ASYMPTOTIC BEHAVIOR OF A SPLINE ESTIMATE OF A DENSITY FUNCTION

KEH-SHIN LI and M. ROSENBLATT
University of California, San Diego, La Jolla, California 92037, U.S.A.

Communicated by B. R. Ainger

(Received December 1974)

Abstract—A class of cubic spline estimators of probability density functions over a finite interval are considered in this paper. The precise asymptotic behavior of the bias and covariance of such estimators is obtained in the interior of the interval. The estimators are shown to be asymptotically normally distributed. The properties of these estimators are compared with those of kernel estimators. The kernel and spline estimators are compared in some Monte Carlo simulations as well as in the analysis of some data obtained in turbulent wind flow.

1. INTRODUCTION

Particular examples of spline functions have been used for years but it was I. J. Shoenberg who singled out their particular structure for special study and named them in the middle of 1940's [6]. From 1960 on they have attracted wide attention. After more than ten years in the hands of analysts, the idea of splines has been generalized and applied in many fields such as interpolation, approximation and differential equations. In 1970, Boneva, Kendall and Stefanov proposed "histosplines" in a density estimation [2]. This appears to have been the first time spline-like techniques have been proposed in density estimation.

Here we propose a rather traditional approach to density estimation via splines. We restrict our attention to the equal-spaced cubic spline for its simplicity. It can be generalized to quintic splines but the analysis of these would require more labor. In Section 2, a statement of the results on bias, covariance and asymptotic distribution is given. The derivation of these results is given in Section 3. And in Section 4 we try to compare spline estimation with kernel type estimation both theoretically and numerically in terms of a number of examples. The results obtained are new and give detailed information about the fine structure of the spline estimates due to "binning".

2. STATEMENT OF THE RESULTS ON BIAS, COVARIANCE AND ASYMPTOTIC DISTRIBUTION

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

\[
y_k = F_n \left( \frac{k}{N} \right), \quad k = 0, 1, \ldots, N = \frac{1}{h}
\]

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 \( x_j = j/N, \ j = 0, 1, \ldots, N \) with \( f(0) = s_n(0) = y_0 \), \( f(1) = s_n(1) = y_N \). Notice that \( s_n(x) \) also depends on \( N \). The derivative of the spline interpolator

\[
s'_n(x) = -M_{i-1} \frac{(x-x_i)^2}{2h} + M_i \frac{(x-x_{i-1})^2}{2h} - \frac{h}{6} (M_i - M_{i-1}) + \frac{1}{h} (y_i - y_{i-1})
\]

when \( x \in [x_{i-1}, x_i] \) where

\[
M_i = s''_n(x_i)
\]

(see [1] p. 10), and \( s'_n(x) \) is used as the estimator of \( f(x) \). At this point we shall give statements of a number of the results on bias and covariance. The proofs given later are for the boundary conditions for the spline \( f(0) = s'_n(0), f(1) = s'_n(1) \) cited above. However, simple modifications of
the proof show that the results are still valid with any other conventional boundary conditions for cubic splines, for example, if

$$s''_x(0) = 0 = s''_x(1)$$

or if one has periodic boundary conditions for the cubic spline (see "Asymptotics and representations of cubic splines" an ONR report of Rosenblatt).

The first result concerns bias and is a direct consequence of a result of Rosenblatt [4]. We introduce the following notation for convenience

$$\sigma = \sqrt{3} - 2$$

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

**Theorem 1:** Let \( f \in C^3[0, 1] \) (continuously differentiable up to third order). Then the bias

$$b_n(x) = E s'_x(x) - f(x) = \frac{f^{(3)}(x)}{4!} \left\{ \frac{1}{h} [(x_i - x)^4 - (x_{i-1} - x)^4] - h [(x_i - x)^2 - (x_{i-1} - x)^2] + o(h^3) \right\}$$

$$= \frac{f^{(3)}(x)}{4!} h^3 ((1 - r)^4 - r^4 - (1 - r)^2 + r^2 + o(1))$$

if \( 0 < x < 1 \) is fixed and \( x \in [x_{i-1}, x_i] \) (that is, \( x_{i-1} = [N x]/N \) where \([y]\) is the greatest integer less than or equal to \( y \)) as \( N \to \infty \). The following result on asymptotic behavior of the variance is obtained.

**Theorem 2.** Let \( f \) be continuous on \([0, 1]\). The variance of the spline estimator \( s'_x(x) \) of \( f(x) \) is then given by

$$\frac{f(x)}{nh} A(r) + O \left( \frac{h}{n} \right).$$

if \( 0 < x < 1 \) is fixed and \( nh \to \infty, h \to 0 \).

Here

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

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

(7)

Notice that \( A'(0) = A'(1) = 0 \).

In Fig. 1, the function \((1 - r)^4 - r^4 - (1 - r)^2 + r^2\) that describes the asymptotic behavior of the bias is given over the range \( 0 \leq r \leq 1 \). The function \( A(r) \) that determines the asymptotic behavior of the variance is represented in Fig. 2.

Fig. 1. Bias and the function \((1 - r)^4 - r^4 - (1 - r)^2 + r^2\).
Asymptotic behavior of a spline estimate of a denisty function

One can show that \( s_n(x), s_n(y) \) are asymptotically uncorrelated if \( x, y \) are fixed and not equal when \( nh \to \infty, h \to 0 \). However, a more detailed result on the behavior of the covariance is given in the following theorem.

**Theorem 3.** Consider \( x, y \) fixed with \( 0 < x, y < 1 \). Let \( x_{i-1} = \lfloor Nx \rfloor / N, x_i = \lfloor Ny \rfloor / N \) and \( d = i - j \). Set \( r_1 = h^{-1} (x - x_{i-1}), r_2 = h^{-1} (y - x_{i-1}) \).

Then

\[
\text{cov}[s_n(x), s_n(y)] = \frac{1}{nh} \frac{3}{4} f(x) \left[ \left( \frac{r_1^2 - 1}{3} \right) \left( \frac{r_2^2 - 1}{3} \right) \right.
+ \left( \frac{1}{3} - (1 - r_1)^2 \right) \left( \frac{1}{3} - (1 - r_2)^2 \right) \left[ 6(d|\sigma^{id} - \frac{12\sigma^{id+1}}{1 - \sigma^2}|) \right]
+ \left( r_1^2 - \frac{1}{3} \right) \left( \frac{1}{3} - (1 - r_2)^2 \right) \left[ 6(d + 1)|\sigma^{id+1} - \frac{12\sigma^{id+1}}{1 - \sigma^2}| \right]
+ \left( r_2^2 - \frac{1}{3} \right) \left( \frac{1}{3} - (1 - r_1)^2 \right) \left[ 6(d - 1)|\sigma^{id-1} - \frac{12\sigma^{id-1}}{1 - \sigma^2}| \right]
\]

\[
+ \frac{f(y) \sqrt{3}}{nh} \frac{\sqrt{3}}{2} \left[ \left( r_1^2 - \frac{1}{3} \right) (\sigma^{id+1} - \sigma^{id}) + \left( \frac{1}{3} - (1 - r_1)^2 \right) (\sigma^{id} - \sigma^{id+1}) \right]
+ \frac{f(x) \sqrt{3}}{nh} \frac{\sqrt{3}}{2} \left[ \left( r_2^2 - \frac{1}{3} \right) (\sigma^{id-1} - \sigma^{id}) + \left( \frac{1}{3} - (1 - r_2)^2 \right) (\sigma^{id} - \sigma^{id+1}) \right]
+ \delta_{d,0} \frac{1}{nh} f(x) + o\left( \frac{1}{n} \right)
\]

as \( nh \to \infty, h \to 0 \).

**Theorem 4.** Given a finite number of fixed distinct points \( 0 < x_1, x_2, \ldots, x_m < 1 \), the estimates \( s_n(x_1), \ldots, s_n(x_m) \) are jointly asymptotically normal and independent with bias given by (4) and variance given by (8) if \( nh \to \infty, h \to 0 \).

**3. PROOFS OF THE RESULTS ON ASYMPTOTIC BEHAVIOR OF THE ESTIMATES**

We shall first give the proof of theorem 2 which describes the asymptotic behavior of the variance of \( s_n(x) \). The computations involved in the proof of Theorem 3 (involving the covariance) are similar but somewhat more involved.

**Proof of Theorem 2:** We first note that

\[
M_i = s_n^2(x_i) = \sum_{j=0}^{N} A_{ij} d_j
\]
where
\[ d_i = \frac{3}{h} (y_{i+1} - 2y_i + y_{i-1}), \quad j = 1, \ldots, N - 1 \]
\[ d_0 = \frac{6}{h} \left( \frac{y_1 - y_0}{h} - \frac{y_N - y_{N-1}}{h} \right) \]
\[ d_N = \frac{6}{h} \left( \frac{y_0 - y_N}{h} - \frac{y_N - y_{N-1}}{h} \right) \] (10)

and \( A_{ij} \) is the \((i, j)\) entry of \( A^{-1} \) with the

\[
A = \begin{bmatrix}
2 & 1 & 0 & 0 & 0 \\
1 & 2 & 1 & 0 & 0 \\
0 & 1 & 2 & 1 & 0 \\
0 & 0 & 1 & 2 & 1 \\
0 & 0 & 0 & 1 & 2 \\
\end{bmatrix}
\] (11)

(see [1] p. 11). Notationally let
\[ \Delta F_{n,k-1} = y_k - y_{k-1}. \] (12)

Then
\[ M_i = \sum_{j=0}^{N} A_{ij} \frac{3}{h^3} (\Delta F_{n,j} - \Delta F_{n,j-1}) \] (13)

where it is understood that
\[ y_{-1} = y_1 - 2hy_0 \]
\[ y_{N+1} = y_{N-1} + 2hy_N. \] (14)

To further simplify notation let
\[ a_{ij} = \frac{h}{2} \left( r^2 - \frac{1}{3} \right) A_{ij} + \frac{h}{2} \left( \frac{1}{3} - (1 - r)^2 \right) A_{i-1,j}. \] (15)

Then using (12), (13) and (15) we find that if \( x \in [x_{i-1}, x_i] \)

\[
\text{var} (s(x)) = \text{var} \left( \frac{3}{h^3} \sum_{j=0}^{N} a_{ij} (\Delta F_{n,j} - \Delta F_{n,j-1}) + \frac{1}{h} \Delta F_{n,i-1} \right) \\
= \frac{9}{h^4} \left[ \sum_{j=0}^{N} a_{ij}^2 \text{var} (\Delta F_{n,j}) + \text{var} (\Delta F_{n,i-1}) \right] \\
+ 2 \sum_{0 \leq j < k < N} a_{ij}a_{ik} \left\{ \text{cov} (\Delta F_{n,j}, \Delta F_{n,k}) + \text{cov} (\Delta F_{n,j-1}, \Delta F_{n,k-1}) \right\} \\
+ \frac{1}{h^3} \text{var} (\Delta F_{n,i}) - \frac{18}{h^4} \left[ \sum_{j=0}^{N-1} a_{ij}a_{i,j+1} \text{var} (\Delta F_{n,j}) \right] \\
+ 2 \sum_{0 \leq j < k < N-1} (a_{ij}a_k + a_{i,j}a_{i,k}) \text{cov} (\Delta F_{n,j}, \Delta F_{n,k}) \\
+ \sum_{j=0}^{N} a_{i0}a_{ij} \text{cov} (\Delta F_{n,-1}, \Delta F_{n,j}) + \sum_{j=0}^{N} a_{i,j}a_{i,j} \text{cov} (\Delta F_{n,N}, \Delta F_{n,j-1}) \\
+ \frac{6}{h^3} \left[ \sum_{j=1}^{N} a_{ij} \text{cov} (\Delta F_{n,j}, \Delta F_{n,j-1}) - \sum_{j \neq 1} a_{ij} \text{cov} (\Delta F_{n,j-1}, \Delta F_{n,j-1}) \right] \\
+ (a_{i,i-1} - a_{i,i}) \text{var} (\Delta F_{n,i-1}) \right]. \] (16)
However, only the terms in the ungainly expression above involving variances make a major contribution asymptotically as one can see from the following remarks on the sample distribution. Since

$$\text{cov} (F_n(x), F_n(x')) = \frac{1}{n} [F(\min (x, x')) - F(x)F(x')]$$

we have

$$\text{cov} (\Delta F_{ni}, \Delta F_{nj}) = \frac{1}{n} \text{DF} \left( \frac{i}{N} \right) \Delta F \left( \frac{j}{N} \right) = O \left( \frac{h^2}{n} \right) \quad \text{if} \quad i \neq j$$

and

$$\text{var} (\Delta F_{ni}) = \frac{1}{n} \Delta F \left( \frac{i}{N} \right) + O \left( \frac{h^2}{n} \right).$$

Here it is to be understood that $F(x) = \int f(u)du$ and $\Delta F(i/N) = F(i/N) - F((i - 1)/N)$.

Using (16), this implies that

$$\text{var} (s_n(x)) = \frac{9}{N} \sum_{j=0}^{N} (a_{ij} - a_{ij-1})^2 \frac{1}{n} \Delta F \left( \frac{i}{N} \right) + \frac{1}{n^2} \Delta F \left( \frac{i-1}{N} \right)$$

$$+ \frac{6}{h^3} \sum_{j=0}^{N} |a_{ij-1} - a_{ij}| \frac{1}{n} \Delta F \left( \frac{i-1}{N} \right) + O \left( \frac{h^2}{n} \right).$$

The elements of $A^{-1}$ (see [1], p. 38) are given by

$$A_{ij} = \frac{\sigma^{i-j}(1+\sigma^{2i})(1+\sigma^{2N-2i})}{(2+\sigma)(1-\sigma^{2N})} \quad \text{if} \quad 0 < i \leq j < N$$

$$A_{i,N} = \frac{(N+1-i)(1+\sigma^{2i})}{(2+\sigma)(1-\sigma^{2N})} \quad \text{if} \quad 0 < i \leq N$$

$$A_{0,i} = \frac{2\sigma^{i}(1+\sigma^{2N-2i})}{(2+\sigma)(1-\sigma^{2N})} \quad \text{if} \quad 0 < j < N$$

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

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

with $\sigma = \sqrt{3} - 2$, $A_{ij}^{-1} = A_{ji}^{-1}$; $0 < i, j < N$ and $A_{1,N}^{-1} = A_{N-1,N}^{-1}$. For all $i,j$, $|A_{ij}^{-1}| \leq 0.3^{1-i-j}$ and for any given $\epsilon > 0$ and

$$\epsilon < \frac{i}{N}, \frac{j}{N} < 1 - \epsilon$$

it follows that $A_{ij}^{-1}$ is to the first order

$$\sigma^{i-j}(2+\sigma)$$

where

$$\sum_i \frac{\sigma^{i-j}}{2+\sigma} = \frac{1}{3}.$$

Using (23), for $i,j$ in the range (22)

$$a_{ij} = a_{ij-1}$$
\[
\begin{align*}
&= \begin{cases}
\frac{h}{2} \left( r^2 - \frac{1}{3} \right) \frac{\sigma(1 - \sigma^{-1})}{2 + \sigma} + \frac{h}{2} \left( 1 - r^2 \right) \frac{1 - \sigma}{2 + \sigma} & \text{if } j = i - 1 \\
\frac{h}{2} \left( r^2 - \frac{1}{3} \right) \frac{\sigma^{j-i}(1 - \sigma^{-1})}{2 + \sigma} + \frac{h}{2} \left( 1 - r^2 \right) \frac{\sigma^{j-i+1}(1 - \sigma)}{2 + \sigma} & \text{if } j \geq i \\
\frac{h}{2} \left( r^2 - \frac{1}{3} \right) \frac{\sigma^{j-i}(1 - \sigma^{-1})}{2 + \sigma} + \frac{h}{2} \left( 1 - r^2 \right) \frac{\sigma^{j-i+1}(1 - \sigma^{-1})}{2 + \sigma} & \text{if } j \leq i - 2.
\end{cases}
\end{align*}
\]

Since
\[
\Delta F \left( \frac{i - 1}{N} \right) = f(x)h + O(h^2)
\]
we have
\[
\begin{align*}
\text{var} (s'(x)) &= \frac{9}{h^4} \left( 1 - \sigma^2 \right)^2 \left[ \frac{1}{3} - (1 - r)^2 - \left( r^2 - \frac{1}{3} \right) \right]^2 \frac{f(x)h}{n} \\
&\quad + \frac{9}{h^4} \left( 1 - \sigma^2 \right)^2 \left[ \left( r^2 - \frac{1}{3} \right) + \frac{1}{3} (1 - r)^2 \right] \sum_{i=0}^{N} \frac{\sigma^{2i+1}}{n} \Delta F \left( \frac{i}{N} \right) \\
&\quad + \frac{9}{h^4} \left( 1 - \sigma^2 \right)^2 \left[ \left( r^2 - \frac{1}{3} \right) + \frac{1}{3} (1 - r)^2 \right] \sum_{i=0}^{N} \frac{\sigma^{2i-1}}{n} \Delta F \left( \frac{i}{N} \right) \\
&\quad + \frac{1}{h^4} \frac{hf(x)}{n} + \frac{6}{h^4} \frac{1 - \sigma}{2 + \sigma} \left[ \frac{1}{3} - (1 - r)^2 - \left( r^2 - \frac{1}{3} \right) \right] \frac{hf(x)}{n} + O \left( \frac{h}{n} \right).
\end{align*}
\]

Since \( \sigma^{2k} \) damps out exponentially and \( f \) is continuous, by using (24) in the summations in formula (26) we obtain the final result (6).

**Proof of Theorem 3:** In looking at the asymptotic covariance of \( s'(x), s'(y) \), consider \( x \in [x_{i-1}, x_i], \ y \in [x_{i-1}, x_i] \) with \( i - k = d \) with \( d \) fixed as \( nh \to \infty, h \to 0 \). This is reasonable, since we shall see, \( s'(x), s'(y) \) are asymptotically uncorrelated as \( nh \to \infty, h \to 0 \), if \( x, y \) are fixed with \( x \neq y \). Using arguments similar to those given in Theorem 2 one has
\[
\begin{align*}
\text{cov} \left( s'(x), s'(y) \right) &= \frac{9}{h^4} \sum_{i} (a_{i,j} - a_{i,j+1}) (a_{i,j} - a_{i,j+1}) \text{var} \left( \Delta F_{n,i} \right) \\
&\quad + \frac{1}{h^4} \text{cov} \left( \Delta F_{n,i-1}, \Delta F_{n,k-1} \right) + \frac{3}{h^2} \left( (a_{k-1} - a_{k}) \text{var} \left( \Delta F_{n,k} \right) \right) \\
&\quad + (a_{k-1} - a_{k}) \text{var} \left( \Delta F_{n,k-1} \right) + O \left( \frac{h}{n} \right).
\end{align*}
\]

By (25) it is seen that
\[
(a_{i,j} - a_{i,j+1}) (a_{i,j} - a_{i,j+1}) = \frac{h^2}{4(2 + \sigma)^2} \left\{ \left( r^2 - \frac{1}{3} \right) \left( r^2 - \frac{1}{3} \right) (\sigma^{i-j} - \sigma^{i-j+1}) (\sigma^{i-k} - \sigma^{i-k+1}) \\
&\quad + \left( \frac{1}{3} - (1 - r)^2 \right) \left( \frac{1}{3} - (1 - r)^2 \right) (\sigma^{i-j+1} - \sigma^{i-j+2}) (\sigma^{i-k+1} - \sigma^{i-k+2}) \\
&\quad + \left( r^2 - \frac{1}{3} \right) \left( \frac{1}{3} - (1 - r)^2 \right) (\sigma^{i-j+1} - \sigma^{i-j+2}) (\sigma^{i-k+1} - \sigma^{i-k+2}) \\
&\quad + \left( r^2 - \frac{1}{3} \right) \left( \frac{1}{3} - (1 - r)^2 \right) (\sigma^{i-j+1} - \sigma^{i-j+2}) (\sigma^{i-k+1} - \sigma^{i-k+2}) \right\}.
\]

Let
\[
A_i(i, k) = (\sigma^{i-j} - \sigma^{i-j+1}) (\sigma^{i-k} - \sigma^{i-k+1}).
\]

Then
\[
B(d) = \sum_{i} A_i(i, k) = B(-d) = |d| (\sigma^d (1 - \sigma^{-1}) (1 - \sigma) + 2 \sigma^d (1 - \sigma^2) (1 - \sigma^2)).
\]
We similarly have

\[ \sum_j A_j(i-1, k-1) = \sum_j A_j(i, k) = B(d) \]

and

\[ \sum_j A_j(i-1, k) = B(d-1). \]

By using the fact that

\[ a_{i-1, k} - a_{i, k} = \frac{h}{2(2+\sigma)} \left\{ \left( r_i^2 - \frac{1}{3} \right) (\sigma^{[k-1-i]} - \sigma^{[k-i]}) + \left( \frac{1}{3} - (1-r_i)^2 \right) (\sigma^{[k-i]} - \sigma^{[k-i+1]}) \right\} \]

and previous estimates the conclusion of Theorem 3 follows.

Let us now establish the asymptotic normality of the estimates \( s^*(x) \).

**Proof of Theorem 4:** For convenience let

\[ I_k(u, v) = \begin{cases} 1 & \text{if } u < X_k \leq v \\ 0 & \text{otherwise.} \end{cases} \] (31)

Then

\[ s^*(x) = \frac{1}{h} (y_i - y_{i-1}) + \sum_{i=0}^{N} \frac{3a_u}{h^2} [y_{i+1} - 2y_i + y_{i-1}] \]

\[ = \frac{1}{nh} \sum_{i=1}^{n} \left\{ I_k(i/N, i/N) + \sum_{j=0}^{N} \frac{3}{h} a_u \left[ I_k(j/N, i/N) - I_k(i/N) \right] \right\}. \]

The terms involving \( y_{i-1}, y_{i+1} \) are negligible asymptotically compared to the others and that will be made use of in the following discussion.

Let

\[ Z_k = \frac{1}{nh} \left\{ I_k(i/N, i/N) + \sum_{j=0}^{N} \frac{3}{h} a_u \left[ I_k(j/N, i/N) - I_k(i/N) \right] \right\}. \]

Then

\[ s^*(x) = \sum_{k=1}^{n} Z_k \] (32)

with the \( Z_k \)'s independent and identically distributed. Now

\[ E[Z_k - EZ_k]^3 = E[Z_k]^3 + 3EZ_k^2[EZ_k] + 4EZ_k^3 \]

and

\[ Z_k^2 \leq \left( \frac{1}{nh} \right)^2 \left\{ I_k(i/N, i/N) + 24 I_k(i/N, i/N) + 36 \sum_{i=0}^{N} \frac{1}{h^2} (a_{i-1} + a_{i+1}) I_k(i/N, i/N) \right\} \]

since

\[ \left| \frac{1}{h} a_u \right| \leq 2 \]

(see formula (15) and \( A_{i-1}^{-1} \) as explicitly given in [1] p. 39).

Therefore

\[ EZ_k^2 \leq \left( \frac{1}{nh} \right)^2 o(h) = O\left( \frac{1}{nh^2} \right) \] (34)
and

\[ |Z_k|^3 \leq Z_k^2 O\left( \frac{1}{nh} \right) \]

since \( |Z_k| \) is bounded by a constant multiple of \( 1/nh \). Hence

\[ E|Z_k|^3 = O\left( \frac{1}{nh^2} \right) \]

and this together with (33), (34) imply

\[ \sum_{k=1}^{n} E|Z_k - EZ_k|^3 = O\left( \frac{1}{nh^2} \right) \]

Since \( \sigma^2[s'(x)] = C/nh \) for some constant \( C \),

\[ \sum_{k=1}^{n} \frac{E|Z_k - EZ_k|^3}{\sigma[s'(x)]^3} = o(1) \]

and an application of Liapounov's central limit theorem gives us the desired result for \( s'(x) \). A trivial modification of this argument yields the result for asymptotic joint normality and independence of \( S'(X_1), \ldots, S'(X_m) \).

4. COMPARISON OF KERNEL AND SPLINE ESTIMATES

Let us briefly recall some properties of kernel estimates. Assume that \( w \) is a bounded integrable weight function with integral one and that \( X_1, \ldots, X_n \) are independent, identically distributed random variables with common density \( f(x) \). A kernel estimate based on \( w \) with bandwidth \( h(n) \) is given by

\[ f_n(x) = \frac{1}{nh(n)} \sum_{i=1}^{n} w\left( \frac{x - X_i}{h(n)} \right) = \int \frac{1}{h(n)} w\left( \frac{x - u}{h(n)} \right) dF_n(u) \]  
(35)

where \( F_n \) is the sample distribution function. If \( f \) is bounded and continuous the variance

\[ \sigma^2(f_n(x)) = \frac{f(x)}{nh(n)} \int w(u)^2 \, du \]  
(36)

as \( nh(n) \to \infty \) and \( h(n) \to 0 \). Further, if \( f \) is twice continuously differentiable with bounded derivatives and \( \int u^2 |w(u)| \, du < \infty \) then the bias

\[ b_n(x) = Ef_n(x) - f(x) = \frac{h(n)^2}{2} f''(x) \int u^2 w(u) \, du + o\{h(n)^2\}. \]  
(37)

Also if \( f(x) > 0 \), \( f_n(x) - f(x) \) is asymptotically normally distributed with mean zero and variance given by (36) (see [5] for details).

Notice that in terms of the order of magnitude of the variance (and covariance) spline and kernel estimates have the same behavior as can be seen from formulas (6) and (36). The bias of the spline (4) appears to be of smaller magnitude than that of the kernel estimate. But we can think of formula (37) being characteristic for non-negative weight functions \( w \). If we allow weight functions \( w \) for which the first three moments are zero (\( w \) must then take on some negative values), we will get an asymptotic behavior of the bias, if \( f \) is smooth enough, as least as good as that of the spline estimate. Notice that splines may take on negative values occasionally (when data is sparse).

However, one should carefully notice that the spline estimate has a detailed fine structure [see formulas (4), (6) and (8)] even asymptotically. From that point of view its asymptotic structure is
Asymptotic behavior of a spline estimate of a density function

more complicated than that of the kernel estimate. Let us be explicit in the simplest case of the bias of the spline. In formula (4) of Theorem 1, the function

\[(1-r)^4 - r^4 - (1-r)^2 + r^2\]

where

\[r = \frac{1}{h} (x - x_{i-1}), \quad 0 \leq r \leq 1,\]

for \(x \in [x_{i-1}, x_i]\) appears as the fine structure factor. Clearly, in formulas (6) and (8) of Theorems 2 and 3 the fine structure factors are more complicated.

Often one takes the weight function \(w\) of a kernel estimate to be bandlimited, say to \((-1/2, 1/2)\), and then the estimates \(f_n(x), f_n(y)\) are asymptotically independent if separated by more than \(h(n)\). However, one needs a somewhat broader separation in the case of a spline estimate for this to happen as can be seen from formula (8) and Table 1. This, of course, reflects the fact that the spline estimate smooths the data in a more global manner than does such a kernel estimate (also see Boneva, Kendall and Stefanov [2]).

A number of Monte Carlo simulations were carried out using P. Lewis' package for generating pseudo-random numbers [3]. In each of these simulations 401 random numbers were generated. We give examples of generation from

1. uniform distribution
2. normal distribution
3. two superimposed triangular distributions.

In each of these we generate a kernel estimate of the density function based on Epanechnikov's weight function

\[w(u) = \begin{cases} 
3(l-x^2)/4 & \text{if } |x| \leq 1 \\
0 & \text{otherwise}
\end{cases} \]

and a spline estimate of the density function. These are given in Figs. 3–6. Notice that generally in these graphs the fluctuations of the kernel estimate and the spline estimate in magnitude are the same. However, the spline estimate is smoother over a range larger than the bandwidth. Spline

Table 1. The covariance function \(D = i - j, R_i = (x_i - x_j)/h, R_j = (y_i - y_j)/h\).

<table>
<thead>
<tr>
<th>(D)</th>
<th>(R_1 = 0.1667)</th>
<th>(R_2 = 0.3333)</th>
<th>(R_3 = 0.5000)</th>
<th>(R_4 = 0.6667)</th>
<th>(R_5 = 0.8333)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.03937E 00</td>
<td>1.12466E 00</td>
<td>1.10179E 00</td>
<td>9.67651E 00</td>
<td>7.30235E 00</td>
</tr>
<tr>
<td>1</td>
<td>3.03161E 02</td>
<td>2.23076E 01</td>
<td>6.83757E 02</td>
<td>1.33974E 01</td>
<td>4.19658E 01</td>
</tr>
<tr>
<td>2</td>
<td>9.85160E 02</td>
<td>1.33974E 01</td>
<td>1.33974E 01</td>
<td>1.33974E 01</td>
<td>1.33974E 01</td>
</tr>
<tr>
<td>3</td>
<td>2.59117E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
</tr>
<tr>
<td>4</td>
<td>8.35160E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
</tr>
<tr>
<td>5</td>
<td>2.59117E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
<td>1.03937E 00</td>
</tr>
</tbody>
</table>

C.A.M.W.A., Vol. 1, No. 2-F
Fig. 3. Monte Carlo simulation from a uniform distribution. Kernel estimate (a) with bin size 1/10. Spline estimate (b) with cell size 1/5. $y^1$, $y^2$ estimated by two cell width.

Fig. 4. Monte Carlo simulation from a standard normal on $(-1, 1)$. Kernel and spline estimates as in Fig. 3 except that $y^1$, and $y^2$ are specified.

Fig. 5. Monte Carlo simulation from half the sum of two triangular densities with base length 2 and 0.4, respectively. Spline $S$ and kernel $\square$ estimates of the density function.
estimates do seem to work very effectively in smoothing data that is already fairly regular. However, splines seem to have a marked disadvantage relative to a bandlimited kernel in following rapidly changing data as is indicated in Figs. 5 and 6 of the estimates of a superimposed triangular. Notice the undershooting lobes in the spline estimate on either side of the peak in the density function. This might be explained by the character of the covariance function of the spline (8).

We have also analyzed some data that was very kindly given to us by Dr. Wyngaard. The data was already considered in an article of Tennekes and Wyngaard [7]. It consists of readings of the derivative of a turbulent wind velocity (in a fixed direction) taken on the Kansas plains and sampled 3200 times per second for roughly an hour. We estimated the probability density of the wind velocity derivative using kernel and spline estimates and in this way it was different from that given in [7]. The data was already binned. The graph of estimates are given for the left tail of the density of \(- (\partial u / \partial t)\) (\(u\) the velocity and \(t\) the time) in three sections, Figs. 7–9. Notice that even where the spline is a good estimate there appears to be a series of small oscillations which may be due to the fine structure. It is curious that here the kernel estimate (based on a triangular-like weight function) appears to be smoother than the spline. Also, as we get far out into the tail (and closer to zero) one has the impression that the spline can show much greater instability than the kernel estimate.

We give a more detailed description of the turbulence data. The original data is formed into a histogram by combining 12 horizontal units into one bin. The left tail contains data from the -31st bin to the -337th bin. The horizontal scale has the standard deviation \(\sigma = 114.65193175\) units or \(\sigma / 12\) bins. The vertical scale is the usual one after taking a logarithmic transformation of the data. If the observation or the fitted value is less than or equal to zero, we set it equal to 9.1. The histogram is denoted by +. The kernel type fit is the piecewise linear curve denoted by X. The weight function used here is

\[
w(x) = \begin{cases} 
\frac{1}{3} & \text{if } |x| \leq \frac{1}{3} \\
\frac{1}{3} & \text{if } \frac{1}{3} \leq |x| \leq \frac{1}{3} \\
0 & \text{otherwise}
\end{cases}
\]
Figs. 7-9. Estimation of left tail of the probability density of turbulent wind velocity derivative. Turbulent Reynolds' number = 8000. Histogram, kernel and two spline estimates.
and the bandwidth is 3 bins. The spline fit with cell-width equal to one bin is the oscillatory curve without any symbol on it. The boundary condition is of Boneva–Kendall–Stefanov type, that is $y_0^L = y_0^R = 0$. The spline fit with cell-width equal to 2 bins is the thick curve denoted by 5. We also try to fit the data with an exponential function $f(x) = A e^{-B|x|^C}$ by a least squares technique using the data from -32nd bin to -192nd bin. The procedure is as follows. Given

$$
\log f(x) = \log A - B|x|^C
$$

$$
\frac{d}{dx} \log f(x) = -BC|x|^{C-1}
$$

minimize $\Sigma [-a \log f(x_i) + |x_i| (d/dx)(\log f(x_i)) + \beta]^2$ and take $\alpha$ as the estimate of $C$, $\beta$ as the estimate of $C \log A$, and $B$ as estimated by

$$
\frac{\Sigma |x_i|^\alpha (\log A - \log f(x_i))}{\Sigma (|x_i|^{\alpha})^2}
$$

where we approximate $(d/dx)(\log f(x_i))$ by $[\log f(x_{i+1}) - \log f(x_{i-1})]/2$. The smooth curve denoted by 8 is the fitted curve $\log_{10}(A e^{-B|x|^C})$. The estimated values of $A$, $B$ and $C$ are $A = 0.74$, $B = 4.2$, $C = 0.41$.

Kolmogorov and Obukhov suggested that the rate of energy dissipation in high Reynolds's number turbulence would have a log normal distribution and this was later predicted in a statistical model of Yaglom[8]. The interest in estimating the probability density of the velocity derivative is in part motivated by a desire to see to what extent this prediction is consistent with experimental data.

**Acknowledgement**—Research supported in part by the office of Naval Research and the National Science Foundation.

**REFERENCES**


