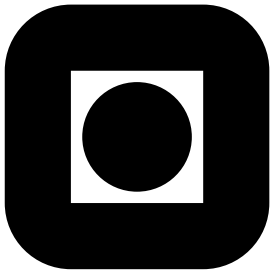
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Some Inequalities for the Mean Integrated Squared Error of Multivariate Kernel Density Estimators

by

Nikolai Ushakov and Vladimir Ushakov

PREPRINT
STATISTICS NO. 5/2009



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2009/S5-2009.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.

Some Inequalities for the Mean Integrated Squared Error of Multivariate Kernel Density Estimators

Nikolai Ushakov
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway
and
Vladimir Ushakov
Department of Mathematical Statistics, Moscow State University, Moscow, Russia

Abstract

Some upper bounds for MISE of multivariate kernel density estimators are obtained. It is shown, in particular, that under some regularity conditions, the actual error is always less than the asymptotic error.

Key words: Density estimation, multivariate kernel estimator, mean integrated squared error, inequalities for characteristic functions, empirical characteristic function

1. Introduction

The most used measure of performance of kernel density estimators as well as the basis of the choice of the smoothing parameter, bandwidth, is the mean intergated squared error (MISE) of the estimator. Practically it is usually replaced by its asymptotic approximation. That is, MISE is represented as the sum of the main term AMISE (asymptotic MISE), having a relatively simple form, and the remainder R such that

$$\text{MISE} \sim \text{AMISE}, \quad \text{R} = o(\text{MISE}) \quad \text{as} \quad n \to \infty,$$

and then the evaluation of the actual error and the bandwidth selection are performed on the basis of AMISE.

Wand and Jones (1995) discovered that, at least for convensional (nonnegative) kernels, AMISE is always strictly greater than MISE. In addition, it turns out (see for example Glad et al., 2007) that the ratio R/MISE can tend to zero very slowly, so that the difference AMISE–MISE can be substantial even for quite large sample sizes ($10^5 - 10^6$). For moderate and small sample sizes this difference is typically so large that it seems to be unreasonable to replace MISE by AMISE. This is corroborated by Figures 2,3 of Section 4. Here are AMISE (solid line) and MISE (dashes) for five different densities (normal mixtures). In Figure 3. they are functions of the sample size when the bandwidth is chosen to be AMISE-optimal. In Figure 2 they are functions of h when the sample size is fixed: $n = 100$. The densities are represented in Figure. Further examples can be found in Marron and Wand (1992).

This makes reasonable to try to find upper bounds for MISE lieing between MISE and AMISE. In the univariate case, a number of such inequalities was obtained in Glad et al. (2007). They can give a substantial gain. For example upper bounds for MISE, given by Theorem 1 of Glad et al. (2007) are represented in Figures 2,3 (dots). In this work, some upper bounds for MISE of multivariate kernel density estimators are derived. It is proved, in particular, that the Maron-Wand inequality (MISE<AMISE) holds also in the multivariate case.

2. Main result

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be independent identically distributed d-dimensional random vectors with density $f(\mathbf{x})$, $\mathbf{x} = (x_1, ..., x_d)^T \in R^d$. Throughout this article, $|\mathbf{A}|$ denotes the determinant of the square matrix $\mathbf{A}$, $\int_{R^d}$ is shorthand for $\int \cdots \int_{R^d}$ and $d\mathbf{x}$ is shorthand for $dx_1 \cdots dx_d$. The general form of the kernel estimator is (Wand and Jones, 1995)

$$f_n(\mathbf{x}; \mathbf{H}) = \frac{1}{n} \sum_{j=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_j),$$

where $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$, $K(\mathbf{x})$ is a multivariate kernel, and $\mathbf{H}$ is a symmetric positive definite $d \times d$ matrix — the bandwidth matrix. In this work we will suppose that $K(\mathbf{x})$ is a symmetric probability density function i.e. it is nonnegative, integrates to one, and $K(\mathbf{x}) = K(-\mathbf{x})$. The mean integrated squared error (MISE) of the estimator is

$$\text{MISE}(f_n(\mathbf{x}; \mathbf{H})) = \text{E} \int_{R^d} [f_n(\mathbf{x}; \mathbf{H}) - f(\mathbf{x})]^2 d\mathbf{x}.$$

Let us introduce the following conditions.

(i) All second derivatives

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

exist and are square integrable.

(ii) The kernel $K(\mathbf{x})$ is square integrable. All second order moments of $K(\mathbf{x})$ are finite.

Denote the covariance matrices of $K(\mathbf{x})$ and $f(x)$ by $\mathbf{\Sigma}_K$ and $\mathbf{\Sigma}_f$, respectively, and entries of the matrix $\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}$ by c_{ij}. Also, let $\mathbf{S}(\mathbf{x})$ be the Hessian matrix of the density $f(\mathbf{x})$, that is the $d \times d$ matrix having (i, j) entry equal to

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}).$$

Finally, denote the trace of the matrix $\mathbf{A}$ by tr$\mathbf{A}$ and the integral of the squared kernel by $R(K)$:

$$R(K) = \int_{R^d} (K(\mathbf{x}))^2 d\mathbf{x}.$$

Theorem 1. *Let conditions* (i) *and* (ii) *be satisfied. Then*

$$\mathrm{MISE}(f_n(\mathbf{x}; \mathbf{H})) < \frac{1}{4} \int_{R^d} \mathrm{tr}^2(\mathbf{H}^{1/2}\mathbf{\Sigma}_K\mathbf{H}^{1/2}\mathbf{S}(\mathbf{x}))d\mathbf{x} + \frac{R(K)}{n|\mathbf{H}|^{1/2}} - \frac{C(d)}{n|\mathbf{\Sigma}|^{1/2}}, \tag{1}$$

where

$$\mathbf{\Sigma} = 2(\mathbf{\Sigma}_f + \mathbf{H}^{1/2}\mathbf{\Sigma}_K\mathbf{H}^{1/2}) \tag{2}$$

and

$$C(d) = [2^{d-1}\pi^{d/2}(d+2)\Gamma(d/2+1)]^{-1}. \tag{3}$$

Proof. Denote characteristic functions of $f(\mathbf{x})$ and $K(\mathbf{x})$ by $\varphi(\mathbf{t})$ and $\psi(\mathbf{t})$ respectively. Let $\varphi_n(\mathbf{t})$ be the empirical characteristic function based on the sample $\mathbf{X}_1, ..., \mathbf{X}_n$, that is

$$\varphi_n(\mathbf{t}) = \frac{1}{n} \sum_{j=1}^{n} e^{i\mathbf{t}^T\mathbf{X_j}}.$$

Then (Parseval identity)

$$\mathrm{MISE}(f_n(\mathbf{x}; \mathbf{H})) = \frac{1}{(2\pi)^d} \mathrm{E} \int_{R^d} [\varphi_n(\mathbf{t})\psi(\mathbf{H}^{1/2}\mathbf{t}) - \varphi(\mathbf{t})]^2 d\mathbf{t}.$$

Transforming the right hand side and taking into account that $\mathrm{E}\varphi_n(\mathbf{t}) = \varphi(\mathbf{t})$ and

$$\mathrm{E}|\varphi_n(\mathbf{t})|^2 = \frac{1}{n} + \left(1 - \frac{1}{n}|\varphi(\mathbf{t})|\right),$$

we obtain the following representation

$$\mathrm{MISE}(f_n(\mathbf{x}; \mathbf{H})) = \frac{1}{(2\pi)^d} \int_{R^d} |\varphi(\mathbf{t})|^2 (1 - \psi(\mathbf{H}^{1/2}\mathbf{t}))^2 d\mathbf{t} +$$

$$+ \frac{1}{n} \cdot \frac{1}{(2\pi)^d} \int_{R^d} |\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t} + \frac{1}{n} \cdot \frac{1}{(2\pi)^d} \int_{R^d} |\varphi(\mathbf{t})\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t}.$$

Denote

$$S_1 = \frac{1}{(2\pi)^d}\int_{R^d} |\varphi(\mathbf{t})|^2(1-\psi(\mathbf{H}^{1/2}\mathbf{t}))^2 d\mathbf{t},$$

$$S_2 = \frac{1}{(2\pi)^d}\int_{R^d} |\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t},$$

$$S_3 = \frac{1}{(2\pi)^d}\int_{R^d} |\varphi(\mathbf{t})\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t}.$$

We derive now upper estimates for S_1 and S_2 and lower estimate for S_3.

The covariance matrix of the distribution, corresponding to the characteristic function $\psi(\mathbf{H}^{1/2}\mathbf{t})$, is $\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}$, therefore (Ushakov, 1999, Theorem 2.7.8)

$$(1-\psi(\mathbf{H}^{1/2}\mathbf{t}))^2 \le \frac{1}{4}\Big(\mathbf{t}^T\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}\mathbf{t}\Big)^2.$$

From this inequality, using the Plancherel formula, we obtain

$$S_1 \le \frac{1}{4}\cdot\frac{1}{(2\pi)^d}\int_{R^d}(\mathbf{t}^T\mathbf{H}^{1/2}\mathbf{\Sigma}\mathbf{H}^{1/2}\mathbf{t})^2|\varphi(\mathbf{t})|^2 d\mathbf{t} =$$

$$= \frac{1}{4}\cdot\frac{1}{(2\pi)^d}\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d\sum_{l=1}^d c_{ij}c_{kl}\int_{R^d} t_it_jt_kt_l|\varphi(\mathbf{t})|^2 d\mathbf{t} =$$

$$= \frac{1}{4}\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d\sum_{l=1}^d c_{ij}c_{kl}\int_{R^d}\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i\partial x_j}\cdot\frac{\partial^2 f(\mathbf{x})}{\partial x_k\partial x_l}\right)d\mathbf{x} =$$

$$= \frac{1}{4}\int_{R^d}\left(\sum_{j=1}^d\sum_{k=1}^d c_{jk}\frac{\partial^2 f(\mathbf{x})}{\partial x_j\partial x_k}\right)^2 d\mathbf{x} = \frac{1}{4}\int_{R^d}\mathrm{tr}^2(\mathbf{H}^{1/2}\Sigma\mathbf{H}^{1/2}\mathbf{S}(\mathbf{x}))d\mathbf{x}.$$

Further,

$$S_2 = \frac{1}{(2\pi)^d}\int_{R^d}|\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t} = |\mathbf{H}|^{-1/2}\frac{1}{(2\pi)^d}\int_{R^d}|\psi(\mathbf{t})|^2 d\mathbf{t} =$$

$$= |\mathbf{H}|^{-1/2}\int_{R^d}(K(\mathbf{x}))^2 d\mathbf{x} = \frac{1}{n}|\mathbf{H}|^{-1/2}R(K).$$

Finally, the covariance matrix of the distribution, corresponding to the characteristic function $|\varphi(\mathbf{t})\psi(\mathbf{H}^{1/2}\mathbf{t})|^2$, is $\mathbf{\Sigma} = 2(\mathbf{\Sigma}_f + \mathbf{H}^{1/2}\mathbf{\Sigma}_K\mathbf{H}^{1/2})$ therefore, using again Theorem 2.7.8 from Ushakov (1999), obtain

$$S_3 \ge \frac{1}{(2\pi)^d}\int_{\{\mathbf{t}:\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}}|\varphi(\mathbf{t})\psi(\mathbf{H}^{1/2}\mathbf{t})|^2 d\mathbf{t} \ge \int_{\{\mathbf{t}:\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}}(1-\mathbf{t}^T\mathbf{\Sigma}\mathbf{t})d\mathbf{t} =$$

$$= \frac{[2^{d-1}\pi^{d/2}(d+2)\Gamma(d/2+1)]^{-1}}{|\mathbf{\Sigma}|^{1/2}}.$$

Now (1) follows from the obtained bounds for S_1, S_2 and S_3.

3. Some special cases

Suppose now that one more condition is satisfied.

(iii)

$$\int_{R^d} \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \mu_2(K)\mathbf{I}$$

where $\mu_2(K) = \int_{R^d} x_j^2 K(\mathbf{x})d\mathbf{x}$ is independent of j, and $\mathbf{I}$ is the $d \times d$ identity matrix.

Under this condition $\mathbf{H}^{1/2}\mathbf{\Sigma}_K\mathbf{H}^{1/2} = \mu_2(K)\mathbf{H}$, and from Theorem 1 we obtain the following

Theorem 2. *Let conditions* (i), (ii) *and* (iii) *be satisfied. Then*

$$\mathrm{MISE}(f_n(\mathbf{x};\mathbf{H})) < \frac{1}{4}\mu_2(K)^2 \int_{R^d} \mathrm{tr}^2(\mathbf{HS})d\mathbf{x} + \frac{R(K)}{n|\mathbf{H}|^{1/2}} - \frac{C(d)}{n|\mathbf{\Sigma}|^{1/2}}, \tag{4}$$

where $\mathbf{\Sigma} = 2(\mathbf{\Sigma}_f + \mu_2(K)\mathbf{H})$ *and* $C(d)$ *is given by (3).*

Corollary. *Let conditions* (i), (ii) *and* (iii) *be satisfied, and the bandwidth matrix depends on the sample size* n *in such a way that* $n^{-1}|\mathbf{H}|^{-1/2}$ *and all entries of the matrix* $\mathbf{H}$ *tend to zero as* $n \to \infty$. *Then*

$$\mathrm{MISE}(f_n(\mathbf{x};\mathbf{H})) < \mathrm{AMISE}(f_n(\mathbf{x};\mathbf{H})). \tag{5}$$

Since under conditions of the Corollary,

$$\mathrm{AMISE}(f_n(\mathbf{x};\mathbf{H})) = \frac{1}{4}\mu_2(K)^2 \int_{R^d} \mathrm{tr}^2(\mathbf{HS})d\mathbf{x} + \frac{R(K)}{n|\mathbf{H}|^{1/2}}$$

(Wand and Jones, 1995, p. 97), (5) immediately follows from (4).

In conclusion we consider two special, perhaps the most frequently used, smoothing parametrizations. In the first case the estimator has form

$$f_n(\mathbf{x};h_1,...,h_d) = \frac{1}{nh_1\cdots h_d}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1},...,\frac{x_d - X_{id}}{h_d}\right) \tag{6}$$

where $\mathbf{X}_i = (X_{i1},...,X_{id})^T$. In this case h_j can be considered as the smoothing parameter associated with the j-th coordinate direction. For this parametrization we will use only kernels with a diagonal covariance matrix. Denote

$$\mu_2^{(j)}(K) = \int_{R^d} x_j^2 K(\mathbf{x})d\mathbf{x}, \quad \sigma_j^2 = \mathrm{Var}X_{ij}, \quad j = 1,...,d.$$

Theorem 3. *Let conditions* (i) *and* (ii) *be satisfied, the estimator have form* (6), *and* $\mathbf{\Sigma}_K$ *be a diagonal matrix. Then*

$$\mathrm{MISE}(f_n(\mathbf{x};h_1,...,h_d)) < \frac{1}{4}\int_{R^d}\left(\sum_{j=1}^{d} h_j^2\mu_2^{(j)}(K)\frac{\partial^2 f(\mathbf{x})}{\partial x_j^2}\right)^2 d\mathbf{x} +$$

$$+\frac{R(K)}{n}\left(\prod_{j=1}^{d} h_j\right)^{-\frac{1}{2}} - \frac{C(d)}{n\sqrt{2}}\left(\prod_{j=1}^{d}(\sigma_j^2 + \mu_2^{(j)}(K)h_j)\right)^{-\frac{1}{2}}, \tag{7}$$

where $C(d)$ *is defined by (3).*

Proof. It is easy to see that under conditions of the theorem, the first two summands in the right hand side of (7) coinside with the first two summands in the right hand side of (1), therefore to prove (7) it is sufficient to show that

$$|\mathbf{\Sigma}| \leq 2 \prod_{j=1}^{d} (\sigma_j^2 + \mu_2^{(j)}(K) h_j), \tag{8}$$

where $\mathbf{\Sigma} = 2(\mathbf{\Sigma}_f + \mathbf{H}^{1/2} \mathbf{\Sigma}_K \mathbf{H}^{1/2})$. Note that under assumptions we made, $\mathbf{H}^{1/2} \mathbf{\Sigma}_K \mathbf{H}^{1/2}$ is the diagonal matrix with diagonal elements $\mu_2^{(j)}(K) h_j$, $j = 1, ..., d$, therefore diagonal elements of $\mathbf{\Sigma}_f + \mathbf{H}^{1/2} \mathbf{\Sigma}_K \mathbf{H}^{1/2}$ are $\sigma_j^2 + \mu_2^{(j)}(K) h_j$, and (8) follows since for a nonnegative definite matrix, the determinant is less than or equal to the product of diagonal elements (see for example Bellman, 1960).

Finally consider the following simplest parametrization. There is a single scalar smoothing parameter h and the kernel estimator is of the form

$$f_n(\mathbf{x}; h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_j}{h} \right). \tag{9}$$

In this case we will use only kernels satisfying condition (iii)

Theorem 4. *Let conditions* (i), (ii), (iii) *be satisfied, and the estimator have form* (9). *Then*

$$\mathrm{MISE}(f_n(\mathbf{x}; h)) < \frac{1}{4} h^4 \mu_2(K)^2 \int_{R^d} (\bigtriangledown^2 f(\mathbf{x}))^2 d\mathbf{x} + \frac{R(K)}{nh^d} - \frac{C(d)}{n\sqrt{2}} \left(\prod_{j=1}^{d} (\sigma_j^2 + \mu_2(K) h) \right)^{-\frac{1}{2}}$$

where

$$\bigtriangledown^2 f(\mathbf{x}) = \sum_{j=1}^{d} \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2}.$$

The theorem follows from Theorem 3.

4. Examples

In the one-dimensional case ($d = 1$) the result given by Theorem 4 of this work is a little inferior to the result given by Theorem 1 of Glad et al. (2007). However the upper bound of Theorem 4 is still a substantial improvement of AMISE. In this section we present several examples. MISE, AMISE, the upper bound of Theorem 1 of Glad et al. (2007) (UB1), and the upper bound of Theorem 4 of this work (UB2) are calculated for five different densities. The densities are

#1. Normal

$$N(0, 1).$$

#2. Plateau

$$\frac{1}{2}(N(-1, 1) + N(1, 1)).$$

#3. Symmetric bimodal

$$\frac{1}{2}(N(-1.5, 1) + N(1.5, 1)).$$

#4. Asymmetric bimodal

$$0.3N(-1.5,1)+0.7N(1.5,1).$$

#5. Kurtotic

$$\frac{1}{2}(N(0,0.1)+N(0,3)).$$

These five densities are represented in Figure 1. Results are presented in Figures 2,3. In Figure 3 AMISE (solid line), MISE (dashes), UB1 (dots), UB2 (dashes-dots) are functions of the sample size n while h is chosen to be AMISE-optimal. In Figure 2 they are functions of h while n is fixed: $n = 100$.

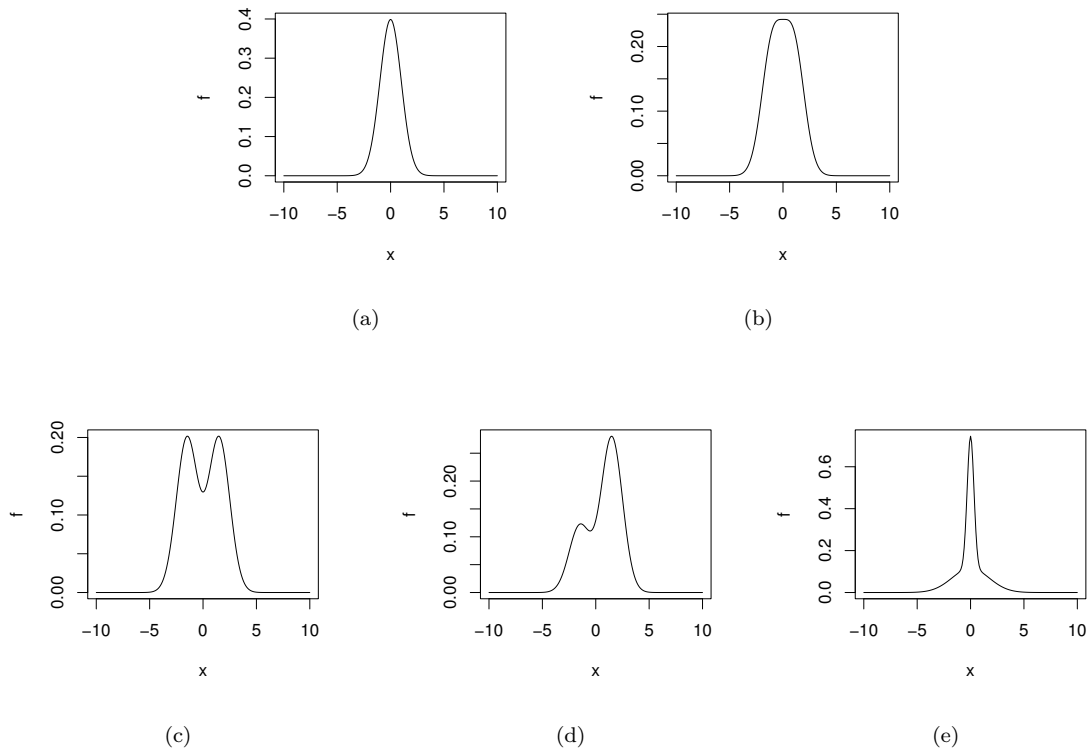
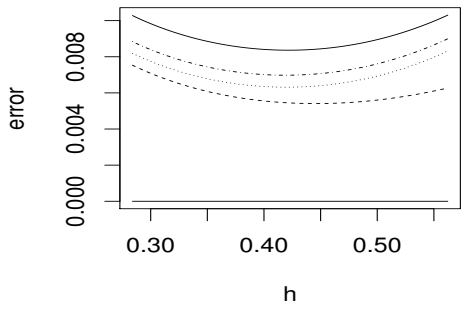


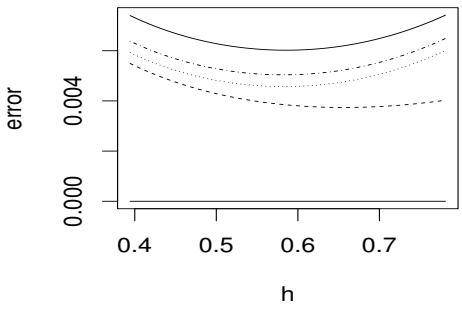
Figure 1: The densities.

5. Conclusions

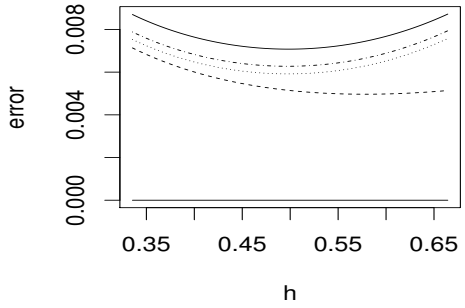
The main (and the most difficult) problem in kernel density estimation is the choice of the smoothing parameter, bandwidth. Upper bounds, contained in Glad et al. (2007) and in this work, give new resources for solving this problem. In bandwidth selection MISE is usually replaced by AMISE; which then is approximately (since it contains parameters of the unknown density) minimized. But any function, situated between AMISE and one of the obtained upper bounds, is a better approximation of MISE than AMISE is. One can choose therefore a curve from this stripe, having desirable properties, and use it for the selection of the bandwidth. For example, it is known that always $h_{\mathrm{AMISE}} < h_{\mathrm{MISE}}$,



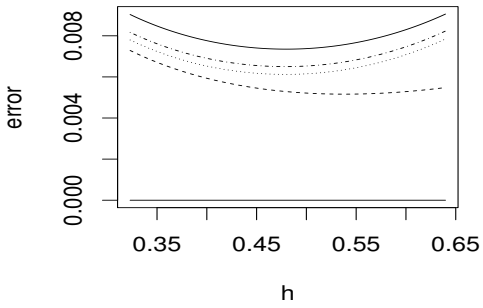
(a)



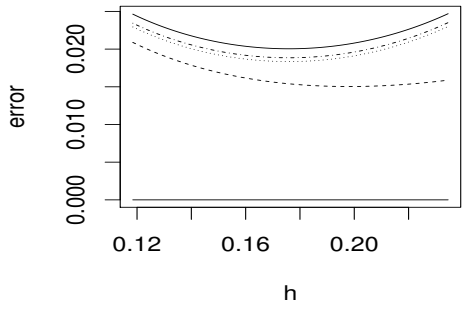
(b)



(c)



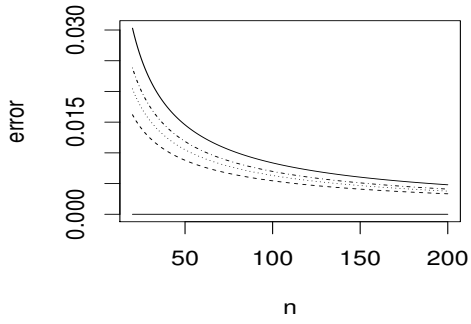
(d)



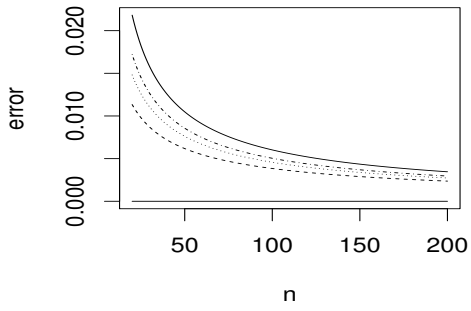
(e)

Figure 2: The results.

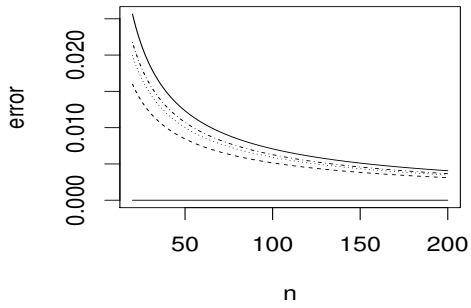
where h_{AMISE} and h_{MISE} are AMISE-optimal and MISE-optimal values of the smoothing parameter, respectively, see Marron and Wand (1992). It is sensible therefore to choose a function from the stripe, which, on the one hand is as simple as AMISE and, on the other hand, has its minimum to the right of h_{AMISE}. A separate work will be devoted to investigation of these potentialities.



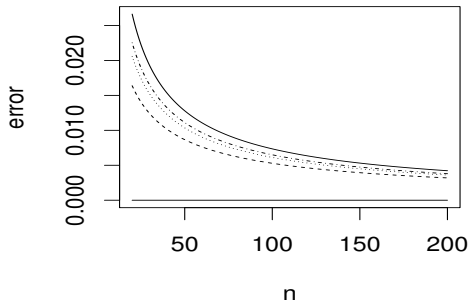
(a)



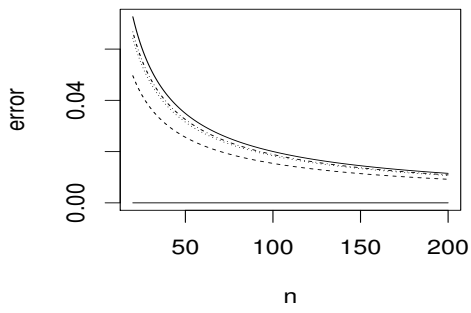
(b)



(c)



(d)



(e)

Figure 3: The results

References

Bellman R. (1960) *Introduction to matrix analysis.* McGraw-Hill book company, New York.

Glad I.K., Hjort N.L. and Ushakov N.G. (2007). Mean-squared error of kernel estimators for finite values of the sample size. *J. Math. Sci. (N. Y.)*, Vol. 146, no. 4, 5977-5983.

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

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

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