QUANTITATIVE METHODS IN ECONOMICS
Vol. XVI, No. 1, 2015, pp. 7 – 16

BIAS REDUCTION IN KERNEL ESTIMATOR OF DENSITY FUNCTION IN BOUNDARY REGION

Aleksandra Baszczyńska
Department of Statistical Methods
University of Lodz
e-mail: albasz@uni.lodz.pl

Abstract: The properties of the classical kernel estimator of density function deteriorate when the support of density function is bounded. The use of classical form of kernel estimator causes the increase of the bias estimator, particularly in the so-called boundary region, close to end of support. It can also lead to undesirable situation where density function estimator has a different support than the density function. The paper presents selected bias reduction procedures, such as reflection method and its modification. An example is presented with an attempt to compare considered procedures.

Keywords: kernel estimator, density function, bias reduction, reflection method

INTRODUCTION

When the density function satisfies certain smoothness criteria (e.g. existing and being continuous of the density derivatives of appropriate orders over the entire real line), the kernel density estimator is characterized by some useful properties, such as: unnecessity of assuming that density belongs to a parametric family of distributions, its calculation is easy and it is asymptotically unbiased and is consistent estimator of unknown density function. The problems may arise for users when these smoothness conditions are not fulfilled, as in the case of some commonly known densities. E.g. when the density function of exponential distribution is being estimated, the kernel estimator is trying to estimate relatively high density for positive values of random variable, whereas for negative values the estimator is aiming to estimate zero. The discontinuity in the function results in the bias increasing of the estimator [Wand, Jones 1995].

The next situation when the properties of the kernel density estimator deteriorates is the bounded domain of definition of a density being estimated. In practical problems such a situation occurs often as many random variables considered in the problems of economic, technical or natural sciences are characterized by bounded support on one or both sides. In most situations left boundary equals zero when the data under consideration are measurements of positive quantities. In different analyses random variables with non-negative values are considered (duration of unemployment, the stock price, time of performing specific technical operation, the amount of inventory in the warehouse, time of growing plants, amount of atmospheric fall). The use of classical form of kernel estimator causes the increase of the bias estimator, particularly in the so-called boundary region, close to end of support. It is possible both when the kernel function is unbounded, and when the kernel function is bounded but partially is ejected out of the density function support. It can also lead to undesirable situation where density function estimator has a different support than the density function [c.f. Jones 1993]. Moreover, in presentation of the data for which the estimation is giving, the situation when any weight is assigned to the negative numbers is treated as unacceptable [Silverman 1986].

Modification of classical kernel estimator is needed to improve the estimator properties. It should be used especially in the situation when the integral of the kernel estimator is not 1 in appropriate support or estimator is not consistent for some observations.

Let density function f be continuous on interval $[0,\infty)$ and be 0 for $x<0$. For smoothing parameter h: interval $[0,h)$ is called boundary region and interval $[h,\infty)$ is interior region.

Note that for interior region it is possible to use the classical form of kernel density estimator. For boundary region information interval $[x-h,x+h]$ may locate outside the support what may cause that some of the observations are not used in construction of the density estimator [Albers 2012]. Estimation is based on reduced information, the bias is large resulting in poor estimation.

CLASSICAL KERNEL DENSITY ESTIMATOR

Function $K_{\nu,k}$ with support $[-1,1]$ is defined as kernel function of degree (ν,k), for $\nu \le k-1$ ($\nu,k \in N$), if it fulfils the following property [c.f. Horová et al. 2012]:

$$\int_{-1}^{1} x^j K_{\nu,k}(x)dx = \begin{cases} 0 \quad \text{for} \quad 0 \le j \le k-1,\ j \ne \nu, \\ (-1)^{\nu}\nu! \quad \text{for} \quad j = \nu, \\ \kappa_k \ne 0 \quad \text{for} \quad j = k, \end{cases} \tag{1}$$

where κ_k is kth moment of the kernel $K_{\nu,k}$.

For $\nu = 0$ and $k = 2$ kernel function $K_{0,2}(\cdot)$ is symmetric function around zero and $\int_{-1}^{1} K_{0,2}(x)dx = 1$. Any density function with support $[-1,1]$ with mean zero is kernel function of degree $(0,2)$ and in most cases they are used in construction of classical kernel density estimators.

Density kernel estimator based on sample $X_1, X_2, ..., X_n$ with kernel $K_{0,2}$ symmetric around zero with support $[-1,1]$ can be written as [Wand, Jones 1995], [Silverman 1996], [Domański et al. 2014]:

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K_{0,2}\left(\frac{x - X_i}{h_n}\right), \tag{2}$$

where h_n is a smoothing parameter, such as $h_n > 0$, $h_n = h(n) : \{h(n)\}$, $h_n \underset{n\to\infty}{\to} 0$, $h_n n \underset{n\to\infty}{\to} \infty$.

Classical kernel density estimator is consistent for continuous $f(x)$ and for $h_n > 0$, $h_n \underset{n\to\infty}{\to} 0$ and $h_n n \underset{n\to\infty}{\to} \infty$. Moreover, it is a density function (is nonnegative and integrates to 1).

Kernel estimator of the ν-derivative of density function (assuming that derivatives exist and are continuous) is:

$$\hat{f}_n^{(\nu)}(x) = \frac{1}{nh_n^{\nu+1}}\sum_{i=1}^{n} K_{\nu,k}\left(\frac{x - X_i}{h_n}\right), \tag{3}$$

where $K_{\nu,k}$ is kernel function of degree (ν,k). For $\nu = 0$, $k = 2$ one can get (2).

Kernel estimator of the ν-derivative of density function for the appropriate kernel function is consistent in points of continuity of derivative.

BIAS REDUCTION OF KERNEL DENSITY ESTIMATOR

Let:

- $X_1, X_2, ..., X_n$ be a random sample drawn from a population with random variable X with density function f with support $[0,\infty)$ ($f(x)=0$ for $x<0$ and $f(x)>0$ for $x \geq 0$);
- $f^{(2)}$ be a second derivative of density function which is continuous away from $x=0$;
- function $K_{0,2}(\cdot)$ be symmetric and smooth kernel function of degree $(0,2)$ with support $[-1,1]$;
- $\hat{f}_n(x)$ be the kernel density function (2) with the smoothing parameter h_n.

Boundary behavior of the kernel estimator can be observed taking into regard its asymptotic properties at a sequence of points which is within one bandwidth of the boundary. Taking $x = ch_n$ for $c \in [0,1)$, kernel density estimator for point x, is defined as: $\hat{f}_n(ch_n) = \frac{1}{nh_n}\sum_{i=1}^{n} K_{0,2}\left(\frac{ch_n - X_i}{h_n}\right)$. For $x > h_n\,(c>1)$, kernel estimator of density function is asymptotically unbiased and consistent. Its expected value is the following:

$$E\left[\hat{f}_n(x)\right] = \frac{1}{nh_n}\sum_{i=1}^{n} E\left[K_{0,2}\left(\frac{X_i - x}{h_n}\right)\right] = f(x) + \frac{1}{2!} f^{(2)}(x) h^2 \kappa_2 + o\left(h^2\right), \qquad (4)$$

where κ_2 is defined in (1). For $0 \leq c \leq 1$ when $\int_{-1}^{c} K_{0,2}(u)du \neq 1$ in general, kernel estimator of density function is not consistent. Its expected value is:

$$E\left[\hat{f}_n(x)\right] = f(x)\int_{-1}^{c} K_{0,2}(u)du + o(1). \qquad (5)$$

It is possible to use an appropriate modification of the kernel estimator in the vicinity of the known boundary. It results in a family of boundary kernels $K_{0,2}^{L}(u,c)$ and the achieving $O\left(h^2\right)$ bias is possible. For different kernel functions and different values of c kernel density estimators based on kernel function from a family of boundary kernels improve the performance of estimator in the boundary region [Wand, Jones 1995].

Simple method used in bias reduction of kernel estimator is based on the estimator calculation only for positive values ignoring the boundary region and then setting kernel estimator to zero for negative values. It causes that the estimator

is zero for negative values but on the other hand the integral of the estimator is not 1 [Jones, Foster 1996].

Another approach uses the reflections of all the points in the boundary that results in a set $\{X_1,-X_1,X_2,-X_2,...\}$. Under the assumption that kernel function is symmetric and differentiable, the resulting estimator has zero derivative at the boundary.

This reflection method can be used directly in the kernel estimator by using appropriate modification of the kernel function outside the interval $[0,\infty)$, for example, symmetric reflection about zero, where parts of kernel function outside $[0,\infty)$ are deleted and next placed in the neighbour of zero in interval $[0,\infty)$.

Kernel estimator using reflection method is the following [Kulczycki 2005]:

$$\hat{f}_{nR}(x)=\frac{1}{nh_n}\sum_{i=1}^{n}\left[K_{0,2}\left(\frac{x-X_i}{h_n}\right)+K_{0,2}\left(\frac{x+X_i}{h_n}\right)\right]. \tag{6}$$

Estimator (6) is consistent estimator of function f but for x close to zero the bias is $O(h)$.

The Karunamuni and Alberts generalized reflection method improves the bias with low variance. The generalized reflection estimator is [Karunamuni, Alberts 2005]:

$$\hat{f}_{nGR}(x)=\frac{1}{nh_n}\sum_{i=1}^{n}\left[K\left(\frac{x-g_1(X_i)}{h_n}\right)+K\left(\frac{x+g_2(X_i)}{h_n}\right)\right], \tag{7}$$

where g_1 and g_2 are some transformation functions (e.g. cubic polynomials with coefficients ensuring criteria for the order of estimators $O(h^2)$).

SIMULATION STUDY

The simulation study was conducted to analyze the properties of chosen methods of the bias reduction of kernel density estimator.

The populations with density functions of bounded support $[0,\infty)$ were taken into consideration, particularly populations of two-parameters Weilbull distribution $W(0,\delta,\gamma)$, where δ is a scale parameter and γ is a shape parameter. The populations were regarded with the following parameters:

W1: $\delta=1,\gamma=0.1$,

W2: $\delta=1,\gamma=0.5$,

W3: $\delta=1,\gamma=1$,

W4: $\delta=1,\gamma=2$ (Rayleigh distribution),

W5: $\delta = 1, \gamma = 3.4$,
W6: $\delta = 1, \gamma = 5$,
W7: $\delta = 4, \gamma = 1$,
W8: $\delta = 4, \gamma = 2$.

The parameters of Weilbull distributions were chosen in such a way that it is possible to analyze the broad range of distributions with bounded supports. The populations are heterogeneous looking from e.g. measure of location, spread or asymmetry.

To extend the study and indicate the area of application of regarded methods, one more population was considered, the measure of agricultural productivity – agriculture value added per worker for countries in 2013. Data are in constant 2005 U.S. dollars. Source of the data is:
http://data.worldbank.org/indicator/EA.PRD.AGRI.KD [18.06.2015].

From each population the samples were chosen where $n = 10,20,...,100$. For each sample, the classical kernel density estimator and kernel density estimator with reflection were calculated using Gaussian kernel function and the reference rule or biased cross validation (in the case of W1) as the most popular methods of choosing the smoothing parameter. The chosen descriptive statistics calculated for samples ($n = 50$) from populations W1-W8 are presented in Table 1.

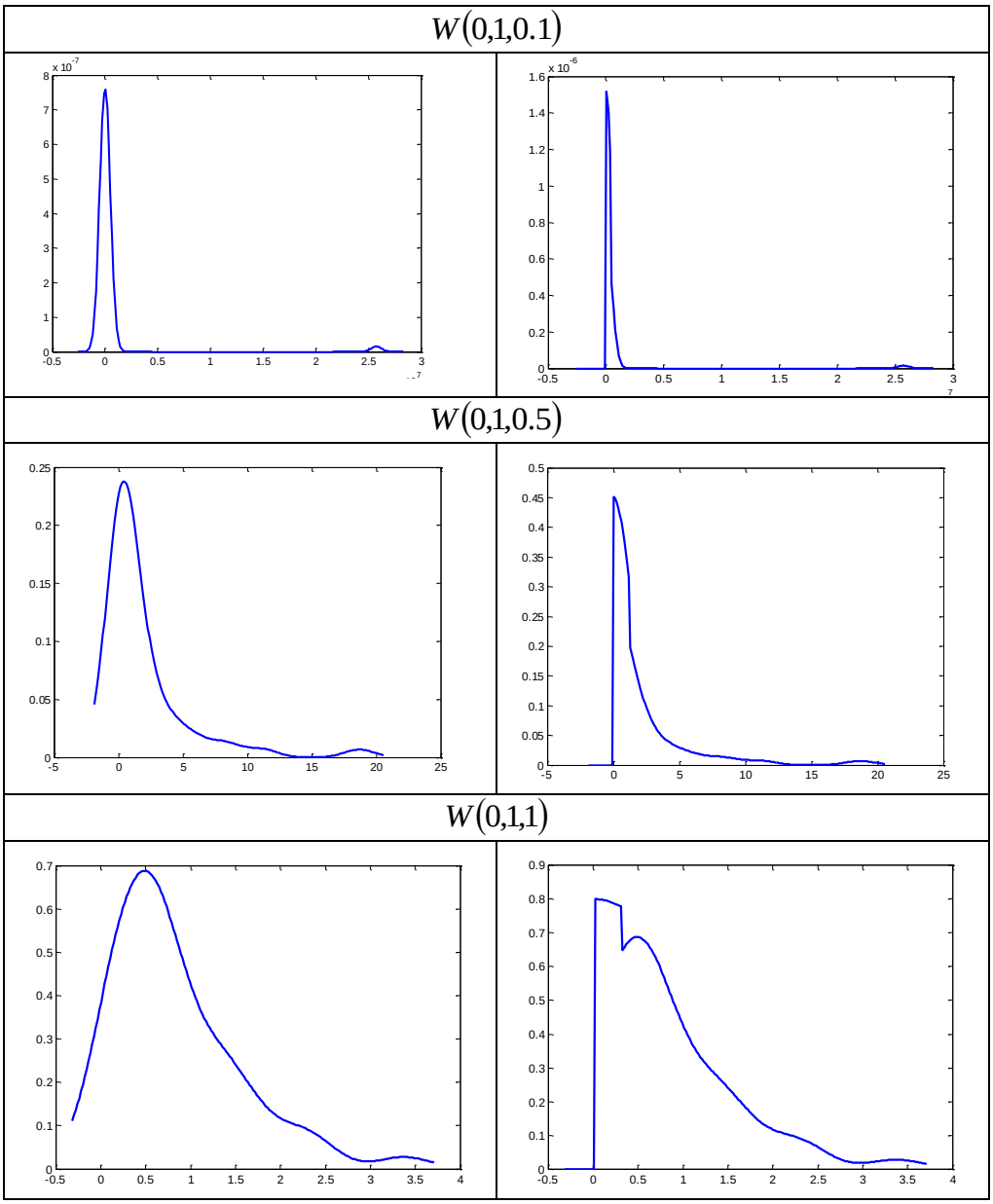
Table 1. Chosen descriptive statistics for samples from populations W1-W8 (n=50)

Sample from population $Wi(0,\delta,\gamma)$	Maximal value	Mean	Median	Standard deviation	Asymmetry	Kurtosis
$i = 1$ $\delta = 1,\ \gamma = 0.1$	2.5772	0.0518	0.0000	0.3644	0.0000	0.000
$i = 2$ $\delta = 1,\ \gamma = 0.5$	18.6900	1.9088	0.4418	3.4493	3.0234	13.3488
$i = 3$ $\delta = 1, \gamma = 1$	3.3774	0.8321	0.6525	0.7073	1.4203	5.1301
$i = 4$ $\delta = 1, \gamma = 2$	1.9582	0.8640	0.8397	0.3911	0.7166	3.2806
$i = 5$ $\delta = 1, \gamma = 3.4$	1.5101	0.8271	0.8312	0.2858	0.0493	2.5580
$i = 6$ $\delta = 1, \gamma = 5$	1.2919	0.9249	0.9549	0.2336	-0.3033	2.0451
$i = 7$ $\delta = 4, \gamma = 1$	17.4815	3.9782	2.5848	3.7848	1.4699	4.9171
$i = 8$ $\delta = 4, \gamma = 2$	7.9888	3.3012	2.7184	2.0423	0.5840	2.2218

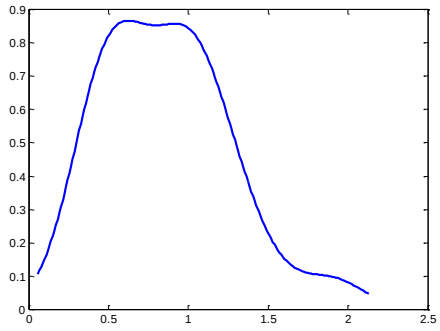
Source: own calculations

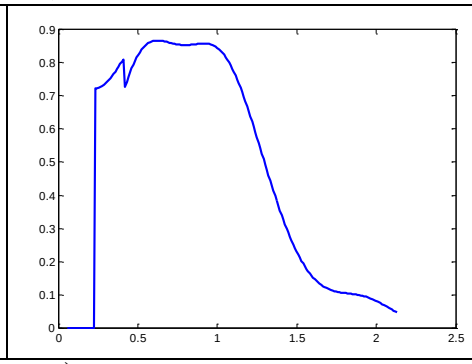
Exemplary results for sample size $n = 50$ are presented in Figures 1-2 with classical kernel density estimator (on the left) and kernel density estimator with reflection (on the right).

Figure 1. Classical kernel density estimator and kernel density estimator with reflection for populations W1-W8

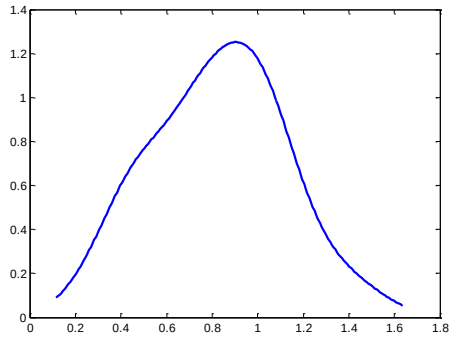


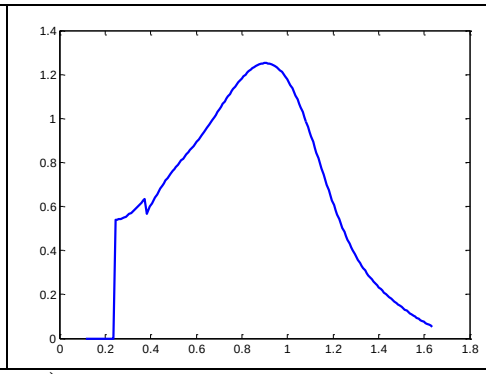
$W(0,1,2)$



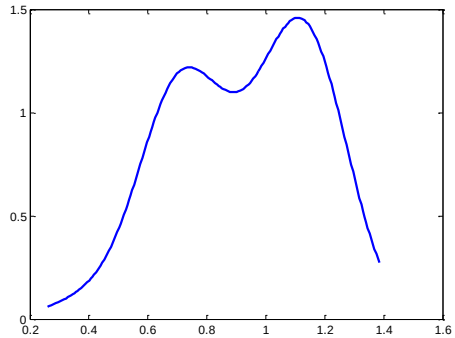


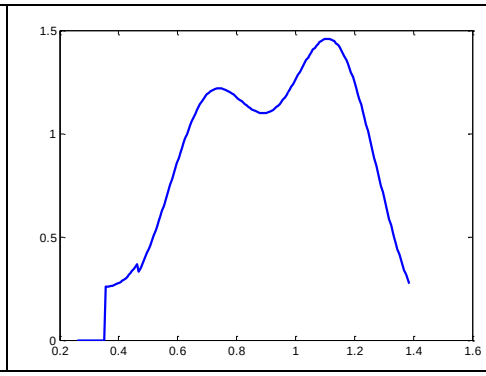
$W(0,1,3.4)$

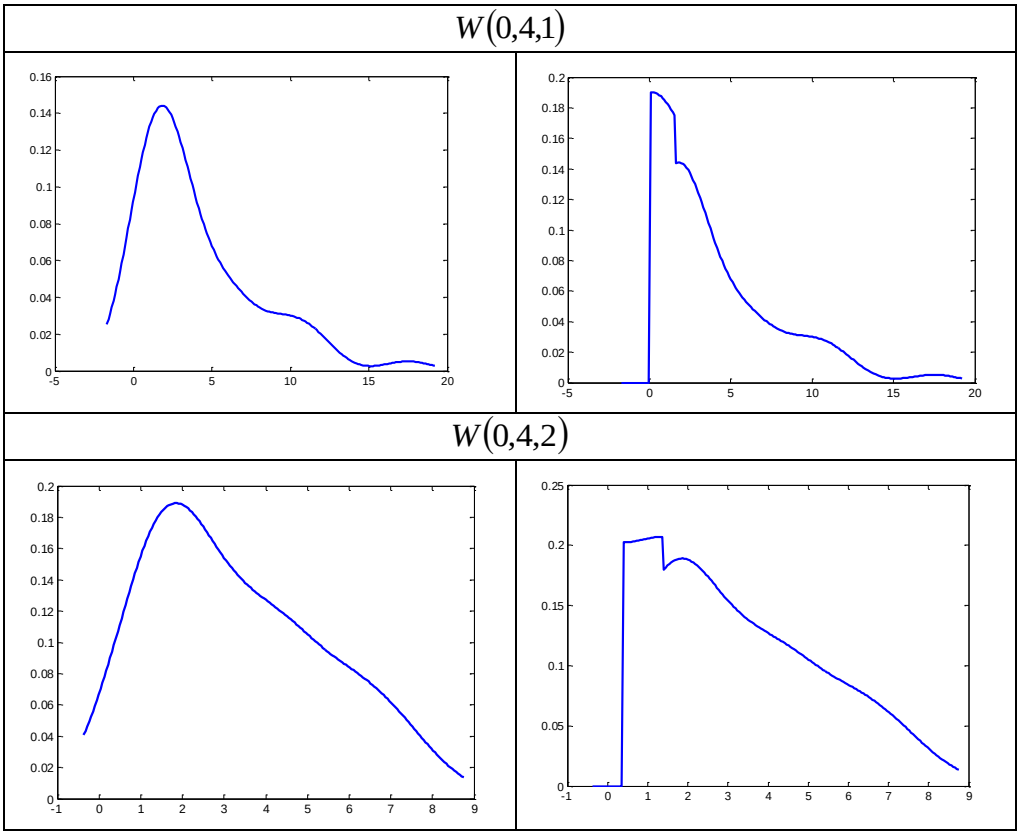




$W(0,1,5)$

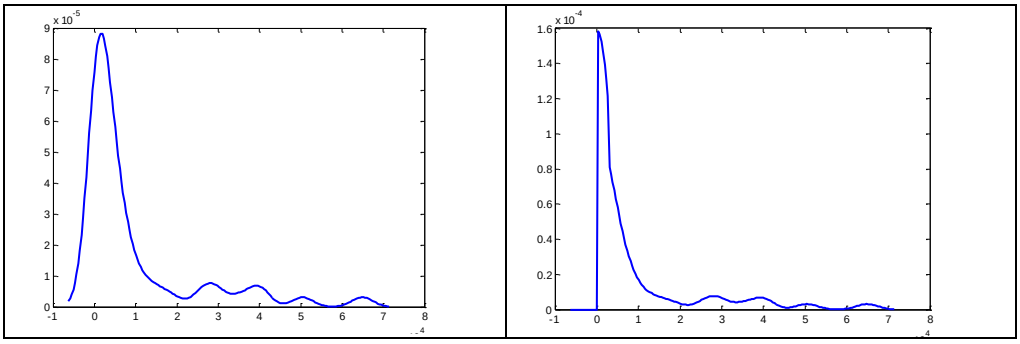






Source: own calculations

Figure 2. Classical kernel density estimator and kernel density estimator with reflection for agriculture value added per worker for countries in the world in 2013



Source: own calculations

SUMMARY

For many random variables considered in practical applications their density functions, by definition, are characterized by bounded support. Sometimes the estimator, e.g. kernel density estimator based on the samples from these populations, has different support than the density function. Such situation was observed for samples from Weilbull distribution, especially with small values of shape parameter. Modification of kernel estimator using the method of reflection ensure the users that the estimator is constructed only for non-negative values. But applying the reflection kernel estimator without initial stage of analyse the classical estimator may cause unnecessary limitation of the support (as it was in the case of W5). Further deeper analysis is needed for indicating such modifications of classical kernel estimator that the estimator will be with the same support as density function and there will be no lack of points of discontinuity. Such modification is necessary especially for practical implementations of regarded methods.

REFERENCES

Albers G. M. (2012) Boundary Estimation of Densities with Bounded Support, Swiss Federal Institute of Technology, Zurich, https://stat.ethz.ch/research/mas_theses/2012/Martina_Albers [18.06.2015]

Domański C., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014) Testy statystyczne w procesie podejmowania decyzji, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Jones M. C. (1993) Simple Boundary Correction for Kernel Density Estimation, Statistics and Computing, 3, pp. 135-146.

Jones M. C., Foster P. J. (1996) A Simple Nonnegative Boundary Correction Method for Kernel Density Estimation, Statistica Sinica, 6, pp. 1005-1013.

Karunamuni R. J., Alberts T. (2005) On Boundary Correction in Kernel Density Estimation, Statistical Methodology, 2, pp. 191-212.

Kulczycki P. (2005) Estymatory jądrowe w analizie systemowej, Wydawnictwa Naukowo-Techniczne, Warszawa.

Horová I., Koláček J., Zelinka J. (2012) Kernel Smoothing in MATLAB. Theory and Practice of Kernel Smoothing, World Scientific, New Jersey.

Silverman B.W. (1996) Density Estimation for Statistics and Data Analysis, Chapman & Hall, London.

Wand M. P., Jones M.C. (1995) Kernel Smoothing, Chapman & Hall, London.