I N S T I T U T D E
S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

0308

Bernstein Estimator For Unbounded Density Function

T.BOUEZMARNI and J.M.ROLIN

http://www.stat.ucl.ac.be

Bernstein Estimator For Unbounded Density Function

T.Bouezmarni* and J.M.Rolin*

February 19, 2003

Abstract

The nonparametric estimation for an unknown probability density function f with an known compact support $[0, 1]$ not necessarily bounded at $x = 0$ is considered. For such class of density functions, we consider the Bernstein estimator. The uniform weak consistency and the uniform strong consistency on each compact I in $(0, 1)$ are established for the Bernstein estimator. We prove also the almost sure convergence to infinity at $x = 0$ of the Bernstein estimator when the density function f is unbounded at $x = 0$. To select the optimal bandwidth parameter of the Bernstein estimator, the least squares cross-validation and the likelihood cross-validation methods are developed.

Key words and phrases. Bernstein polynomial, Weak Consistency, Unbounded density, Beta density.
JEL Classification : C13,C14. *MSC 2000* : 62G07, 62G08

Acknowledgements Comments by Ingrid Van Keilegom are gratefully acknowledged; they have been most useful for the improvement of the exposition. Support from the contract"Projet d'Actions de Recherche Concertées" $n^0\, 98/03 - 217$, and from the IAP research network n^0 P5/24 of the Belgian state are also gratefully acknowledged.

*Institute of Statistics, Université Catholique de Louvain.

1 Introduction

Let $X_1, \cdots, X_n$ be a random sample from a distribution function F with an unknown probability density function f which has a known compact support. We suppose in this paper that the support is $[0,1]$ and f is unbounded at $x = 0$ (which implies a concentration of observations near zero).

The Bernstein estimator $\hat{f}_k$ for f is considered:

$$\hat{f}_k(x) = \sum_{i=0}^{k} \omega_{i,k} B(i+1, k-i+1)(x) \tag{1}$$

where the integer k is the smoothing bandwidth, the weights $\omega_{i,k}$ are random and defined by

$$\omega_{i,k} = F_n\left(\frac{i+1}{k+1}\right) - F_n\left(\frac{i}{k+1}\right) \tag{2}$$

with F_n the empirical distribution function and B represents the value of the beta density function.

The Bernstein estimator $\hat{f}_k$ is then a mixture of $(k+1)$ beta densities with random coefficients $\omega_{i,k}$. Based on the classical theorem of Weierstrass on the approximation of a continuous function by means of Bernstein polynomials, the estimator (1) was introduced by Vitale(1973) for densities with the compact support $[0,1]$. He showed that the Bernstein estimator is free of boundary bias and has mean integrated squared error of $O(n^{-4/5})$ for f in $C^2([0,1])$. When the density function f is continuous on $[0,1]$, then bounded on $[0,1]$, the uniform weak consistency is proved by Tenbusch(1994) by using the results of the oscillation behavior of the empirical processes.

The Bernstein estimator is also of the form

$$\hat{f}_k(x) = (k+1) \sum_{i=0}^{k} \omega_{i,k}\, p_{ki}(x), \tag{3}$$

where $p_{ki}(.)$ is the Binomial distribution with parameter k and x, namely:

$$p_{ki}(x) = \left(\begin{array}{c} i \\ k \end{array} \right) x^i \, (1-x)^{k-i} \qquad i = 0, \cdots, k.$$

The Bernstein estimator $\hat{f}_k$ with a slight modification of the scaling parameter is a special case of the *smooth histogram* density estimator developed by Grawronski and Stadtm$\ddot{u}$ller(1981). Grawronski and Stadtm$\ddot{u}$ller(1981) found also that the *smooth histogram* density estimators are free of boundary bias and the rate of convergence of the mean integrated squared error is $O(n^{-4/5})$ for f in $C^2([0,1])$. By using the Hadamard product technique, Stadtm$\ddot{u}$ller(1983) proves the uniform weak consistency for this estimator when the density function f is continuous on $[0,1]$.

This paper considers the Bernstein estimator to estimate a density function which is not necessarily bounded at $x = 0$. We show that the Bernstein estimator has the same behavior as the underlying density at the boundary points even if the density is unbounded at this point. In Section 2, we prove the uniform weak consistency and the uniform strong consistency, of this estimator, on each compact I in $(0,1)$ if we choose the bandwidth parameter in a certain way. The almost sure convergence to infinity at $x = 0$, of the Bernstein estimator, is proved when the density function is unbounded at $x = 0$. As the Kernel estimator, the smoothing parameter k has an influence on the quality of the Bernstein estimates. In Section 3, we develop the least-squares cross-validation and the likelihood cross-validation methods to select the optimal smoothing bandwidth parameter k. The behavior of the optimal bandwidth parameter when the density function converges faster to infinity at $x = 0$ is shown.

2 Main Results

The problem of the estimation of the density function f which has a bounded support by the standard Kernel method is the existence of the bias at the border even when the underlying density is bounded. To remove this problem of boundary bias many authors, Rice (1984), Schuster (1985), M$\ddot{u}$ller(1991) and Chen(1999) have suggested different methods. All these methods are for the class of bounded density functions. When the density is unbounded at the border points, from simulations, the kernel estimator and the true density have different behaviors at the boundary, as shown in Figure(1). The data (n=200) are drawn from the beta(0.7,6) density function. We use the standard normal kernel with the optimal bandwidth parameter $h = 0.15$ which minimizes the mean integrated squared error. Recently, two methods have been shown to accomodate the case of an unbounded density: the complicated P and PD algorithms developed by Marron and Ruppert (1994), and Beta kernel estimators by Bouezmarni and Rolin (2001). (the P, resp. PD, algorithm further requires the presence of poles at both boundaries, resp. one boundary).

In this section, we show that the Bernstein estimator $\hat{f}_k$ for unbounded densities at one boundary or both boundaries has the same asymptotic behavior as the true density f, as shown in Figure(1) (we take $k = 45$ as a bandwidth parameter). In fact, in the first step we prove the uniform weak consistency and the uniform strong consistency of the Bernstein estimator $\hat{f}_k$, on each compact I in $(0, 1)$, if we choose the bandwidth parameter in a certain way. In the second step, and without loss of generality we consider the left-hand boundary, we prove the almost sure convergence of this estimator to infinity at $x = 0$ when the underlying density function f is unbounded at this point.

We define the function f_k as follows

$$\begin{aligned} f_k(x) &= I\!E(\hat{f}_k(x)) \\ &= \sum_{i=0}^{k} \left\{ F\left(\frac{i+1}{k+1}\right) - F\left(\frac{i}{k+1}\right) \right\} B(i+1, k-i+1)(x). \end{aligned} \tag{4}$$

We note that $f_k(x)$ is the derivative of the Bernstein polynomial of order $(k + 1)$ of F. The Bernstein polynomial is introduced by Bernstein(1912) to give a simple proof of the Weierstrass theorem on the approximation of a continuous function by means of polynomials. Then the idea of the Bersntein estimator $\hat{f}_k$ of the density function is to replace, in the formula of $f_k(x)$, the unknown distribution function F by the maximum likelihood estimator F_n (the empirical distribution function) which is a consistent estimator. Therefore the Bernstein estimator is the (generalized) maximum likelihood estimator of f_k, the derivative of the Bernstein polynomial of the distribution function F. The Bernstein estimator $\hat{f}_k$ is an unbiased estimator of f_k.

Proposition 2.1 *Let f be a continuous function on the interval $(0, 1)$, i.e. $f \in C((0, 1))$. I is a compact in $(0, 1)$ and f_k defined in (4).*
Then

$$\sup_{x \in I} \mid f_k(x) - f(x) \mid \longrightarrow 0, \quad as \quad k \longrightarrow \infty. \tag{5}$$

■

Proposition(**??**) states the convergence of the bias of Bernstein estimator $\hat{f}_k$ to zero as the parameter k converges to infinity on each compact in $(0, 1)$. From Proposition(**??**) and Scheffé's theorem (1947), the integrated absolute error converges to zero almost surely and in the mean even if the density is unbounded at $x = 0$ or/and $x = 1$. To prove this proposition we need the following lemma.

Lemma 2.1 *Let S_k be a random variable with Binomial distribution $Bi(k, p)$, then we have for any positive* $0 < \delta < 1 - p$,

$$\mathbb{P}\left(\frac{S_k}{k} - p \geq \delta\right) \leq e^{-2k\delta^2}.$$

and for any $0 < \delta < p$,

$$\mathbb{P}\left(\frac{S_k}{k} - p \leq -\delta\right) \leq e^{-2k\delta^2}.$$

■

This inequality is called the inequality for the probability of large deviations for the Binomial distribution.

Proof of Lemma(2.1).

Using the Cramer-Chernoff computations, we get

$$\mathbb{P}(S_k - kp \geq k\delta) \ \leq \ e^{-k\psi_p(p+\delta)},$$

where

$$\psi_p(s) \ = \ s \ln\left(\frac{s}{p}\right) + (1 - s)\ln\left(\frac{1 - s}{1 - p}\right) \quad \text{for} \quad 0 < s < 1.$$

Now.

$$\psi_p'(s) = \ln\left(\frac{s(1 - p)}{p(1 - s)}\right) \qquad \psi_p''(s) \ = \ \frac{1}{s(1 - s)} \ \geq \ 4.$$

Using Taylor expansion,

$$\psi_p(s) \quad = \quad \psi_p(p) \ + \ (s - p)\psi_p'(p) + \frac{1}{2}(s - p)^2\psi_p''(p + \lambda(s - p)),$$

where $\lambda \in (0, 1)$, we have

$$\psi_p(s) \geq 2(s - p)^2.$$

The second inequality is similary obtained.

■

Proof of Proposition(??).

Let $x \in I$ and, without loss of generality, we suppose that $I = [\eta, 1-\eta]$, $\eta > 0$.

$$\begin{aligned}
\mid f_k(x) - f(x) \mid &= \left|(k+1)\sum_{i=0}^{k} p_{ki}(x)\int_{\frac{i}{k+1}}^{\frac{i+1}{k+1}} \{f(t)-f(x)\}\, dt\right| \\
&= \left|(k+1)\mathbb{E}_x\left(\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} \{f(t)-f(x)\}\, dt\right)\right| \\
&\leq (k+1)\mathbb{E}_x\left(\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} |f(t)-f(x)|\, dt\right)
\end{aligned}$$

where $S_k \sim Bi(k,x)$.

Let δ be a small positive number, we have

$$\begin{aligned}
\mid f_k(x) - f(x) \mid &\leq (k+1)\mathbb{E}_x\left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} |f(t)-f(x)|\, dt \ ;\ \frac{S_k}{k} - x > \delta\right] \\
&+ (k+1)\mathbb{E}_x\left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} |f(t)-f(x)|\, dt \ ;\ \frac{S_k}{k} - x < -\delta\right] \\
&+ (k+1)\mathbb{E}_x\left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} |f(t)-f(x)|\, dt \ ;\ \mid \frac{S_k}{k} - x \mid \leq \delta\right] \\
&= I + II + III. \qquad (6)
\end{aligned}$$

On one hand, using Lemma(2.1) and with some computations we get,

$$\begin{aligned}
\mid I \mid &\leq (k+1)\mathbb{E}_x\left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} f(t)\, dt \ ;\ \frac{S_k}{k} - x > \delta\right] + f(x)\mathbb{P}\left(\frac{S_k}{k} - x > \delta\right) \\
&\leq \{(k+1) + \parallel f \parallel_I\}\, \mathbb{P}\left(\frac{S_k}{k} - x > \delta\right) \\
&\leq \{(k+1) + \parallel f \parallel_I\}\, exp(-2k\delta^2). \qquad (7)
\end{aligned}$$

And

$$\mid II \mid \leq (k+1)\mathbb{E}_x\left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} f(t)\, dt \ ;\ \frac{S_k}{k} - x > \delta\right] + f(x)\mathbb{P}\left(\frac{S_k}{k} - x < -\delta\right)$$

$$
\begin{aligned}
&\leq \{(k+1)+ \| f \|_I\} \, I\!\!P_x \left(\frac{S_k}{k} - x < -\delta \right) \\
&\leq \{(k+1)+ \| f \|_I\} \, exp(-2k\delta^2). \qquad (8)
\end{aligned}
$$

The inequalities (7) and (8) are satisfied for all x in $I = [\eta, 1-\eta]$ if $\delta < \eta$. Then the first terms (I) and (II) in the inequality (6) converge to zero as the bandwidth parameter k converges to infinity if $\delta < \eta$.

On the other hand,

$$
\begin{aligned}
\mid III \mid &= (k+1) I\!\!E_x \left[\int_{\frac{S_k}{k+1}}^{\frac{S_k+1}{k+1}} \mid f(t) - f(x) \mid \, dt; \mid \frac{S_k}{k} - x \mid \leq \delta \right] \\
&= I\!\!E_x \left[\int_0^1 \mid f(\frac{S_k+s}{k+1}) - f(x) \mid \, ds; \mid \frac{S_k}{k} - x \mid \leq \delta \right]
\end{aligned}
$$

f is uniformly continuous on $[\frac{\eta}{2}, 1-\frac{\eta}{2}]$, i.e.

For $\epsilon > 0$, there exists $\delta(\epsilon) > 0$, such that

$$
\mid f(t') - f(t) \mid < \epsilon, \quad \text{for} \quad \mid t' - t \mid \leq \delta(\epsilon) \quad \text{and} \quad t, t' \in [\frac{\eta}{2}, 1-\frac{\eta}{2}].
$$

If we choose δ and k such that:

$$
\delta < \min \left(\frac{\eta}{4}, \frac{\delta(\epsilon)}{2} \right) \qquad \text{and} \qquad k > \max \left(\frac{4}{\eta}, \frac{2}{\delta(\epsilon)} \right). \qquad (9)
$$

Then

$$
\left| \frac{S_k+s}{k+1} - x \right| \leq \delta(\epsilon), \qquad \text{for} \qquad \mid \frac{S_k}{k} - x \mid \leq \delta \quad \text{and} \quad 0 < s < 1
$$

and

$$
\frac{S_k+s}{k+1} \in [\frac{\eta}{2}, 1-\frac{\eta}{2}]
$$

Indeed

$$
\left| \frac{S_k+s}{k+1} - x \right| = \left| \frac{k}{k+1} \left(\frac{S_k}{k} - x \right) + \frac{1}{k+1}(s-x) \right|
$$

$$
\begin{aligned}
&\leq \frac{k}{k+1}\left|\frac{S_k}{k}-x\right|+\frac{1}{k+1}\,|s-x| \\
&\leq \frac{k}{k+1}\,\delta+\frac{1}{k+1} \\
&\leq \delta(\epsilon)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{S_k+s}{k+1} &\leq x+\frac{k}{k+1}\,\delta+\frac{1}{k+1} \\
&\leq 1-\eta+\frac{k}{k+1}\,\delta+\frac{1}{k+1} \\
&\leq 1-\frac{\eta}{2}
\end{aligned}
$$

The same arguments prove that:

$$
\frac{S_k+s}{k+1} \geq \frac{\eta}{2}
$$

Therefore

$$
\begin{aligned}
\mid III \mid &\leq I\!\!E_x\left[\int_0^1 \mid f(\frac{S_k+s}{k+1})-f(x)\mid\, ds;\ \mid \frac{S_k}{k}-x\mid \leq \delta\right] \\
&\leq \epsilon.
\end{aligned}
$$

■

For the remainder of the section, we suppose that the bandwidth parameter $k=k_n$ depend on the sample size. we state, in the following proposition, the uniform weak consistency of Bernstein estimator under two conditions on the bandwidth parameter.

Proposition 2.2 *Assume that $X_1,\cdots,X_n$ are i.i.d observations with a continuous distribution function F, having a density function, $f \in C((0,1))$. Let I be a compact in $(0,1)$.*
Then

$$
\sup_{x\in I}\left|\hat{f}_{k_n}(x)-f(x)\right| \xrightarrow{P} 0 \quad as \quad n \longrightarrow +\infty.
$$

Under the following conditions:

$$
\lim_n k_n=+\infty \qquad and \qquad \lim_n nk_n^{-2}=+\infty.
$$

■

Proof of Proposition(??)

We begin by the the usual inequality:

$$\sup_{x\in I} \mid \hat{f}_{k_n}(x) - f(x) \mid \leq \sup_{x\in I} \mid \hat{f}_{k_n}(x) - f_{k_n}(x) \mid + \sup_{x\in I} \mid f_{k_n}(x) - f(x) \mid \quad (10)$$

The second term in the upperbound is non-stochastic and tends to zero by Proposition(**??**), then it is sufficient to prove that

$$\sup_{x\in I} \mid \hat{f}_{k_n}(x) - f_{k_n}(x) \mid \xrightarrow{P} 0 \quad \text{as} \quad n \longrightarrow +\infty.$$

For $x \in I$ we have

$$\begin{aligned} \mid \hat{f}_{k_n}(x) - f_{k_n}(x) \mid &= \left| (k_n+1) \sum_{j=0}^{k} \left(F_n(A_{k_nj}) - F(A_{k_nj})\right) p_{k_nj}(x) \right| \\ &\leq 2(k_n+1) \sup_{t\in[0,1]} |F_n(t) - F(t)|, \qquad (11) \end{aligned}$$

where

$$A_{k_nj} = \left(\frac{j}{k_n+1}, \frac{j+1}{k_n+1} \right] \qquad j = 0, \cdots, k.$$

Using the Dvoretzky-Kiefer-Wolfowitz(1956) and Massart(1990) inequality we get for all $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(\sup_{x\in I} \mid \hat{f}_{k_n}(x) - f_{k_n}(x) \mid > \varepsilon \right) &\leq \mathbb{P}\left(\sup_{x\in[0,1]} \mid F_n(x) - F(x) \mid > \frac{\varepsilon}{2(k_n+1)} \right) \\ &\leq 2 \exp\left(-\frac{1}{2}\epsilon^2 \, n(k_n+1)^{-2} \right). \end{aligned}$$

That concludes the proof of the proposition.

■

We arrive now at the most important results of the paper wich prove the performance of Bernstein estimator for unbounded density function. In fact, in the first step we state the uniform strong consistency on each compact I in $(0,1)$ under two conditions on the bandwidth parameter, i.e the Bernstein estimator has the same asymptotic behavior as the underlying density at the interior points. This is done in Theorem(2.1). In the second step, we prove the almost convergence to infinity of Bernstein estimator at $x = 0$.

Theorem 2.1 *(Uniform Strong Consistency) Assume that $X_1, \cdots, X_n$ are i.i.d observations with a continuous distribution function F, having a density function, $f \in C((0,1))$. Let I be a compact in $(0,1)$.*

Under the two conditions:

$$\lim_n k_n = +\infty \qquad \textit{and} \qquad \lim_n \frac{k_n^2 \log n}{n} = 0.$$

we have

$$\sup_{x\in I} \left| \hat{f}_{k_n}(x) - f(x) \right| \xrightarrow{a.s} 0 \quad \textit{as} \quad n \longrightarrow +\infty.$$

■

Proof of Theorem(2.1)

Indeed, from Proposition(**??**), it suffices to prove

$$\sup_{x\in I} \mid \ \hat{f}_{k_n}(x) - f_{k_n}(x) \ \mid \xrightarrow{a.s} 0 \quad \text{as} \quad n \longrightarrow +\infty.$$

In fact, from inequality(11), for $\epsilon_n > 0$

$$I\!\!P \left(\sup_{x\in I} \mid \ \hat{f}_{k_n}(x) - f_{k_n}(x) \ \mid > \epsilon_n \right) \ \leq \ 2 \exp\left(-\frac{1}{2} \epsilon_n^2 n (k_n+1)^{-2} \right)$$

Let $\epsilon_n = \alpha(k_n+1) \left(\dfrac{\log n}{n} \right)^{1/2}$ $(\alpha > \sqrt{2})$ Then

$$\sum_{n\geq 1} \exp\left(-\frac{1}{2} \epsilon_n^2 n (k_n+1)^{-2} \right) < \infty$$

Therefore, from Borel-Cantelli lemma we get

$$\limsup_n \frac{1}{\epsilon_n} \sup_{x\in I} \left| \hat{f}_{k_n}(x) - f_{k_n}(x) \right| \leq 1 \quad a.s. \tag{12}$$

which concludes the proof of Theorem(2.1).

■

Note that the inequality(12) is true for $I \in [0,1]$ even for unbounded density function. The most interesting feature of Bernstein estimator $\hat{f}_k$ is its almost sure convergence to $f(0)$ even if the density function f is unbounded at $x = 0$. We show this in the following theorem.

Theorem 2.2 *Assume that $X_1, \cdots, X_n$ are i.i.d observations with a continuous distribution function F, having a density function $f \in C((0,1))$ and unbounded at $x = 0$. Under the two conditions in Theorem(2.1) we get*

$$\hat{f}_{k_n}(0) \xrightarrow{a.s} +\infty \quad as \quad n \longrightarrow +\infty.$$

■

Proof of Theorem(2.2)

Indeed, from the proof of Theorem(2.1) we have for $\alpha > \sqrt{2}$,

$$\limsup_n \frac{\left|\hat{f}_{k_n}(0) - f_{k_n}(0)\right|}{(k_n+1)\left(\dfrac{\log n}{n}\right)^{1/2}} \leq \alpha \quad a.s$$

Then it remains to prove that

$$f_{k_n}(0) \longrightarrow +\infty \quad \text{as} \quad k_n \longrightarrow \infty. \tag{13}$$

In fact, for $C > 0$, there exists $\delta(C)$, such that,

$$\text{for} \quad |\, x \,| < \delta(C), \quad \text{we have} \quad f(x) > C$$

Then,

$$\begin{aligned} f_k(0) &= (k_n+1)\int_0^{1/(k_n+1)} f(t)\,dt \\ &> C, \end{aligned} \tag{14}$$

if

$$\frac{1}{k_n+1} < \delta(C).$$

That concludes the proof.

■

Remark 1 *The distribution function F, for $K = [0,1]$, can be estimated by:*

$$\begin{aligned}\hat{F}_k(x) &= \int_0^x \hat{f}_k(t)\,dt \\ &= \sum_{i=0}^{k+1} F_n(\frac{i}{k+1}) p_{(k+1)i}(x).\end{aligned}$$

Where $p_{(k+1)i}(x)$ is a Binomial distribution with parameter $(k+1)$ and x.

■

The estimator $\hat{F}_k(x)$ is an unbiased estimator of the Bernstein polynomial of order $(k+1)$ of the distribution function F. In Figure(2), the distribution function beta(0.7,6) and the Bernstein estimator $\hat{F}_k(x)$ with $k = 45$ are plotted. The data are the same as in Figure(1).

3 Bandwidth Selection

Nonparametric density estimation requires the specification of the smoothing parameter. Much theoretical and practical studies have been carried out for choosing the optimal smoothing parameter for the standard kernel estimators.
For Bernstein polynomial estimators, with a large value of the bandwidth parameter k, we undersmooth the density and for a small value of k we oversmooth the density and the Bernstein estimator becomes close to the uniform on $[0,1]$, as shown in Figure (3), with a sample of size equal to 200 from the mixture of two beta densities:

$$f^*(x) = 0.4\,B(0.7,6)(x) + \ 0.6\,B(4,2)(x).$$

Then the choice of the optimal bandwidth is very important. In this section, we use the least-squares cross-validation method for f in L_2 and the likelihood cross-validation method for any density f, already used for the Kernel estimator, to choose the smoothing parameter for the Bernstein estimator. When the density function converges very fast to infinity at $x = 0$, the behaviors of the L_1 error and the optimal bandwidth parameter of the Bernstein density estimator are shown in this section.

For this section we use the following notations:

$$m_{kj}^{(-i)} = \#\,\{s,\ X_i \neq X_s \ \text{ and } \ X_s \in A_{kj}\}$$

and

$$\hat{f}_k^{(-i)}(x) = \frac{1}{n-1} \sum_{j=0}^{k} m_{kj}^{(-i)} \, B(j+1, k-j+1)(x) \tag{15}$$

which is the Bernstein estimator constructed from all data points except X_i.

3.1 Least-Squares Cross-Validation Method

The method was suggested by Rudemo(1982) and Bowman(1984). For the class of density function f in L_2, an optimal bandwidth is the integer k which minimizes the mean integrated squared error MISE(k):

$$\begin{aligned} MISE(k) &= \mathbb{E}\left\{\int_0^1 \left(\hat{f}_k(x) - f(x)\right)^2 dx\right\} \\ &= \mathbb{E}\left(\int_0^1 \hat{f}_k^2\,(x)\,dx\right) - 2\mathbb{E}\left(\int_0^1 \hat{f}_k(x)f(x)\,dx\right) + \int_0^1 f^2(x)\,dx. \end{aligned}$$

The last term is unknown but does not depend on k, hence minimizing the mean integrated squared error is equivalent to minimizing the following quantity

$$R(\hat{f}_k) = \mathbb{E}\left\{\int_0^1 \hat{f}_k^2\,(x)\,dx \; - 2\int_0^1 \hat{f}_k(x)f(x)\,dx\right\}.$$

The idea of the least-squares cross-validation is to construct an estimator of $R(\hat{f}_k)$ from the data themselves and then to minimize this estimate over k to get the optimal choice of the bandwidth parameter. An estimator of this quantity is as follows

$$LSCV(k) = \int_0^1 \hat{f}_k^2\,(x)\,dx \; - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_k^{(-i)}(X_i)$$

The optimal bandwidth parameter by the least-squares cross-validation method is

$$k_{LSCV(k)} = argmin_k \, LSCV(k).$$

$LSCV(k)$ is an **unbiased** estimator of $R(\hat{f}_k)$. To prove this, it is sufficient to prove that

$$\mathbb{E}\left(\hat{f}_k^{(-i)}(X_i)\right) = \mathbb{E}\left(\int_0^1 \hat{f}_k(t)f(t)dt\right).$$

In fact,

$$\begin{aligned}
\mathbb{E}\left(\hat{f}_k^{(-i)}(X_i)\right) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{j=0}^{k} m_{kj}^{(-i)}\,\beta(X_i, j+1, k-j+1)\right) \\
&= \sum_{j=0}^{k} \mathbb{E}\left(\frac{m_{kj}^{(-i)}}{n-1}\right) \mathbb{E}(\beta(X_i, j+1, k-j+1)) \\
&= \sum_{j=0}^{k} F(A_{kj}) \int_0^1 \beta(t, j+1, k-j+1) f(t) dt \\
&= \int_0^1 \mathbb{E}(\hat{f}_k(t)) f(t) dt \\
&= \mathbb{E}\left(\int_0^1 \hat{f}_k(t) f(t) dt\right).
\end{aligned}$$

■

After some computations, it can be proved that

$$LSCV(k) = \frac{1}{n^2}\sum_{i=0}^{k}\sum_{j=0}^{k} m_{ki} m_{kj} d_{ij} - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_k^{(-i)}(X_i),$$

where

$$d_{ij} = \frac{\beta(i+j+1, 2k-i-j+1)}{\beta(i+1, k-i+1)\beta(j+1, k-j+1)}$$

and,

$$\beta(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}.$$

This formula is very helpful for practical simulations.

3.2 Likelihood cross-validation Method

The following method of selection of the optimal bandwidth parameter is for any density function defined on $[0,1]$. The optimal choice of the bandwidth parameter by the likelihood cross-validation method is, for the given data, the integer which maximizes the following quantity:

$$LCV(k) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\hat{f}_k^{(-i)}(X_i)\right) \tag{16}$$

where $\hat{f}_k^{(-i)}$ is as defined in (14) above. The likelihood cross-validation quantity $LCV(k)$ was introduced by Habbema, Hermans and van Der Broeck (1974) and Duin (1976). For more detail of the likelihood cross-validation method see Silverman(1986).

Remark 2 *An explicit form of the optimal bandwidth parameter of the least squares cross-validation and likelihood cross-validation method does not exist. A numerical solution is necessary for the choice of the optimal bandwidth parameter.*

For the simulations we take the same data of size $n = 200$, in Figure (3), drawn from the mixture of beta densities

$$f^*(x) = 0.4\, B(0.7, 6)(x) + \ 0.6\, B(4, 2)(x)$$

The least-squares cross-validation quantity $LSCV$ is plotted in Figure (4) and the likelihood cross-validation quantity LCV is plotted in Figure (5). The integer k which minimizes least squares cross-validation $LSCV$ quantity is $k_{LSCV} = 33$ and the integer which maximizes the likelihood cross-validation quantity LCV is $k_{LCV} = 27$. The least squares cross-validation $LSCV$ quantity is more variable in comparison with the likelihood cross-validation quantity LCV. In Figure (6), the true density f^* and the Bernstein estimator for f^* with the k_{LSCV} and k_{LCV} are plotted. From Figure (6), we can conclude that the Bernstein estimator is quite the same for the two values of k, k_{LSCV} and k_{LCV}.

3.3 Simulation Results

In this section we present results of a simulation designed to show the influence of the bandwidth parameter on the performance in $L_1[0, 1]$ of the Bernstein density estimator. In particular for an unbounded density function, the simulation gives an idea on the behavior of the optimal bandwidth parameter when the density function converge very fast to infinity at $x = 0$.

A mixture of beta density functions is considered:

$$f(x) = 0.4B(\alpha, 6)(x) + 0.6B(4, 2)(x)$$

We choose four value for α: $\alpha = 3$ (bounded density function), $\alpha = 0.7$, $\alpha = 0.5$ and $\alpha = 0.3$ (unbounded density functions). Note that the density

function f with $\alpha = 0.3$ converges faster to infinity than with $\alpha = 0.7$. We take different values of the bandwidth parameter k for the Bernstein density estimator: $k = 10, 20, 25, 30, 35, 50, 75, 100, 125, 150, 175, 200, 250$.
For each value of α and the bandwidth parameter k, $N = 500$ random samples of size $n = 200$ were generated from the density function f. For each simulated sample we calculate the integrate absolute deviation, called IAD, for the Bernstein estimator:

$$IAD = \int_0^1 \mid \hat{f}_k(t) - f(t) \mid \, dt.$$

As a measure of performance for the Bernstein density estimator the mean integrated absolute deviations is calculated. The results of the simulation are in Table (1). The first remark is, that for each fixed value of the bandwidth parameter k, the error in L_1 of Bernstein density estimator increases when the parameter α decreases. The second remark is, that for each fixed value of α,the error in L_1 of Bernstein density estimator is almost the same in the neighborhood of the optimal bandwidth parameter k. The last and the more important remark is that the optimal bandwidth parameter k increases when the parameter α decreases, i.e., when the density function converges faster to infinity.

References

Bernstein, S. (1912). *Démonstration du théorème de Weierstrass, fondée sur le calcul des probabilités.* Commun. Soc. Math. Kharkow **(2)**: 1-2.

Bouezmarni, T. and Rolin, J.M. (2001). *Consistency of Beta kernel Density Function Estimator.* Institut de Statistique, Université Catholique de Louvain, *Discussion paper* # 0108. <http://www.stat.ucl.ac.be>.

Brown, B.M and Chen, S.X (1999). *Beta-Bernstein Smoothing for Regression Curves with Compact Supports.* Scand. J. Statist, **26**: 47-59.

Chen, S.X (1999). *A beta Kernel estimation for density functions* . Computational Statistics and Data analysis. **31**: 131-145.

Duin, R.P.W (1976). *On the choice of the smoothing parameters for Parzen estimators of probability density functions.* IEEE Trans. Comput., **C-25**: 1175-1179.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J (1956). *Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator.* The Annals of Mathematical Statistics, **1**: 444-453.

Grawronski, N, Stadtmüller, U. (1981). *Smoothing histograms by lattice and continuous distributions.* Metrika **28**: 155-164.

Grawronski, N, Stadtmüller, U. (1984). *Lineair Combinations of Iterated Generalised Bernstein Function with an Application to Density Estimation.* Acta Sci. Math. **47**: 205-221.

Lorentz, GG. (1986). *Bernstein Polynomials.* Chelsea Publishing Company, New York.

Massart, P. (1990). *The Tight Constant in Dvoretzky-Kiefer-Wolfowitz Inequality.* The Annals of Probability, **18**: 1269-1283.

Rice. J. A.(1984). *Boundary Modification for Kernel Regression.* Commun. Statist. Theor. Meth. , **13**: 893-900.

Rosenblatt, M. (1956). *Remarks on Some Nonparametric Estimates of a Density Function.* Ann. Math. Statist. **27**: 832-837.

Scheffé. H. (1947). *A useful convergence theorem for probability distributions.* Ann. of Math. Statist., **18**: 434-458.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stadtmüller, U. (1983). *Asymptotic Distribution of Smoothed Histograms.* Metrika **30**: 145-158.

Tenbusch, A. (1994). *Two-Dimentional Bernstein Polynomial Density Estimation.* Metrika **41**: 233-253.

Tenbusch, A. (1998). *Nonparametric Curve Estimation Whith Bernstein Polynomial.* Universitätsverlag Rasch, Osnabrück.

Vitale, R.A. (1973). *A Bernstein Polynomial Approach to Density Estimation.* Comm Stat **2**: 493-506.

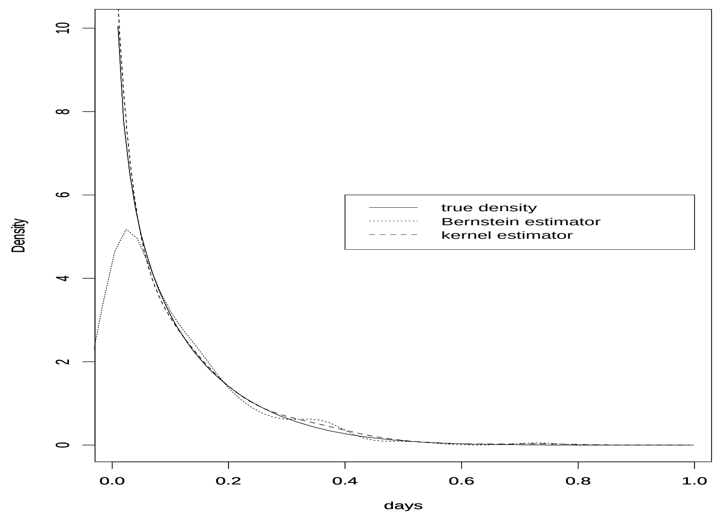

Figure 1: The Bernstein and the kernel estimator of density $f(x) = \beta(0.7, 6)$ (n = 200)

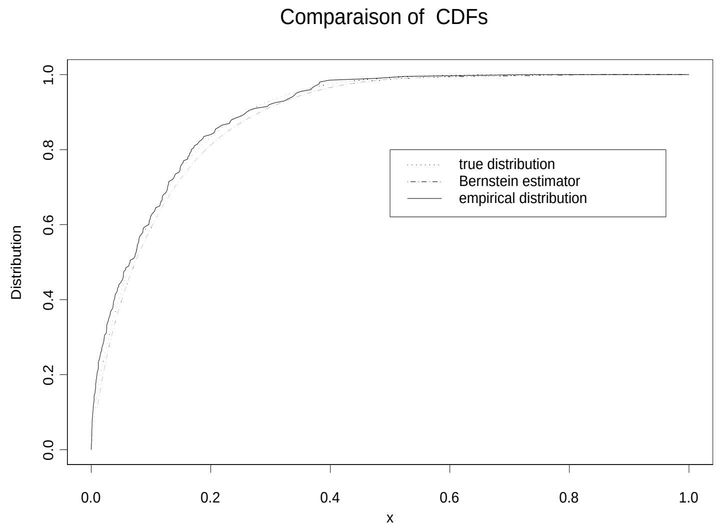

Figure 2: The true distribution and the Bernstein polynomial estimator for the distribution function $\beta(0.7, 6)(n = 200)$

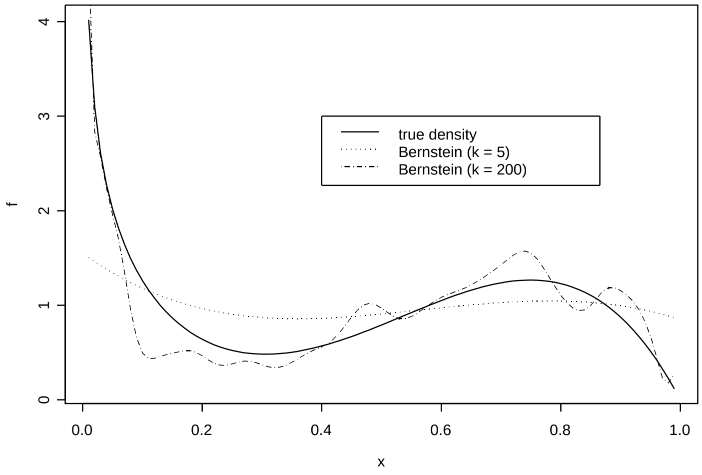

Figure 3: The Bernstein estimator of the density $f^*(x) = 0.4\,\beta(0.7, 6) + \; 0.6\,\beta(4, 2)$ (n = 200)

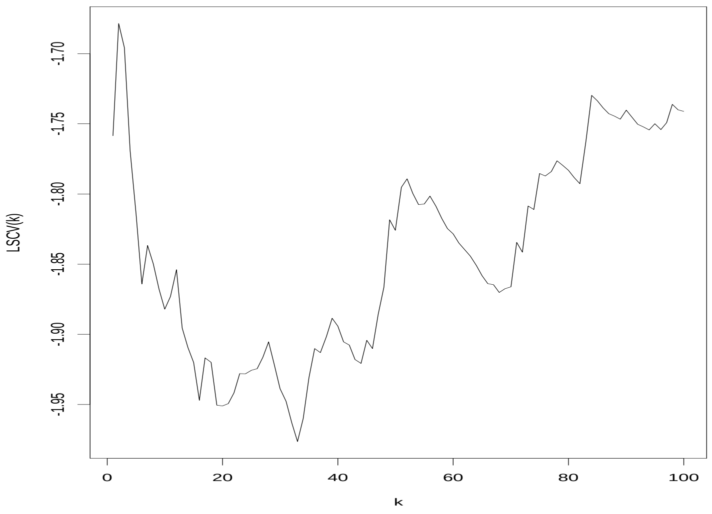

Figure 4: The least-squares cross-validation quantity LSCV(k). The data are from $f^*(x) = 0.4\,\beta(0.7, 6) + \; 0.6\,\beta(4, 2)$ (n = 200)

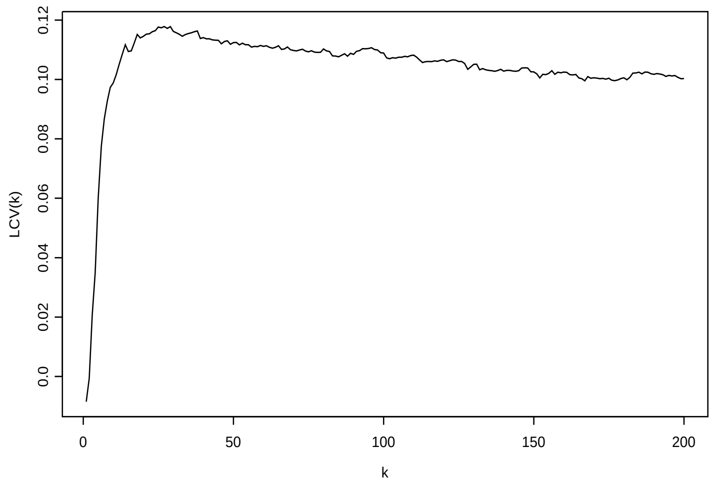

Figure 5: The likelihood cross-validation quantity LCV(k). The data are from $f^*(x) = 0.4\,\beta(0.7, 6) + \; 0.6\,\beta(4, 2)$ (n = 200)

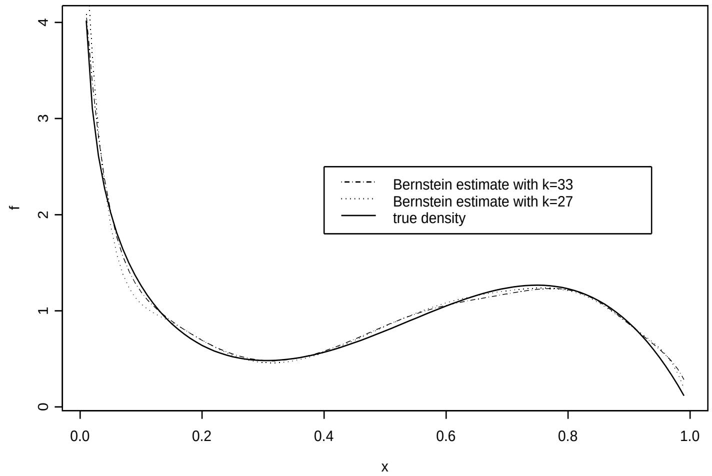

Figure 6: The Bernstein polynomial estimator with the $k_{\mathrm{LSCV}} = 33$ and the $k_{\mathrm{LCV}} = 27$ for the density $f^*x) = 0.4\,\beta(0.7, 6) + \; 0.6\,\beta(4, 2)$ (n = 200)

	$\alpha = 3$	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.3$
$k = 10$	0.1214938	0.2239964	0.2639941	0.4210704
$k = 20$	$\Rightarrow$ <u>0.1007924</u>	0.1574274	0.1804325	0.3345240
$k = 25$	0.1030814	0.1499197	0.1702740	0.3159052
$k = 30$	0.1039102	0.1444219	0.1647330	0.3025242
$k = 35$	0.1052087	$\Rightarrow$ <u>0.1414991</u>	0.1638277	0.2946617
$k = 50$	0.113193	0.1444268	$\Rightarrow$ <u>0.1614005</u>	0.2807913
$k = 75$	0.1282887	0.1536246	0.1685675	0.2704440
$k = 100$	0.1379182	0.1617323	0.1729462	$\Rightarrow$ <u>0.2664806</u>
$k = 125$	0.1542741	0.1695478	0.1811452	0.2671203
$k = 150$	0.1553925	0.1773798	0.1892006	0.2681968
$k = 175$	0.1625965	0.1838027	0.1925876	0.2697780
$k = 200$	0.1697385	0.189443	0.1989246	0.2717065
$k = 250$	0.1795354	0.1991608	0.2084664	0.2750491

Table 1: The L_1 distance between the density function $0.4\beta(\alpha, 6) + 0.6\beta(4, 2)$ and the Bernstein estimator.