The Annals of Statistics
1978, Vol. 6, No. 5, 1138–1148

A GLOBAL MEASURE OF A SPLINE DENSITY ESTIMATE

BY KEH-SHIN LII

Northwestern University

A spline density function estimate is considered. Limit theorems are obtained for a quadratic norm of the normalized deviation of the estimate from its expected value. Results obtained can be used in a test of goodness of fit. Some of the commonly used boundary conditions for the spline function are considered. It appears that in some situations certain boundary conditions are undesirable.

1. Introduction. Let $f(x)$ be a continuous density function on $[0, 1]$. Suppose that $X_1, X_2, \cdots, X_n$ are independent, identically distributed random variables with density f. Consider

$$y_k = F_n(k/N) \qquad k = 0, 1, \cdots, N = 1/h$$

where $F_n(x)$ is the sample distribution function and $h = 1/N$ is the bin size. Let $S_N(x)$ be the cubic spline interpolator of F_n with knots at the points $x_j = j/N$, $j = 0, 1, \cdots, N$. The derivative of $S_N(x)$ which is given by (see [1])

$$S_N'(x) = -M_{j-1}\frac{(x_j - x)^2}{2h} + M_j\frac{(x - x_{j-1})^2}{2h} + \frac{y_j - y_{j-1}}{h} - \frac{M_j - M_{j-1}}{6}h$$
$$\text{when} \quad x \in [x_{j-1}, x_j]\,, \quad j = 1, \cdots, N\,,$$
$$\text{with} \quad M_j = S_N''(x_j)\,, \quad j = 0, 1, \cdots, N\,,$$

will be used as the estimator of $f(x)$.

The continuity requirements at the knots imply that

$$\frac{h}{6}M_{j-1} + \frac{2h}{3}M_j + \frac{h}{6}M_{j+1} = \frac{1}{h}(y_{j+1} - 2y_j + y_{j-1}) \qquad j = 1, \cdots, N-1\,.$$

We need two more conditions to determine the $N+1$ unknown M_j's for $j = 0, \cdots, N$. We consider a few such conditions here:

(I) Specify the second derivative at the boundaries

$$M_0 = S_N''(0) = y_N''\,, \quad M_N = S_N''(1) = y_N''\,.$$

(II) A special case of (I):

$$M_0 = 0 = M_N\,.$$

(III) Specify the derivative at the boundaries

$$S_N'(0) = y_0' = f(0)\,, \qquad S_N'(1) = y_N' = f(1)\,.$$

Received February 1977; revised October 1977.

AMS 1970 *subject classifications.* Primary 62G05; Secondary 62E15.

Key words and phrases. Density function estimate, spline function, boundary condition, global measure of deviation, asymptotic distribution, goodness-of-fit test, Gaussian process.

We remark that the periodic spline is not realistic here. If we use boundary condition (III) and solve the system of linear equations we get

$$M_i = \textstyle\sum_{j=0}^{N} A_{i,j}^{-1} d_j$$

where $A_{i,j}^{-1}$ is asymptotically equal to

$\sigma^{|i-j|}/(2+\sigma)$ to the first order with $\sigma = 3^{\frac{1}{2}} - 2$ when M_i is a fixed distance away from the boundary,

and $d_j = 3(y_{j+1} - 2y_j + y_{j-1})/h^2$, $j = 1, \cdots, N-1$,

$$d_0 = \frac{6}{h}\left(\frac{y_1 - y_0}{h} - y_0'\right)$$

$$d_N = \frac{6}{h}\left(y_N' - \frac{y_N - y_{N-1}}{h}\right)$$

(see [1], [5]). The precise asymptotic behavior of the bias and covariance of $S_N'(x)$ in the interior of the interval was obtained in [5]. The estimator is shown to be asymptotically normally distributed there.

2. A quadratic measure of deviation. Our main object is to get a global measure of how good $s_N'(x)$ is as an estimate of $f(x)$. In particular, the asymptotic distribution of the functional

$$\int_0^1 \frac{[s_N'(x) - f(x)]^2}{f(x)}\, dx\,,$$

with proper normalization, is our main concern. The basic technique in obtaining the results is that of approximating the normalized and centered sample distribution function by an appropriate Brownian motion process on a convenient probability space by using a result due to Komlós, Major and Tusnády [4].

Rewrite $s_N'(x)$ for $x \in [x_{i-1}, x_i]$, as

$$s_N'(x) = \frac{1}{h}(y_i - y_{i-1}) + \frac{h}{2}\textstyle\sum_{j=1}^{N-1} C_{i,j} d_j + \frac{h}{2}(C_{i,0} d_0 + C_{i,N} d_N)$$

where

$$C_{i,j} = A_{i-1,j}^{-1}(\tfrac{1}{3} - (1-r)^2) + A_{i,j}^{-1}(r^2 - \tfrac{1}{3})$$

for $i = 1, 2, \cdots, N$, $j = 0, 1, \cdots, N$, and $r = (x - x_{i-1})/h$. Usually the values of y_0' and y_N' are not known. If one uses a simple and rather crude estimate of y_0' and y_N',

$$y_0^{\#} = \frac{y_1 - y_0}{h}\,, \qquad v_N^{\#} = \frac{y_N - y_{N-1}}{h}\,, \tag{2.1}$$

then

$$d_0 = d_N = 0\,.$$

We will need a better estimate later and will come back to this point. Boundary condition (2.1) is used just for convenience. All the following derivations will go through if we replace (2.1) by any of the boundary conditions mentioned earlier. Now

$$\begin{aligned} s_N'(x) &= \frac{1}{h}(y_i - y_{i-1}) + \frac{h}{2}\sum_{j=1}^{N-1} C_{i,j} d_j \\ &= \frac{1}{h}(y_i - y_{i-1}) + \frac{h}{2}\sum_{j=1}^{N-1} C_{i,j} \frac{3}{h^2}(y_{j+1} - 2y_j + y_{j-1}) \\ &= \frac{1}{h}\left[F_n\left(\frac{i}{N}\right) - F_n\left(\frac{i-1}{N}\right)\right] + \frac{1}{h}\sum_{j=0}^{N-1} D_{i,j}\left[F_n\left(\frac{j+1}{N}\right) - F_n\left(\frac{j}{N}\right)\right] \end{aligned}$$

where

$$\begin{aligned} D_{i,j} &= -\tfrac{3}{2}C_{i,1} && \text{if } j = 0 \\ &= \tfrac{3}{2}(C_{i,j} - C_{i,j+1}) && \text{if } j = 1, 2, \cdots, N-2 \\ &= \tfrac{3}{2}C_{i,N-1} && \text{if } j = N-1 \,. \end{aligned}$$

To simplify the notation further we set $\Delta F_{n,j} = y_{j+1} - y_j$; then

$$\begin{aligned} s_N'(x) &= \frac{1}{h}\sum_{j=0}^{N-1} E_{i,j}\Delta F_{n,j} \\ &= \frac{1}{h}\int_0^1 W_N(x, y)\, dF_n(y) \end{aligned}$$

where

$$\begin{aligned} E_{i,j} &= D_{i,j} && \text{if } j \neq i-1 \\ &= D_{i,j} + 1 && \text{if } j = i-1 \end{aligned}$$

and the step function

$$W_N(x, y) \equiv W_{i,j,N}(x, r, y) = E_{i,j}$$

when $y \in [x_j, x_{j+1}]$, $x \in [x_{i-1}, x_i]$ and $r = (x - x_{i-1})/h$.

REMARK 1. Let $D(x, y) =$ cell distance of x and y; i.e.,

$$\begin{aligned} D(x, y) &= d && \text{if } x \text{ and } y \text{ are not knots} \\ &= d - 2 && \text{if } x \text{ or } y \text{ are knots} \end{aligned}$$

where $d = |i - j|$ if $x \in [x_i, x_{i+1}]$ and $y \in [x_j, x_{j+1}]$. Then

(1) $W_N(x, y)$ is a step function of y for each x.

(2) $W_N(x, y)$ is a continuous differentiable function of x for each y if we let $D(x, y)$ be constant.

(3) $|W_N(x, y)| \leqq k_1\sigma^{k_2 D(x,y)}$ for some positive constants k_1, k_2 from the expression of $A_{i,j}^{-1}$ (see [1]).

We would like to make the following assumptions:

A1: The density function $f(x) \in C^3[0, 1]$.

A2: $F'(x) = f(x) \neq 0$ on $[0, 1]$ and f is bounded.

For convenience, we introduce the following notation:

$$Z_n^0(t) = n^{\frac{1}{2}}(F_n^*(t) - t)$$

where $F_n^* = F_n(F^{-1})$ is the sample distribution of $F(X_1), \cdots, F(X_n)$. $Z_n^0(t)$ will be approximated by $Z^0(t)$, the Brownian bridge, given by

$$Z^0(t) = Z(t) - tZ(1)\,, \qquad t \in [0, 1]$$

where Z is a standard Wiener process on [0, 1]. Let

$$\begin{aligned} Y_n(t) &= (hf(t))^{-\frac{1}{2}} \textstyle\int_0^1 W_N(t, y)\, dZ_n^0(F(y)) \\ &= (nhf^{-1}(t))^{\frac{1}{2}}(s_N'(t) - Es_N'(t))\,. \end{aligned}$$

This process is central to our discussion. We also introduce the following approximations.

$$\begin{aligned} {}_0Y_n(t) &= (hf(t))^{-\frac{1}{2}} \textstyle\int_0^1 W_N(t, y)\, dZ^0(F(y))\,. \\ {}_1Y_n(t) &= (hf(t))^{-\frac{1}{2}} \textstyle\int_0^1 W_N(t, y)\, dZ(F(y))\,. \\ {}_2Y_n(t) &= (hf(t))^{-\frac{1}{2}} \textstyle\int_0^1 W_N(t, y) f^{\frac{1}{2}}(y)\, dZ(y)\,. \end{aligned}$$

All the integrals with respect to $dZ^0(F(\cdot))$, $dZ(F(\cdot))$, $dZ(\cdot)$, and $dZ^0(\cdot)$ are taken in the L^2 sense. For convenience, suppose all our processes are realized as random elements taking their values in the space $D[0, 1]$ (cf. [3]).

For $x \in D[0, 1]$, let

$$\|x\| = \sup_{0 \leq t \leq 1} |x(t)|\,.$$

Notice that ${}_1Y_n(t)$ and ${}_2Y_n(t)$ have the same covariance structure.

Our approximations start with the following theorem of Komlós, Major and Tusnády [4].

THEOREM 2.1. *There exists a probability space* (Ω, A, P) *on which one can construct versions of* Z_n^0 *and* Z^0 *such that*

$$\|Z_n^0 - Z^0\| = O_P(n^{-\frac{1}{2}} \log n)\,.$$

From this we can derive the following lemma.

LEMMA 2.1. *If the processes* Z_n^0 *and* Z^0 *are constructed as above and* A2 *holds, then*

$$\|Y_n - {}_0Y_n\| = O_P(h^{-\frac{1}{2}} n^{-\frac{1}{2}} \log n)$$

as $nh \to \infty$, $h \to 0$.

PROOF.

$$\begin{aligned} Y_n(t) &= (hf(t))^{-\frac{1}{2}} \textstyle\int_0^1 W_N(t, y)\, dZ_n^0(F(y)) \\ &= (hf(t))^{-\frac{1}{2}} \textstyle\sum_{j=0}^{N-1} E_{i,j} \int_{[j/N,(j+1)/N)} dZ_n^0(F(y)) \end{aligned}$$

and

$${}_0Y_n(t) = (hf(t))^{-\frac{1}{2}} \textstyle\sum_{j=0}^{N-1} E_{i,j} \int_{[j/N,(j+1)/N)} dZ^0(F(y))\,.$$

Hence

$$\begin{aligned} |Y_n(t) - {}_0Y_n(t)| &\leq (hf(t))^{-\frac{1}{2}} \cdot 2 \textstyle\sum_{j=0}^{N-1} |E_{i,j}| \|Z_n^0 - Z^0\| \\ &= O_P(h^{-\frac{1}{2}} n^{-\frac{1}{2}} \log n) \end{aligned}$$

since

$$\textstyle\sum_{j=0}^{N-1} |E_{i,j}| \leqq 16 \quad \text{in any case.}$$

LEMMA 2.2. *If* A2 *holds then*

$$\|{}_0Y_n - {}_1Y_n\| = O_P(h^{\frac{1}{2}}) . \tag{2.2}$$

PROOF.

$$\begin{aligned} |{}_0Y_n(t) - {}_1Y_n(t)| &= |Z(1)|(hf(t))^{-\frac{1}{2}}|\textstyle\int_0^1 W_N(t, y)f(y)\,dy| \\ &\leqq |Z(1)|(hf(t))^{-\frac{1}{2}} \textstyle\sum_{j=0}^{N-1} |E_{i,j}| \int_{[j/N,(j+1)/N)} f(y)\,dy . \end{aligned}$$

Since

$$\textstyle\int_{[j/N,(j+1)/N)} f(y)\,dy = O(h) ,$$

(2.2) follows.

For convenience, we further introduce the following functionals:

$$\begin{aligned} T_n &= nh \textstyle\int_0^1 [s_N'(x) - E(s_N'(x)]^2 a(x)\,dx \\ &= \textstyle\int_0^1 L_n^2(x)a(x)\,dx \end{aligned}$$

where $L_n = f^{\frac{1}{2}}Y_n$ and $a(x)$ is a bounded piecewise smooth function. Set

$${}_iL_n = f^{\frac{1}{2}}{}_iY_n , \qquad i = 0, 1, 2 .$$

From now on, we assume A1 and A2 hold through this section.

LEMMA 2.3. *If*

$$h = n^{-\delta} \qquad \textit{with} \quad \delta < \tfrac{1}{2} \tag{2.3}$$

then

$$|T_n - \textstyle\int_0^1 {}_0L_n^2(x)a(x)\,dx| = o_p(h^{\frac{1}{2}}) .$$

PROOF.

$$\begin{aligned} |T_n - \textstyle\int_0^1 {}_0L_n^2(x)a(x)\,dx| &= \textstyle\int_0^1 f(x)[Y_n^2(x) - {}_0Y_n^2(x)]a(x)\,dx \\ &\leqq \sup_x |f(x)| \sup |Y_n - {}_0Y_n| \sup |Y_n + {}_0Y_n| \textstyle\int_0^1 |a(x)|\,dx \\ &= O_P(h^{-\frac{1}{2}}n^{-\frac{1}{2}} \log n(\log N)^{\frac{1}{2}}) , \end{aligned}$$

because

$${}_0Y_n(t) = [hf(t)]^{-\frac{1}{2}} \textstyle\sum_{j=0}^{N-1} E_{i,j} \int_{[j/N,(j+1)/N)} dZ^0(F(y))$$

and

$$\textstyle\int_{[j/N,(j+1)/N)} dZ^0(F(y)) = \int_{[j/N,(j+1)/N)} dZ(F(y)) + Z(1) \int_{[j/N,(j+1)/N)} dF(y) .$$

The variance of

$$\textstyle\int_{[j/N,(j+1)/N)} dZ(F(y))$$

is

$$F\left(\frac{j+1}{N}\right) - F\left(\frac{j}{N}\right) = O(h) .$$

Thus

$$\max_{0 \leqq j \leqq N-1} \{\textstyle\int_{[j/N,(j+1)/N)} dZ(F(y))\} = O_p(h^{\frac{1}{2}}(2 \log N)^{\frac{1}{2}})$$

and it is obvious that

$$Z(1) \int_{[j/N,(j+1)/N)} dF(y) = O(h) .$$

Therefore

$$|{}_0Y_n(t)| = O_p((2 \log N)^{\frac{1}{2}}) .$$

By Lemma 2.1, we have $|Y_n| = O_p((2 \log N)^{\frac{1}{2}})$. Hence

$$\begin{aligned} |T_n - \textstyle\int_0^1 {}_0L_n^2(x)a(x)\,dx| &= O_p(h^{-\frac{1}{2}}n^{-\frac{1}{2}} \log n \cdot (2 \log N)^{\frac{1}{2}}) \\ &= o_p(h^{\frac{1}{2}}) \end{aligned}$$

if

$$h = n^{-\delta} \qquad \text{with} \quad \delta < \tfrac{1}{2}$$

or

$$n^{-\frac{1}{2}} \log n(2 \log N)^{\frac{1}{2}} = o(h) .$$

LEMMA 2.4.

$$|\textstyle\int_0^1 ({}_1L_n^2(x) - {}_0L_n^2(x))a(x)\,dx| = o_p(h^{\frac{1}{2}}) .$$

PROOF.

$$\begin{aligned} &|\textstyle\int_0^1 ({}_1L_n^2(x) - {}_0L_n^2(x))a(x)\,dx| \\ &\quad = |\textstyle\int_0^1 f(x)({}_1Y_n^2(x) - {}_0Y_n^2(x))a(x)\,dx| \\ &\quad = |\textstyle\int_0^1 f(x)h^{-1}f^{-1}(x)[(\int_0^1 W_N(x, y)\,dZ^0(F(y)))^2 \\ &\qquad - (\textstyle\int_0^1 W_N(x, y)\,dZ(F(y)))^2]a(x)\,dx| \\ &\quad = h^{-1}|\textstyle\int_0^1 \{(Z(1) \int_0^1 W_N(x, y)f(y)\,dy)^2 \\ &\qquad - 2Z(1) \textstyle\int_0^1 W_N(x, y)\,dZ(F(y)) \int_0^1 W_N(x, y)f(y)\,dy\}a(x)\,dx| \\ &\quad \leqq h^{-1}|Z(1)|^2 \textstyle\int_0^1 |a(x)|\,dx|\int_0^1 W_N(x, y)f(y)\,dy|^2 \\ &\qquad + 2h^{-1}|Z(1)|O(h)|\textstyle\int_0^1 \int_0^1 W_N(x, y)g(x)\,dZ(F(y))a(x)\,dx| \end{aligned}$$

where

$$hg(x) = \int_0^1 W_N(x, y)f(y)\,dy + o(h)$$

and clearly $g(x)$ is bounded in view of A1 and Remark 1, (3).

$$\begin{aligned} &|\textstyle\int_0^1 \int_0^1 W_N(x, y)\,dZ(F(y))g(x)a(x)\,dx| \\ &\quad = |\textstyle\int_0^1 \sum_{j=0}^{N-1} E_{i,j} \int_{[j/N,(j+1)/N)} dZ(F(y))g(x)a(x)\,dx| \\ &\quad = |\textstyle\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \int_{[i/N,(i+1)/N)} \int_{[j/N,(j+1)/N)} E_{i,j}\,g(x)a(x)\,dx\,dZ(F(y))| \qquad (2.4) \\ &\quad = |\textstyle\sum_{j=0}^{N-1} \int_{[j/N,(j+1)/N)} \sum_{i=0}^{N-1} \int_{[i/N,(i+1)/N)} E_{i,j}\,g(x)a(x)\,dx\,dZ(F(y))| \\ &\quad = |\textstyle\sum_{j=0}^{N-1} \int_{[j/N,(j+1)/N)} \bar{g}(y)\,dZ(F(y))| \\ &\quad = |\textstyle\int_0^1 \bar{g}(y)\,dZ(F(y))| \end{aligned}$$

where

$$\bar{g}(y) = \sum_{i=0}^{N-1} \int_{[i/N,(i+1)/N)} E_{i,j}\,g(x)a(x)\,dx$$

is of order h. Also

$$\operatorname{Var}(\textstyle\int_0^1 \bar{g}(y)\,dZ(F(y))) = \int_0^1 \bar{g}^2(y)f(y)\,dy = O(h^2) .$$

Hence (2.4) equals to $O_p(h)$. Obviously,

$$\int_0^1 W_N(x, y)f(y)\,dy = O(h)\,.$$

Therefore the right-hand side of the inequality in the proof is $O_p(h) = o_p(h^{\frac{1}{2}})$. Hence the result follows.

LEMMA 2.5. *If* (2.3) *holds then*

$$|T_n - \int_0^1 {}_1L_n^2(x)a(x)\,dx| = o_p(h^{\frac{1}{2}})\,,$$
$$|T_n - \int_0^1 {}_2L_n^2(x)a(x)\,dx| = o_p(h^{\frac{1}{2}})\,.$$

PROOF. These are just simple corollaries of previous results. Now

$${}_2L_n(x) = h^{-\frac{1}{2}} \int_0^1 W_N(x, t)f^{\frac{1}{2}}(t)\,dZ(t)\,.$$

For $0 < x, y < 1$, $x \in [i/N, (i+1)/N]$, $y \in [k/N, (k+1)/N]$ and $d = |i-k|$,

$$\mathrm{Cov}\,({}_2L_n(x), {}_2L_n(y)) = \mathrm{Cov}\,[h^{-\frac{1}{2}} \sum_{j=0}^{N-1} \int_{[j/N,(j+1)/N)} E_{i,j} f^{\frac{1}{2}}(t)\,dZ(t)\,,$$
$$h^{-\frac{1}{2}} \sum_{j=0}^{N-1} \int_{[j/N,(j+1)/N)} E_{k,j} f^{\frac{1}{2}}(t)\,dZ(t)] = h^{-1} \sum_{j=0}^{N-1} \int_{[j/N,(j+1)/N)} E_{i,j}E_{k,j}f(t)\,dt$$
$$= \sum_{j=0}^{N-1} E_{i,j}E_{k,j}f(jh) + O(h) \sum_{j=0}^{N-1} E_{i,j}E_{k,j}\,.$$

But we know that

$$\mathrm{Cov}\,(s_N'(x), s_N'(y)) = \mathrm{Cov}\left(\frac{1}{h} \sum_{j=0}^{N-1} E_{i,j}\Delta F_{n,j}, \frac{1}{h} \sum_{j=0}^{N-1} E_{k,j}\Delta F_{n,j}\right)$$
$$= \frac{1}{h^2} \sum_{j=0}^{N-1} E_{i,j}E_{k,j}\,\mathrm{Var}\,(\Delta F_{n,j}) + o\left(\frac{1}{n}\right)$$
$$= \frac{1}{h^2} \sum_{j=1}^{N-1} E_{i,j}E_{k,j}f(jh)\frac{h}{n} + o\left(\frac{1}{n}\right)$$
$$= \frac{1}{nh} \sum_{j=0}^{N-1} E_{i,j}E_{k,j}f(jh) + o\left(\frac{1}{n}\right)\,.$$

Hence

$$\begin{aligned}\mathrm{Cov}\,({}_2L_n(x), {}_2L_n(y)) &\equiv r(x, y) \equiv r_d(r_1, r_2) \\ &= nh\,\mathrm{Cov}\,(s_N'(x), s_N'(y)) + o(h) \qquad (2.5)\\ &= f(x)B_d(r_1, r_2) + o(h)\end{aligned}$$

asymptotically by Theorem 3 in [5].

The following lemma, due to Bickel and Rosenblatt [2], enables us to determine the characteristic function of a quadratic functional

$$Z = \int Y^2(x)a(x)\,dx \qquad (2.6)$$

of a Gaussian process $Y(x)$, under certain conditions.

LEMMA 2.6. *Let $Y(x)$, $EY(x) \equiv 0$ be a Gaussian process with bounded, uniformly continuous covariance function $r(x, y)$. If $a(x)$ is a piecewise smooth integrable function, the quadratic functional* (2.6) *has characteristic function formally given by*

$$E(e^{itZ}) = \exp\{\sum_{k=1}^{\infty} 2^{k-1}(it)^k C_k/k\} \qquad (2.7)$$

with

$$C_k = \int \cdots \int r(x_1, x_2)r(x_2, x_3) \cdots r(x_k, x_1)a(x_1) \cdots a(x_k)\,dx_1 \cdots dx_k\,.$$

The representation (2.7) *is valid for* $|t| < 1/2K$ *where* $K = \|r\| \int |a(t)|\,dt$ *and* $(k-1)!\,2^{k-1}C_k$ *are the cumulants of* (2.6).

REMARK 2. The simple estimate of the derivative at the endpoints from (2.1) is too crude to give a uniform estimate of order of magnitude sufficiently small for the bias on the interval [0, 1]. We know that boundary conditions (I), (II), and (III) will give a uniform estimate of order of magnitude for the bias on [0, 1] although we do not have asymptotically precise information in the immediate neighborhood of 0 and 1 [5]. The same situation occurs for the covariance, i.e., $B_d(r_1, r_2)$ in (2.5) is not known if x or y is close to the boundary. In the following derivation, we will use boundary condition (III) which specifies the derivatives at the endpoints. In the proof of Theorem 3 in [5] we see obviously that the order of magnitude is uniform on [0, 1] for variance and covariance.

LEMMA 2.7.

$$E(\int_0^1 {}_2L_n^2(x)a(x)\,dx) = \int_0^1 B(x)\,dx \int_0^1 f(x)a(x)\,dx + o(h^{\frac{1}{2}})$$

where $B(x) = B_0(r_1, r_1)$.

PROOF. Set $\frac{1}{2} < q < 1$. Then, asymptotically, (2.5) is true for x, y in $[h^q, 1 - h^q]$. For convenience, we extend (2.5) to [0, 1] formally. The error is of smaller order of magnitude. Then from Lemma 2.6 and Remark 2, we have

$$\begin{aligned} E(\int_0^1 {}_2L_n^2(x)a(x)\,dx) &= \int_0^1 f(x)B(r)a(x)\,dx + O(h) + O(h^q) \\ &= \sum_{j=0}^{N-1} \int_{jh}^{(j+1)h} f(x)B(r)a(x)\,dx + O(h^q) \\ &= \sum_{j=0}^{N-1} f(jh)a(jh) \int_{jh}^{(j+1)h} B(x)\,dx + O(h^q) \\ &= \sum_{j=0}^{N-1} f(jh)a(jh)h \int_0^1 B(x)\,dx + O(h^q) \\ &= \int_0^1 B(x)\,dx \cdot \int_0^1 f(x)a(x)\,dx + o(h^{\frac{1}{2}})\,. \end{aligned}$$

One should keep in mind that $r = (1/h)[x - x_i]$ if $x \in [x_i, x_{i+1}]$. Also, it is understood that the error term is carried through each step implicitly. Similarly we have

LEMMA 2.8.

$$\begin{aligned} &\operatorname{Var}(\int_0^1 {}_2L_n^2(x)a(x)\,dx) \\ &\qquad = 2h \int_0^1 f^2(x)a^2(x)\,dx \cdot \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d^2(x_1, x_2)\,dx_1\,dx_2 + o(h)\,. \end{aligned}$$

PROOF.

$$\begin{aligned} &\operatorname{Var}[\int_0^1 {}_2L_n^2(x)a(x)\,dx] \\ &\quad = 2 \int_0^1 \int_0^1 f(x_1)f(x_2)B_d(r_1, r_2)B_d(r_2, r_1)a(x_1)a(x_2)\,dx_1\,dx_2 + O(h^{2q}) \\ &\quad = 2 \sum_i \sum_j \int_{ih}^{(i+1)h} \int_{jh}^{(j+1)h} f(x_1)f(x_2)a(x_1)a(x_2)B_d^2(r_1, r_2)\,dx_1\,dx_2 + O(h^{2q}) \\ &\quad = 2 \sum_{j=0}^{N-1} f^2(jh)a^2(jh)h^2 \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d^2(x_1, x_2)\,dx_1\,dx_2 + O(h^{2q}) \end{aligned}$$

(since $B_d(r_1, r_2)$ damps out exponentially as $d \to \infty$ we can use this approximation)

$$= 2h \int_0^1 f^2(x)a^2(x)\,dx \cdot \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d{}^2(x_1, x_2)\,dx_1\,dx_2 + o(h)$$

to the first order. Similarly, the kth cumulant of $\int {}_2L_n{}^2(x)a(x)\,dx$ is to the first order

$$(k-1)!\; 2^{k-1}h^{k-1} \int_0^1 f^k(x)a^k(x)\,dx \cdot V^k + o(h^{k/2})$$

where

$$V^k = \sum_{d_1} \cdots \sum_{d_k} \int \cdots \int B_{d_1}(x_1, x_2)B_{d_2}(x_2, x_3) \cdots B_{d_k}(x_k, x_1)\,dx_1 \cdots dx_k\,.$$

As a result, we have the following theorem.

THEOREM 2.2. *Let* A1, A2 *hold and suppose that* $a(x)$ *is integrable piecewise smooth and bounded. Suppose further that* (2.3) *holds. Then*

$$h^{-\frac{1}{2}}\left[T_n - \int_0^1 B(x)\,dx \int_0^1 f(x)a(x)\,dx\right]$$

is asymptotically normally distributed with mean zero and variance

$$2 \int_0^1 f^2(x)a^2(x)\,dx \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d{}^2(x_1, x_2)\,dx_1\,dx_2\,.$$

The statistic

$$\tilde{T}_n = nh \int_0^1 [s_N'(x) - f(x)]^2 a(x)\,dx$$

probably is of greater interest than T_n.

Before we go on to discuss the statistic $\tilde{T}_n$, we make a further comment on the boundary condition. In many practical situations the "true" derivatives y_0' and y_N' at the endpoints are not known. To compute a spline estimate, one would either set certain boundary conditions based on some a priori knowledge or estimate boundary conditions as well as possible from the collected data. The estimates in (2.1) are known to be too crude. In order to maintain a bias error, in terms of order of magnitude, as small as in the interior, we look at the following estimates:

$$\text{(2.8)} \qquad y_0{}^* = \frac{1}{h}\left(\tfrac{1}{3}y_3 - \tfrac{3}{2}y_2 + 3y_1 - \tfrac{11}{6}y_0\right)$$

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

Then

$$Ey_0{}^* - f(0)$$

$$= \frac{1}{h}\left[\tfrac{1}{3}F(3h) - \tfrac{3}{2}F(2h) + 3F(h) - \tfrac{11}{6}F(0)\right] - f(0)$$

$$= \frac{1}{h}\Big[\tfrac{1}{3}\Big(F(0) + (3h)F'(0) + \tfrac{1}{2}(3h)^2F''(0) + \frac{1}{3!}(3h)^3F'''(0) + O(h^4)\Big)$$

$$- \tfrac{3}{2}\Big(F(0) + (2h)F'(0) + \tfrac{1}{2}(2h)^2F''(0) + \frac{1}{3!}(2h)^3F'''(0) + O(h^4)\Big)$$

$$+ 3\Big(F(0) + hF'(0) + \tfrac{1}{2}h^2F''(0) + \frac{1}{3!}h^3F'''(0) + O(h^4)\Big) - \tfrac{11}{6}F(0)\Big] - f(0)$$

$$= F'(0) - f(0) + O(h^3) = O(h^3)\,.$$

Similarly

$$Ey_N^* - f(1) = O(h^3) .$$

Hence the bias $b_N(x)$ has the same order of magnitude for every $x \in [0, 1]$ under the "boundary conditions" (2.8). The order of magnitude for the variance of $s_N'(x)$ with (2.8) is uniform also as can be seen easily.

THEOREM 2.3. *Assume* A1, A2 *hold. If*

$$h = o(n^{-\frac{2}{13}}) \qquad \text{and} \qquad n^{-\frac{1}{2}}(\log n)(\log N)^{\frac{1}{2}} = o(h) \tag{2.9}$$

as $n \to \infty$, *then*

$$h^{-\frac{1}{2}}[nh \int_0^1 [s_N'(x) - f(x)]^2 a(x)\, dx - \int_0^1 f(x)a(x)\, dx \int_0^1 B(x)\, dx]$$

is asymptotically normally distributed with mean zero and variance

$$2 \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d^2(x_1, x_2)\, dx_1\, dx_2 \int_0^1 f^2(x)a^2(x)\, dx$$

as $n \to \infty$ *where* $a(x)$ *is a piecewise smooth integrable function.*

PROOF.

$$\begin{aligned}
\tilde{T}_n &= nh \int_0^1 [s_N'(x) - f(x)]^2 a(x)\, dx \\
&= nh\{\int_0^1 [s_N'(x) - Es_N'(x)]^2 a(x)\, dx \\
&\quad + 2 \int_0^1 [s_N'(x) - Es_N'(x)][Es_N'(x) - f(x)]a(x)\, dx \\
&\quad + \int_0^1 [Es_N'(x) - f(x)]^2 a(x)\, dx\} \\
&= T_n + 2nh \int_0^1 [s_N'(x) - Es_N'(x)][Es_N'(x) - f(x)]a(x)\, dx \\
&\quad + nh \int_0^1 [Es_N'(x) - f(x)]^2 a(x)\, dx .
\end{aligned}$$

Since Var $(s_N'(x)) = O(1/nh)$ and bias $b_N(x) = O(h^3)$ are uniform on [0, 1] (see [5]),

$$2nh \int_0^1 [s_N'(x) - Es_N'(x)][Es_N'(x) - f(x)]a(x)\, dx$$

is asymptotically normal with mean zero and variance

$$(nh)^2 \cdot O\left(\frac{1}{nh}\right) \cdot (O(h^3))^2 = O(nh^7) .$$

Also

$$nh \int_0^1 [Es_N'(x) - f(x)]^2 a(x)\, dx = nhO(h^6) = O(nh^7)$$

if we require $nh^{\frac{13}{2}} \to 0$ as $n \to \infty$ then $h^{-\frac{1}{2}}[\tilde{T}_n - T_n] = o_p(1)$. This and Theorem 2.2 yield Theorem 2.3.

REMARK. The statistic in Theorem 2.3 has an obvious application in the test of goodness-of-fit.

We can test $H: f = f_0$ at a given level α by calculating $\tilde{T}_n$ for $f = f_0$ and reject H when $T_n \geqq d(\alpha)$ where by Theorem 2.3

$$d(\alpha) = \int f_0(x)a(x)\, dx \int B(x)\, dx + h^{\frac{1}{2}}\Phi^{-1}(1 - \alpha)/SD$$

where $SD = [2 \sum_{d=-\infty}^{\infty} \int\int B_d^2(x_1, x_2)\,dx_1\,dx_2 \int f_0^2(x)a^2(x)\,dx]^{\frac{1}{2}}$. If we can take $a(x) = f_0^{-1}(x)$ then this is distribution free. The rounded off numerical values of mean and variance can be obtained from

$$\int_0^1 B(x)\,dx = 1.166$$

$$2 \sum_{d=-\infty}^{\infty} \int_0^1 \int_0^1 B_d^2(x_1, x_2)\,dx_1\,dx_2 = 2.7434\,.$$

Acknowledgment. I would like to thank Professor Murray Rosenblatt for his helpful advice and encouragement. This is part of the author's doctoral dissertation.

REFERENCES

[1] Ahlberg, J. H., Nilson, E. N. and Walsh, J. L. (1967). *Theory of Splines and Their Application.* Academic Press.

[2] Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.

[3] Billingsley, P. (1968). *Convergence of Probability Measure.* Wiley, New York.

[4] Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent rv's and the sample df. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **32** 111-131.

[5] Lii, K. S. and Rosenblatt, M. (1975). Asymptotic behavior of a spline estimate of a density function. *Comput. Math. Appl.* **1** 223-235.

Department of Mathematics
Northwestern University
Evanston, Illinois 60201