Open Journal of Statistics, 2011, 1, 157-160
doi:10.4236/ojs.2011.13019 Published Online October 2011 (http://www.SciRP.org/journal/ojs)

A Note on Spline Estimator of Unknown Probability Density Function

Muhanmadjon S. Muminov[1*], Kh. Soatov[2]
[1]*Institute of Mathematics and Information Technologies*, *Tashkent University*, *Tashkent*, *Uzbekistan*
[2]*Tashkent University of Information Technologies*, *Tashkent University*, *Tashkent*, *Uzbekistan*
E-mail: [*]*m.muhammad@rambler.ru*
Received March 30, 2011; *revised May* 4, 2011; *accepted May* 13, 2011

Abstract

In the present paper as estimation of unknown pdf derivative of a spline function is suggested. It is studied its some statistical properties which are used to approximate maximal deviation of the spline estimation from pdf with maximum of nonstationary gaussian process.

Keywords: Spline-Estimator, Empirical Distribution, Gauss Process

1. Introduction

The construction of a confidence interval for unknown probability density function (pdf) trough histogram for the first time has been suggested by Smirnov [1]**.** Bikel and Rosenblatt [2], Rosenblatt [3] have considered analogues problem using of Parsen-Rosenblatt's estimation. The problem of construction of a confidence interval for unknown pdf trough spline-function was studied by Muminov and Khashimov [4]. Recently for unknown multidimensional distribution density function the kernel estimation is constructed and similar problem is studied by Muminov [5,6].

Several authors have considered the rate of convergence of the distribution of the maximum of difference between Parsen-Rosenblat's estimator and unknown pdf, see, for example, Konakov and Piterbarg [7-9]. Nevertheless there is no such kind of result for the spline-estimators. The results obtained in this work help to approximate the deviation of spline estimation of unknown density by Gaussian process.

It should be noted that in the works of Lii and Rosenblatt [10], Muminov [11] asymptotical unbiasedness and strong state of the spline estimation are proved. Importance of spline-estimation and its application in statistics are given in the works [5,12].

The paper is organized as follows. In Sec. 2 the spline estimation is constructed and some auxiliary results are stated, and also the main theorem is given. The main theorem is proved in Sec. 3.

2. Results

Let $X_1, X_2, \cdots, X_n$ be independent identical distributed random variables (r.v.) with pdf $f(x)$ and let $S_n(x)$ be the cubic spline-function which do interpolation of $y_k = F_n(x_k)$ at the points $x_k = k/N$, $k = 0,1,\cdots,N$ where $N = N(n)$, $F_n(x)$ is the epirical distribution function of the sample $X_1, X_2, \cdots, X_n$. Theboundary condition for $S_n(x)$ are $S_n^{'}(0) = \frac{y_1 - y_0}{h}$, $S_n^{'}(1) = \frac{y_N - y_{N-1}}{h}$, $h = \frac{1}{N}$.

Then the derivative of spline-function $S_n(x)$ is as follows, see Lii [13]

$$S_n^{'}(x) = \frac{1}{h}\int_0^1 W_N(x,y)\,\mathrm{d}F_n(y),$$

where $W_N(x,y) = W_{N,i}(x,y) = E_{i,j}(x)$, $x \in [x_{i-1}, x_i]$, $y \in [x_j, x_{j+1}]$, $i = \overline{1,N}$, $j = \overline{0,N-1}$

$$E_{i,j} = E_{i,j}(x) = \begin{cases} D_{i,j}(x) \text{ if } j \ne i-1 \\ D_{i,j}(x) + 1 \text{ if } j = i-1 \end{cases}$$

$$D_{i,j} = D_{i,j}(x) = \begin{cases} -\frac{3}{2}C_{i,1}(x) \text{ if } j = 0 \\ \frac{3}{2}\left[C_{i,j}(x) - C_{i,j+1}(x)\right] \text{ if } j = \overline{1,N-2} \\ \frac{3}{2}C_{i,N-1}(x) \text{ if } j = N-1 \end{cases}$$

$$C_{i,j} = C_{i,j}(x) = \left[\frac{1}{3} - (1-r)^2\right] A_{i-1,j}^{-1} + \left(r^2 - \frac{1}{3}\right) A_{i,j}^{-1},$$

$$r = \frac{x - x_{i-1}}{h}$$

$$A_{i,j}^{-1} = \frac{\sigma^{j-i}\left(1+\sigma^{2i}\right)\left(1+\sigma^{2N-2j}\right)}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \text{ for } 0 < i \le j < N,$$

$$A_{i,N}^{-1} = \frac{\sigma^{N-i}\left(1+\sigma^{2i}\right)}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \text{ for } 0 < i \le N,$$

$$A_{0,j}^{-1} = \frac{2\sigma^{j}\left(1+\sigma^{2N-2j}\right)}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \text{ for } 0 < j < N,$$

$$A_{0,N}^{-1} = \frac{2\sigma^N}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \quad A_{0,0}^{-1} = \frac{2-\sigma^{2N-1}(1+\sigma)^2}{2(2+\sigma)\left(1-\sigma^{2N}\right)}$$

$\sigma = \sqrt{3} - 2$, $A_{i,j}^{-1} = A_{j,i}^{-1}$,for $0 < i < N$, $0 < j < N$ and $A_{i,j}^{-1} = A_{N-i,N-j}^{-1}$ for the other values of $I, j,\ 1 \le j \le i < N$.

We take the statistic $S_n^{'}(x)$ as estimator of pdf $f(x)$. We define r.v. $f(x)$ by the following equality

$$\xi_n = \sqrt{nh} \max_{0\le x\le 1} \left| \frac{S_n^{'}(x) - f(x)}{\sigma_N(x)\sqrt{f(x)}} \right|,$$

where $\sigma_N^2(x) = \frac{1}{h}\int_0^1 W_N^2(x,y)\mathrm{d}y$.

R.v. ξ_n is interesting with point of view of solution of the following problems:

1) to find a confidential strip for $f(t)$, $t \in [0,1]$ on given coefficient of trust α $(0 < \alpha < 1)$;

2) to construct criterion for test of null hypothesis $H_0 : f(t) = f_0(t)$ on given significance level $\beta(0 < \beta < 1)$.

Our main goal in the sequel is: to solve the problems 1) and 2). For this we have to find limit distribution of r.v. ξ_n. The results, obtained in this work, allow to approximate distribution of r.v. ξ_n with distribution of maximum of Gaussian process.

Let $F_n^*(x)$ be an empirical distribution function of the sample $F(X_1), \cdots, F(X_n)$, $\{\omega_n(t),\ t \in [0,1]\}$ be a sequence of Wiener process. Set

$$Y_n(t) = \sqrt{n}\left[F_n^*(t) - t\right], \quad t \in [0,1]$$

$$B_n(t) = \omega_n(t) - t\omega_n(1), \quad t \in [0,1]$$

$$\xi_n^*(x) = \sqrt{nh}\, \frac{S_n^{'}(x) - ES_n^{'}(x)}{\sigma_N(x)\sqrt{f(x)}}$$

$$\xi_n^{(1)}(x) = \frac{1}{\sigma_N(x)\sqrt{hf(x)}} \int_0^1 W_N(x,y)\mathrm{d}\omega_n\left(F(y)\right)$$

$$\xi_n^{(2)}(x) = \frac{1}{\sigma_N(x)\sqrt{hf(x)}}$$

$$\int_0^1 W_N(x,y)\left[Y_n\left(F(y)\right) - B_n\left(F(y)\right) - \omega_n(1)F(y)\right],$$

$$\eta_n(x) = \frac{1}{\sigma_N(x)\sqrt{h}} \int_0^1 W_N(x,y)\mathrm{d}\omega_n(y)$$

$$\eta_n^{(1)}(x) = \frac{1}{\sigma_N(x)\sqrt{h}} \int_0^1 W_N(x,y)\left[\sqrt{\frac{f(y)}{f(x)}} - 1\right]\mathrm{d}\omega_n(y).$$

It is evident that

$$\xi_n^*(x) = \frac{1}{\sigma_N(x)\sqrt{hf(x)}} \int_0^1 W_N(x,y)\mathrm{d}Y_N\left(F(y)\right) = \xi_n^{(1)}(x) + \xi_n^{(2)}(x)$$

and the structure of co-variations of the Gaussian processes $\eta_n(x) + \eta_n^{(1)}(x)$ and $\xi_n^{(1)}(x)$ is coincided.

We assume that $nh \to \infty$, $h \to 0$ as $n \to \infty$ and the following conditions are fulfilled:

1) $f(x) \ge C_0 > 0$, $\forall x \in [0,1]$,

2) The pdf $f(x)$ continuously differentiable in the interval [0, 1].

In what follows *C* and *c* with or without index is universal positive number.

Theorem. Suppose that the conditions 1) and 2) are satisfied. Then for arbitrary $\varepsilon > 0$ one has

$$P\left(\max_{0\le x\le 1}\left|\eta_n^{(1)}(x)\right| > \varepsilon\right) \le \frac{C_1 h}{\log N} + \frac{\sqrt{h}}{\varepsilon} C_2 \exp\left\{-C_3 h^{-1}\varepsilon^2\right\} \tag{1}$$

Also there is *C* such that

$$\max_{0\le x\le 1}\left|\xi_n^{(2)}(x)\right| \le C \max\left(\frac{\log n}{\sqrt{nh}}, \sqrt{h\log n}\right) \tag{2}$$

with probability equal to 1. The following assertion is proved by Komlosh *et al.* [14].

Lemma 1. There exist a probabilistic space (Ω, F, P) where it is possible to define version of the $F_n^*(t)$ and the sequence of Brownian bridge $B_n(t)$ such that for all *x*.

$$P\left(\sup_{0\le t\le 1}\left|n\left[F_n^*(t) - t\right] - \sqrt{n}B_n(t)\right| > c_1 \log n + x\right) \le c_2 \mathrm{e}^{-c_3 x}.$$

Lemma 2. Let $1 \le i \le N$. $0 \le j \le N-1$ For all *l* and *J* such that

$$|i - j| \ge \frac{\alpha \log N}{|\log 0{,}3|} + 1$$

where $\alpha > 0$, one has

$$\max_{0\le x\le 1}\left|E_{i,j}(x)\right| \le \frac{16}{N^\alpha}.$$

Also for any $i \in \{1, 2, \cdots, N\}$ and $x \in \left[x_{i-1}, x_i\right]$ the following holds

$$\sum_{j=0}^{N-1}\left|E_{i,j}(x)\right|\le 16 .$$

the following Lemma 3 is proved in the book of Lamperty [15].

Lemma 3. Let $X_1, X_2, \cdots$ be a sequence of standard normal distributed r.v.s then

$$P\left(\left|X_n\right| = O\left(\sqrt{\log n}\right)\right) = 1$$

3. Proofs of the Main Results

The proof of Lemma 2 is simple and hence it is omitted. The proof of the main theorem. We have

$$\left|\xi_n^{(2)}(x)\right| \le h^{-1/2}\left[\sigma_N(x)\sqrt{f(x)}\right]^{-1}\left\{\left|\omega_n(1)\right|\left|\int_0^1 W_N(x,y)\mathrm{d}F(y)\right|\right.$$

$$\left.+\left|\int_0^1 W_N(x,y)d\left[Y_n(F(y))-B_n(F(y))\right]\right|\right\}.$$

Hence

$$\max_{0\le x\le 1}\left|\xi_n^{(2)}(x)\right| \le \sqrt{h}C_6\left|\omega_n(1)\right| + h^{-\frac{1}{2}}C_7\max_{0\le t\le 1}\left|Y_n(t)-B_n(t)\right|$$

because $\inf_{0\le x\le 1}\sigma_N(x) = C_5 > 0$. From Lemma 1 and 3 it follows that as $n\to\infty$

$$\sup_{0\le t\le 1}\left|Y_n(t)-B_n(t)\right| = 0\left(\frac{\log n}{\sqrt{n}}\right) \text{ and } \omega_n(1) = O(\sqrt{\log n})$$

with probability equal to 1. The relation (2) follows.

Let $x\in\left[x_{i-1}, x_i\right]$. Then

$$\eta_n^{(1)}(x) = J_1(x) + J_2(x)$$

where

$$J_1(x) = \left[\sigma_N(x)\sqrt{hf(x)}\right]^{-1}\sum_{j=0}^{i-1}E_{i,j}(x)$$

$$\int_{x_j}^{x_{j+1}}\left[\sqrt{f(y)}-\sqrt{f(x)}\right]\mathrm{d}\omega_n(y)$$

$$J_2(x) = \left[\sigma_N(x)\sqrt{hf(x)}\right]^{-1}\sum_{j=i}^{N-1}E_{i,j}(x)$$

$$\int_{x_j}^{x_{j+1}}\left[\sqrt{f(y)}-\sqrt{f(x)}\right]\mathrm{d}\omega_n(y)$$

for $i = N$ we suppose $\sum_N^{N-1}\cdot = 0$. Set

$$J_i^1(x) = \sum\nolimits_j^1 E_{i,j}(x)\left[\sqrt{f(x)}-\sqrt{f\left(x_j\right)}\right]\left[\omega_n\left(x_{j+1}\right)-\omega_n\left(x_j\right)\right]$$

$$J_i^2(x) = \sum\nolimits_j^2 E_{i,j}(x)\left[\sqrt{f(x)}-\sqrt{f\left(x_j\right)}\right]\left[\omega_n\left(x_{j+1}\right)-\omega_n\left(x_j\right)\right]$$

where $\sum_j^1$ and $\sum_j^2$ are denoted a summation over all $i\in\{1,2,\cdots,N\}$ and $j\in\{0,1,\cdots,i-1\}$ satisfying the inequalities

$$i-j < \frac{2\log N}{\left|\log 0,3\right|}+1 \quad (3)$$

and

$$i-j \ge \frac{2\log N}{\left|\log 0,3\right|}+1 \quad (4)$$

respectively. Integrating by part we find

$$\left|J_1(x)\right| \le C_8\sqrt{h}\max_{0\le t\le 1}\left|\omega_n(t)\right| + C_9 h^{-\frac{1}{2}}\left[\left|J_i^1(x)\right|+\left|J_i^2(x)\right|\right]. \quad (5)$$

Put

$$\xi_j^* = \frac{\omega_n\left(x_{j+1}\right)-\omega_n\left(x_j\right)}{\sqrt{h}}, \quad \bar{\xi}_N = \max_{0\le j\le N-1}\left|\xi_j^*\right|.$$

Since (3)

$$x-x_j = \left(x-x_{i-1}\right)+\left(x_{i-1}-x_j\right) \le h + \frac{i-j-1}{N} \le \frac{2\log N}{N\left|\log 0,3\right|}$$

According to conditions a), b), also Lagrange's mean-value theorem and form the last inequality we have

$$\left|J_i^1(x)\right| \le \sum_j{}^1\left|E_{i,j}(x)\right|\left|x-x_j\right|\left|\frac{f'(\bar{x})}{2\sqrt{f'(\bar{x})}}\right|\bar{\xi}_N$$

$$\le \bar{C}_9\frac{\log N}{N}\sum_j{}^1\left|E_{i,j}(x)\right|\bar{\xi}_N$$

where $x_j < \bar{x} < x_i$. Here we take into account $x - x_j \le \frac{2\log N}{N\left|\log 0,3\right|}$ too.

From lemma 2 we obtain $\sum_j^1\left|E_{i,j}(x)\right|\le 16$. Combining above-mentioned we obtain

$$\max_{1\le i\le N}\max_{x_{i-1}\le x\le x_i}\left|h^{-\frac{1}{2}}J_i^1(x)\right| \le C_{10}\frac{\log N}{N}\bar{\xi}_N. \quad (6)$$

Similarly $J_i^1(x)$ we have

$$\left|J_i^2(x)\right| \le \sum\nolimits_j^2\left|E_{i,j}(x)\right| 2\sup_{0\le x\le 1}\sqrt{f(x)}\bar{\xi}_N.$$

By virtue of lemma 2 when (4) is fulfilled the following is true

$$\left|E_{i,j}(x)\right| \le \frac{16}{N^2}, \quad \forall x\in[x_{i-1}, x_i].$$

As a result we have

$$\max_{1\le i\le N}\max_{x_{i-1}\le x\le x_i}\left|h^{-\frac{1}{2}}J_i^2(x)\right| \le C_{11}\frac{1}{N}\bar{\xi}_N \quad (7)$$

Reasoning alike presented at p. 410 of Cramér [16] we find

$$\bar{\xi}_N = \sqrt{2\log N} - \frac{\log\log N + \ln 4\pi}{2\sqrt{2\log N}} - \frac{\log \xi/2}{\sqrt{2\log N}} + 0\left(\frac{1}{\log N}\right) \tag{8}$$

where ξ is a random variable with pdf

$$g(x) = \begin{cases} \left(1 - \dfrac{x}{N}\right)^{N-1} & \text{if } x \in [0, N], \\ 0 & \text{if } x \notin [0, N]. \end{cases}$$

It is known (see, (29.2) of Skorohod [17]) that for arbitrary $\varepsilon > 0$

$$P\left(\sup_{0\le t\le h} |\omega_n(t)| > \varepsilon\right) \le 2P\left(\sup_{0\le t\le h} \omega_n(t) > \varepsilon\right) = \frac{1}{h\sqrt{2\pi}} \int_{\varepsilon}^{\infty} \mathrm{e}^{-\frac{x^2}{2h}} \mathrm{d}x .$$

Use this, (5)-(8) and Chebishev's inequality to get

$$P\left(\max_{0\le x\le 1} |J_1(x)| > \frac{\varepsilon}{2}\right) \le C_{12} \frac{h}{\log N} + C_{13} \frac{\sqrt{h}}{\varepsilon} \times \exp\left\{-\frac{C_{14}\varepsilon^2}{h}\right\} \tag{9}$$

By same way we can find that

$$P\left(\max_{0\le x\le 1} |J_2(x)| > \frac{\varepsilon}{2}\right) \le C_{15} \frac{h}{\log N} + C_{16} \frac{\sqrt{h}}{\varepsilon} \times \exp\left\{-\frac{C_{17}\varepsilon^2}{h}\right\} \tag{10}$$

The inequality (1) follows from (9) and (10). The proof of Theorem is completed.

The theorem allows to approximate the distribution of r.v. ξ_n by distribution of the maximum of Gaussian process $\eta_n(x)$.

4. References

[1] N. B. Smirnov, "On Construction of a Confidence Interval for the Probability Density Function," *Soviet Reports*, Vol. 74, 1959, pp. 1189-1191.

[2] P. J. Bikel and M. Rosenblatt, "On Some Global Measures of the Deviations of Density Functions Estimates," *The Annals of Statistics*, Vol. 1, No. 6, 1973, pp. 1071-1095.

[3] M. Rosenblatt, "On the maximal deviation of k-dimensional density estimates", *Annals of Probability*, Vol. 4, No. 6, 1976, pp. 1009-1015. doi:10.1214/aop/1176995945

[4] M. S. Muminov and Sh. A. Khashimov, "On Limit Distribution of the Maximal Deviation of Spline Density Estimators," FAN, Tashkent, 1986.

[5] M. S. Muminov, "On a Limit Distribution of the Maximal Level of Empirical Distribution Density and the Regression Function. I," *Theory Probability and Its Application*, Vol. 55, No. 3, 2010, pp. 582-590.

[6] M. S. Muminov, "On a Limit Distribution of the Maximal Level of Empirical Distribution Density and the Regression Function. II," *Theory Probability and Its Application*, Vol. 56, No. 1, 2011, pp. 162-173.

[7] V. D. Konakov and V. I. Piterbarg, "On the Convergence Rate of Maximal Deviations Distribution for Kernel Regression Estimates," *Journal of Multivariate Annalysis*, Vol. 15, No. 3, 1984, pp. 279-294. doi:10.1016/0047-259X(84)90053-8

[8] V. D. Konakov and V. I. Piterbarg, "High Level Excursions of Gaussian Fields and the Weakly Optimal Choice of the Smoothing Parameter. I," *Mathematical Methods of Statistics*, Vol. 4, 1995, pp. 481-434.

[9] V. D. Konakov and V. I. Piterbarg, "High Level Excursions of Gaussian Fields and the Weakly Optimal Choice of the Smoothing Parameter. II," *Mathematical Methods of Statistics*, Vol. 1, 1997, pp. 112-124.

[10] K. S. Lii and M. Rosenblatt, "Asymptotic Behavior of a Spline of a Density Function," *Computters & Mathematics with Applications*, No. 1, 1975, pp. 223-235.

[11] M. S. Muminov, "On Statistical Estimation of the Probability Density Function by LineFunctions," Ph.D. Thesis, Tashkent, p. 110.

[12] M. S. Muminov, "On Approximating the Probability of a Large Excursion a Nonstationary Gaussian Process," *Siberian Mathematical Journal*, Vol. 51, No. 1, 2010, pp. 175-195. doi:10.1007/s11202-010-0015-6

[13] K. S. Lii, "A Global Measure of a Spline Density Estimate," *The Annals of Statistics*, Vol. 6, No. 5, 1978, pp. 1138-1148. doi:10.1214/aos/1176344316

[14] Y. Komlos, P. Major and G. Tusnady, "An Approximation of Partial Sums of Independent RV's and the Sample DF. I," *Probability Theory and Related Fields*, Vol. 32, No. 1-2, 1975, pp.111-131.

[15] G. Lamperty, "Probability," Nauka, Moscow, 1973.

[16] G. Cramér, "The Mathematical Method in Statistics," Mir, Moscow, 1976.

[17] A. V. Skorohod, "The Random Processes with Independent Increments," Nauka, Moskov, 1964.