Open Journal of Statistics, 2016, 6, 373-379
Published Online April 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.62032

Strong Consistency of the Spline-Estimation of Probabilities Density in Uniform Metric

Mukhammadjon S. Muminov[1], Khaliq S. Soatov[2]

[1]Institute of Mathematics, National University of Uzbekistan, Tashkent, Uzbekistan
[2]Tashkent University of Information Technologies, Tashkent, Uzbekistan
Email: m.muhammad@rambler.ru, kh.soatov@mail.ru

Received 5 December 2015; accepted 24 April 2016; published 27 April 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

In the present paper as estimation of an unknown probability density of the spline-estimation is constructed, necessity and sufficiency conditions of strong consistency of the spline-estimation are given.

Keywords

Strong Consistency, Spline-Estimation, Probability Density in Uniform Metric, Uniform Metric, Soatov, Muminov, Tashkent University, Institute of Mathematics

1. Introduction

We assume that on the interval $[a,b]$, $a,b \in (-\infty,+\infty)$, $a < b$. The following mesh

$$\Delta_N : a = x_0 < x_1 < \cdots < x_N = b, \quad (1)$$

is given, where N is a natural number. Let P_k be the set of polynomials of degree $\le k$ and $C_k[a, b]$ be the set of continuous on the $[a, b]$ functions having continuous derivative of order k, $k = 1, 2, \cdots$. In the book of Stechkin and Subbotin [1] the following is given.

Definition. The function $S_N(x) = S_N(x, F)$ is called by interpolation cubic spline with respect to the mesh (1) for the function $F(x)$, if:

a) $S_N(x) \in P_3,\ x \in [x_{i-1}, x_i], i = \overline{1, N}$,

b) $S_N(x) \in C_2[a,b]$,

c) $S_N(x_i) = t_i = \overline{0, N}, N \ge 2.$

Here $t_i = F(x_i) \cdot i = \overline{0, N}$.

The points $\{x_i\}$ are called by the nodes of the spline.

How to cite this paper: Muminov, M.S. and Soatov, K.S. (2016) Strong Consistency of the Spline-Estimation of Probabilities Density in Uniform Metric. *Open Journal of Statistics*, **6**, 373-379. http://dx.doi.org/10.4236/ojs.2016.62032

Later on for convenience we let $[a,b]=[0,1]$ and the obtained results will remain valid for any finite interval $[a, b]$.

Let $X_1, X_2, \cdots, X_N$ be independent identical distributed random variables with unknown density distribution $f(x)$ concentrated and continuous on the interval [0, 1], and $S_N(x)$ be cubic spline interpolating the values $y_k = F_n(x_k)$ in the points $x_k = kh$, $k=\overline{0,N}$, $N=N_{(n)}$ with "boundary conditions"

$$S'_N(a)=a_N, \quad S'_N(b)=b_N.$$

Here $F_n(x)$ is the empirical function of the distribution of the sample $X_1, X_2, \cdots, X_N$, $h=\frac{1}{N}$ and $nh \to \infty$, $h \to 0$ as $n \to \infty$, a_N and b_N are given real numbers. Concrete choice of these numbers depends on the considered problem.

As estimation of an unknown probability density we take the statistics $S'_N(x)$.

In the present work as estimation of the unknown density $f(x)$ we take the statistics $S'_n(x)$ defined as in Theorem 1 and in Theorem 2 as well.

It is clear that, in Theorems 1 and 2 spline estimations are constructed with different boundary conditions.

Theorem 3 is devoted to asymptotic unbiasedness of the spline estimation. Also for completeness of the results the dispersion and the covariance of the spline-estimation are given.

In the main Theorem 4 necessity and sufficiency conditions for strong consistency of the spline-estimation are given.

Similar result for the Persen-Rozenblatt estimation is obtained in the book of Nadaraya (1983) [2].

More detailed review on spline estimation is given in works of Wegman, Wright [3], Muminov [4].

2. Auxiliary Results

Using the results of the work Lii [5] the following theorems are easily proved.

2.1. Theorem 1

Let $F_n(x)$ be empirical function of the distribution constructed by simple sample $X_1, X_2, \cdots, X_N$ and $S_N(x)$ be cubic spline interpolating the values $F_n(x_k)$ in the nodes of the mesh (1). *If we choose the boundary conditions for $S_N(x)$ in the form*

$$a_N = \frac{y_1 - y_0}{h}, \quad b_N = \frac{y_N - y_{N-1}}{h}$$

then the derivative $S'_N(x)$ of the spline function is defined by the equality

$$S'_N(x) = \frac{1}{h}\int_0^1 W_N(x,y)\,\mathrm{d}F_N(y).$$

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

$$E_{i,j}(x) = \begin{cases} D_{i,j}(x), & j \neq i-1 \\ D_{i,j}(x)+1, & j = i-1 \end{cases}$$

and

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

$C_{i,j}(x)$ are defined by the following relations:

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

$$r = \frac{x-x_{i-1}}{h}, \quad i=\overline{1,N}, \quad j=\overline{0,N-1},$$

where

$$A^{-1}_{i,j} = \frac{\sigma^{j-1}\left(1+\sigma^{2i}\right)\left(1+\sigma^{2N-2j}\right)}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \quad 0<i\le j<N,$$

$$A^{-1}_{i,N} = \frac{\sigma^{N-i}\left(1+\sigma^{2i}\right)}{(2+\sigma)\left(1-\sigma^{2N}\right)}, \quad 0<i\le N,$$

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

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

$$\sigma = \sqrt{3}-2, \quad A^{-1}_{i,j} = A^{-1}_{N-1,N-j} \quad \textit{for the other } i \textit{ and } j.$$

2.2. Theorem 2

Let $F_n(x)$ be empirical function of the distribution constructed by simple sample $X_1, X_2, \cdots, X_n$ and $S_N(x)$ be cubic spline interpolating the values $F_n(x_k)$. in the mesh (1). *If we choose the boundary conditions for $S_N(x)$ in the form*

$$\alpha_N = \frac{1}{h}\left(\frac{1}{3}y_3 - \frac{3}{2}y_2 + 3y_1 - \frac{11}{6}y_0\right),$$

$$b_N = \frac{1}{h}\left(\frac{11}{6}y_N - 3y_{N-1} + \frac{3}{2}y_{N-2} - \frac{1}{3}y_{N-3}\right).$$

Then the derivative $S'_N(x)$ of the spline function is defined by the equality

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

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

$$\hat{E}_{i,j}(x) = \begin{cases} \hat{D}_{i,j}(x) & j\ne i-1 \\ \hat{D}_{i,j}(x)+1 & j=i-1 \end{cases}$$

$$\hat{D}_{i,0} = -\frac{3}{2}C_{i,1} - \frac{5}{2}C_{i,0}, \quad \hat{D}_{i,1} = \frac{3}{2}\left(C_{i,1}-C_{i,2}\right) + \frac{7}{2}C_{i,0},$$

$$\hat{D}_{i,2} = \frac{3}{2}\left(C_{i,2}-C_{i,3}\right) - C_{i,0}, \quad \hat{D}_{i,j} = \frac{3}{2}\left(C_{i,j}-C_{i,j+1}\right), \quad j=3,4,\cdots,N-4,$$

$$\hat{D}_{i,N-3} = \frac{3}{2}\left(C_{i,N-3}-C_{i,N-2}\right) + C_{i,N}, \quad \hat{D}_{i,N-2} = \frac{3}{2}\left(C_{i,N-2}-C_{i,N-1}\right) - \frac{7}{2}C_{i,N},$$

$$\hat{D}_{i,N-1} = \frac{3}{2}C_{i,N-1} + \frac{5}{2}C_{i,N},$$

and $C_{i,j}$ *are defined by formula* (2).

We introduce the following denotations:

$X_1, X_2, \cdots, X_n$ is the simple sample from the general population

$$F(t) = \int_{-\infty}^{t} f(x)\,dx;$$

$F_n^*(t) = F_n(F^{-1}(t))$ is empirical function of distribution of the sample $F(X_1), F(X_2), \cdots, F(X_n)$;

$Y_n(t) = \sqrt{n}\left[F_n^*(t) - t\right], t \in [0,1]$ is the empirical process;

$\{\omega_n(t), t \in [0,1]\}$ is the sequence of wiener processes;

$B_n(t) = \omega_n(t) - t\omega_n(1), t \in [0,1]$ is the brownian bridge.

We give the auxiliary lemmas.

2.3. Lemma 1 [6]

There exists a probability space (Ω, F, P).

On which it can be defined version $F_n^*(t)$ *and the sequence of Brownian bridges* $B_n(t)$ *such that for all* $x > 0$

$$P\left(\sup_{0\le t\le 1}\left|n\left(F_n^*(t) - t\right) - \sqrt{n}B_n(t)\right| > ax + b\log n + c\log 2\right) \le e^{-x},$$

where $a = 3.26$, $b = 4.86$, $c = 2.70$.

2.4. Lemma 2 [7]

Let ω *be modulus of continuity of the brownian bridge* $B_n(t)$,

$$p(u) = \begin{cases} \sqrt{u(1-u)}, & \text{if } 0 \le u \le 1/2, \\ 1/2, & \text{if } u > 1/2 \end{cases}$$

and $q(u) = \int_0^u \sqrt{\ln(1/v)}\,dp(v)$. *Then with probability* 1 ω *does not exceed the quantity* $16\left(p\sqrt{\ln v_\varepsilon} + q\sqrt{2}\right)$. *Here* v_ε *is the random variable which is not less than* 1 *almost everywhere and* $Mv_\varepsilon < 4\sqrt{2}$.

3. Main Results and Proofs

The following theorem characterizes the asymptotic behavior of the bias, the covariance and the dispersion of the spline estimation.

3.1. Theorem 3

Let $S'_N(x)$ *be the spline estimation.*

1) *If* $f \in C_k[0,1], k = 0,1,2$ *and* $S'_N(x)$ *are defined as in Theorem* 2, *then for* $n \to \infty$

$$MS'_N(x) = f(x) + o(h^k).$$

2) *If* $f \in C[0,1]$ *and* $S'_N(x)$ *are defined as in Theorem* 1, *then*

$$\sup_{0\le x\le 1}\left|MS'_N(x) - f(x)\right| \to 0,\ n \to \infty,$$

$$DS'_N(x) = \frac{f(x)}{nh}A(r) + O(h/n),\ n \to \infty,$$

where $0 < x < 1$,

$$A(r)=-\frac{3(1-\sigma)}{2+\sigma}\left(2r^2-2r+\frac{1}{3}\right)+\frac{9}{4}\left(\frac{1-\sigma}{2+\sigma}\right)^2$$

$$\times\left\{\left(2r^2-2r+\frac{1}{3}\right)^2+\left[\left(r^2-\frac{1}{3}\right)+\sigma\left(\frac{1}{3}-(1-r)^2\right)\right]^2\frac{1}{1-\sigma^2}\right.$$

$$\left.+\left[\left(r^2-\frac{1}{3}\right)+\frac{1}{\sigma}\left(\frac{1}{3}-(1-r)^2\right)\right]^2\frac{\sigma^2}{1-\sigma^2}\right\},$$

$$\sigma=\sqrt{3}-2,\ \ r=\frac{x-x_{i-1}}{h},\ \ x_{i-1}=\frac{[\![N_x]\!]}{N},$$

$[y]$ *is the integer part of the number* y.

3) *Suppose* $0<x,y<1$, $x_{i-1}=\frac{[\![N_x]\!]}{N}$, $x_{j-1}=\frac{[\![N_y]\!]}{N}$, $d=i-j$, $r=\frac{x-x_{i-1}}{h}$ *and* $r_2=\frac{y-x_{j-1}}{h}$, *then for* $n\to\infty$

$$\operatorname{cov}\left[S_N'(x),S_N'(y)\right]$$

$$=\frac{1}{nh}\cdot\frac{3}{4}f(x)\left\{\left[\left(r_1^2-\frac{1}{3}\right)\left(r_2^2-\frac{1}{3}\right)+\left(\frac{1}{3}-(1-r_1)^2\left(\frac{1}{3}-(1-r_2)\right)^2\right)\right]\left(6|d|\sigma^{|d|}-\frac{12\sigma^{|d+1|}}{1-\sigma^2}\right)\right.$$

$$+\left(r_1^2-\frac{1}{3}\right)\left(\frac{1}{3}-(1-r_2)^2\right)\left(6|d+1|\sigma^{|d+1|}-\frac{12\sigma^{|d+1|+1}}{1-\sigma^2}\right)$$

$$\left.+\left(r_2^2-\frac{1}{3}\right)\left(\frac{1}{3}-(1-r_1)^2\right)\left(6|d-1|\sigma^{|d-1|}-\frac{12\sigma^{|d-1|+1}}{1-\sigma^2}\right)\right\}$$

$$+\frac{f(y)}{nh}\cdot\frac{\sqrt{3}}{2}\left[\left(r_1^2-\frac{1}{3}\right)\left(\sigma^{|d+1|}-\sigma^{|d|}\right)+\left(\frac{1}{3}-(1-r_1)^2\right)\times\left(\sigma^{|d|}-\sigma^{|d-1|}\right)\right]$$

$$+\frac{f(x)}{nh}\cdot\frac{\sqrt{3}}{2}\left[\left(r_2^2-\frac{1}{3}\right)\left(\sigma^{|d-1|}-\sigma^{|d|}\right)+\left(\frac{1}{3}-(1-r_2)^2\right)\left(\sigma^{|d|}-\sigma^{|d+1|}\right)\right]+\frac{f(x)}{nh}\delta_{d,0}+0\left(\frac{1}{n}\right).$$

Proof. By virtue of $MS_N'(x)=\left(MS_N'(x)\right)'$, Theorems 9, 11, 12 from Stechkin and Subbotin [1] and Theorems 1 from Lii [5] follows the first statement of Theorem 3. The second and the third statement of Theorem 3 are proved in Lii [5].

3.2. Theorem 4

Suppose $\frac{\ln n}{nh}\to 0$ *as* $n\to\infty$. *Then in order with probability* 1

$$\sup_{0\le x\le 1}\left|S_N'(x)-g(x)\right|\to 0\ \text{ as } n\to\infty,$$

it is necessary and sufficient that the function $g(x)$ *is the density of the distribution* $F(x)$ *concentrated and continuous on the interval* $[0,1]$ *with respect to Lebesgue measure*.

Proof. *Sufficiency*. It is clear that

$$\sup_{0\le x\le 1}\left|S_N'(x)-f(x)\right|\le\varepsilon_N+\delta_N, \tag{3}$$

where

$$\varepsilon_N = \sup_{0\le x\le 1}\left|S'_N(x) - MS'_N(x)\right|, \quad \delta_N = \sup_{0\le x\le 1}\left|MS'_N(x) - f(x)\right|.$$

First we estimate the term ε_N in the right hand part of (3). We have

$$\varepsilon_N \le \frac{32}{\sqrt{nh}}\left[\sup_{0\le x\le 1}\left|Y_n(t) - B_n(t)\right| + \frac{1}{2}\max_{1\le i\le N}\left|B_n(F(x_i)) - B_n(F(x_{i-1}))\right|\right]. \quad (4)$$

From Lemma 1 it follows that with probability 1 for $n\to\infty$

$$\sup_{0\le x\le 1}\left|Y_n(t) - B_n(t)\right| = 0\left(\frac{\ln n}{\sqrt{n}}\right). \quad (5)$$

If we denote the modulus of continuity $B_n(t)$ by $\theta(h)$ then from Lemma 2

$$\left|B_n(F(x_i)) - B_n(F(x_{i-1}))\right| \le (1+B)\theta(h) \quad (6)$$

where

$$B = \sup_{0\le t\le 1} f(t)$$

$$\theta(h) \le 16\sqrt{h}\left[\sqrt{\ln v_n} + \sqrt{2}\left(\sqrt{\ln N} + \sqrt{2\pi/\ln 2}\right)\right],$$

with probability $1 v_n \ge 1$ and $Mv_n < 4\sqrt{2}$.

This, combining (3)-(6) and using Theorem 3 we get the sufficiency condition of Theorem 4.

Necessity. Let with probability 1

$$\sup_{0\le x\le 1}\left|S'_N(x) - g(x)\right| \to 0 \text{ as } n\to\infty.$$

Hence, from continuity of $S'_N(x)$ it follows continuity of $g(x)$ on the interval [0, 1].

Therefore, the sequence random variables

$$\tau_n = \sup_{0\le x\le 1}\left|S'_N(x) - g(x)\right|, n = 1, 2, \cdots$$

are uniformly integrable. Therefore according to Theorem 5 from Shiryaev [8] and the inequalities

$$\sup_{0\le x\le 1}\left|MS'_N(x) - g(x)\right| = \sup_{0\le x\le 1}\left|M\left(S'_N(x) - g(x)\right)\right|$$

$$\le \sup_{0\le x\le 1} M\left|S'_N(x) - g(x)\right| \le M\sup_{0\le x\le 1}\left|S'_N(x) - g(x)\right|$$

it follows that for $n\to\infty$

$$\sup_{0\le x\le 1}\left|MS'_N(x) - g(x)\right| \to 0. \quad (7)$$

By virtue of (7) it is easy to see that the sequence of functions

$$g_n(x) = \frac{1}{h}\int_0^1 W_N(x,y)\,\mathrm{d}F(y)$$

uniformly converges to some continuous function $g_0(x)$, *i.e.* for $n\to\infty$

$$\sup_{0\le x\le 1}\left|g_n(x) - g_0(x)\right| \to 0. \quad (8)$$

We show now continuity of $F(x)$ on the interval [0, 1].

We assume the inverse that there exists a point x_0, $x_0\in[0,1]$ such that $P(X_1 = x_0) = p_0 > 0$. Then by virtue of (8) and

$$\frac{p_0}{h}\sup_{0\le x\le 1}\left|W_n(x,x_0)\right| \le \sup_{0\le x\le 1}\left|g_n(x)\right| \le \frac{1}{h}\sup_{0\le x\le 1}\left|W_n(x,y)\right|$$

it follows continuity of $F(x)$ on the interval [0, 1].

By (8) for all $0 \le x, y \le 1$

$$\lim_{n\to\infty} \int_y^x MS_N'(t)\,\mathrm{d}t = \int_y^x g(t)\,\mathrm{d}t \tag{9}$$

$$\int_y^x MS_N'(t)\,\mathrm{d}t = \int_y^x \mathrm{d}\left(MS_N(t)\right) = MS_N(x) - MS_N(y). \tag{10}$$

From another side, according to Theorem 11 from Stechkin and Subbotin (1976)

$$\lim_{n\to\infty} MS_N(x) = F(x). \tag{11}$$

By virtue of (9)-(11)

$$F(x) - F(y) = \int_y^x g(t)\,\mathrm{d}t.$$

Theorem 4 is proved.

References

[1] Stechkin, S.B. and Subbotin, Y.N. (1976) Splines in Computational Mathematics. Moscow, Nauka, 272 p.

[2] Nadaraya, E.A. (1983) Nonparametric Estimation of Probability Density and Regression Curve. Tbilisi University, Tbilisi, 195 p.

[3] Wegman, E.J. and Wright, I.W. (1983) Splinesin Statistics. *Journal of the American Statistical Association*, **78**, 351-365. http://dx.doi.org/10.1080/01621459.1983.10477977

[4] Muminov, M.S. (2010) On Appoximation of the Probability of the Lagre Outlier of Nonstationary Gauss Process. *Siberian Mathematical Journal*, **51**, 175-195. http://dx.doi.org/10.1007/s11202-010-0015-6

[5] Lii, K.S. (1978) A Global Measure of a Spline Density Estimate. *Annals of Statistics*, **6**, 1138-1148. http://dx.doi.org/10.1214/aos/1176344316

[6] Rio, E. (1994) Local Invariance Principles and Application to Density Estimation. *Probability Theory and Related Fields*, **98**, 21-26. http://dx.doi.org/10.1007/BF01311347

[7] Garsia, F. (1970) Continuity Properties of Gaussian Processer with Multidimensional Time Parameter. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 369-374.

[8] Shiryaev, A.N. (1982) Probability. Moscow, Nauka, 576 p.