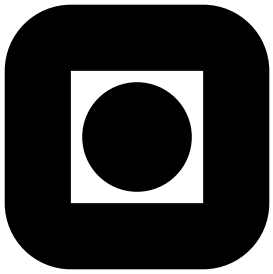
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Calculation of Lp errors of kernel density estimators

by

H. Kile and N.G. Ushakov

PREPRINT
STATISTICS NO. 10/2010



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2010/S10-2010.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.

Calculation of L^p errors of kernel density estimators

Håkon Kile and Nikolai Ushakov
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Abstract

Explicit formulas for calculating L^p and integrated L^p errors of kernel density estimators are obtained. Numerical realisation is discussed. Some practical recommendations are given.

Key words: Kernel estimation; Mean absolute error; Mean integrated absolute error; Characteristic function

1. Introduction

Let $X_1, ..., X_n$ be independent and identically distributed random variables from an absolutely continuous distribution with probability density $f(x)$. The kernel density estimator of $f(x)$ is defined as

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

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

Construction of estimators and analysis of their properties are based on some error criteria. One of the main three approaches is used: asymptotic analysis, simulation or numerical calculation (the three approaches are discussed in particular in Marron and Wand (1992)). In case of kernel density estimation, the overwhelming majority of works uses the mean squared error (MSE) and mean integrated squared error (MISE) as error criteria. This is primarily because of their mathematical simplicity compared for example with the absolute, or other L^p, errors. MSE and MISE are defined as

$$\mathrm{MSE}(f_n(x;h)) = \mathrm{E}[f_n(x;h) - f(x)]^2$$

and

$$\mathrm{MISE}(f_n(\cdot;h)) = \mathrm{E}\int_{-\infty}^{\infty}[f_n(x;h) - f(x)]^2 dx$$

and are represented in terms of $f(x)$ and $K(x)$ as follows (see for example Wand and Jones (1995))

$$\mathrm{MSE}(f_n(x;h)) = \frac{1}{n}[(K_h^2 * f)(x) - (K_h * f)^2(x)] + [(K_h * f)(x) - f(x)]^2, \tag{1}$$

$$\mathrm{MISE}(f_n(\cdot;h)) = \frac{1}{n}\int_{-\infty}^{\infty}[(K_h^2 * f)(x) - (K_h * f)^2(x)]dx + \int_{-\infty}^{\infty}[(K_h * f)(x) - f(x)]^2 dx, \tag{2}$$

where $K_h(x) = h^{-1}K(x/h)$ and $*$ denotes the convolution. The presence of formulas (1) and (2) makes it possible to effectively use asymptotic analysis and numerical calculation. The absence of such formulas for other L^p errors leads to that practically only simulation can be used (for L^1 is asymptotic analysis also possible). The goal of this paper is to (at least partially) fill this gap. We derive explicit formulas for the mean L^p and mean integrated L^p errors in terms of $K(x)$ and $f(x)$.

Mean L^p error (MLPE) and mean integrated L^p error (MILPE) are defined as

$$\mathrm{MLPE}(f_n(x;h)) = \mathrm{E}\,|f_n(x;h) - f(x)|^p$$

and

$$\mathrm{MILPE}(f_n(\cdot;h)) = \mathrm{E}\int_{-\infty}^{\infty}|f_n(x;h) - f(x)|^p dx,$$

respectively. In this work we consider only $0 < p < 2$.

2. Main Results

Let $\Re z$ denote the real part of the complex number z and let $i = \sqrt{-1}$.

Theorem 1.

$$\mathrm{MLPE}(f_n(x;h)) =$$
$$= C(p)\int_{-\infty}^{\infty}\frac{1}{|t|^{p+1}}\left[1-\Re\left(e^{-itf(x)}\left(\int_{-\infty}^{\infty}\exp\left(\frac{it}{nh}K\left(\frac{x-y}{h}\right)\right)f(y)dy\right)^n\right)\right]dt \tag{3}$$

where

$$C(p) = -\frac{\Gamma(p+1)\cos((p+1)\pi/2)}{\pi}. \tag{4}$$

Theorem 2.

$$\mathrm{MILPE}(f_n(\cdot;h)) =$$
$$= C(p)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{|t|^{p+1}}\left[1-\Re\left(e^{-itf(x)}\left(\int_{-\infty}^{\infty}\exp\left(\frac{it}{nh}K\left(\frac{x-y}{h}\right)\right)f(y)dy\right)^n\right)\right]dtdx \tag{5}$$

where $C(p)$ is given by (4).

Proofs of the theorems are based on the following

Lemma 1. *Let Y be a random variable with characteristic function $\phi(t)$. If $\mathrm{E}\,|Y|^p < \infty$, $0 < p < 2$, then*

$$\mathrm{E}\,|Y|^p = C(p)\int_{-\infty}^{\infty}\frac{1}{|t|^{p+1}}(1-\Re\phi(t))dt \tag{6}$$

where $C(p)$ is given by (4).

Proof. Let $0 < p < 2$. Denote the distribution function of Y by $G(y)$. Since

$$\int_{-\infty}^{\infty}\frac{1-\cos t}{|t|^{p+1}}dt = -\frac{\pi}{\Gamma(p+1)\cos((p+1)\pi/2)} = \frac{1}{C(p)},$$

it is easy to see that

$$C(p)\int_{-\infty}^{\infty}\frac{1-\cos(yt)}{|t|^{p+1}}dt = |y|^p.$$

Hence

$$\mathrm{E}\,|Y|^p = C(p)\int_{-\infty}^{\infty}|y|^p dG(y) = C(p)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1-\cos(yt)}{|t|^{p+1}}dtdG(y)$$
$$= C(p)\int_{-\infty}^{\infty}\frac{1}{|t|^{p+1}}\left(\int_{-\infty}^{\infty}(1-\cos(yt))dG(y)\right)dt = C(p)\int_{-\infty}^{\infty}\frac{1}{|t|^{p+1}}(1-\Re\phi(t))dt.$$

□

Proof of Theorem 1. Let x be fixed. Consider the random variable $Y = Y(x) = f_n(x;h) - f(x)$. Since

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

the characteristic function of Y is

$$\phi_Y(t) = \mathrm{E}\exp\left[it\left(f_n(x;h) - f(x)\right)\right] = e^{-itf(x)}\,\mathrm{E}\exp\left[\frac{it}{nh}\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)\right]$$
$$= e^{-itf(x)}\prod_{j=1}^{n}\mathrm{E}\exp\left[\frac{it}{nh}K\left(\frac{x-X_j}{h}\right)\right] = e^{-itf(x)}\left(\int_{-\infty}^{\infty}\exp\left[\frac{it}{nh}K\left(\frac{x-y}{h}\right)\right]f(y)dy\right)^n.$$

The result now follows from Lemma 1. □

Theorem 2 follows from Theorem 1 by integration with respect to x.

3. Numerical realisation

The calculation of the right hand side of (3) has some specific features which should be taken into account. In this section we briefly consider the problem of numerical realisation and give some practical recommendations. Calculation of MLPE consists of two numerical integrations, with respect to y and t.

Integration limits

The two integrals are from $-\infty$ to ∞, so the first question arising in the numerical integration, is the choice of integration intervals. For this choice it is useful to take the following into account. In the inner integral (with respect to y) the absolute value of the function under the integral sign does not exceed $f(y)$. In the external integral (with respect to t) the function decreases quite slowly, and for large values of t it practically coincides with $|t|^{-p-1}$. Thus the integration intervals should be chosen in such a way that, according to the desirable accuracy, the functions $f(y)$ and $|t|^{-p-1}$ can be considered as negligible outside their respective intervals.

Oscillations

The magnitude of t/n determines the frequency of the trigonometric functions

$$\cos\left(\frac{t}{n}K_h(x-y)\right) \quad \text{and} \quad \sin\left(\frac{t}{n}K_h(x-y)\right)$$

in (3) (here $K_h(x) = h^{-1}K(x/h)$). If $t \gg n$, the trigonometric functions oscillate with a high frequency and this has to be taken into account, in particular in the choice of number of integration points. Generally we recommend to use Composite Simpson's Rule for the evaluation of

$$\int_{-\infty}^{\infty}\left[\cos\left(\frac{t}{n}K_h(x-y)\right)+i\sin\left(\frac{t}{n}K_h(x-y)\right)\right]f(y)dy$$

Integration near 0

Consider (6). When t is close to 0, both the numerator and denominator under the integral sign are close to 0. Let $\mu_2 = \mathrm{E}\,Y^2 < \infty$. Then

$$\Re\phi(t) = 1 - \frac{\mu_2 t^2}{2} + o(t^2), \quad t \to 0,$$

and we can approximate the integral over a small neighbourhood of 0 by

$$\int_{-\varepsilon}^{\varepsilon}\frac{1-\Re\phi(t)}{|t|^{p+1}}dt \approx \frac{\mu_2}{2-p}\varepsilon^{2-p}.$$

Using inequalities

$$1 - \frac{\mu_2 t^2}{2} \le \Re\phi(t) \le 1 - \frac{\mu_2 t^2}{2} + \frac{\mu_4 t^4}{4!}$$

where $\mu_4 = \mathrm{E}\,Y^4$ (see for example Ushakov (1999)), we can obtain an upper bound for the error of this approximation:

$$\left|\int_{-\varepsilon}^{\varepsilon}\frac{1-\Re\phi(t)}{|t|^{p+1}}dt - \frac{\mu_2}{2-p}\varepsilon^{2-p}\right| \le \frac{\mu_4}{12(4-p)}\varepsilon^{4-p}.$$

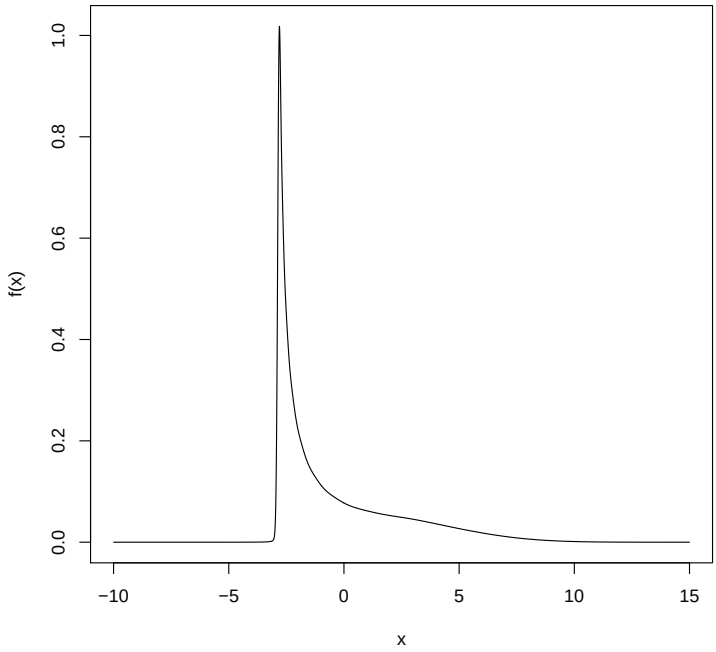


Figure 1: The probability density in our example.

An example

In conclusion we give an example of calculation of mean integrated L^p error for $p = \{0.5, 1, 1.5\}$. The probability density $f(x)$ is a normal mixture density, specifically constructed to approximate the shape of a gamma density, with shape parameter less than 1. The probability density is illustrated in figure 1. The corresponding MILPE's are plotted in figure 2, where the vertical lines are their respective minimisers.

References

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.

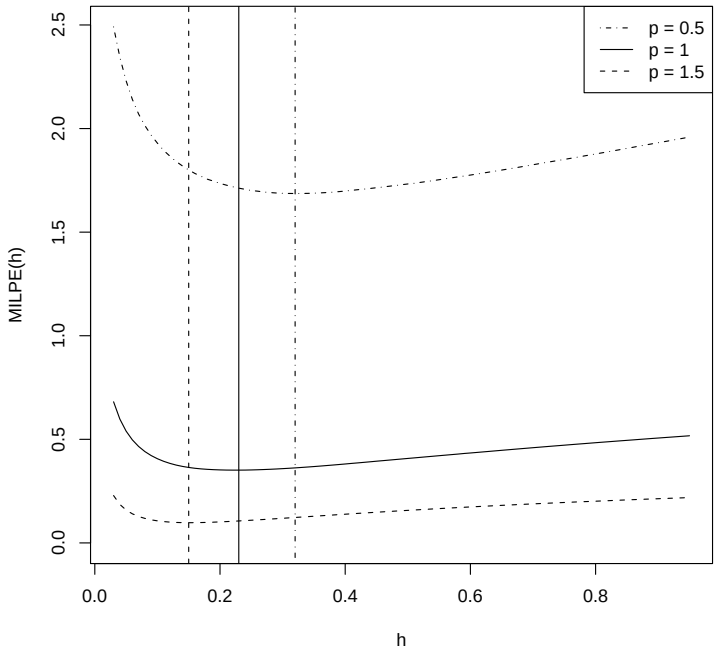


Figure 2: MILPE's for the probability density in figure 1. The vertical lines are their respective minimisers.