)=\sum_{i=1}^{r}\pi_i f_i(x)$ for any natural n will be the following sum:

$$\bar{f}_n\left(x,\sqrt{n}\right)=\frac{3}{4\sqrt{n}}\sum_{i=1}^{n}(1-n(x-X_i)^2)\mathrm{I}_{[|x-X_i|\le\frac{1}{\sqrt{n}}]} \quad (16)$$

and for the $\bar{J}_1(\sqrt{n})=\int_{-\infty}^{\infty}\left|\bar{f}_n(x,\sqrt{n})-\tilde{f}(x)\right|dx$ value, the following estimate is valid:

$$E\bar{J}_1(\sqrt{n})\le\frac{1}{\sqrt[4]{n}}\sqrt{\frac{6}{5\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{0.1}{n}\sum_{i=1}^{r}\pi_i\sup_{a>0}\int_{-\infty}^{\infty}\left|\left(f_i*\varphi_a\right)''(x)\right|dx+\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\frac{1}{\sqrt[4]{n}}\right). \quad (17)$$

Also, if $f_i(x)\in F$, $i=\overline{1,r}$, the following estimate will be valid:

$$E\bar{J}_1(\sqrt{n}) \le \frac{1}{\sqrt[4]{n}}\sqrt{\frac{6}{5\pi}}\sum_{i=1}^{r}\int_{-\infty}^{\infty}\sqrt{\pi_i f_i(x)}dx +$$

$$+\frac{0.1}{n}\sum_{i=1}^{r}\pi_i\int_{-\infty}^{\infty}\left|f_i''(x)\right|dx+\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\sqrt{c_i(\pi,P)}+o\left(\frac{1}{\sqrt[4]{n}}\right). \qquad (18)$$

Therefore, we built nonparametric density estimates with conditionally independent and chain-dependent observations and determined the approximation densities with these estimates.

The method used to validate the theorems is to transit from a mathematical wait to a conditional mathematical wait during the observation of a $\bar{\xi}_{1n}=(\xi_1,\xi_2,...,\xi_n)$ trajectory.

The sum to be considered on fixed trajectory is divided into several summons. One of them is the sum of independent summons on fixed trajectory. After some transformations, it results as a type for which famous results are used from the theory of building of estimations with independent observation [6]. For estimations of the rest of the sums inequalities of Fubini and Helder are and the characteristics of the governing sequences. For the estimation of all summons are because the $\nu_n(i)$ and $f(\nu_n(i))$ functions (when $f\in L_2(-\infty,\infty)$ are measures toward σ-algebra induced by the division of the Ω space generated by the fixation of the ξ_{1n} trajectory [12]. Because of this, they are taken out of the conditional mathematical waiting mark.

The method used allows for the future to consider the estimation of various parameters by dependent observations.

მათემატიკა

სიმკვრივის დამოკიდებული დაკვირვებებით აგებული გულოვანი შეფასებები და მათი მიახლოების სიზუსტე L_1 მეტრიკით

ზ. ქვათაძე* და ბ. ფარჯიანი*

* *საქართველოს ტექნიკური უნივერსიტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის ე. ნადარაიას მიერ)

პირობითად დამოუკიდებელი და ჯაჭვურად დამოკიდებული დაკვირვებებით აგებულია განაწილების უცნობი სიმკვრივის როზენბლატ-პარზენის ტიპის გულოვანი შეფასებები. დადგენილია აგებული შეფასებებით სიმკვრივის მიახლოების ზედა საზღვრები L_1 მეტრიკით. მიღებული შედეგები დაზუსტებულია ბარტლეტის გულის შემთხვევაში და გაგლუვების $a_n=\sqrt{n}$ კოეფიციენტისათვის.

REFERENCES

1. Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist*., **27**: 832-637. Chicago, USA.
2. Parzen E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.,* 33: 1065-1076. Stanford, USA.
3. Mania G.M. (1974) Statisticheskoe otsenivanie raspredeleniia veroiatnostei, p. 238. Tbilisi (in Russian).
4. Mnatsakanov R.M., Khmaladze E.M. (1981) Ob L_1-skhodimosti statisticheskikh iadernykh otsenok plotnostei raspredelenii. *Dokl. Akad. Nauk SSSR* **258,** 5: 1052-1055. M. (in Russian).
5. Nadaraya E.A. (1983) Neparametricheskoe otsenivanie plotnosti veroiatnostei i krivoi regressii, p. 194. Tbilisi (in Russian).
6. Devroye L., Györfi L. (1985) Nonparametric density estimation: the L_1 view. Wiley series in probability and mathematical statistics, p. 367. Canada, USA.
7. Yakowitz Sidney (1989) Nonparametric density and regression estimation for Markov sequences without mixing assumptions. 85721–*Journal of Multivariate Analysis*, 30: 124-136. Arisona, USA.
8. Kvatadze Z., Pharjiani B. (2018) Nonparametric estimates of a distribution density constructed by dependent observations and their approximation accuracy. *Bull. Georg. Natl. Acad. Sci.,* **13**, 2: 7-11. Tbilisi (in Georgian).
9. Kvatadze Z., Pharjiani B. (2019) On the exactness of distribution density estimates constructed by some class of dependent observations. *Mathematics and Statistics,* **7**(4): 135-145. San Jose, USA.
10. Kvatadze Z., Shervashidze T. (2008) Some limit theorems for IID and conditionally independent random variables. *The second international Conference, Problems of Cybernetics and Informatics*. September 10-12, **II**: 217-219. Baku (Azerbaijan).
11. Kvatadze Z., Shervashidze T. (1986) On limit theorems for conditionally independent random variable controlled by a finite Markov chain. Probability theory and mathematical stastistics. Proc. 5th Japan-USSR Symposium on Probability Theory. *Lecture Notes in Mathematics,* **1299**: 250-259. Springer-Verlag. Berlin (Germany).
12. Kvatadze Z., Kvatadze TS., Maisuradze A. (2019) Limiting distribution of a sequence of functions defined on a Markov chain. XXXIII Enlarged Sessions of the Seminar of Ilia Vekua Institute of Applied Mathematics (VIAM), (TSU). April, 23-25, Book of Abstracts, p. 82. Tbilisi (in Georgian).

Received November, 2019