

ON A ROLE OF A SCALE PARAMETER WHEN USING THE GAUSSIAN KERNEL DENSITY ESTIMATOR

Takeuchi, Hiroyuki
Department of Economics, Tokyo International University

<https://doi.org/10.5109/13450>

出版情報 : Bulletin of informatics and cybernetics. 28 (1), pp.1-13, 1996-03. Research
Association of Statistical Sciences
バージョン :
権利関係 :



ON A ROLE OF A SCALE PARAMETER WHEN USING THE GAUSSIAN KERNEL DENSITY ESTIMATOR

By

Hiroyuki TAKEUCHI *

Abstract

It seems that the curve of the Gaussian kernel density estimate is often flat as compared with that of true p.d.f., although its window width has been chosen suitably. This flatness sometimes causes a bad kernel estimate of density. In this paper we shall introduce a scale parameter into true p.d.f. to show the flatness of the curve of Gaussian kernel density estimate. We study the asymptotic properties of the scale parameter to propose a modified kernel density estimator. Some simulations show a superiority of the estimator.

1. Introduction

In kernel density estimation, estimate of window width has received much attention by many investigators. See for example, Park and Marron [8] or Jones and Kappenman [4]. They have concentrated on studying window width to reduce mean integrated squared error (*MISE*). However in recent years, these papers which deal with not only window width but also a transformation of the data appeared. It is called a data-transformed density estimator. Wand *et al.* [13] proposed, so called, *back-transform* method and Park *et al.* [7] suggested a modification of the shifted power transformation in Wand *et al.* [13]. Ruppert and Wand [9] applied the method to correct for kurtosis before using a global window width to the kernel method. The back-transform method has been succeeded in a sense however, generally speaking, these data transformation methods have critical problems. They are, mainly, “when we should transform the data” and “how we estimate the parameters in the transform function”. See the comments in Wand *et al.* [13].

In this paper we shall consider a little passive but essential scale transformation of random variables. Let X_1, X_2, \dots, X_n be a sample from a distribution which has a p.d.f. $f(x)$. And let $K(y)$ and h be a kernel and a window width, respectively. The kernel density estimator is written in

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^n K_h(x - X_j) \quad (1.1)$$

* Department of Economics, Tokyo International University, 1-13-1 Matobakita, Kawagoe, Saitama 350-11, Japan

where $K_h(y) = h^{-1}K(y/h)$. The characteristic function of $f(x)$ is denoted by $\varphi_f(t)$, and let $c_n(t) = n^{-1} \sum_{j=1}^n e^{itX_j}$ be the empirical characteristic function (e.c.f.) based on the X_j 's. On the analogy of Fourier analysis, we shall use terms time and frequency domain as the domain of p.d.f. and of its characteristic function, respectively. Suppose that the kernel satisfies $K(\cdot) \in L^1$ then the characteristic function of (1.1) is

$$\varphi_{\hat{f}_n}(t) = \varphi_K(ht)c_n(t) \quad (1.2)$$

In general, $|\varphi_K(ht)|$ converges to 1 from below for each $t \in R$, as $h \downarrow 0$. We have $\varphi_K(0) = 1$, and $\lim_{|t| \rightarrow \infty} \varphi_K(ht) = 0$ for each $h > 0$ by Riemann-Lebesgue's lemma. And since from Feuererger and Mureika [2]

$$\lim_{n \rightarrow \infty} \sup_{|t| \leq T} |c_n(t) - \varphi_f(t)| = 0, \quad a.s.$$

for any fixed $T > 0$, we may say that the characteristic function of $K_h(y)$, that is the $\varphi_K(ht)$ in (1.2), is diminishing the tail of the e.c.f. That is to say $\varphi_K(ht)$ is reducing the high frequency component of $c_n(t)$. Therefore in the physical sense, window width controls the elimination of the noise in the e.c.f. But it is doubtful whether the control, by the window width alone, is sufficient or not. If the reducing of the high-frequency component of e.c.f. is too much in the frequency domain, the curve of kernel density estimate will be flat as compared with that of true p.d.f. in the time domain. To investigate this matter, under the conditions of $\hat{f}_n(\cdot), f(\cdot) \in L^2$, we define $ISE_n(b)$ as

$$ISE_n(b) \equiv \int \left| \hat{f}_n(x) - \frac{1}{b} f\left(\frac{x}{b}\right) \right|^2 dx = \frac{1}{2\pi} \int \left| \varphi_{\hat{f}_n}(t) - \varphi_f(bt) \right|^2 dt. \quad (1.3)$$

The right hand side of (1.3) is justified by Parseval's formula, see Lukacs [5]. Given a set of data, let b_n minimize (1.3) with respect to b and let h_{opt} be a minimizer of the $ISE_n(b = 1)$. Since the results of the paper shows that b_n with nonzero probability, it is concluded that there is a flatness in kernel estimators with constant window width. And this is caused by that the reduction of the high frequency component by the $\varphi_K(h_{opt}t)$ is too much in (1.2).

If the conjecture described above is justified, we should modify the (1.2) by introducing a scale parameter s as $\varphi_{\hat{f}_n(\cdot; s)}(t) = \varphi_K(ht)c_n(t/s)$ to pass the high frequency component of the e.c.f. more. In this case the kernel density estimator is written in

$$\hat{f}_n(x; s) = \frac{1}{n} \sum_{j=1}^n K_h(x - X_j/s). \quad (1.4)$$

Silverman and Young [11] proposed to use a shrunk kernel estimate for the smoothed bootstrap. It is easy to verify that their estimate is asymptotically equivalent to (1.4) with $s = \sqrt{1 + h^2}$. (1.4) will be effective with respect to avoid the flatness and it may contribute to lowering the integrated squared error. Hereafter we call the scale parameter s as the *shrink parameter* to distinguish it from the parameter b . As described above,

the shrink parameter s plays a role that it controls the high frequency component of the e.c.f. that the window width can not have controled well.

In what follows, we shall illustrate this more concretely by an example. When $f(x)$ is the p.d.f. of the standard normal and $K(y)$ is the Gaussian kernel, Fryer [3] calculated the $MISE$ for this case. Here the $MISE(b)$, the expectation of the (1.3), for $\hat{f}_n(x)$ with respect to the $b^{-1}f(x/b)$ is shown to be

$$2\sqrt{\pi}MISE(b) = \frac{1}{nh} + \left(1 - \frac{1}{n}\right) \frac{1}{\sqrt{\sigma^2 + h^2}} - \frac{2\sqrt{2}}{\sqrt{(b^2 + 1)\sigma^2 + h^2}} + \frac{1}{\sigma b}.$$

It is easy to check that the minimizer b_n of this $MISE(b)$ is $b_n = \sqrt{1 + (h/\sigma)^2}$ and that $MISE(b_n) < MISE(1)$. These facts tell us two points. One is that the case we don't consider the scale parameter, i.e. $b \equiv 1$, the curve of the Gaussian kernel density estimate $\hat{f}_n(x)$ is flat as compared with that of true p.d.f. in the $MISE$ sense. The other is that if we are to introduce the scale parameter b , we may set its value greater than 1. In fact our simulation in section 5 shows that the event $\{b_n > 1\}$ occurs very often for the standard normal, some of normal mixtures and double exponential distribution, in the integrated squared error sense. If we use a super-kernel, a symmetric and bounded continuous function $K(y)$ with satisfying $\int K(y)dy = 1$ and $\int y^r K(y)dy = 0$ ($r = 1, 2, \dots, k$) for some positive integer $k \geq 2$, we may not need the shrink parameter. But in practice one uses the Gaussian kernel very often, because it is easy to compute and the super-kernel may cause the estimate to take a negative value in its tails.

In section 2, some of mathematical assumptions are stated, and an invariant property of the scale parameter b is shown. This remarkable property will be of use when we propose an estimate of b_n which is the minimizer of the $ISE_n(b)$ defined in (1.3). We also show the asymptotic distribution of the b_n . In section 3, an existence of the b_n in the open interval $(1, \infty)$ is shown under some restricted conditions. The law of iterated logarithm of the minimizer b_n is proved in section 4, and we shall propose an estimate of the b_n by using this convergence rate. Finally in section 5, we conducted some simulations to show how our estimate $\hat{f}_n(x; s)$ reduces integrated squared error. Even if we employ a very simple and naive estimate \hat{s}_n , proposed from (5.1), it would be worth while to use $\hat{f}_n(x; \hat{s}_n)$ instead of the ordinary Gaussian kernel density estimator.

2. Assumptions and asymptotic distribution of the minimizer b_n

Here we assume that $Var X = \sigma^2$ where X has a p.d.f. $f_\sigma(x)$. The b_n is given as the minimizer of $ISE_n(b, h; \sigma)$:

$$ISE_n(b, h; \sigma) = \int \left| \frac{1}{n} \sum_{j=1}^n K_h(x - X_j) - \frac{1}{b} f_\sigma\left(\frac{x}{b}\right) \right|^2 dx \quad (2.1)$$

with respect to $b > 0$. Note that b_n is a random variable. If c is a positive constant then

$$ISE_n(b, ch; c\sigma) = \frac{1}{c} ISE_n(b, h; \sigma).$$

Thus the scale parameter b is invariant with respect to the constant multiplication of the data and the window width. Therefore we may propose an estimate of the b_n without considering the constant multiplication. (2.1) is also invariant for location shift of the data, however the integrated squared error between (1.4) and $f(\cdot)$ is not, if $s \neq 1$. So in the practical use of the shrink parameter, it is important to normalize the data.

Let (Ω, \mathcal{F}, P) be a probability space and X_1, X_2, \dots, X_n be i.i.d. random variables defined on it. And as described above, X_1 has a p.d.f. $f(x)$. We suppose that following (A.1) to (A.3).

$$(A.1) \quad f(\cdot) \in L^2, \quad E|X_1|^3 < \infty, \quad \int |\varphi_f(t)| dt < \infty, \quad \int |t\varphi'_f(t)| dt < \infty, \\ \int |t\varphi'_f(t)|^2 dt < \infty, \quad \int t^2 |\varphi''_f(t)| dt < \infty.$$

$$(A.2) \quad K(\cdot) \text{ is a continuous function on } R \text{ such that}$$

$$K(\cdot) \in L^2, \quad K(y) \geq 0, \quad K(-y) = K(y) \text{ for any } y \in R, \text{ and } \int K(y) dy = 1.$$

$$(A.3) \quad g(\cdot) \text{ is a real valued function on } R \text{ such that}$$

$$\exists m > 0 : g(u) > m \text{ for any } u \in R,$$

$$g'(\cdot) \text{ is continuous on } R, \quad \exists M > 0 : 0 < g'(u) < M \text{ for any } u \in R,$$

$$\text{and } g(0) = g'(0) = 1.$$

These conditions will be used everywhere in the proof without notice.

The scale parameter b must take positive value on R . Thus we write $b = g(\beta)$ by using the function $g(u)$ defined in (A.3) to deal with β ($-\infty < \beta < \infty$) instead of the b . An example of $g(u)$ is $g(u) = (2\eta/\pi) \arctan\{(\pi/2\eta)u\} + 1$, where η is a constant with $0 < \eta < 1$. Finally, we define $\beta_n = g^{-1}(b_n)$ where b_n is the minimizer of $ISE_n(b)$ in (1.3). Now we shall show the almost sure convergence of the b_n .

THEOREM 2.1. *The minimizer b_n converges to 1 with probability one.*

PROOF. Firstly we require to show the existence of the minimizer b_n . Let B_1 and B_2 be fixed numbers with satisfying $0 < B_1 < 1 < B_2$. The $ISE_n(b)$ defined by (1.3) is differentiable with respect to $b > 0$. Hence there exists a b_n such that

$$ISE_n(b_n) = \min_{B_1 \leq b \leq B_2} ISE_n(b) \quad (2.2)$$

for each $n \in N$ and $\omega \in \Omega$. Secondly, by Parseval's formula and dominated convergence theorem, we have

$$\lim_{n \rightarrow \infty} \{ISE_n(b) - ISE_n(1)\} = \frac{1}{2\pi} \int |\varphi_f(t) - \varphi_f(bt)|^2 dt, \quad a.s. \quad (2.3)$$

for each $b > 0$. Note that (2.3) takes 0 or a positive value if and only if $b = 1$ or $b \neq 1$, respectively with probability one. Finally, we assume that $\lim_{n \rightarrow \infty} b_n \neq 1$ with some positive probability. Then from the definition we have,

$$ISE_n(b_n) < ISE_n(1),$$

for infinitely many $b_n \neq 1$ with some positive probability. This contradicts (2.3) and completes the proof of the theorem.

From this theorem, we can say that the effectiveness of the scale parameter b will vanish when the sample size is large enough. And as in section 5, our simulation shows that the estimator $\hat{f}_n(x; s)$ reduces the integrated squared error more than the case $b \equiv 1$, even if sample size is small to moderately large. Then we show the asymptotic distribution of the b_n .

THEOREM 2.2. *If the window width satisfies (2.4)*

$$\int |t\varphi'_f(t)| |\varphi_K(ht) - 1| dt = o(n^{-\frac{1}{2}}) \quad (2.4)$$

for large n , then we have

$$\sqrt{n}(b_n - 1) \xrightarrow{\mathcal{D}} N(0, \sigma_f^2),$$

as $n \rightarrow \infty$, where σ_f^2 is

$$\sigma_f^2 = E \left| \operatorname{Re} \left[\int t \overline{\varphi'_f(t)} (e^{iX_1 t} - \varphi_f(t)) dt \right] \right|^2 / \left(\int |t\varphi'_f(t)|^2 dt \right)^2.$$

PROOF. Define a real valued function $\rho_n(\cdot, \cdot)$ on R^2 as follows

$$\begin{aligned} & \frac{d}{d\beta} ISE_n(g(\beta)) \\ &= -\frac{g'(\beta)}{n\pi} \sum_{j=1}^n \operatorname{Re} \left[\int t \overline{\varphi'_f(g(\beta)t)} (e^{iX_j t} \varphi_K(ht) - \varphi_f(g(\beta)t)) dt \right] \\ &= -\frac{g'(\beta)}{n\pi} \sum_{j=1}^n \rho_n(X_j, g(\beta)). \end{aligned}$$

The β_n is given as a solution of the equation $\sum_{j=1}^n \rho_n(X_j, g(\beta)) = 0$, with respect to β .

Taylor expansion yields that

$$\begin{aligned} 0 &= \sum_{j=1}^n \rho_n(X_j, g(\beta_n)) \\ &= \sum_{j=1}^n \rho_n(X_j, 1) + \beta_n \sum_{j=1}^n \frac{d}{d\beta} \rho_n(X_j, g(\beta)) \Big|_{\beta=\theta\beta_n} \end{aligned}$$

for some $\theta \in (0, 1)$. So we get

$$\sqrt{n}\beta_n = -\frac{1}{\sqrt{n}} \sum_{j=1}^n \rho_n(X_j, 1) / \frac{1}{n} \sum_{j=1}^n \frac{d}{d\beta} \rho_n(X_j, g(\beta)) \Big|_{\beta=\theta\beta_n} \quad (2.5)$$

It is easy to check that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \rho_n(X_j, 1) \xrightarrow{\mathcal{D}} N\left(0, E \left| \operatorname{Re} \left[\int t \overline{\varphi'_f(t)} (e^{iX_1 t} - \varphi_f(t)) dt \right] \right|^2\right) \quad (2.6)$$

as $n \rightarrow \infty$. And since β_n converges to 0 almost surely by Theorem 2.1., we have

$$\left. \frac{1}{n} \sum_{j=1}^n \frac{d}{d\beta} \rho_n(X_j, g(\beta)) \right|_{\beta=\theta\beta_n} \longrightarrow \int |t\varphi'_f(t)|^2 dt, \quad a.s. \quad (2.7)$$

as $n \rightarrow \infty$. Together with (2.5), (2.6) and (2.7) we have $\sqrt{n}g^{-1}(b_n) \xrightarrow{\mathcal{D}} N(0, \sigma_f^2)$ by Slutsky's theorem. Finally by the theorem 3.1.A in Serfling [10], we get the conclusion.

The condition (2.4) may seem to be technical, but we can show that it is not so strong, as follows. We assume that the kernel is Gaussian and that the derivative of the characteristic function of true p.d.f. satisfies $|\varphi'_f(t)| \leq Ae^{-\tau|t|}$ for some $A, \tau > 0$ that are independent of t . If the order of the window width is $h_n = o(n^{-\frac{1}{4}-\epsilon})$, then (2.4) is satisfied for any $\epsilon > 0$. See Takeuchi [12] for the proof. We should remark that the asymptotic variance of $\sqrt{n}(b_n - 1)$ does not depend on the kernel and it only depends on true p.d.f., in other words, the asymptotic variance is inherent in the underlying density function $f(x)$.

3. Existence of the scale parameter in the open interval $(1, \infty)$

In section 1, we conjectured that the scale parameter b_n which minimizes (1.3) would be greater than 1 with large probability. We shall state Proposition 3.1. to estimate this probability. Here we use the notation of e.c.f. as $c_n(t, \omega)$, where ω denotes an element of the set Ω , to distinguish that it is a random function.

PROPOSITION 3.1. *The probability that there exists a $b > 1$ such that*

$$ISE_n(b) < ISE_n(1)$$

is not less than

$$P \left\{ \omega : \operatorname{Re} \left[\int t\varphi'_f(-t) (\varphi_f(t) - \varphi_K(ht)c_n(t, \omega)) dt \right] > 0 \right\}, \quad (3.1)$$

for each $n \in N$.

PROOF. We set $DISE_n(b) = ISE_n(1) - ISE_n(b)$, where $b > 0$. Since $DISE_n(1) = 0$, a.s. and $DISE_n(b)$ is differentiable about b , it is immediately that

$$\left\{ \omega : \left. \frac{d}{db} DISE_n(b) \right|_{b=1} > 0 \right\} \subset \{ \omega : \exists b > 1, DISE_n(b) > 0 \}$$

However by the interchange of differentiation and integration, we have

$$\left. \frac{d}{db} DISE_n(b) \right|_{b=1} = \frac{1}{\pi} \operatorname{Re} \left[\int t\varphi'_f(-t) (\varphi_f(t) - \varphi_K(ht)c_n(t, \omega)) dt \right]$$

This completes the proof.

It would be not easy to calculate (3.1) exactly for finite n , however we can evaluate its lower bound under some restricted conditions for $f(x)$. We shall use the following notation for the corollary below.

$$I_{\delta,T} = 2 \int_{|t|>T} |t\varphi'_f(t)| dt / \int_{\delta<|t|\leq T} |t\varphi'_f(t)| dt$$

And define these sets $A_n(\delta, T)$, $B_n(\delta) \in \mathcal{F}$ as follows.

$$\begin{aligned} A_n(\delta, T) &= \left\{ \omega : \inf_{\delta<|t|\leq T} (\varphi_f(t) - \varphi_K(ht) \operatorname{Re}[c_n(t, \omega)]) > I_{\delta,T} \right\} \\ B_n(\delta) &= \left\{ \omega : \inf_{|t|\leq \delta} (\varphi_f(t) - \varphi_K(ht) \operatorname{Re}[c_n(t, \omega)]) \geq 0 \right\} \end{aligned}$$

COROLLARY 3.2. *Suppose (A.1) holds and that*

$$f(-x) = f(x) \quad \text{and} \quad \varphi'_f(t) \leq 0, \quad \text{for} \quad t \geq 0.$$

Then (3.1) is not less than

$$\sup_{0<\delta<T<\infty} P \{A_n(\delta, T) \cap B_n(\delta)\} \quad (3.2)$$

for each $n \in N$.

PROOF. Write

$$\begin{aligned} &\operatorname{Re} \left[\int t\varphi'_f(-t) \varphi_K(ht) c_n(t, \omega) dt \right] \\ &= \int_{|t|\leq \delta} + \int_{\delta<|t|\leq T} + \int_{|t|>T} t\varphi'_f(-t) \varphi_K(ht) \operatorname{Re}[c_n(t, \omega)] dt \\ &= I_1 + I_2 + I_3 \end{aligned}$$

say, where $0 < \delta < T < \infty$. If $\omega \in B_n(\delta)$ then

$$I_1 \leq \int_{|t|\leq \delta} t\varphi'_f(-t) \varphi_f(t) dt. \quad (3.3)$$

For $\omega \in A_n(\delta, T)$, we have

$$\begin{aligned} I_2 &< \int_{\delta<|t|\leq T} t\varphi'_f(-t) (\varphi_f(t) - I_{\delta,T}) dt \\ &= \int_{\delta<|t|\leq T} t\varphi'_f(-t) \varphi_f(t) dt - 2 \int_{|t|>T} |t\varphi'_f(t)| dt. \end{aligned} \quad (3.4)$$

And with probability one, we get

$$I_3 \leq \int_{|t|>T} t\varphi'_f(-t) \varphi_f(t) dt + 2 \int_{|t|>T} |t\varphi'_f(t)| dt. \quad (3.5)$$

Together with (3.3), (3.4) and (3.5), we have

$$\operatorname{Re} \left[\int t \varphi'_f(-t) \varphi_K(ht) c_n(t, \omega) dt \right] < \int t \varphi'_f(-t) \varphi_f(t) dt,$$

that is

$$\left. \frac{d}{db} \operatorname{DISE}_n(b) \right|_{b=1} > 0,$$

with probability

$$P \{A_n(\delta, T) \cap B_n(\delta)\}.$$

Finally, it is obvious that from the definition of $A_n(\delta, T)$ and $B_n(\delta)$, the value of (3.1) is not less than supremum of this probability. This completes the proof of the corollary.

So far the author has been unable to find explicit estimate of the (3.2). But from the definition of the sets $A_n(\delta, T)$ and $B_n(\delta)$, $\varphi_K(ht)$ should be bounded away from 1 for each $t \in R$ to make the probability $P \{A_n(\delta, T) \cap B_n(\delta)\}$ large. Thus we may conjecture that the probability of the event $\operatorname{ISE}_n(b) < \operatorname{ISE}_n(1)$ with $b > 1$ will be large when the window width is comparatively large, i.e. when the sample size is not so large. This agrees with our conjecture in section 2.

4. Estimate of the minimizer b_n

In this section we shall show the law of iterated logarithm of the minimizer b_n . This leads to an estimate of b_n which will be used in the simulations. Firstly we state a lemma without proof for it is obvious from law of iterated logarithm.

LEMMA 4.1. *Let $\{Y_i, Y_{i,n} : i, n \in N\}$ be a sequence of random variables defined on the (Ω, \mathcal{F}, P) . Suppose that*

- (i) Y_1, Y_2, \dots are i.i.d. random variables,
- (ii) $EY_1 = 0, \operatorname{Var} Y_1 = \sigma^2 < \infty$,
- (iii) $\max_{1 \leq i \leq n} |Y_{i,n} - Y_i| = o \left(\left(\frac{1}{n} \log \log n \right)^{\frac{1}{2}} \right), \quad a.s.$

Then we have

$$\limsup_{n \rightarrow \infty} \frac{1}{\sqrt{2n \log \log n}} \sum_{i=1}^n Y_{i,n} = \sigma, \quad a.s.$$

THEOREM 4.2. *If the window width satisfies the following*

$$\int |t \varphi'_f(t)| |\varphi_K(ht) - 1| dt = o \left(\left(\frac{1}{n} \log \log n \right)^{\frac{1}{2}} \right),$$

then we have

$$\limsup_{n \rightarrow \infty} \left(\frac{n}{\log \log n} \right)^{\frac{1}{2}} g^{-1}(b_n) = \sqrt{2} \sigma_f, \quad a.s., \quad (4.1)$$

where σ_f is defined in Theorem 2.2.

PROOF. Let random variables Y_j and $Y_{j,n}$ be

$$\begin{aligned} Y_j &= \operatorname{Re} \left[\int t \overline{\varphi'_f(t)} (e^{iX_j t} - \varphi_f(t)) dt \right], \\ Y_{j,n} &= \rho_n(X_j, 1), \end{aligned}$$

respectively, where the function $\rho_n(\cdot, \cdot)$ is defined in the proof of Theorem 2.2. We shall show that the sequence of the r.v.'s $\{Y_j, Y_{j,n} : j, n \in N\}$ satisfy these conditions (i), (ii) and (iii) of Lemma 4.1. (i) is obvious and it is easy to check that $EY_1 = 0$, and $\operatorname{Var} Y_1 \equiv s^2 < \infty$. Then by the condition, (iii) is shown in the following way.

$$\begin{aligned} \max_{1 \leq j \leq n} |Y_{j,n} - Y_j| &= \max_{1 \leq j \leq n} \left| \operatorname{Re} \left[\int t \overline{\varphi'_f(t)} e^{iX_j t} (\varphi_K(ht) - 1) dt \right] \right| \\ &\leq \int |t \varphi'_f(t)| |\varphi_K(ht) - 1| dt, \quad a.s. \end{aligned}$$

Therefore we have

$$\limsup_{n \rightarrow \infty} \frac{1}{\sqrt{(2n \log \log n)}} \sum_{j=1}^n \rho_n(X_j, 1) = s, \quad a.s.$$

However from the proof of Theorem 2.2 and (2.5),

$$\limsup_{n \rightarrow \infty} \left(\frac{n}{\log \log n} \right)^{\frac{1}{2}} \beta_n = \sqrt{2}s / \int |t \varphi'_f(t)|^2 dt, \quad a.s.$$

Hence from the relation $g(\beta_n) = b_n$, we have (4.1).

In what follows we propose an estimate of the b_n which always takes a value greater than 1. If this estimate \hat{b}_n satisfies $ISE_n(\hat{b}_n) < ISE_n(1)$ then the curve of kernel density estimate is flat as compared with that of true p.d.f. From Theorem 4.2, we may propose a natural estimate of β_n as $\hat{\beta}_n = \sqrt{2}\hat{\sigma}_f \left(\frac{1}{n} \log \log n \right)^{\frac{1}{2} + \delta}$, where $\hat{\sigma}_f$ is an estimate of σ_f and δ is a positive constant. By the assumption (A.3), $g(u)$ can be approximated by $1 + u$ when $|u|$ is sufficiently small. Therefore from the relation $b_n = g(\beta_n)$, we may propose an estimate \hat{b}_n of the minimizer b_n as

$$\hat{b}_n = 1 + \sqrt{2}\hat{\sigma}_f \left(\frac{1}{n} \log \log n \right)^{\frac{1}{2} + \delta}. \quad (4.2)$$

It is noted that from the definition of the σ_f in Theorem 2.2, σ_f is invariant with respect to the constant multiplication of the standard deviation of $f(x)$. This fact agrees with the contents described in section 2.

As we saw in Theorem 2.1, the difference between $ISE_n(b_n)$ and $ISE_n(1)$ will vanish if sample size n tends to infinity. So, we may say that the scale parameter b is more effective when n is not so large. This will be also confirmed by our simulation in section 5, however, the vanishing speed of the difference is evaluated as follows.

THEOREM 4.3. *Let $\hat{\beta}_n$ be any estimator for the β_n such that it converges to 0 with probability one, as n tends to infinity. Then we have*

$$ISE_n \left(g \left(\hat{\beta}_n \right) \right) - ISE_n(1) = O \left(\hat{\beta}_n \right), \quad a.s.$$

for large n .

PROOF. By Taylor expansion and from (1.3), straight forward calculation yields that

$$\begin{aligned} & \left| ISE_n \left(g \left(\hat{\beta}_n \right) \right) - ISE_n(1) \right| \\ &= \frac{1}{2\pi} \left| 2 \operatorname{Re} \left[\int \left\{ \varphi_f(t) - \varphi_f \left(g \left(\hat{\beta}_n \right) t \right) \right\} \varphi_K(ht) \overline{c_n(t)} dt \right] + \int \left| \varphi_f \left(g \left(\hat{\beta}_n \right) t \right) \right|^2 dt \right. \\ & \quad \left. - \int |\varphi_f(t)|^2 dt \right| \\ &\leq \frac{1}{\pi} \left| \int \hat{\beta}_n g' \left(\theta \hat{\beta}_n \right) t \varphi_f' \left(g \left(\theta \hat{\beta}_n \right) t \right) \varphi_K(ht) \overline{c_n(t)} dt \right| \\ & \quad + \frac{1}{2\pi} \left| \left(\hat{\beta}_n g' \left(\theta \hat{\beta}_n \right) \right)^2 \int \left| t \varphi_f' \left(g \left(\theta \hat{\beta}_n \right) t \right) \right|^2 dt \right. \\ & \quad \left. + 2 \hat{\beta}_n g' \left(\theta \hat{\beta}_n \right) \operatorname{Re} \left[\int t \varphi_f' \left(g \left(\theta \hat{\beta}_n \right) t \right) \overline{\varphi_f(t)} dt \right] \right| \\ &\leq \frac{2M}{\pi m^2} \left| \hat{\beta}_n \right| \int |t \varphi_f'(t)| dt + \frac{M^2}{2\pi m^3} \hat{\beta}_n^2 \int |t \varphi_f'(t)|^2 dt \\ &= O \left(\hat{\beta}_n \right), \quad a.s. \end{aligned}$$

where θ satisfies $0 < \theta < 1$. Hence we get the conclusion.

5. Simulation study

In this section we conduct some simulations to see how the scale parameter works for these distributions of the standard normal, mixture of some normals and double exponential. Marron and Wand [6] provided the exact mean integrated squared errors of some mixtures of normal densities. Since any density may be well approximated by the normal mixture, we employ some of mixtures in Marron and Wand [6] for the testing densities. Hereafter we rewrite the left hand side of (1.3) as $ISE_n(b, h)$. The optimal window width h_{opt} is defined by $ISE_n(1, h_{opt}) = \min_{h>0} ISE_n(1, h)$. There are several methods to estimate the optimal window width, however, it seems that Bowman's [1] cross-validation has a consistently moderate performance compared with other methods (Jones and Kappenman [4]). Thus we use his estimate here and denote it as \hat{h}_{cv} . The scale parameter b_n is the minimizer of $ISE_n(b, h_{opt})$ with respect to b . And as we have mentioned in section 2, the data is previously normalized. From (4.2), we define \hat{b}_n as

$$\hat{b}_n = 1 + \alpha \left(\frac{1}{n} \log \log n \right)^{0.6}, \quad (5.1)$$

where α is a positive constant. Each simulation is replicated 200 times to compute the sample relative efficiency:

$$Ref_n(b, h) = \frac{1}{200} \sum_{i=1}^{200} \frac{ISE_n(b, h)}{ISE_n(1, h)} \quad (5.2)$$

where n is a sample size ($n = 25, 50, 100, 200$). We use $\phi_{(\mu, \sigma)}$ as the p.d.f. of the normal $N(\mu, \sigma^2)$, and $\phi(p_1, \mu_1, \sigma_1; \dots; p_m, \mu_m, \sigma_m)$ represents the mixture of these normals such that $\sum_{i=1}^m p_i \phi_{(\mu_i, \sigma_i)}$, where $\sum_{i=1}^m p_i = 1$, $p_1, p_2, \dots, p_m \geq 0$. The underlying distributions are $f_1 : \phi_{(0,1)}$ the standard normal, $f_2 : \phi(\frac{1}{5}, -1, 1; \frac{1}{5}, \frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{13}{12}, \frac{5}{9})$ left skewed density, $f_3 : \phi(\frac{1}{5}, -\frac{1}{2}, 1; \frac{1}{5}, -\frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{3}{2}, \frac{5}{9})$ asymmetric bimodal density, $f_4 : \phi(\frac{9}{20}, -\frac{6}{5}, \frac{3}{5}; \frac{9}{20}, \frac{6}{5}, \frac{3}{5}; \frac{1}{10}, 0, \frac{1}{4})$ trimodal density and $f_5 : \frac{1}{2}e^{-|x|}$ double exponential density. We take the Gaussian kernel $K = \phi_{(0,1)}$ and we set the α in (5.1) as 0.3 by an empirical choice. The results are displayed in terms of $Ref_n(b, h) \times 10^2$ in the Table 1. Each component consists of the three values: $Ref_n(b_n, h_{opt})$, $Ref_n(\hat{b}_n, h_{opt})$ and $Ref_n(\hat{b}_n, \hat{h}_{cv})$. The number in parentheses shows the numbers of the times that the event $\{b_n > 1\}$ did occur among the 200 replications.

Table 1. Values of $Ref_n(b, h) \times 10^2$.

sample	f_1	f_2	f_3	f_4	f_5
n=25	65.8(195)	72.1(176)	79.3(165)	83.5(151)	66.2(199)
	87.5	90.2	94.4	95.8	88.7
	90.7	92.2	94.9	95.7	91.4
n=50	67.7(200)	74.9(175)	80.5(147)	84.2(141)	68.4(197)
	88.2	90.8	95.3	96.6	90.6
	89.4	93.2	95.9	97.1	92.4
n=100	67.7(200)	75.5(175)	81.6(149)	86.6(140)	73.9(194)
	86.4	91.5	96.0	97.4	92.2
	90.4	93.3	96.4	97.3	93.2
n=200	71.4(200)	76.1(184)	82.7(153)	87.9(147)	77.4(188)
	87.6	91.6	95.2	97.9	93.8
	91.0	93.5	96.0	98.0	94.6

In the Table 1, almost all cases the numbers in parentheses are more than 150, except for f_4 . And in all cases the sample relative efficiencies are smaller than 100, especially for f_1 and f_5 cases are remarkable. The smaller the sample size is, the smaller value the sample relative efficiency takes. These results agree with the contents of Theorem 4.3.

From the above results, we may employ the modified kernel density estimate $\hat{f}_n(x; s)$ defined in (1.4). To see the performance of the integrated squared error of the (1.4), we

define (5.3) as

$$ISE_n^*(s, h) = \int \left| \hat{f}_n(x; s) - f(x) \right|^2 dx. \quad (5.3)$$

We also define the sample relative efficiency in the same way as in (5.2).

$$Ref_n^*(s, h) = \frac{1}{200} \sum_{i=1}^{200} \frac{ISE_n^*(s, h)}{ISE_n^*(1, h)} \quad (5.4)$$

Note that the s_n is the minimizer of the $ISE_n^*(s, h_{opt})$ in this case. As the \hat{s}_n , an estimate of the s_n , we also employ the estimate \hat{b}_n in (5.1), with $\alpha = 0.2$ by an empirical choice.

Table 2. Values of $Ref_n^*(s, h) \times 10^2$.

sample	f_1	f_2	f_3	f_4	f_5
n=25	72.0(188)	77.5(161)	82.4(126)	85.5(117)	77.9(191)
	92.1	94.9	98.0	99.0	94.8
	95.4	97.0	98.9	99.3	96.9
n=50	73.0(193)	78.1(163)	81.9(126)	85.5(122)	78.0(186)
	92.2	94.5	97.5	99.0	95.5
	94.0	96.4	98.2	99.0	96.8
n=100	72.0(198)	79.8(157)	83.4(118)	87.6(124)	81.2(187)
	91.1	95.4	98.6	99.0	96.0
	94.4	97.0	99.2	99.2	96.7
n=200	75.0(199)	78.9(170)	84.2(140)	88.8(135)	79.9(178)
	92.0	94.9	97.2	99.0	97.0
	94.5	96.3	97.9	99.3	96.9

In the Table 2, each component also consists of the three values: $Ref_n^*(s_n, h_{opt})$, $Ref_n^*(\hat{s}_n, h_{opt})$ and $Ref_n^*(\hat{s}_n, \hat{h}_{cv})$. The number in parentheses shows the numbers of the times that the event $\{s_n > 1\}$ did occur among the 200 replications. The same tendencies are recognized as the Table 1. All values are greater than those of the corresponding ones in the Table 1, but the values in the parentheses are opposite. We should remark that if there exists a better estimate \hat{h} than Bowman's cross-validation method for window width, the sample relative efficiency $Ref_n^*(\hat{s}_n, \hat{h})$ can attain up to the second row values in each component of the table. In this way, we can reduce the sample relative efficiency with such a naive and crude estimate \hat{s}_n of the shrink parameter s_n . Thus it is of use to employ the modified density estimate (1.4) with the Gaussian kernel. The author believes that the shrink parameter will be effective for these kernels which have not large amount of high frequency component in the frequency domain such as the Gaussian.

Acknowledgements

The author would like to express his thanks to the editor and referees for their valuable comments, and to Prof. M. Sibuya with the members of his seminar of Keio University for their critical comments and suggestions. Thanks are also due to Mr. K. Nittono for his comments on the computer programming.

References

- [1] Bowman, A. : An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71** (1984), 353-360.
- [2] Feuerverger, A. and Mureika, R. A. : The empirical characteristic function and its applications, *Ann. Statist.*, **5** (1977), 88-97.
- [3] Fryer, M. J. : Some errors associated with the nonparametric estimation of density functions, *J. Inst. Maths. Applics.*, **18** (1976), 371-380.
- [4] Jones, M. C. and Kappenman, R. F. : On a class of kernel density estimate bandwidth selectors, *Scand. J. Statist.*, **19** (1991), 337-349.
- [5] Lukacs, E. : *Characteristic functions*, 2nd. ed. Griffin, London, (1970).
- [6] Marron, J. S. and Wand, M. P. : Exact mean integrated squared error, *Ann. Statist.*, **20** (1992), 712-736.
- [7] Park, B. U., Chung, S. S. and Seog, K. H. : An empirical investigation of the shifted power transformation method in density estimation, *Comp. Statist. & Data Anal.*, **14** (1992), 183-191.
- [8] Park, B. U. and Marron, J. S. : Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Ass.*, **85** (1990), 66-72.
- [9] Ruppert, D. and Wand, M. P. : Correcting for kurtosis in density estimation, *Austral. J. Statist.*, **34** (1992), 19-29.
- [10] Serfling, R. J. : *Approximation Theorems of Mathematical Statistics*, Wiley, New York, (1980).
- [11] Silverman, B. W. and Young, G. A. : The bootstrap : To smooth or not to smooth, *Biometrika*, **74** (1987), 469-479.
- [12] Takeuchi, H. : A location shift problem in nonparametric density estimation, *Bull. Inf. Cybernet.*, **25** (1993), 195-212.
- [13] Wand, M. P., Marron, J. S. and Ruppert, D. : Transformations in density estimation (with discussions), *J. Amer. Statist. Ass.*, **86** (1991), 343-361.

Received May 10, 1995

Revised january 24, 1996