Mathematica Pannonica
10/1 (1999), 17–28

ON THE CHOICE OF THE SPLINE DENSITY ESTIMATORS

Zbigniew **Ciesielski**

Instytut Matematyczny PAN, ul. Abrahama 18, 81–825 Sopot, Poland

Anna **Kamont**

Instytut Matematyczny PAN, ul. Abrahama 18, 81–825 Sopot, Poland

Dedicated to Professor Ferenc Schipp on his 60th birthday

Received: November 1998

MSC 1991: 65 D 07, 62 G 07

Keywords: Density estimator, B-splines.

Abstract: Given a simple sample on $\mathbb{R}$, an estimator for an apriori density is defined as a suitable linear combination of rescaled cardinal B-splines. The scaling is such that the knots of the B-splines are equally spaced at the distance h. It is shown that there is exactly one h such that the mean and the variance of the estimator are equal to the empirical mean and to the unbiased empirical variance, respectively. Extensions of this result to apriori densities with supports in bounded or unbounded intervals are treated as well.

1. Introduction

We start with the following setup: an apriori density f on $\mathbb{R}$ is assumed to be given. We also assume that we are given a sequence $X_1, X_2, \ldots$ of independent identically distributed, according to f, random variables. For given N and a window parameter $h > 0$, the $X_1, \ldots, X_N$ can be used to construct a density $f_{h,N}$, called an estimator, which should approximate f as $h \to 0$ and $N \to \infty$. In particular, a class of

often treated estimators are the kernel estimators, see e.g. [4]. Investigations of this type belong to the asymptotic part of the estimation theory, and this will not be discussed here.

In this note we consider a vector $\mathbf{X} = (X_1, \dots, X_N) \in \mathbb{R}^N$ as a realization of the first N random variables mentioned above. It is called a *simple sample of size* N. Using $\mathbf{X}$ and B-splines with equally spaced knots at the distance h, we construct an estimator $f_{h,N}$. Asymptotic behaviour of such estimators was discussed in [3] and [5]. Here, we are looking for h_0 such that the difference between the empirical mean and the mean with respect to $f_{h,N}$, and also the difference between the unbiased empirical variance and the variance with respect to $f_{h,N}$, are minimal. This approach is discussed in three different cases: $X_j \in (a, b)$ for $j = 1, \dots, N$ with (i) $(a, b) = \mathbb{R}$, (ii) (a, b) – a bounded interval and (iii) (a, b) – a half-line. In the last section, we outline the main steps of an algorithm for calculating the optimal density estimator.

Note that the argument in Section 3 is in spirit probabilistic, while the one in Section 4 is purely analytic. The first one is presented here because it is simple and beautiful, and the second one because there is no simple probabilistic proof of Ths. 4.1 and 4.2.

A different approach to the choice of the optimal density estimator, based on approximation ideas, is presented in [3].

2. Preliminaries

The estimators investigated in this paper are spline functions of given order $r \geq 1$. The knots of the corresponding splines are equidistant and simple in case of the real line, i.e. when there are no restrictions for the support of the apriori density. In case the support of the apriori density is known to be contained in an interval, the knots are again equidistant, and the finite endpoints of that interval are assumed to be knots of multiplicity r.

It is well-known that splines are linear combinations of B-splines. Therefore, we recall the properties of the B-splines needed in the proofs below. For more details, we refer to [1], [6], [7].

For given $r \geq 1$, let $\Pi = \{\tau_i, i \in \mathbb{Z}\}$ be a sequence of reals such that $\tau_i \leq \tau_{i+1}$, $\tau_i < \tau_{i+r}$, $\lim_{i\to-\infty} \tau_i = -\infty$ and $\lim_{i\to+\infty} \tau_i = +\infty$. Such Π is a sequence of knots admitting multiplicities up to r. Now, B-splines $\{N_i^{(r)}, i \in \mathbb{Z}\}$ of order r with knots Π are defined by the formula

(2.1) $$N_i^{(r)}(x) = (\tau_{i+r} - \tau_i)[\tau_i, \dots, \tau_{i+r}; (\cdot - x)_+^{r-1}],$$

where $[\tau_i, \dots, \tau_{i+r}; f]$ is the divided difference of order r of the function f. In addition, define $M_i^{(r)} = \frac{r}{\tau_{i+r} - \tau_i} N_i^{(r)}$. It follows from (2.3) below that

(2.2) $$M_i^{(r)}(x) \geq 0 \quad \text{and} \quad \int_{\mathbb{R}} M_i^{(r)}(x)\, dx = 1.$$

The following properties of B-splines are needed in the sequel:

(2.3) $N_i^{(r)}(x) \geq 0$, supp $N_i^{(r)} = [\tau_i, \tau_{i+r}]$ and $\int_{\mathbb{R}} N_i^{(r)}(x) dx = \frac{\tau_{i+r} - \tau_i}{r}$.

(2.4) $$\sum_{i \in \mathbb{Z}} N_i^{(r)}(x) = 1 \quad \text{for each } x \in \mathbb{R}.$$

Let $m \in \mathbb{N}$, $m < r$. Then

(2.5) $$x^m = \sum_{i \in \mathbb{Z}} \left(\frac{1}{\binom{r-1}{m}} \sum_{0 < j_1 < \dots < j_m < r} \tau_{i+j_1} \cdots \tau_{i+j_m} \right) N_i^{(r)}(x).$$

Let $m \in \mathbb{N}$. Then

(2.6) $$\int_{\mathbb{R}} x^m M_i^{(r)}(x) dx = \frac{1}{\binom{r+m}{m}} \sum_{0 \leq j_1 \leq \dots \leq j_m \leq r} \tau_{i+j_1} \cdots \tau_{i+j_m},$$

In particular, we get from (2.5), (2.6) for $r \geq 3$:

$$x = \sum_{i \in \mathbb{Z}} \alpha_i N_i^{(r)}(x), \qquad x^2 = \sum_{i \in \mathbb{Z}} \beta_i N_i^{(r)}(x),$$

with

(2.7) $$\alpha_i = \frac{\tau_{i+1} + \dots + \tau_{i+r-1}}{r-1},$$

(2.8) $$\beta_i = \frac{(\tau_{i+1} + \dots + \tau_{i+r-1})^2 - (\tau_{i+1}^2 + \dots + \tau_{i+r-1}^2)}{(r-1)(r-2)},$$

(2.9) $$w_i := \int_{\mathbb{R}} x M_i^{(r)}(x)\, dx = \frac{\tau_i + \dots + \tau_{i+r}}{r+1},$$

(2.10) $$z_i := \int_{\mathbb{R}} x^2 M_i^{(r)}(x)\, dx = \frac{(\tau_i + \dots + \tau_{i+r})^2 + (\tau_i^2 + \dots + \tau_{i+r}^2)}{(r+1)(r+2)}.$$

Now, we let us specify what we mean by a simple sample $\mathbf{X}$: a *simple sample* $\mathbf{X} = (X_1, \dots . X_N)$ of size N is just a vector in $\mathbb{R}^N$. In what follows it is assumed that $\mathbf{X}$ is given and fixed, and it is not trivial, i.e. we exclude the case of $X_1 = \dots = X_N$. The *empirical distribution* corresponding to $\mathbf{X}$ is denoted by $F_{\mathbf{X}}$ and it is defined by the formula

(2.11) $$F_{\mathbf{X}}(x) = \frac{\#\{i : X_i < x,\ 1 \leq i \leq N\}}{N},$$

where $\#A$ is the cardinality of the set A. We also recall that the *empirical mean* m_N and the *unbiased empirical variance* s_N^2 are given by

$$(2.12)\quad m_N = \frac{X_1 + \cdots + X_N}{N},\quad s_N^2 = \frac{(X_1 - m_N)^2 + \cdots + (X_N - m_N)^2}{N-1}.$$

The spline density estimators based on the sample $\mathbf{X}$ are defined in the following sections, separately in case of the whole line and in case of intervals.

3. Estimators on $\mathbb{R}$

On the real line, we use $B^{(r)}$, the symmetric cardinal B-spline of order r with simple knots $\{i + r/2, i \in \mathbb{Z}\}$. The simplest way of introducing $B^{(r)}$ is probabilistic. Indeed, $B^{(r)}$ is the density of the sum of r independent uniformly distributed on $[-1/2, 1/2]$ random variables $(U_1, \ldots, U_r)$ i.e.

$$(3.1)\qquad P\{U_1 + \cdots + U_r < t\} = \int_{-\infty}^{t} B^{(r)}(s)\, ds.$$

It should be noted that $B^{(r)}$ is identical with $N_0^{(r)}$ corresponding to knots $\tau_i = -r/2 + i, i \in \mathbb{Z}$. Therefore, $B^{(r)}$ is defined on $\mathbb{R}$, it is nonnegative, supp $B^{(r)} = [-r/2, r/2]$, it is a polynomial of degree $r-1$ on each interval $[j-r/2, j+1-r/2]$ with $j \in \mathbb{Z}$, $j = 0, \ldots, r-1$. Moreover, for $r \geq 2$, $B^{(r)}$ is $r-2$ times continuously differentiable. In case $r = 1$ it is understood to be left-continuous step function. For the properties of the cardinal B-splines we refer to [6].

Now, for given value $h > 0$ of the window parameter, the *density estimator* is defined as follows

$$(3.2)\qquad f_{h,N}(x) = f_{h,N}(x, \mathbf{X}) = \sum_{i \in \mathbb{Z}} \frac{1}{h} \int_{\mathbb{R}} B_{h,i}^{(k)} dF_{\mathbf{X}} \cdot B_{h,i}^{(r)}(x),$$

where the k and r are given orders of the B-splines and

$$(3.3)\qquad B_{h,i}^{(k)}(x) = B^{(k)}\left(\frac{x}{h} - i\right).$$

It is easy to see that $f_{h,N}$ is a density. Asymptotic statistical behavior of the estimator $f_{h,N}(x, \mathbf{X})$ as $N \to \infty$ for a given apriori density was treated in [3] and [5] and it will be not touched here. Instead, our goal is to find h such that

$$(3.4)\qquad \int_{\mathbb{R}} x f_{h,N}(x)\, dx = m_N,$$

$$(3.5)\qquad \int_{\mathbb{R}} (x - m_N)^2 f_{h,N}(x)\, dx = s_N^2,$$

where m_N is the empirical mean and s_N^2 is the unbiased empirical variance as given in (2.12).

The main result of this section is the following

Theorem 3.1. *Assume that we are given a simple sample* $\mathbf{X} \in \mathbb{R}^N$ *and positive integers* r *and* k *with* $k \geq 3$. *Then there is exactly one* $h = h_0$ *satisfying equations* (3.4) *and* (3.5), *and it is given by formula*

$$h_0 = \sqrt{\frac{12}{k+r}}\frac{s_N}{\sqrt{N}}. \tag{3.6}$$

The following lemma, collecting well-known properties of cardinal B-splines, is basic for the proof of Th. 3.1. Although these properties can be derived e.g. from (2.3) – (2.6), we have decided to present here a simple probabilistic proof.

Lemma 3.2. *For each positive integer* k *and for* $y \in \mathbb{R}$ *we have the following identities*

$$\int_{\mathbb{R}} B^{(k)}(x)\,dx = 1, \tag{3.7}$$

$$\int_{\mathbb{R}} xB^{(k)}(x)\,dx = 0, \tag{3.8}$$

$$\int_{\mathbb{R}} x^2 B^{(k)}(x)\,dx = \frac{k}{12}, \tag{3.9}$$

$$\sum_{i\in\mathbb{Z}} B^{(k)}(y-i) = 1, \tag{3.10}$$

$$\sum_{i\in\mathbb{Z}} iB^{(k)}(y-i) = y \quad \text{for} \quad k \geq 2, \tag{3.11}$$

$$\sum_{i\in\mathbb{Z}} i^2 B^{(k)}(y-i) = y^2 + \frac{k}{12} \quad \text{for} \quad k \geq 3. \tag{3.12}$$

Proof. Property (3.7) follows from (3.1), (3.8) follows by symmetry of the B-spline $B^{(k)}$ and we obtain (3.9) calculating the variance of the random variable $U_1 + \cdots + U_k$. Now, formula (3.10) is clearly satisfied in case $k = 1$. On the other hand (3.1) implies for $k \geq 2$

$$P\{U_1 + \cdots + U_k < t\} = \int_{-1/2}^{1/2} P\{U_1 + \cdots + U_{k-1} < t - s\}\,ds$$

whence by differentiation we get

$$B^{(k)}(t) = \int_{-1/2}^{1/2} B^{(k-1)}(t-s)\,ds. \tag{3.13}$$

Thus, by induction we obtain (3.10). Formula (3.13) implies also that

$$\frac{d}{dt}B^{(k)}(t) = B^{(k-1)}(t+1/2) - B^{(k-1)}(t-1/2). \tag{3.14}$$

For $k = 2$ the B-splines are continuous and piece-wise linear with knots at the points $i + 1/2$, and consequently formula (3.11) holds. For $k > 2$

formula (3.11) follows again from (3.13) by induction. It remains to check (3.12). Differentiation of the left hand side of (3.12), and then application of (3.14), (3.10) and (3.11) give

$$\frac{d}{dy}\sum_{i\in\mathbb{Z}} i^2 B^{(k)}(y-i) = 2y, \tag{3.15}$$

and consequently for some constant c_k

$$\sum_{i\in\mathbb{Z}} i^2 B^{(k)}(y-i) = y^2 + c_k. \tag{3.16}$$

In particular we obtain

$$\sum_{i\in\mathbb{Z}} i^2 B^{(k)}(i) = c_k. \tag{3.17}$$

Integration of (3.16) over the interval [-1/2,1/2] and formulas (3.13) and (3.17) give $c_{k+1} = \frac{1}{12} + c_k$. It follows however by (3.13) that $B^{(3)}(-1) = = 1/8 = B^{(3)}(1)$ and $B^{(3)}(0) = 3/4$. Thus, according to (3.17) $c_3 = 1/4$ and by (3.16) the formula (3.12) follows, and this completes the proof. $\Diamond$

Proof of Theorem 3.1. We observe first that by definition (3.2), properties (3.7), (3.8) and (3.11) we obtain (3.4) for each $h > 0$. This means that equation (3.4) puts no restrictions on h and that h_0 should be determined by equation (3.5). Taking into account (3.4) we get

$$\int_{\mathbb{R}} (x - m_N)^2 f_{h,N}(x)\,dx = \int_{\mathbb{R}} x^2 f_{h,N}(x)\,dx - m_N^2. \tag{3.18}$$

Now, like in the mean value case we have

$$\int_{\mathbb{R}} x^2 f_{h,N}(x)\,dx = \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} B_{h,i}^{(k)} dF_{\mathbf{X}} \cdot \int_{\mathbb{R}} x^2 B_{h,i}^{(r)}(x)\,dx. \tag{3.19}$$

Using properties (3.7), (3.8) and (3.9) we find that

$$\int_{\mathbb{R}} x^2 B_{h,i}^{(r)}(x)\,dx = h^3\left(\frac{r}{12} + i^2\right). \tag{3.20}$$

Substituting (3.20) into (3.19) and using properties (3.7) and (3.12) we get

$$\int_{\mathbb{R}} x^2 f_{h,N}(x)\,dx = h^2\left(\frac{r}{12} + \sum_{i\in\mathbb{Z}} \int_{\mathbb{R}} i^2 B_{h,i}^{(k)} dF_{\mathbf{X}}\right) = \tag{3.21}$$

$$= h^2\left(\frac{r}{12} + \int_{\mathbb{R}} \left(\sum_{i\in\mathbb{Z}} i^2 B_{h,i}^{(k)}\right) dF_{\mathbf{X}}\right) = h^2\left(\frac{r}{12} + \frac{k}{12}\right) + \int_{\mathbb{R}} x^2\,dF_{\mathbf{X}}(x).$$

This and (3.18) give

$$\int_{\mathbb{R}} (x - m_N)^2 f_{h,N}(x)\, dx = h^2 \frac{r+k}{12} + \frac{N-1}{N} s_N^2.$$

Comparing this with (3.5) find the final equation for h, i.e.

$$h^2 = \frac{12}{r+k} \frac{s_N^2}{N},$$

and this completes the proof. $\Diamond$

4. Estimators on intervals

Consider now the bounded closed interval $I = [a, b]$ and space of splines of order r with equally spaced knots $\Pi_n = \{t_{n,i} : -r < i < n+r\}$, with the endpoints as knots of multiplicity r each, i.e.

$$t_{n,i} = \begin{cases} a & \text{for i=-r+1, \dots , 0} \\ a + \frac{i}{n}(b-a) & \text{for i= 1, \dots , n-1} \\ b & \text{for i = n, \dots , n+r-1.} \end{cases} \tag{4.1}$$

Let $N_{n,i}^{(r)}$, $i = -r+1, \dots, n-1$, be the B-spline of order r with the knots $t_{n,i}, \dots, t_{n,i+r}$. Note that these are all the B-splines with support contained in $[a, b]$.

For a simple sample $\mathbf{X} = (X_1, \dots, X_N)$ with $X_j \in (a, b)$ and $n \geq r$, we consider the following spline density estimator

$$f_{n,N}(x) = f_{n,N}(x, \mathbf{X}) = \sum_{i=-r+1}^{n-1} \int_{[a,b]} N_{n,i}^{(r)} dF_{\mathbf{X}} \cdot M_{n,i}^{(r)}(x). \tag{4.2}$$

The parameter n corresponds to the window parameter $h = \frac{b-a}{n}$.

Let μ_n, σ_n^2 be the mean value and the variance of the density $f_{n,N}$, i.e.

$$\mu_n = \int_a^b x f_{n,N}(x) dx, \quad \sigma_n^2 = \int_a^b (x - \mu_n)^2 f_{n,N}(x) dx. \tag{4.3}$$

Contrary to the case of the whole line, now μ_n need not be equal to the empirical mean m_N, as given in (2.12). Therefore, looking for the optimal n, we consider both $\mu_n - m_N$ and $\sigma_n^2 - s_N^2$. For later convenience, introduce

$$P_n(\mathbf{X}) = \mu_n - m_N, \quad R_n(\mathbf{X}) = \sigma_n^2 - s_N^2. \tag{4.4}$$

For a given continuous and strictly increasing functions $u, v : [0, \infty) \to$ $\to [0, \infty)$, such that $u(0) = v(0) = 0$ and $\lim_{t\to\infty} u(t) = \lim_{t\to\infty} v(t) =$ $= \infty$, we define

$$G_n(\mathbf{X}) = u(|P_n(\mathbf{X})|) + v(|R_n(\mathbf{X})|). \tag{4.5}$$

The value n_0 is the optimal n if $G_{n_0}(\mathbf{X}) = \inf_{n \geq r} G_n(\mathbf{X})$. The functions $u(x), v(x)$ can be taken e.g. $u(x) = x^{p_1}, v(x) = x^{p_2}$ with $p_1, p_2 > 0$.

The main result of this section is the existence of such n_0:

Theorem 4.1. *Let $r \geq 3$. Let $\mathbf{X} = (X_1, \dots, X_N)$ be a simple sample such that $X_j \in (a,b)$ for each j, $1 \leq j \leq N$, and let $G_n(\mathbf{X})$ be given as in* (4.5). *Then there is $n_0 \in \mathbb{N}$ such that*

$$G_{n_0}(\mathbf{X}) = \inf_{n \geq r} G_n(\mathbf{X}).$$

Before we proceed with the proof, we have to calculate the coefficients $\alpha_{n,i}$, $\beta_{n,i}$, $w_{n,i}$, $z_{n,i}$ appearing in the formulas (with $r \geq 3$)

$$x = \sum_{i=-r+1}^{n-1} \alpha_{n,i} N_{n,i}^{(r)}(x), \quad x^2 = \sum_{i=-r+1}^{n-1} \beta_{n,i} N_{n,i}^{(r)}(x), \tag{4.6}$$

$$w_{n,i} = \int_a^b x M_{n,i}^{(r)}(x)\,dx, \quad z_{n,i} = \int_a^b x^2 M_{n,i}^{(r)}(x)\,dx. \tag{4.7}$$

For simplicity, we use below the symbol $(n)_k = n(n-1)\cdot\ldots\cdot(n-k+1)$. Using (2.7) – (2.10) and the particular form of the knots (4.1), we find by elementary calculation the explicit formulas for $\alpha_{n,i}, \beta_{n,i}, w_{n,i}, z_{n,i}$:

For $-r < i < 0$:

$$\begin{aligned}
\alpha_{n,i} &= a + \frac{(b-a)(i+r)_2}{2n(r-1)}, \\
\beta_{n,i} &= a^2 + \frac{a(b-a)(i+r)_2}{n(r-1)} + \frac{(b-a)^2(i+r)_3(3i+3r-1)}{12n^2(r-1)_2}, \\
w_{n,i} &= a + \frac{(b-a)(i+r+1)_2}{2n(r+1)}, \\
z_{n,i} &= a^2 + \frac{a(b-a)(i+r+1)_2}{n(r+1)} + \\
&\qquad + \frac{(b-a)^2(i+r+2)_3(3i+3r+1)}{12n^2(r+2)_2}.
\end{aligned} \tag{4.8}$$

For $0 \leq i \leq n-r$:

$$\begin{aligned}
\alpha_{n,i} &= w_{n,i} = a + \frac{(b-a)i}{n} + \frac{(b-a)r}{2n}, \\
\beta_{n,i} &= a^2 + \frac{a(b-a)(2i+r)}{n} + \frac{(b-a)^2(3(2i+r)^2 - r)}{12n^2}, \\
z_{n,i} &= a^2 + \frac{a(b-a)(2i+r)}{n} + \frac{(b-a)^2(3(2i+r)^2 + r)}{12n^2}.
\end{aligned} \tag{4.9}$$

For $n-r<i<n$:

$$\alpha_{n,i}=b-\frac{(b-a)(n-i)_2}{2n(r-1)},$$

$$\beta_{n,i}=b^2-\frac{b(b-a)(n-i)_2}{n(r-1)}+\frac{(b-a)^2(n-i)_3(3n-3i-1)}{12n^2(r-1)_2},$$

$$w_{n,i}=b-\frac{(b-a)(n-i+1)_2}{2n(r+1)}, \tag{4.10}$$

$$z_{n,i}=b^2-\frac{b(b-a)(n-i+1)_2}{n(r+1)}+$$

$$+\frac{(b-a)^2(n-i+2)_3(3n-3i+1)}{12n^2(r+2)_2}.$$

Now, we can write (cf. (4.4))

$$P_n(\mathbf{X})=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=-r+1}^{n-1}(w_{n,i}-\alpha_{n,i})N_{n,i}^{(r)}(X_j), \tag{4.11}$$

$$R_n(\mathbf{X})= \tag{4.12}$$

$$=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=-r+1}^{n-1}(z_{n,i}-\beta_{n,i})N_{n,i}^{(r)}(X_j)-P_n(\mathbf{X})(2m_N+P_n(\mathbf{X}))-\frac{s_N^2}{N}.$$

Thus, for given functions u, v, formulas (4.8) – (4.10) allow us to calculate explicitly $G_n(\mathbf{X})=u(|P_n(\mathbf{X})|)+v(|R_n(\mathbf{X})|)$.

Proof of Theorem 4.1. For given $\mathbf{X}=(X_1,\dots X_N)$ with $X_j\in(a,b)$, let

$$\delta=\min\{X_j-a,b-X_j:1\le j\le N\},\quad n_\delta=\min\{n\in N:\frac{r(b-a)}{n}<\delta\}.$$

Now, for each $n\ge n_\delta$ and $-r<i<0$, or $n-r<i<n$, and $1\le j\le N$ we have $N_{n,i}^{(r)}(X_j)=0$. Using this, (4.11) and formulas (4.9) for $\alpha_{n,i}, w_{n,i}$ we get for $n\ge n_\delta$

$$P_n(\mathbf{X})=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=0}^{n-r}(w_{n,i}-\alpha_{n,i})N_{n,i}^{(r)}(X_j)=0.$$

Similarly, we have by (4.9) and by the partition of unity property (2.4)

$$\frac{1}{N}\sum_{j=1}^{N}\sum_{i=-r+1}^{n-1}(z_{n,i}-\beta_{n,i})N_{n,i}^{(r)}(X_j)=\frac{r(b-a)^2}{6n^2},$$

hence by (4.12)

$$R_n(\mathbf{X}) = \frac{r(b-a)^2}{6n^2} - \frac{s_N^2}{N}.$$

Let $h > 0$ be such that $\frac{r(b-a)^2}{6}h^2 = \frac{s_N^2}{N}$, and let $l \in N$ be such that $\frac{1}{l} \leq h < \frac{1}{l-1}$. Observe that for $n \geq l$ we have $|R_n(\mathbf{X})| = \frac{s_N^2}{N} - \frac{r(b-a)^2}{6n^2}$, which is increasing in n. Thus, for $n \geq \max(n_\delta, l)$ we have $G_n(\mathbf{X}) = v(|R_n(\mathbf{X})|)$, and $G_n(\mathbf{X})$ is incresing in n. Therefore

$$\inf_{n \geq r} G_n(\mathbf{X}) = \min_{r \leq n \leq \max(n_\delta, l)} G_n(\mathbf{X}),$$

and now it is enough to find n_0, $r \leq n_0 \leq \max(n_\delta, l)$ such that

$$G_{n_0}(\mathbf{X}) = \min_{r \leq n \leq \max(n_\delta, l)} G_n(\mathbf{X}). \quad \Diamond \tag{4.13}$$

4.1. Estimators on half-lines

Let us discuss briefly the case of $[a, \infty)$. Now, it is assumed that the sample $\mathbf{X} = (X_1, \ldots, X_N)$ satisfies $X_j \in (a, \infty)$, $j = 1, \ldots, N$.

In this case, for given $h > 0$ and $r \geq 3$, we consider the following knots on $[a, \infty)$:

$$t_{h,i} = \begin{cases} a & \text{for i=-r+1, } \ldots \text{, 0} \\ a + ih & \text{for i= 1, 2, } \ldots \text{,} \end{cases}$$

i.e. a is the knot of multiplicity r. Now, we take the respective B-splines $N_{h,i}^{(r)}$ with $i = -r+1, \ldots$, and the corresponding estimator

$$f_{h,N}(x) = f_{h,N}(x, \mathbf{X}) = \sum_{i=-r+1}^{\infty} \int_{[a,\infty]} N_{h,i}^{(r)} dF_{\mathbf{X}} \cdot M_{h,i}^{(r)}(x). \tag{4.14}$$

Like in the case of the finite interval, we are looking for h_0 such that

$$G_{h_0}(\mathbf{X}) = \inf_{h>0} G_h(\mathbf{X}),$$

where $G_h(\mathbf{X})$ is defined by means of (4.3) – (4.5), with $P_n, R_n, \mu_n, \sigma_n^2$, $f_{n,N}, [a, b]$ replaced by $P_h, R_h, \mu_h, \sigma_h^2, f_{h,N}, [a, \infty)$, respectively. We have also in this case the following theorem:

Theorem 4.2. *Let $r \geq 3$. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be a simple sample such that $X_j \in (a, \infty)$ for each j, $1 \leq j \leq N$. Then there is $h_0 > 0$ such that*

$$G_{h_0}(\mathbf{X}) = \inf_{h>0} G_h(\mathbf{X}).$$

The **proof** follows the same lines as the proof of Th. 4.1. It should only be noted that formulas (4.6), (4.7) take now the form

$$x = \sum_{i=-r+1}^{\infty} \alpha_{h,i} N_{h,i}^{(r)}(x), \qquad x^2 = \sum_{i=-r+1}^{\infty} \beta_{h,i} N_{h,i}^{(r)}(x),$$

$$w_{h,i} = \int_a^\infty x M_{h,i}^{(r)}(x)\,dx,\ z_{h,i} = \int_a^\infty x^2 M_{h,i}^{(r)}(x)\,dx.$$

This time, the coefficients $\alpha_{h,i}, \beta_{h,i}, w_{h,i}, z_{h,i}$ for $-r < i < 0$ are calculated by formulas analogous to (4.8), and for $i \geq 0$ by formulas analogous to (4.9), with $\frac{b-a}{n}$ replaced by h. Using these formulas, one can check that $\lim_{h\to\infty} |\mu_h - m_N| = \infty$. For small h, the argument used in the proof of Th. 4.1 can be repeated. The remaining natural changes in the proof are omitted. $\Diamond$

Finally, let us note that the case of the interval $(-\infty, b]$ can be handled by treating the sample $\mathbf{Y} = -\mathbf{X}$ in the interval $[a, \infty)$ with $a = -b$.

5. Algorithm for the optimal estimator

There are two parts of the algorithm:

(i) Calculating linear combinations of B-splines at a given point. For this, we refer to [2] or [7]. This allows us, for given sample and window parameter h (or n), to calculate the coefficients in (3.2), (4.2), (4.14) and the estimators themselves.

(ii) Calculating the window parameter:

(a) In case of the whole real line, the optimal value of the window parameter is given directly by formula (3.6).

(b) In case of bounded interval, given the functions u, v, the optimal n can be found by (4.13). Moreover, it follows from the proof of Th. 4.1, in the notation of that proof, that to find

$$\min_{r\leq n\leq \max(n_\delta, l)} G_n(\mathbf{X})$$

in case $n_\delta < l$, it is enough to calculate

$$G_l(\mathbf{X}),\quad G_{l-1}(\mathbf{X}) \quad \text{and} \quad \min_{r\leq n\leq n_\delta} G_n(\mathbf{X}).$$

Indeed, we have for $n_\delta \leq n < l$

$$|R_n(\mathbf{X})| = \frac{r(b-a)^2}{6n^2} - \frac{s_N^2}{N},$$

and in this range of n's $|R_n(\mathbf{X})|$ is decreasing in n. Thus, $G_n(\mathbf{X}) = v(|R_n(\mathbf{X})|)$ is also decreasing on $n_\delta \leq n < l$, which implies that in this case

$$\min_{r\leq n\leq \max(n_\delta, l)} G_n(\mathbf{X}) = \min_{r\leq n\leq l} G_n(\mathbf{X}) = \min\{G_l(\mathbf{X}), G_{l-1}(\mathbf{X}), \min_{r\leq n\leq n_\delta} G_n(\mathbf{X})\}.$$

To calculate $G_n(\mathbf{X})$, we combine formulas (4.4) - (4.5), (4.11) - (4.12), (4.8) - (4.10), and then apply the algorithm mentioned in (i).

(c) In case of half-line, given u, v and $h > 0$, the value of $G_h(\mathbf{X})$ can be calculated using the algorithm mentioned in (i). Then, one should apply any algorithm for finding minimum of a function.

References

[1] DE BOOR, C.: Splines as linear combinations of B-splines, in Approximation Theory II. C. K. Chui, J. D. Ward, L.L. Schumaker (Eds.), Academic Press, New York 1976, 1–47.

[2] DE BOOR, C.: A Practical Guide to Splines, Springer-Verlag, New York, 1978.

[3] CIESIELSKI, Z.: Asymptotic nonparametric spline density estimation, *Prob. and Math. Stat.* **12/1** (1991), 1–24.

[4] DEVROYE, L. and GYÖRFI, L.: Nonparametric density estimation, The L_1 View. J. Wiley, New York, 1985.

[5] KRZYKOWSKI, G.: Equivalent conditions for the consistency of nonparametric spline density estimation, *Prob. and Math. Stat.* **13**/2 (1992), pp. 269–276.

[6] SCHOENBERG, I.J.: Cardinal spline interpolation, Regional Conference Series in Appl. Math. **12** SIAM Publ., Philadelphia, 1973.

[7] SCHUMAKER, L.L.: Spline functions: Basic theory, J. Wiley, New York, 1981.