WSB University in Wroclaw
Research Journal
ISSN 1643-7772 I eISSN 2392-1153
Vol. 16 I No. 3

Zeszyty Naukowe Wyższej Szkoły
Bankowej we Wrocławiu
ISSN 1643-7772 I eISSN 2392-1153
R. 16 I Nr 3

Boundary Effect Reduction in Kernel Estimation of Chosen Functional Characteristics of Random Variable

Author: Aleksandra Baszczyńska

Abstract

For a random variable with bounded support, the kernel estimation of functional characteristics may lead to the occurrence of the so-called boundary effect. In the case of the kernel density estimation it can mean an increase of the estimator bias in the areas near the ends of the support, and can lead to a situation where the estimator is not a density function in the support of a random variable. In the paper the procedures for reducing boundary effect for kernel estimators of density function, distribution function and regression function are analyzed. Modifications of the classical kernel estimators and examples of applications of these procedures in the analysis of the functional characteristics relating to gross national product per capita are presented. The advantages of procedures are indicated taking into account the reduction of the bias in the boundary region of the support of the random variable considered.

Keywords: kernel estimation, boundary effect, reflection method, gross national product per capita
JEL: C13, C14

History: Otrzymano 2015-11-21, poprawiono 2016-06-30, zaakceptowano 2016-07-05

Introduction

In statistical analyses concerned with economic, medical, social or technical issues the random variables under discussion may be characterized by having a bounded support. Bounding the support of a random variable to specified intervals, for example: $[a;\infty)$, $(-\infty; b]$, $[a; b]$ results from the specificity of those variables. What follows is that some economic indicators, while describing relationships between economic sizes and being widely used in economic situation analyses and in predictions of future economic changes, are characterized by having bounded support type $[1;0]$ (Gini coefficient, corruption perception index) or $[0;\infty)$ (research and development expenditures of companies, the number of dwellings completed).

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

Other examples of indicators used in statistical analyses of random variables with bounded support include: disease entity duration, diagnostic indicators for specific disease entities, social indicators of marginalization and social exclusion and efficiency index of technical equipment.

Classical procedures which take into account the assumption that the form of the functional characteristics is known, being defined as parametric procedures, are fairly often applied in practice mainly on account of their theoretical and computational simplicity and availability through suitable tools in statistics and econometrics packages. However, in many research situations risk associated with adopting the assumption on the specified form of the characteristics analyzed constitutes a serious argument in the decision-making process on the nature of statistical procedures for the benefit of the nonparametric procedures.

The estimation of functional characteristics of random variables with bounded support can lead to the occurrence of the so called boundary effect, specified as the lack of estimation consistency for x which are near the ends of the support, that is for x belonging to the so called boundary region. In the boundary region there are fewer observations subject to averaging, which has an impact on the variance and estimator bias. The boundary effect plays a particularly important role for small and medium-sized samples, for then a significant part of observations may be influenced by the boundary effect (cf. Härdle, 1994, pp. 159-162). This problem concerns, in general, the group of estimation methods described as smoothing methods, and, especially, nonparametric methods, including the kernel ones applied in the estimation of such functional characteristics like density function, distribution function and regression function.

The approach presented here of the kernel estimation of density function and distribution function of random variables with bounded support focuses on one-dimensional random variable while a multi-dimensional analysis is a natural extension of those procedures.

If we determine kernel estimator with an unknown density function of the population from which the sample $x_1, x_2, ..., x_n$ is drawn, we need to adopt the assumption on appropriate smoothness degree of the unknown density function, at the least the existence of a second continuous derivative of that function. In the kernel density estimation, the occurrence of the boundary effect can lead to the discordance between the support of the random variable and that of the density estimator, which has large practical implications, particularly for graphical presentation of the nonparametric estimation results. This discordance may result in mistaken interpretation of a specific estimator of the functional characteristics of the random variable. For random variables which are economic in nature and which often take on only nonnegative values ($[0;\infty)$) even a properly constructed kernel density estimator can take on values other than zero, also on $(-\infty;0)$. Not only is this possible when the kernel function with an unbounded support is applied in the construction of the kernel density estimator, but even when the kernel function support is bounded (cf. Kulczycki, 2005, pp. 94-97). The approach involving cutting the estimator at point 0 and assuming that $\hat{f}(x) = 0$ for $x < 0$ has the effect that the estimator does not satisfy the condition of integrability to unity in the support of the random variable.

Point consistence for the kernel density estimation for one-dimensional random variable X with support $[0;1]$ is discussed, for example,

in the work of Qi Li and Jeffrey Scott Racine (Li, Racine, 2007, pp. 30-32). It can be shown that for $x \in [0;1]$ there is $\hat{f}(x) - f(x) = o_p(1)$[1], whereas for x belonging to the boundary region of the support of the random variable the mean squared error of the kernel density estimator does not satisfy the condition $MSE\left[\hat{f}(x)\right] = o_p(1)$. For $x = 0$ and for $f(0) > 0$ the expected value and the kernel density estimator bias are as follows:

$$E\left[\hat{f}(0)\right] = \frac{f(0)}{2} + O(h)^2,$$

$$B\left[\hat{f}(0)\right] = E\left[\hat{f}(0)\right] - f(0) = \frac{-f(0)}{2} + O(h),$$

where h is a smoothing parameter in the kernel density estimator.

It is therefore necessary to introduce suitable modifications to the classical kernel methods in the nonparametric estimation, so that the kernel estimator is consistent.

The modifications of the classical kernel method may involve data transformation, and the application of a pseudo-data method, local linear method or jackknife method. However, the most frequently used methods are those consisting in employing the so called boundary functions of the kernel and the reflection method.

The kernel density estimator with the boundary kernel function is of the form (assuming that $x \in [0;1]$):

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x, x_i) \quad , \quad (1)$$

where $x_1, x_2, ..., x_n$ is a random sample chosen from the population with an unknown density function $f(x)$, h is a smoothing parameter, while $K_h(x,x_i)$ a boundary kernel function of the form:

$$K_h(x,x_i) = \begin{cases} \dfrac{\frac{1}{h}K\left(\frac{x_i - x}{h}\right)}{\int_{-\frac{x}{h}}^{\infty} K(u)\,du} & \text{for } x \in [0;h), \\ \frac{1}{h}K\left(\frac{x_i - x}{h}\right) & \text{for } x \in [h;1-h], \\ \dfrac{\frac{1}{h}K\left(\frac{x_i - x}{h}\right)}{\int_{-\infty}^{\frac{1-x}{h}} K(u)\,du} & \end{cases} \quad (2)$$

where $K(\cdot)$ is the kernel function of a second-order, that is, satisfying the following conditions:

$$\begin{cases} \int_{-\infty}^{\infty} K(u)du = 1, \\ \int_{-\infty}^{\infty} uK(u)du = 0, \\ \int_{-\infty}^{\infty} u^2 K(u)du = \kappa_2 > 0 \end{cases} \quad (3)$$

It can be shown that for the random variable with support $[0;1]$ and for x from the boundary region $x \in [0;h]$ the expected value and kernel density estimator bias (1) are as follows:

$$E\left[\hat{f}(x)\right] = f(x) + O(h)$$

$$B\left[\hat{f}(x)\right] = O(h)$$

The bias of the kernel estimator (1) approaches zero for $n \to \infty$. Unfortunately, applying the density function estimator with boundary kernel function (2) may lead to situations where the density estimator takes on negative values.

The reflection method is one of the fairly frequently methods applied in practice of bias reduction in the kernel estimation of functional characteristics. The modification of the classical kernel density estimator consists in isolating that part of the kernel function which is outside the interval of the support of the random variable and then on its symmetrical reflection. This reflection is done in relation to the boundary of the support a (in the

[1] For a sequence of real random variables $\{X_n\}_{n=1}^{\infty}$ $X_n = o_p(1)$ if $X_n \xrightarrow{p} 0$.

[2] For a nonnegative constant n, $a_n = O(b_n)$ if $\frac{a_n}{b_n} = O(1)$ $a_n \le Cb_n$ (for a certain constant C and for all sufficiently large n.

case of the left-hand boundary of the support of the random variable $[a;\infty)$) or b (in the case of the right-hand boundary of the support $(-\infty;b]$).
It can be shown (Kulczycki, 2005, pp. 94-97) that the estimator taking into account the reflection method of the kernel function has a support which is the same as the support of the random variable. After having complemented any derivative at point a (for the left-hand boundary of the support of the random variable $[a;\infty)$) or b (for the right-hand boundary of the support $(-\infty;b]$) with the null value, the derivative becomes continuous. The estimator taking into account the reflection method has a continuous derivative of a specific order, if the kernel function has a continuous derivative of a specific order.
The data transformation method, pseudo-data method and local linear method have been outlined in the works of, for example: Bernard Silverman (Silverman, 1986, pp. 29-32), Matt Wand and Chris Jones (Wand, Jones, 1995, pp. 46-49), Chris Jones (Jones, 1993), Chris Jones and P. Foster (Jones, Foster, 1996) and Ivanka Horova, Jan Koláčk and Jiři Zelinka (Horová, Koláček, Zelinka, 2012, pp. 39-41). The application of the jackknife method in the nonparametric kernel estimation of regression function has been demonstrated, among others, by Wolfgang Härdle (Härdle 1994, pp. 159-162) and Herman Bierens (Bierens, 1987, pp. 99-144).

Estimation of Density Function

The Rosenblatt-Parzen's classical kernel density estimator based on a random sample $x_1, x_2, ..., x_n$ drawn from the population with an unknown density function $f(x)$ is given by the form (Silverman, 1996, pp. 13-19; Wand, Jones, 1995, pp. 11-14):

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where h is a smoothing parameter such that for $n \to \infty$, $h = h(n) \to 0$ and $nh \to \infty$, while $K(\bullet)$ is the kernel function having the properties (3). If the kernel function is, in addition, nonnegative and symmetric about zero, then: $\hat{f}(x) \geq 0$ and $\int_{-\infty}^{\infty} \hat{f}(x)dx = 1$. The properties and the procedures as regards the choice of the smoothing parameter and kernel function are presented, for example, in the work of Czesław Domański, Dorota Pekasiewicz, Aleksandra Baszczyńska and Anna Witaszczyk (Domański et al., 2014).
The kernel density estimator with the reflection kernel function for a random variable with the support is of the form:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad . (5)$$

The generalization of the kernel estimator which takes into account the reflection of the kernel function (5) proposed by Rohan Karunamuni and Tom Alberts (Karunamuni, Alberts, 2005) for a random variable with the support is defined as:

$$\hat{f}_{GR}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x - g_1(x_i)}{h}\right) + K\left(\frac{x + g_2(x_i)}{h}\right)\right] \quad (6)$$

where g_i, $i = 1,2$ are nonnegative, continuous and increasing functions on the interval $[0;\infty)$ (cf. Karunamuni, Zhang, 2005).
It can be noticed that the generalization (6) may be viewed simultaneously as a generalization of the reflection method and of the data transformation method, as the kernel estimator is applied to the set , and transformation

g is so selected that the bias in boundary area is of the order $O(h^2)$. Estimator (6) is a consistent estimator of density function f, having the bias of the order . The analysis of the properties of the kernel density estimator taking into account the reflection method is illustrated in the works of, for example, Matt Jones (Jones, 1993), Matt Jones and Foster (Jones, Foster, 1996), Martina Albers (Albers, 2012) and Aleksandra Baszczyńska (Baszczyńska 2015).

Distribution Function Estimation

Let $X_1, X_2, ..., X_n$ be independent random variables with a distribution function F and density function f. Let $x_1, x_2, ..., x_n$ be a random sample drawn from the population having the distribution function F.
The kernel estimator of the distribution function is of the form:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right) \quad , (7)$$

where the smoothing parameter is specified in the same way as in the kernel density estimation, whereas

$$W(x) = \int_{-1}^{x} K(t)dt \quad \text{for} \quad K(t) \geq 0,$$

being a unimodal and symmetric second-order kernel function having the support $[-1;1]$.
The kernel distribution function estimator taking into account the reflection of the kernel function for the random variable with support and the generalized distribution function estimator are as follows (cf. Koláček, Karunamuni, 2009, 2012):

$$\hat{F}_R(x) = \frac{1}{n}\sum_{i=1}^{n}\left[W\left(\frac{x - x_i}{h}\right) - W\left(\frac{x + x_i}{h}\right)\right], \quad (8)$$

$$\hat{F}_{GR}(x) = \frac{1}{n}\sum_{i=1}^{n}\left[W\left(\frac{x - g_1(x_i)}{h}\right) - W\left(\frac{x + g_2(x_i)}{h}\right)\right], \quad (9)$$

where the smoothing parameter and the distribution function $W(x)$ are specified, as is the case for estimator (7), whereas the functions g_i for $i = 1,2$ are nonnegative, continuous and increasing functions determined on $[0;\infty)$.
It can be shown that the variances of estimators (7) and (9) are of the same order, whereas the bias of estimator (9) is of order $O(h^2)$, which implies that estimator (9) reduces the bias effect in the kernel estimation of the distribution function, while the bias in the boundary region is of the same order as the bias of the estimator in the internal region.

Regression Function Estimation
In the regression model being of the form:
$Y_i = m(x_i) + \varepsilon_i$, for $i = 1,...,n$, $n \in N$, $E(\varepsilon_i) = 0$, $D^2(\varepsilon_i) = \sigma^2 > 0$,
the approximation of an unknown function m can be carried out using the kernel estimation with a smoothing parameter h and kernel function K.

Let $x_i = \frac{i-1}{n}$ for $i = 1,...,n$ on $[0;1]$.

Kernel regression estimators are, among others:
a) Nadaraya-Watson estimator:

$$\hat{m}_{NW}(x) = \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \quad , (10)$$

b) Gasser-Müller estimator:

$$\hat{m}_{GM}(x) = \sum_{i=1}^{n} Y_i \int_{s_i - 1}^{s_i} \frac{1}{h} K\left(\frac{t - x}{h}\right) dt \quad , (11)$$

where:
$s_i = \frac{x_i + x_{i+1}}{2}$, $i = 1,...,n-1$, $s_0 = 0$,
$s_n = 1$,
moreover, the smoothing parameter h and kernel function K are

specified in the same way as for the kernel density function estimation.
The kernel regression estimator taking into account the reflection of the kernel at points x_i , $i = 0,...,n+1$ for the random variable having the support $[0;1]$ is of the form:

$$\hat{m}_{GMR}(x) = \frac{1}{h}\sum_{j=1}^{3n}\overline{Y}_j \int_{s_j-1}^{s_j} K\left(\frac{x-u}{h}\right)du \text{ , (12)}$$

where:

$$s_j = \frac{\overline{x}_i + \overline{x}_{i+1}}{2},\ j = 1,...,3n-1,$$
$$s_0 = -1,\ s_{3n} = 2 .$$

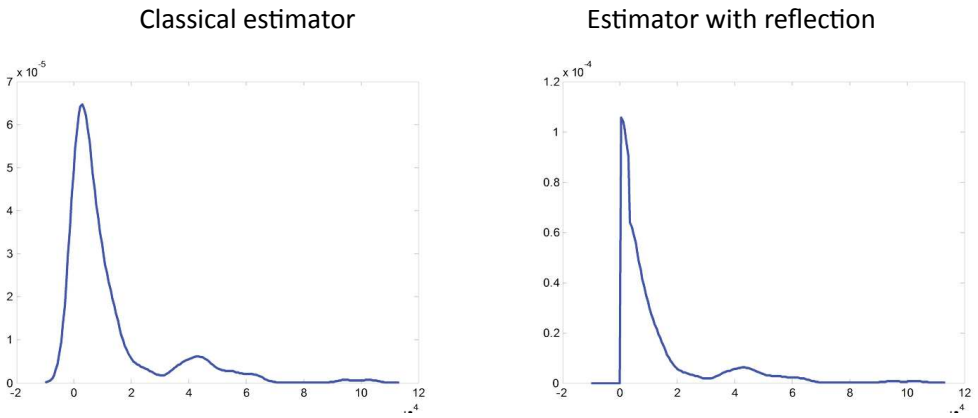
The analysis of the properties of estimator (12) along with the proposal of the optimum smoothing parameter is presented in the works of, among others, Jan Koláček and Jitka Poměnkova (Koláček, Poměnková, 2006).

The Example of the Boundary Reduction Method Application in the Kernel Estimation of Chosen Functional Characteristics of Random Variable

In order to compare the classical procedures with the procedures taking into consideration the reflection method, estimators of the selected functional characteristics of a random variable were determined for the data on gross national product per capita converted to U.S. dollars, which made it possible to make comparisons between a variety of economies. To smooth fluctuations in prices and exchange rates, Atlas method of conversion was employed.
The data stem from the records of the World Bank (http://www.worldbank.org, [10.10.2015]).

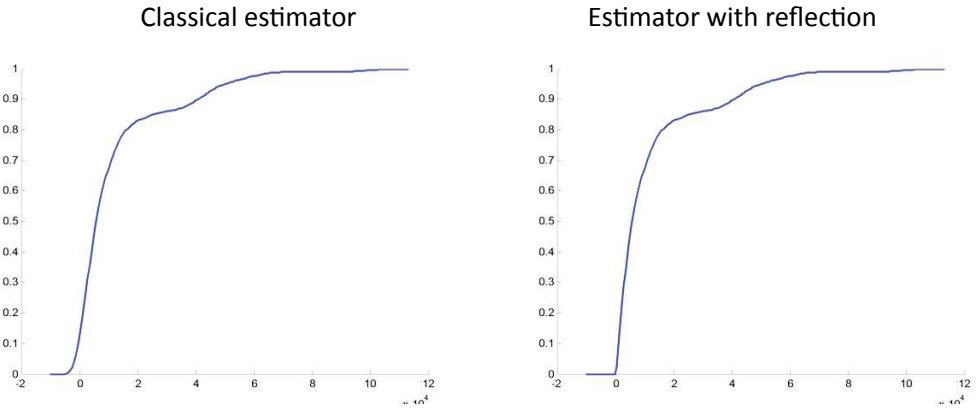
In the kernel estimation concerned with the density function, distribution function and regression function (in the classical approach and estimation with reflection) the second-order kernel functions were used – the Gaussian, Epanechnikov and quartic. The smoothing parameter was determined using the Silverman and cross-validation method. Choosing precisely those parameters of the kernel method (kernel function and smoothing parameter) had its reason in the fact that those are the kernel method parameters which are most frequently used in practice while ensuring that proper results of the kernel estimation procedures are obtained. The application of different parameters of the kernel method provided the opportunity to choose the best method for specific data.
The use of data on gross national product per capita, both cross-sectional for countries across the world and time-based for Poland, may constitute a procedure applied at the preliminary stage of the statistical analysis, forming a starting point for further in-depth studies employing, for example, econometric models.
At the first stage, the kernel density estimators of the gross national product per capita were determined for the year 2014, encompassing 180 countries worldwide. The sample results for Epanecznikov kernel function and Silverman method are illustrated in Figure 1.
The second stage involved determining the distribution function estimators for gross national product per capita in 2014 for 180 countries across the world. The sample results for Epanecznikov kernel function and Silverman method are illustrated in Figure 2.
Classical kernel estimators, both of density function and distribution function, are characterized by some drawbacks. It can be easily inferred that their support is not consistent with

Fig. 1 Kernel density estimators of gross national product per capita in 2014 for 180 countries worldwide



Source: Author's own study

Fig. 2 Kernel estimators of the distribution function of gross national product per capita in 2014 for 180 countries worldwide

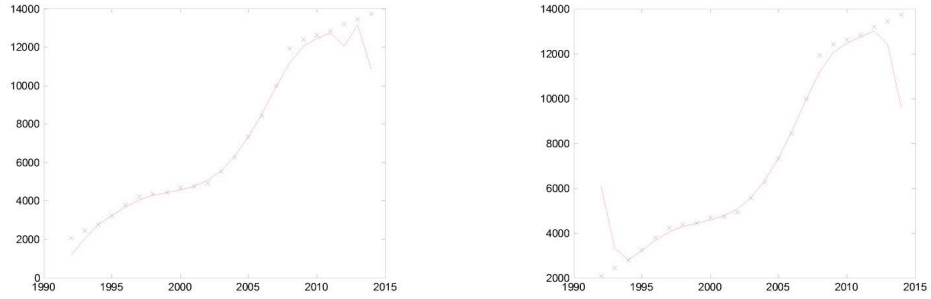


Source: Author's own study

the support of the random variable (estimators on the left-hand side). Introducing a modification consisting in reflecting the kernel function eliminates this drawback to a considerable degree.

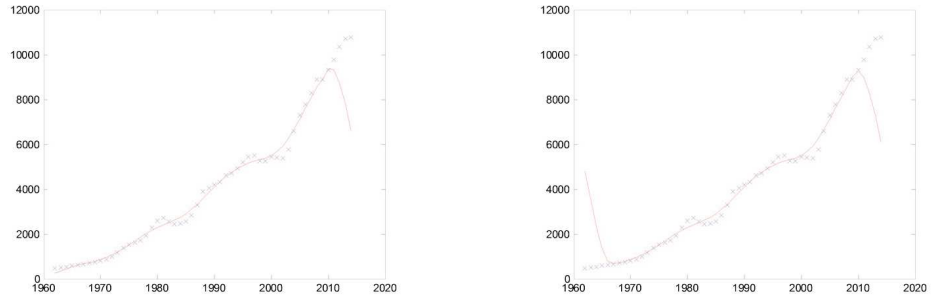
At the third stage, the kernel regression estimators were determined. The sample results of the estimation for the kernel method parameters, the same as in the first and second stage, are shown in Figure 3-4.

Fig. 3 Kernel regression estimators of gross national product per capita for Poland in 1990-2014



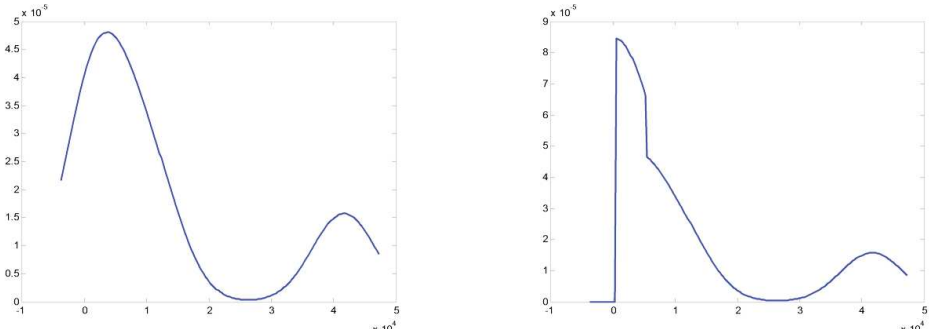
Source: Author's own study

Fig. 4 Kernel distribution function estimators of gross national product per capita worldwide in 1962-2014



Source: Author's own study

Fig. 5 Kernel density estimators of gross national product per capita in 2014 for a 10-element-sample, Epanecznikov kernel function, Silverman method for choosing smoothing parameter.



Source: Author's own study

Introducing the modification of the classical regression estimator consisting in reflecting the kernel function has the effect that the kernel estimator has a different form, which ensures the bias reduction in the region near zero.

In order to assess the impact of the sample size on the results of the kernel estimation, at the fourth stage, the samples comprising 10 elements were drawn providing the basis for determining kernel density estimators for gross national product per capita in 2014 for the countries worldwide. The results of the kernel density estimation are demonstrated in Figure 5.

The impact of the modification taking into account the kernel function reflection is easier to notice in small-sized samples. For small-sized samples, the contribution of observations for which the kernel functions are reflected in the boundary region is relatively larger than for large samples. Hence, there are such big changes in the estimator's shape after having introduced the kernel function reflection.

Summary

Every smoothing method used near the ends of the support of the random variable with bounded support becomes less accurate. In the case of the kernel estimation there is a considerable worsening of the statistical properties of the estimator, which is caused by cutting off the kernel function at the boundary point.

In constructing the modification of the classical estimator consisting in taking into account the reflection of this part of the kernel function which is not in the support of the random variable leads to a considerable improvement of the properties of the estimator. It is of particular relevance for the analyses carried out based on the graphic representation of estimators obtained. The concordance between the support of the estimator and that of the random variable is then provided.

The comparative analysis between the classical estimators and the kernel estimators with reflection

clearly shows that it is the modified procedures that should be commonly applied in practical studies in the situation when a random variable has a bounded support.

However, no unequivocal conclusions can be made as to the impact of the parameters of the kernel method (smoothing parameter and kernel function) on the form of the estimator. The Gaussian kernel function, even though being the kernel function with unbounded support, yields similar effects to those of the kernel function with bounded support.

The size of a sample is the factor which has a large impact on the final form of the estimator.

Therefore further investigation appears necessary, including simulation tests allowing one to indicate what type of random variable distribution and what distribution parameters can influence the estimator to the greatest extent in the analyses of random variables with bounded support.

Bibliography

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 [18.11.2015]

Baszczyńska A., (2015), Bias Reduction in Kernel Estimator of Density Function in Boundary Region, Quantitative Methods in Economics, in the process of being printed.

Bierens H. J. (1987), Kernel Estimators of Regression Functions, w: Truman F. Bewley (ed.), Advances in Econometrics: Fifth World Congress, Cambridge University Press, 99-14.

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

Härdle W. (1994), Applied Nonparametric Regression, Cambridge University Press, Cambridge.

Li Q., Racine J. S. (2007), Nonparametric Econometrics. Theory and Practice, Princeton University Press, Princeton and Oxford.

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

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

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

Karunamuni R. J., Zhang (2008),Some Improvements on a Boundary Corrected Kernel Density Estimator, Statistics and Probability Letters, 78, 497-507.

Koláček J., Karunamuni R. J., (2009), On Boundary Correction in Kernel Estimation of ROC Curves, Australian Journal of Statistics, 38, 17-32.

Koláček J., Karunamuni R. J., (2012), A Generalized Reflection Method for Kernel Distribution and Hazard Function Estimation, Journal of Applied Probability and Statistics, 6, 73-85.

Koláček J., Poměnková J., (2006), A Comparative Study of Boundary Effects for Kernel Smoothing, Australian Journal of Statistics, 35, 281-288.

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 and Hall, London.

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

Redukcja efektu brzegowego w estymacji jądrowej wybranych charakterystyk funkcyjnych zmiennej losowej

Abstrakt

Dla zmiennej losowej o ograniczonym nośniku estymacja jądrowa charakterystyki funkcyjnej może oznaczać wystąpienie tzw. efektu brzegowego. W przypadku estymacji funkcji gęstości oznacza to zwiększenie obciążenia estymatora w obszarze blisko krańców nośnika, jak również prowadzić może do sytuacji, że estymator nie posiada pożądanych własności dla funkcji gęstości w nośniku zmiennej losowej. W pracy poddano analizie procedury redukujące efekt brzegowy estymatora jądrowego funkcji gęstości, dystrybuanty oraz funkcji regresji. Przedstawiono modyfikacje klasycznych estymatorów jądrowych oraz zaproponowano zastosowanie tych procedur w analizie charakterystyk

funkcyjnych dotyczących dochodu narodowego brutto na mieszkańca. Wykazano zalety procedur uwzględniających redukcję obciążenia w obszarze brzegowym nośnika rozważanej zmiennej losowej.

Słowa kluczowe: estymacja jądrowa, efekt brzegowy, metoda odbicia, dochód narodowy brutto na mieszkańca