NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Upper bounds for the MISE of Kernel Estimators of Probability Density Functions and their Derivatives

by

Ingrid K. Glad, Nils Lid Hjort and Nikolai Ushakov

PREPRINT
STATISTICS NO. 3/2007

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL `http://www.math.ntnu.no/preprint/statistics/2007/S3-2007.ps`

Nikolai Ushakov has homepage: `http://www.math.ntnu.no/~ushakov`
E-mail: `ushakov@stat.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491
Trondheim, Norway.

Upper bounds for the MISE of Kernel Estimators of Probability Density Functions and their Derivatives

Ingrid K. Glad, Nils Lid Hjort
Department of Mathematics, University of Oslo, Norway
and
Nikolai Ushakov
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Abstract

The performance of kernel density estimators is usually studied via Taylor expansions and asymptotic approximation arguments, in which the bandwidth parameter tends to zero with sample size tending to infinity. In contrast, this paper focusses directly on the finite-sample situation. Informative upper bounds are derived for the integrated mean squared error function. Results are reached for the traditional case, where the kernel is a probability density function, under various sets of assumptions on the underlying density to be estimated.

Key words: Kernel estimation, meanan integrated squared error, Empirical characteristic function, Finite samples, Inequalities

1. Introduction

Let $X_1, ..., X_n$ be independent and identically distributed random variables with absolutely continuous distribution function $F(x)$ and density function $f(x)$. The kernel density estimator associated with the sample $X_1, ..., X_n$ is defined as

$$f_n(x; h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right),$$

where $K(x)$ is the kernel function with scaled version

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right),$$

and $h = h_n$ is a positive number (depending on n) called the bandwidth or the smoothing parameter.

Kernel estimation of probability density functions is one of the most popular methods having a number of advantages and quite well developed and studied, see, for example, Wand and Jones (1995) and Fan and Gijbels (1996).

Let both the density to be estimated $f(x)$ and the kernel $K(x)$ be m times differentiable. Then $f_n(x; h)$ is also m times differentiable and it is natural to estimate the m-th derivative of $f(x)$ by the m-th derivative of $f_n(x; h)$:

$$f_n^{(m)}(x; h) = \frac{1}{nh^{m+1}} \sum_{j=1}^{n} K^{(m)}\left(\frac{x - X_j}{h}\right). \tag{1}$$

We unite these two cases (estimation of a density and its derivatives) defining, as usually, the 0-th derivative of a function as the function itself: $f^{(0)}(x) = f(x)$. Thus, the main subject of this work is estimator (1) with $m \geq 0$.

Under some standard assumptions, the mean integrated squared error (MISE) of the kernel estimator is represented as the sum of the main term and the remainder

$$\text{MISE} = \text{AMISE} + \text{REM}$$

where the AMISE (asymptotic MISE) is explicitly expressed in terms of a few parameters of the kernel and of the density to be estimated, and

$$\text{REM} = o(\text{AMISE}), \quad h \to 0 \ \ (n \to \infty)$$

(more precise and detailed description of the AMISE is given in the next section). Typically, the remainder is ignored and all procedures (like selection of the smoothing parameter, estimation of the real error — MISE, etc.) are based on the AMISE. This approach, however, does not have an adequate basis because in the considered case "large sample size" means "very large". This is illustrated by examples presented in Marron and Wand (1992). Typically the reminder REM has approximately the same value (near the optimal value of h) as the MISE if $n \leq 100$. It is more than 25% of MISE for $n = 1000$ and almost 10% for $n = 10^6$.

Marron and Wand (1992) established the following remarkable fact. If the kernel is non-negative, then, under some regularity conditions, for $m = 0$,

$$\text{MISE} < \text{AMISE} \tag{2}$$

for all h. Marron and Wand (1992) speculated that (2) was true in general i.e. for kernels taking both positive and negative values. This is not correct (see Appendix 1), but (2) is true for all $m \geq 0$ if the kernel is non-negative (Corollary of Theorem 1 below). For many situations however this beautiful inequality is not of practical interest because of the reason we have just discussed, that is this inequality is very crude, except n is huge.

In this paper some upper bounds for the MISE are obtained which are essentially closer to the MISE than the AMISE is. Analysis of these bounds shows that in many situations the bandwidth selection, based on the AMISE minimization approach, is justified.

2. AMISE definition and MISE representation

In the following we will omit integration limits when the integral is to be taken over the full real line. Let $\hat{f}_n(x)$ be an estimator (not necessarily a kernel estimator) of $f^{(m)}(x)$ associated with the sample $X_1, ..., X_n$. The bias, the mean squared error (MSE) and the mean integrated squared error (MISE) of $\hat{f}_n(x)$ are defined, respectively, as

$$\mathrm{B}(\hat{f}_n(x)) = \mathrm{E}\,\hat{f}_n(x) - f^{(m)}(x),$$

$$\mathrm{MSE}(\hat{f}_n(x)) = \mathrm{E}[\hat{f}_n(x) - f^{(m)}(x)]^2,$$

and

$$\mathrm{MISE}(\hat{f}_n(x)) = \int \mathrm{MSE}(\hat{f}_n(x))dx = \mathrm{E}\int [\hat{f}_n(x) - f^{(m)}(x)]^2 dx.$$

In case of the kernel estimator $f_n^{(m)}(x;h)$, defined by (1), the bias may be expressed via the convolution as

$$B(f_n^{(m)}(x;h)) = (K_h^{(m)} \star f)(x) - f^{(m)}(x) = \int K_h^{(m)}(x-y)f(y)dy - f^{(m)}(x) =$$

$$= (K_h \star f^{(m)})(x) - f^{(m)}(x) = \int K_h(x-y)f^{(m)}(y)dy - f^{(m)}(x). \tag{3}$$

Since convolution is a kind of smoothing, the bias of the kernel estimator is the difference between a smoothed density/derivative and the density/derivative itself. The mean integrated squared error admits a well-known decomposition into integrated variance and integrated squared bias, with consequent representation

$$\mathrm{MISE}(\hat{f}_n(x)) = \int B^2(\hat{f}_n(x))dx + \int \mathrm{Var}(\hat{f}_n(x))dx. \tag{4}$$

Consider the MISE of the kernel estimator (1). The integrated squared bias tends to 0 as $h \to 0$. The integrated variance tends to 0 as $nh^{2m+1} \to \infty$. Suppose that conditions $n \to \infty$, $h \to 0$, and $nh^{2m+1} \to \infty$ are satisfied. Consider the expansion of the integrated squared bias into a series in powers of h, and the expansion of the integrated variance into a series in powers of $1/(nh^{2m+1})$, assuming that necessary derivatives exist. We define the sum of the main terms of the first order of these two expansions as the main term of the MISE and call it asymptotic mean integrated squared error (AMISE). It is clear that MISE $\sim$ AMISE as $n \to \infty$, $h \to 0$, $nh^{2m+1} \to \infty$.

Note that together with MSE and MISE other mausures of deviation may be used. Among them the mean absolute error $\mathrm{E}\,|\hat{f}_n(x) - f^{(m)}(x)|$ and its integral are especially important (see Devroye and Györfi, 1985). In the present article we restrict attention to MSE and MISE, however.

Denote the characteristic function of random variables X_j by $\varphi(t)$ and the empirical characteristic function associated with sample $X_1, ..., X_n$ by $\varphi_n(t)$:

$$\varphi(t) = \mathrm{E}\, e^{itX_j} = \int e^{itx} f(x)dx, \quad \varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}.$$

The general form of the kernel estimator (1) in terms of the empirical characteristic function is

$$f_n^{(m)}(x;h) = \frac{1}{2\pi}\int e^{-itx}(-it)^m \varphi_n(t)\psi(ht)dt,$$

where ψ is the characteristic function of the kernel:

$$\psi(t) = \int e^{itx} K(x)dx.$$

The characteristic function of the estimator $f_n^{(m)}(x;h)$ is $(-it)^m\varphi_n(t)\psi(ht)$.

Lemma 1. *Let $\varphi(t)$ and $\psi(t)$ be characteristic functions of the density $f(x)$ and the kernel $K(x)$ respectively. If the density $f(x)$ and the kernel $K(x)$ are m times differentiable, then*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) = \frac{1}{2\pi}\int t^{2m}|\varphi(t)|^2|1-\psi(ht)|^2dt +$$

$$+\frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}(1-|\varphi(t)|^2)|\psi(ht)|^2dt. \tag{5}$$

Proof. Due to the Parseval-Plancherel identity we have

$$\mathrm{MISE}(f_n^{(m)}(x;h)) = \mathrm{E}\int (f_n^{(m)}(x;h) - f^{(m)}(x))^2dx =$$

$$= \frac{1}{2\pi}\,\mathrm{E}\int t^{2m}|\varphi_n(t)\psi(ht) - \varphi(t)|^2dt =$$

$$= \frac{1}{2\pi}\int t^{2m}\,\mathrm{E}\left[(\varphi_n(t)\psi(ht) - \varphi(t))(\overline{\varphi_n(t)}\;\overline{\psi(ht)} - \overline{\varphi(t)})\right]dt =$$

$$= \frac{1}{2\pi}\int t^{2m}\left[|\psi(ht)|^2\,\mathrm{E}\,|\varphi_n(t)|^2 - \varphi(t)\overline{\psi(ht)}\,\mathrm{E}\,\overline{\varphi_n(t)} - \overline{\varphi(t)}\psi(ht)\,\mathrm{E}\,\varphi_n(t) + |\varphi(t)|^2\right]. \tag{6}$$

Evidently

$$\mathrm{E}\,\varphi_n(t) = \varphi(t) \tag{7}$$

and

$$\mathrm{E}\,\overline{\varphi_n(t)} = \overline{\varphi(t)}. \tag{8}$$

Find $\mathrm{E}\,|\varphi_n(t)|^2$:

$$\mathrm{E}\,|\varphi_n(t)|^2 = \mathrm{E}\left|\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right|^2 = \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\cdot\frac{1}{n}\sum_{k=1}^{n} e^{-itX_k}\right] =$$

$$= \frac{1}{n^2}\left[n + \sum_{j\neq k} e^{it(X_j - X_k)}\right] = \frac{1}{n} + \left(1 - \frac{1}{n}\right)|\varphi(t)|^2. \tag{9}$$

Substituting (7)—(9) to the right hand side of (6) and using the elementary identity $|\psi|^2 - \psi + \overline{\psi} + 1 = |1-\psi|^2$, we obtain (5). ■

Remark. Consider the integrated squared bias. Using (3) and the Parseval-Plancherel identity, we obtain

$$\int B^2(f_n^{(m)}(x;h))dx = \int \left[\int K_h(x-y)f^{(m)}(y)dy - f^{(m)}(x)\right]^2 dx =$$

$$= \frac{1}{2\pi}\int |(-it)^m\psi(ht)\varphi(t) - (-it)^m\varphi(t)|^2\, dt = \frac{1}{2\pi}\int t^{2m}|\varphi(t)|^2|1-\psi(ht)|^2 dt,$$

that is the integrated squared bias is the first summand in representation (5) of Lemma 1. Hence, due to (4), the second summand in the right hand side of (5) is the integrated variance of the estimator.

3. Smooth case

3.1. Main result

For a real valued function $g(x)$ we will use the following notation, provided the integrals exist:

$$\mu_k(g) = \int |x|^k g(x)dx, \quad k = 0,1,2,..., \quad R(g) = \int g^2(x)dx.$$

In this section, we suppose that the function to be estimated (density or derivative) is twice differentiable. More exactly, we suppose that the following conditions are satisfied:

(a) the density $f(x)$ is $m+2$ ($m \geq 0$) times differentiable and the $(m+2)-$nd derivative is square integrable,

(b) the kernel $K(x)$ is a symmetric m times differentiable probability density function with finite second moment.

Under these conditions $\mathrm{MISE}(f_n^{(m)}(x;h))$ is represented in the form (see Wand and Jones, 1995, p. 21)

$$\mathrm{MISE}(f_n^{(m)}(x;h)) = \mathrm{AMISE}(f_n^{(m)}(x;h)) + o\left(h^4 + \frac{1}{nh^{2m+1}}\right)$$

as

$$h \to 0, \quad nh^{2m+1} \to \infty,$$

where AMISE is

$$\mathrm{AMISE}(f_n^{(m)}(x;h)) = \frac{1}{4}h^4\mu_2(K)^2R(f^{(m+2)}) + \frac{1}{nh^{2m+1}}R(K^{(m)}).$$

Denote the variance of the density to be estimated by σ^2 (including $\sigma^2 = \infty$) and put

$$c(h) = \min\left\{\frac{1}{\sigma}, \frac{1}{\sqrt{\mu_2(K)}h}\right\} \tag{10}$$

(if $\sigma^2 = \infty$, then $c(h) = 0$).

Theorem 1. *Let conditions (a) and (b) be satisfied. Then for all m, n and h*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) < \mathrm{AMISE}(f_n^{(m)}(x;h)) -$$

$$-\frac{c(h)^{2m+1}}{\pi(2m+1)n}\left[1-\frac{2m+1}{2m+3}(\sigma^2+\mu_2(K)h^2)c(h)^2+\frac{2m+1}{2m+5}\sigma^2\mu_2(K)h^2c(h)^4\right], \tag{11}$$

where $c(h)$ is defined by (10).

Proof. Since $K(x)$ is symmetric, $\psi(t)$ is real, therefore, due to Lemma 1,

$$\mathrm{MISE}(f_n^{(m)}(x;h)) = \frac{1}{2\pi}\int t^{2m}|\varphi(t)|^2(1-\psi(ht))^2 dt +$$

$$+\frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}(1-|\varphi(t)|^2)|\psi(ht)|^2 dt.$$

We have

$$\big(1-\psi(ht)\big)^2 \le \frac{\mu_2(K)^2h^4t^4}{4}$$

for all t, and

$$|\varphi(t)\psi(ht)|^2 \ge (1-\sigma^2t^2)(1-\mu_2(K)h^2t^2)$$

for $|t| \le c(h)$ (see Ushakov, 1999, p.89) Using these inequalities and the Parseval-Plancherel identity in two forms

$$\int K^{(m)}(x)^2 dx = \frac{1}{2\pi}\int t^{2m}|\psi(t)|^2 dt$$

and

$$\int (f^{(m+2)}(x))^2 dx = \frac{1}{2\pi}\int t^{2(m+2)}|\varphi(t)|^2 dt,$$

we obtain

$$\mathrm{MISE}(f_n(x;h)) \le$$

$$\le \frac{\mu_2(K)^2h^4}{4}\frac{1}{2\pi}\int t^{2(m+2)}|\varphi(t)|^2 dt + \frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}|\psi(ht)|^2 dt -$$

$$-\frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}|\psi(ht)\varphi(t)|^2 dt \le \frac{\mu_2(K)^2h^4}{4}\int \big(f^{(m+2)}(x)\big)^2 dx +$$

$$+\frac{1}{nh^{2m+1}}\int K^{(m)}(x)^2 dx - \frac{1}{\pi n}\int_0^{c(h)} t^{2m}(1-\sigma^2t^2)(1-\mu_2(K)h^2t^2)dt =$$

$$= \frac{h^4\mu_2(K)^2R(f^{(m+2)})}{4} + \frac{1}{nh^{2m+1}}R(K^{(m)}) -$$

$$-\frac{c(h)^{2m+1}}{\pi(2m+1)n}\left[1-\frac{2m+1}{2m+3}(\sigma^2+\mu_2(K)h^2)c(h)^2+\frac{2m+1}{2m+5}\sigma^2\mu_2(K)h^2c(h)^4\right] =$$

$$= \mathrm{AMISE}(f_n^{(m)}(x;h)) -$$

$$-\frac{c(h)^{2m+1}}{\pi(2m+1)n}\left[1-\frac{2m+1}{2m+3}(\sigma^2+\mu_2(K)h^2)c(h)^2+\frac{2m+1}{2m+5}\sigma^2\mu_2(K)h^2c(h)^4\right].$$

■

It is easy to see that the expression in the square brackets in the right hand side of (11) is always positive, therefore we obtain the following extention of the Marron–Wand inequality (2) to arbitrary m.

Corollary. *Let conditions (a) and (b) be satisfied. Then*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) < \mathrm{AMISE}(f_n^{(m)}(x;h))$$

for all n, m and h.

Note that inequality (11) is essentially more precise than (2).

3.2. Bandwidth selection

Majority of methods of bandwidth selection are based on estimating h_{AMISE} — such a value of h which minimizes the AMISE and which in the considered case (the estimated function $f^{(m)}(x)$ is twice differentiable) is

$$h_{AMISE} = \left[\frac{4(2m+1)R(K^{(m)})}{n\mu_2(K)^2 R(f^{(m+2)})}\right]^{1/(2m+5)}.$$

The right hand side of (11) is a more precise approximation of the MISE than the AMISE therefore it seems to be reasonable to estimate h which minimizes the right hand side of (11) instead of AMISE. It turns out that both minimizations practically give the same result. Indeed, consider the reminder in the right hand side of (11)

$$R = \frac{c(h)^{2m+1}}{\pi(2m+1)n}\left[1 - \frac{2m+1}{2m+3}(\sigma^2 + \mu_2(K)h^2)c(h)^2 + \frac{2m+1}{2m+5}\sigma^2\mu_2(K)h^2c(h)^4\right].$$

Except a very irregular density and under a reasonable choice of the kernel, for h to be close to the optimal value, σ is much greater than $\mu_2(K)^{1/2}h$, so in this case, $c(h) = 1/\sigma$, and

$$R \approx \frac{1}{\pi(2m+1)n\sigma^{2m+1}}\left(1 - \frac{2m+1}{2m+3}\right). \tag{12}$$

The right hand side of (12) does not depend on h, i.e. R is almost a constant (with respect to h), and hence minimization of the difference AMISE $-R$ is, roughly speaking, almost the same as minimization of the AMISE.

Thus the bandwidth selection, based on the AMISE minimization approach, is justified (at least for more or less regular densities).

4. Intermediate case

In this section, we assume that the function to be estimated (density or derivative) has the first derivative but does not have the second one. If the function to be estimated $f^{(m)}(x)$ is twice differentiable (with square integrable $f^{(m+2)}(x)$), then, under the optimal choice of the bandwidth h, the order of decrease (as $n \to \infty$) of the MISE of the kernel estimator under consideration (order of consistency) is $n^{-4/(2m+5)}$ independently of further smoothness of $f(x)$. If $f^{(m)}(x)$ has only the first derivative and does not have the second one, then the MISE decreases slower than $n^{-4/(2m+5)}$, one can only guarantee that

$$\inf_{h>0} \mathrm{MISE}(f_n^{(m)}(x;h)) = O\left(n^{-2/(2m+3)}\right), \quad n \to \infty.$$

In fact, for any $\alpha \in [2/(2m+3), 4/(2m+5)]$, there exists a density $f(x)$ (for which $f^{(m+1)}(x)$ exists and $f^{(m+2)}(x)$ does not exist) such that

$$\inf_{h>0} \mathrm{MISE}(f_n^{(m)}(x;h)) = O(n^{-\alpha}), \quad n \to \infty. \tag{13}$$

Existence of such densities is proved in Appendix 2. Exact conditions, under which (13) holds for a given α, are related to fractional derivatives of $f(x)$. Some results, concerning these conditions, will be published elsewhere.

The following theorem gives an upper bound for MISE in the considered case.

Theorem 2. *Let $f^{(m)}(x)$, $m \geq 0$, be differentiable and its derivative be square integrable. If $K(x)$ is a symmetric m times differentiable probability density function with finite first moment, then for all n,*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) < h^2\mu_1(K)^2 R(f^{(m+1)}) + \frac{1}{nh^{2m+1}} R(K^{(m)}).$$

Proof. We have (see Ushakov, 1999, p. 91)

$$(1-\psi(ht))^2 \leq \mu_1(K)^2 h^2 t^2.$$

Using this inequality, Lemma 1 and the Parseval-Plancherel identity, we obtain

$$\mathrm{MISE}(f_n^{(m)}(x;h)) < h^2\mu_1(K)^2 \frac{1}{2\pi}\int t^{2(m+1)}|\varphi(t)|^2 dt +$$

$$+\frac{1}{nh^{2m+1}} \cdot \frac{1}{2\pi}\int t^{2m}|\psi(t)|^2 dt - \frac{1}{2\pi n}\int t^{2m}|\varphi(t)|^2|\psi(ht)|^2 dt <$$

$$< h^2\mu_1(K)^2 R(f^{(m+1)}) + \frac{1}{nh^{2m+1}} R(K^{(m)}).$$

■

Corollary. *Let conditions of Theorem 2 be satisfied. Then*

$$\inf_{h>0} \mathrm{MISE}(f_n^{(m)}(x;h)) <$$

$$< \left(1 + \frac{2}{2m+1}\right) 2^{-\frac{2}{2m+3}} (2m+1)^{\frac{2}{2m+3}} \left[\mu_1(K)^{2(2m+1)} R(K^{(m)})^2 R(f^{(m+1)})^{2m+1}\right]^{\frac{1}{2m+3}} n^{-\frac{2}{2m+3}}.$$

In the proof of the theorem, we ignored the term

$$-\frac{1}{2\pi n}\int t^{2m}|\varphi(t)|^2|\psi(ht)|^2 dt.$$

It can be taken into account in absolutely the same way as in Section 3, that has sense if n is small or moderate. Then we obtain

Theorem 3. *Let conditions of Theorem 3 be satisfied and, in addition, $f(x)$ and $K(x)$ have finite variances. Then*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) < h^2\mu_1(K)^2 R(f^{(m+1)}) + \frac{1}{nh^{2m+1}} R(K^{(m)}) -$$

$$-\frac{c(h)^{2m+1}}{\pi(2m+1)n}\left[1 - \frac{2m+1}{2m+3}(\sigma^2 + \mu_2(K)h^2)c(h)^2 + \frac{2m+1}{2m+5}\sigma^2\mu_2(K)h^2 c(h)^4\right],$$

where $c(h)$ is defined by (10).

Let $m = 0$, i.e. a density is estimated. Using Theorem 2 and its Corollary, we can make comparisons of the kernel estimator to the histogram. Usually these comparisons are made in the smooth case, i.e. when the density to be estimated is at least two times differentiable. In that case, the MISE of the histogram is asymptotically inferior to the kernel density estimator: convergence rate of the MISE is $O(n^{-2/3})$ for the histogram and $O(n^{-4/5})$ for the kernel estimator. However, if the density to be estimated is only one time differentiable, the order of convergence rate is the same for the histogram and for the kernel estimator. Nevertheless, Theorem 2 shows that in this case, the kernel estimator is still better than the histogram, both asymptotically and for finite values of n. Indeed, let $f_H(x; b)$ be the histogram with binwidth b. Then (see for example Wand and Jones (1995), p. 23)

$$\inf_{b>0} \mathrm{MISE}(f_H(x;b)) \sim \frac{1}{4}(36R(f'))^{1/3} n^{-2/3}, \quad n \to \infty.$$

Consider the kernel estimator with, for example, the uniform kernel: $K(x) = 1/2$ for $|x| \le 1$ and $K(x) = 0$ for $|x| > 1$. Then $\mu_1(K) = 1/2$, $R(K) = 1/2$, and therefore, due to Corollary of Theorem 2

$$\inf_{h>0} \mathrm{MISE}(f_n(x;h)) < \frac{3}{4} R(f')^{1/3} n^{-2/3} < \frac{1}{4}(36R(f'))^{1/3} n^{-2/3} \approx \frac{3.3}{4} R(f')^{1/3} n^{-2/3}.$$

5. Non-smooth case

Functions (densities/derivatives), considered in this section, are not supposed to be differentiable and even continuous. We only suppose that they have bounded total variation, which we denote by $\mathrm{V}(\cdot)$ and which is defined for a real-valued function $g(x)$ as

$$\mathrm{V}(g) = \sup \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})|$$

where sup is taken over all n and all collections $x_0, x_1, ..., x_n$ such that $-\infty < x_0 < x_1 < ... < x_n < \infty$. As to the kernel, we suppose that $K(x)$ is a symmetric density function. No assumption is made about moments of the kernel.

It is convenient to use the following parameter of the kernel, which was introduced by Watson and Leadbeter (1963) and which is expressed in terms of its characteristic function $\psi(t)$:

$$P(K) = \frac{1}{2\pi} \int \left(\frac{1 - \psi(t)}{t} \right)^2 dt.$$

Note that $P(K)$ exists (is finite) under very mild conditions, in particular, it is sufficient if the expectation of $K(x)$ exists.

Theorem 4. *Let $K(x)$ be a symmetric, m times differentiable ($m \ge 0$) density function and $f^{(m)}(x)$ have bounded total variation. Then*

$$\mathrm{MISE}(f_n^{(m)}(x;h)) \le h \, \mathrm{V}(f^{(m)})^2 P(K) + \frac{1}{nh^{2m+1}} R(K^{(m)}). \tag{13}$$

Proof. Estimate the first summand in the right hand side of (5) using the inequality

$$|\varphi(t)| \le \frac{\mathrm{V}(f^{(m)})}{|t|^{m+1}}$$

which holds for all t (see Ushakov and Ushakov, 2000). Then

$$\frac{1}{2\pi}\int t^{2m}|\varphi(t)|^2(1-\psi(ht))^2dt \leq$$

$$\leq \mathrm{V}(f^{(m)})^2 h\frac{1}{2\pi}\int\left(\frac{1-\psi(t)}{t}\right)^2 dt = \mathrm{V}(f^{(m)})^2 P(K)h.$$

The second term in the right hand side of (5) is estimated as above (see for example proof of Theorem 1):

$$\frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}(1-|\varphi(t)|^2)|\psi(ht)|^2dt \leq \frac{1}{nh^{2m+1}}R(K^{(m)}).$$

Thus we obtain (13). ■

Corollary. *Let conditions of Theorem 3 be satisfied. Then*

$$\inf_{h>0}\mathrm{MISE}(f_n^{(m)}(x;h)) \leq (2m+1)^{\frac{1}{2m+2}}\left(1+\frac{1}{2m+1}\right)\times$$

$$\times\left(\mathrm{V}(f^{(m)})^{2(2m+1)}R(K^{(m)})P(K)^{2m+1}\right)^{\frac{1}{2m+2}} n^{-\frac{1}{2m+2}}.$$

Appendix 1

In this Appendix we give an example when AMISE < MISE. Let $m=0$ Consider the following kernel:

$$K(x)=\frac{1}{\pi x^5}\left[(2cx^3+4c^3x^3-24cx)\cos(cx)+(24-2x^2-12x^2c^2)\sin(cx)\right],$$

where

$$c=\sqrt{\frac{\sqrt{5}-1}{2}}.$$

This is an even, bounded, continuous, integrable function with

$$\int K(x)dx=1$$

and

$$\mu_2(K)=2$$

i.e. $K(x)$ is a second-order kernel. The characteristic function of $K(x)$ is

$$\psi(t)=\int e^{itx}K(x)dx=\begin{cases}1-t^2-t^4 & \text{for } |t|\leq c,\\ 0 & \text{for } |t|>c.\end{cases}$$

Let $f(x)$ be a four times differentiable density function (with square integrable fourth derivative) whose characteristic function $\varphi(t)$ vanishes for $|t|>1$, examples of such densities can be found in Ramachandran (1997) or in Ushakov (1999). For $h<c$ the MISE is

$$\mathrm{MISE}=\frac{1}{2\pi}\int|\varphi(t)|^2(1-\psi(ht))^2dt+\frac{1}{n}\cdot\frac{1}{2\pi}\int(1-|\varphi(t)|^2)(\psi(ht))^2dt=$$

$$= h^4 R(f^{(2)}) + 2h^6 R(f^{(3)}) + h^8 R(f^{(4)}) + \frac{1}{nh} c_1 -$$

$$-\frac{1}{n}\big[R(h) - 2h^2 R(f') - h^4 R(f^{(2)}) + 2h^6 R(f^{(3)}) + h^8 R(f^{(4)})\big]$$

where

$$c_1 = \frac{1}{\pi}\int_0^c (1 - t^2 - t^4)^2 dt,$$

and the AMISE is

$$\text{AMISE} = h^4 R(f^{(2)}) + \frac{1}{nh} c_1.$$

It is clear that under appropriate choice of h and n the difference

$$\text{MISE} - \text{AMISE} = 2\left(1 - \frac{1}{n}\right) h^6 R(f^{(3)}) + \left(1 - \frac{1}{n}\right) h^8 R(f^{(4)}) -$$

$$-\frac{1}{n}\big[R(f) + 2h^2 R(f') + h^4 R(f^{(2)})\big]$$

can be made positive.

Appendix 2

In this appendix, we prove that for any $\alpha \in (0, 4/(2m+5))$ there exists an m times differentiable density $f(x)$ such that under very mild restrictions on the kernel $K(x)$, the MISE of the kernel estimator of the m-th derivative $f^{(m)}(x)$ satisfies inequalities

$$d_1 n^{-\alpha} \le \inf_{h>0} \text{MISE}(f_n^{(m)}(x;h)) \le d_2 n^{-\alpha},$$

where d_1 and d_2 are some positive constants.

Let us fix an arbitrary α, $0 < \alpha < 4/(2m+5)$, and put

$$\gamma = \frac{2m+1}{2(1-\alpha)}$$

(then $m + 1/2 < \gamma < m + 5/2$) and

$$c = \frac{\gamma}{1+\gamma}.$$

Consider the function

$$\varphi(t) = \begin{cases} 1 - c|t| & \text{for } |t| \le 1, \\ (1-c)|t|^{-\gamma} & \text{for } |t| > 1. \end{cases}$$

This function is symmetric, continuous, decreasing for $t > 0$, convex for $t > 0$ and satisfies conditions $\varphi(0) = 1$ and $\lim_{|t|\to\infty} \varphi(t) = 0$, therefore (see, for example, Feller, 1971) it is the characteristic function of an absolutely continuous distribution. Denote the corresponding density by $f(x)$. Since $\gamma > m + 1/2$, the function $t^{2m}|\varphi(t)|^2$ is integrable, and therefore $f(x)$ is m times differentiable, and $f^{(m)}(x)$ is square integrable.

Let $K(x)$, the kernel, be a symmetric probability density function with finite second moment, m times differentiable with square integrable m-th derivative $K^{(m)}(x)$. Denote its characteristic function by $\psi(t)$ and put

$$a_0 = \frac{1}{2\pi}\int t^{2m}|\psi(t)|^2 dt = \int K^{(m)}(x)^2 dx,$$

$$a_1 = \frac{1}{2\pi}\int_{|t|\ge 1} t^{2m}|\psi(t)|^2 dt,$$
$$c_0 = \int_0^\infty t^{2m-2\gamma}(1-\psi(t))^2 dt,$$
$$c_1 = \int_1^\infty t^{2m-2\gamma}(1-\psi(t))^2 dt$$
(c_0 and c_1 are positive and finite for $m+1/2<\gamma<m+5/2$),
$$\sigma^2 = \int x^2 K(x)dx.$$

Consider the kernel estimator $f_n^{(m)}(x;h)$ of $f^{(m)}(x)$ based on the kernel $K(x)$. According to Lemma 1,
$$\mathrm{MISE}(f_n^{(m)}(x;h)) =$$
$$= \frac{1}{2\pi}\int t^{2m}|\varphi(t)|^2|1-\psi(ht)|^2 dt + \frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}(1-|\varphi(t)|^2)|\psi(ht)|^2 dt =$$
$$= \frac{1}{\pi}\int_0^1 t^{2m}(1-ct)^2|1-\psi(ht)|^2 dt + \frac{1}{\pi}\int_1^\infty t^{2m}\left(\frac{1-c}{t^\gamma}\right)^2|1-\psi(ht)|^2 dt +$$
$$+\frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}(1-|\varphi(t)|^2)|\psi(ht)|^2 dt = S_1+S_2+S_3.$$

Find lower and upper bounds for each of the three summands in the right hand side. Without loss of generality we will suppose that $0<h<1$. We have
$$0 \le S_1 \le \frac{1}{\pi}\int_0^1 (1-\psi(ht))^2 dt \le \frac{1}{4\pi}\sigma^4 h^4 \int_0^1 t^4 dt = \frac{\sigma^4}{20\pi}h^4 \tag{14}$$
(we used the inequality $\psi(t)\ge 1-\sigma^2t^2/2$ that holds for all t). For S_2 we have
$$S_2 = \frac{(1-c)^2}{\pi}h^{2\gamma-2m-1}\int_h^\infty t^{2m-2\gamma}(1-\psi(t))^2 dt$$
therefore, since we assume that $0<h<1$,
$$c_1\frac{(1-c)^2}{\pi}h^{2\gamma-2m-1} \le S_2 \le c_0\frac{(1-c)^2}{\pi}h^{2\gamma-2m-1}. \tag{15}$$
Finally, for S_3 we have
$$S_3 \le \frac{1}{n}\cdot\frac{1}{2\pi}\int t^{2m}|\psi(ht)|^2 dt = \frac{a_0}{nh^{2m+1}}$$
and
$$S_3 \ge \frac{1}{n}\cdot\frac{1}{2\pi}\int_{|t|\ge 1} t^{2m}(1-(1-c)^2)\psi(ht)^2 dt = \frac{c(2-c)}{nh^{2m+1}}\cdot\frac{1}{2\pi}\int_{|t|\ge h} t^{2m}|\psi(t)|^2 dt \ge$$
$$\ge \frac{c(2-c)}{nh^{2m+1}}\cdot\frac{1}{2\pi}\int_{|t|\ge 1} t^{2m}|\psi(t)|^2 dt = \frac{c(2-c)a_1}{nh^{2m+1}}a_1,$$
i.e.
$$\frac{c(2-c)a_1}{nh^{2m+1}} \le S_3 \le \frac{a_0}{nh^{2m+1}}. \tag{16}$$
From (19)—(21) we obtain that (under conditions $m+1/2<\gamma<m+5/2$ and $0<h<1$)
$$b_1 h^{2\gamma-2m-1} + b_2\frac{1}{nh^{2m+1}} \le \mathrm{MISE}(f_n^{(m)}(x;h)) \le B_1 h^{2\gamma-2m-1} + B_2\frac{1}{nh^{2m+1}}$$

with some positive constants b_1, b_2, B_1, B_2. These inequalities imply that, for some positive constants d_1 and d_2,

$$d_1 n^{-1+(2m+1)/(2\gamma)} \le \inf_{h>0} \mathrm{MISE}(f_n^{(m)}(x;h)) \le d_2 n^{-1+(2m+1)/(2\gamma)}$$

i.e.

$$d_1 n^{-\alpha} \le \inf_{h>0} \mathrm{MISE}(f_n^{(m)}(x;h)) \le d_2 n^{-\alpha}.$$

References

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications.* Monograps on Statistics and Applied Probability. Chapman and Hall, London.

Feller, W. (1971). *An introduction to probability theory and its applications.* Vol. 2, 2d ed. Wiley, New York.

Marron, J.S. and Wand, M.P. (1992) Exact mean integrated squared error. *Ann. Statist.,* Vol. 20, 712-736.

Ramachandran, B. (1997). Characteristic functions with some powers real III. *Statist. Probab. Lett.,* Vol. 34, no. 1, 33-36.

Ushakov, N.G. (1999). *Selected Topics in Characteristic Functions.* VSP, Utrecht.

Ushakov, V.G., Ushakov, N.G. (2000). Some inequalities for characteristic functions of densities with bounded variation. *Moscow Univ. Comput. Math. Cybernet.,* no. 3, 45-52.

Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing.* Chapman and Hall, London.

Watson, G.S. and Leadbetter, M.R. (1963) On the estimation of the probability density, I. *Ann. Math. Statist.,* Vol. 34, 480-491.