https://doi.org/10.26485/0459-6854/2018/68.3/4 PL ISSN 0459-6854

B U L L E T I N

DE LA SOCIÉTÉ DES SCIENCES ET DES LETTRES DE ŁÓDŹ

2018 Vol. LXVIII

Recherches sur les déformations no. 3

pp. 51–58

Dedicated to the memory of
Professor Yurii B. Zelinskii

Aleksandra Baszczyńska

SMOOTHING PARAMETER VALUES IN AUTOMATIC CHOICE PROCEDURE AND IN ACCEPTABLE INTERVAL IN THE KERNEL DENSITY ESTIMATION

Summary

Automatic procedure for determining the parameters of kernel method, allows the simultaneous selection of two method parameters: kernel function and smoothing parameter. This approach simplifies the procedure for parameters selection and at the same time provides a good properties of kernel estimators. The second procedure regarded in the paper is the acceptable interval of values of smoothing parameter, allowing for a much more generalized approach in choosing the smoothing parameter in the kernel estimation. The results of the smoothing parameter values comparison, where these values are set in the automatic procedure and the procedure of the acceptable interval of smoothing parameters values in the estimation of density function, are presented in the paper. Comparison of these values is made basing on the results of applying the simulation methods. Basing on simulation studies results new intervals of values of smoothing parameter are proposed and analyzed.

Keywords and phrases: kernel density estimation, smoothing parameter, kernel function, automatic choice

1. Introduction

Kernel method is widely used in estimation of some functional characteristics of random variable, such as density function, cumulative distribution function, receiver operating characteristic or regression function (cf.: [1, 2, 3, 4, 5, 6]) as well as in

estimation of some characteristic parameters moments or positional parameters including quantiles (cf.: [7, 8, 9]). There are also some kernel procedures that can be used in hypothesis testing (cf.:[10, 11]). Kernel method was used for the first time in density estimation. Kernel density estimator, known as Rosenblatt-Parzen estimator is the following (cf.: [12, 13]):

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

where $X_1, X_2, \ldots, X_n$ is random sample drawn from population with unknown density function $f(x)$, n is sample size, h is smoothing parameter, $K(u)$ is kernel function. Smoothing parameter h and kernel function $K(u)$ can be treated as kernel method parameters and some considerations on choosing the appropriate values and forms of them are found in literature ([14, 15, 16, 17, 18]). Kernel function $K(u)$ is characterized by its order k and the smoothness μ, where for nonnegative integers ν and k $(0 \leq \nu < k)$, kernel function $K \in S_{\nu,k}$ and

$$S_{\nu,k} = \{K : K \in Lip[-1,1] \ \wedge \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0 & for \quad 0 \leq j \leq k-1,\ j \neq \nu, \\ (-1)^\nu \nu! & for \quad j = \nu, \\ \kappa_k \neq 0 & for \quad j = k, \end{cases} \tag{2}$$

where $Lip[-1,1]$ denotes such a class of functions that for some constant $L > 0$ the condition $|f(x) - f(y)| \leq L|x - y|$ is fulfilled for every $x, y \in [-1,1]$. The kernel function is smooth of order μ when for $K \in S^{\mu}_{\nu,k}$ the following condition is fulfilled: $S^{\mu}_{\nu,k} = \left\{K : K \in S_{\nu,k} \cap C^{\mu}[-1,1] \wedge K^{(j)}(-1) = K^{(j)}(1) = 0,\ j = 0, \ldots, \mu - 1\right\}$, where $C^{\mu}[-1,1]$ denotes a class of real functions on $[-1,1]$ that are μ-times continuously differentiable.
For smoothing parameter the conditions are: $h = h(n) > 0$, $h \underset{n\to\infty}{\longrightarrow} 0$ and $nh \underset{n\to\infty}{\longrightarrow} \infty$ (cf.: [19, 20]).

2. Automatic procedure for determining the parameters of kernel method

Automatic procedure for determining the parameters of kernel method, allows the simultaneous selection of two method parameters: kernel function and smoothing parameter. This approach simplifies the procedure for parameters selection and at the same time provides a good properties of kernel estimators.

The detailed procedure for finding the order of kernel function, the form of kernel function and value of smoothing parameter is presented in [19]. The main elements in this procedure include the designation of:

a) the optimal kernel function (polynomial of degree k) K_{opt} and its canonical factor $\gamma_{0k} = \left(\frac{D^2(K_{opt})}{\kappa_k^2}\right)^{\frac{1}{2k+1}}$,

b) the optimal smoothing parameter $\hat{h}_{0,k}$ that belongs to the interval (with bounds: minimum distance of all points from sample and maximum smoothing parameter),

c)) the optimal order of the kernel function as the minimization with respect to k of $L(k) = \left(|\kappa_k| D^2(K_{opt})\right)^{\frac{2}{2k+1}} \frac{\gamma_{0k}(2k+1)}{2nkh_{0,k}}$.

3. Acceptable interval of smoothing parameters

The idea of acceptable interval of smoothing parameters comes from the paper of Horová [19]. This interval has the form:

$$H_n = \left[\min_{i \neq j} |X_i - X_j|, \hat{h}_{MS}\right], \tag{3}$$

where: $X_1, X_2, \ldots, X_n$ is random sample drawn from population with unknown density, $\min\limits_{i \neq j} |X_i - X_j|$ denotes the minimal distance between points X_i and X_j, for $i, j = 1, 2, \ldots, n$, $i \neq j$ and $\hat{h}_{MS}$ is described for second order kernel function as:

$$\hat{h}_{MS} = \frac{3}{35^5} \hat{\sigma} \left(\frac{R(K)}{\kappa_2^2}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}, \tag{4}$$

where: $\kappa_k = \int\limits_{-\infty}^{+\infty} u^k K(u) du$, $R(K) = \int\limits_{-\infty}^{+\infty} K^2 dx$.

In particular, for Gaussian and Epanechnikov kernel functions the smoothing parameter selectors are, respectively: $\hat{h}_{MS,G} = 1.144 \hat{\sigma} n^{-\frac{1}{5}}$ and $\hat{h}_{MS,E} = 2.532 \hat{\sigma} n^{-\frac{1}{5}}$, where $\hat{\sigma}$ is the estimate of standard deviation.

4. Results of the simulation study

In the simulation study the comparison between values of smoothing parameters chosen by different methods in kernel density estimation is taken into account. Analyses of smoothing parameter values make possible to indicate some remarks on properties of smoothing parameter selection in kernel density estimation. In addition some new intervals for smoothing parameter values are proposed and analysed.

Two different kinds of populations are regarded:

- populations with normally distributed random variables $X \sim N(10, \sigma)$ where some chosen values of standard deviations are considered $\sigma = 1, 5, 8, 10$. In this way different levels of variability in symmetric populations with infinite support of density function are regarded.
- populations with exponentially distributed random variables $X \sim Exp(0.1)$ and $X \sim Exp(10)$, so asymmetric populations with finite support (left-sided finite support) are taken into account.

From these populations small, moderate and large sample are drawn ($n = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$). Basing on these samples kernel density estimators are constructed with:

- Gaussian kernel function K_G and kernel functions from automatic choice K_A, where $K_A \in S^{\mu}_{\nu,k}$ and $\nu = 0$, $k = 2, 4, 6, 8$, $\mu = 0$, where $\nu = 0$, $k = 2$, $\mu = 0$ indicates Epanechnikov kernel function;
- smoothing parameters from acceptable intervals $\left[\hat{h}_L; \hat{h}_{MS,G}\right]$ and $\left[\hat{h}_L; \hat{h}_{MS,A}\right]$ (where $\hat{h}_L = \min_{i \neq j} |X_i - X_j|$, $\hat{h}_{MS,G}$ denotes maximum smoothing parameter and Gaussian kernel function used in kernel density estimator, $\hat{h}_{MS,A}$ denotes maximum smoothing parameter and kernel function set in automatic choice) and smoothing parameter from automatic choice $\hat{h}_A$.

The number of repetitions is set to 20000.

Table 1. Values of smoothing parameters in kernel density estimation for samples drawn from normally distributed populations $X \sim N(10, 1)$ and $X \sim N(10, 5)$.

Population distribution	Sample size	$\hat{h}_L$	$\hat{h}_{MS,G}$	$\hat{h}_{MS,A}$	$\hat{h}_A$	(ν, k, μ) for $K_A \in S^{\mu}_{\nu,k}$
$X \sim N(10,1)$	10	0.0524	0.4167	0.9224	2.2792	(0,2,0)
	20	0.0237	0.4426	1.6748	2.6026	(0,4,0)
	30	0.0068	0.5271	2.7315	4.2352	(0,6,0)
	40	0.0039	0.5568	2.2415	2.4163	(0,4,0)
	50	0.0374	0.5248	2.8959	3.7519	(0,6,0)
	60	0.0291	0.4317	1.8018	2.1692	(0,4,0)
	70	0.0461	0.4190	1.7310	1.9795	(0,4,0)
	80	0.0497	0.4521	2.6438	3.3279	(0,6,0)
	90	0.0429	0.5054	3.6655	4.9773	(0,8,0)
	100	0.0307	0.4547	2.7328	3.1688	(0,6,0)
$X \sim N(10,5)$	10	1.2227	4.3373	9.6020	15.1086	(0,2,0)
	20	0.3818	3.9022	14.7704	21.0973	(0,4,0)
	30	0.0468	3.3089	12.9804	17.5129	(0,4,0)
	40	0.2372	2.7479	14.7536	21.2017	(0,6,0)
	50	0.0670	2.4289	5.3770		
	60	0.0616	2.0027	11.3027	17.1435	(0,6,0)
	70	0.0238	2.5956	18.1668	24.6826	(0,8,0)
	80	0.1147	2.3558	16.8822	24.4400	(0,8,0)
	90	0.0017	2.2051	15.9912	23.1458	(0,8,0)
	100	0.0253	1.9590	4.3367		

The results for samples from normally distributed populations are presented in Tables 1 and 2, where empty areas denote impossibility of finding parameters in automatic choice because of order of kernel higher than 8.

Table 2. Values of smoothing parameters in kernel density estimation for samples drawn from normally distributed populations $X \sim N(10, 8)$ and $X \sim N(10, 10)$.

Population distribution	Sample size	$\hat{h}_L$	$\hat{h}_{MS,G}$	$\hat{h}_{MS,A}$	$\hat{h}_A$	(ν, k, μ) for $K_A \in S^{\mu}_{\nu,k}$
$X \sim N(10, 8)$	10	3.4447	4.8131	10.6553	16.8165	(0,2,0)
	20	0.0627	5.8941	22.3096	33.4207	(0,4,0)
	30	0.0848	2.4380	5.3972	8.8088	(0,2,0)
	40	0.0361	3.3855	13.6287	16.5713	(0,4,0)
	50	0.0598	3.9427	26.3155	38.2695	(0,8,0)
	60	0.0349	4.2935	29.4163	43.7601	(0,8,0)
	70	0.3011	4.3776	25.1787	31.0347	(0,6,0)
	80	0.1998	3.5868	25.5826	36.3712	(0,8,0)
	90	0.0558	3.9885	23.6614	28.3960	(0,6,0)
	100	0.5360	3.9897	29.3671	39.1810	(0,8,0)
$X \sim N(10, 10)$	10	0.1981	7.3437	16.2574	24.5606	(0,2,0)
	20	0.9087	5.6353	21.3302	31,1485	(0,4,0)
	30	0.0042	5.4565	12.0787	12.6792	(0,2,0)
	40	2.3366	5.1395	20.6898	26.7374	(0,4,0)
	50	2.3015	4.4756	29.8722	46.2420	(0,8,0)
	60	0.1279	4.4420	18.5383	20.3304	(0,4,0)
	70	0.0281	3.6567	15.4714	17.4918	(0,4,0)
	80	0.8842	4.5999	26.8963	33.7427	(0,6,0)
	90	0.1981	4.3650	31.6545	46.76100	(0,8,0)
	100	0.0214	0.4547	2.7328	3.1688	(0,6,0)

The results for samples from exponentially distributed populations are presented in Table 3, where empty areas denote impossibility of finding parameters in automatic choice because of order of kernel higher than 8.

5. Conclusion

Both, in the case of symmetric distribution with infinite support of density function of population from which samples are drawn and in the case of asymmetric distribution with bounded support of density function, the values of smoothing parameters in the procedure of constructing the kernel density estimators, are bigger in these cases

when the population is characterized by high level of dispersion. This remark is connected directly with the process of calculating the kernel density estimator and it does not depend on the method of smoothing parameter selection and the form of kernel function.

Table 3. Values of smoothing parameters in kernel density estimation for samples drawn from exponentially distributed populations.

Population distribution	Sample size	$\hat{h}_L$	$\hat{h}_{MS,G}$	$\hat{h}_{MS,A}$	$\hat{h}_A$	(ν, k, μ) for $K_A \in S^{\mu}_{\nu,k}$
	10	0.0106	0.0447	0.0988	0.1790	(0,2,0)
	20	0.0051	0.0381	0.1441	0.2066	(0,4,0)
	30	0.0084	0.0369	0.0877		
	40	0.0013	0.0392	0.0869	0.0742	(0,2,0)
$X \sim Exp(0.1)$	50	0.0019	0.0503	0.1114	0.0592	(0,2,0)
	60	0.0076	0.0456	0.1010		
	70	0.0016	0.0431	0.0955		
	80	0.0046	0.0417	0.0922		
	90	0.0028	0.0374	0.0828		
	100	0.0004	0.0411	0.0911		
	10	1.4195	5.0052	11.0806	17.5863	(0,2,0)
	20	0.0771	2.3684	5.2431	5.1285	(0,2,0)
	30	0.6970	4.6805	10.3617	9.6334	(0,2,0)
	40	0.2741	4.0106	8.8786	6.4240	(0,2,0)
$X \sim Exp(10)$	50	0.2647	3.9986	8.8521		
	60	0.0018	3.4048	7.5375		
	70	0.1115	2.4637	5.4542		
	80	0.1416	3.2939	7.2921		
	90	0.0001	3.2411	7.1752		
	100	0.1149	3.5167	7.7853		

When the method of acceptable interval is used it can be noticed that right boundary of the interval is smaller when Gaussian kernel is used in the constructing kernel density estimation. It is a rule independently on the form of kernel function set in the automatic method of kernel method parameters selection. It may mean that smoothing parameter that is set in maximum smoothing method is the smallest when Gaussian kernel function (that is kernel with unbounded support) is used, compering to the situation when kernel density estimator is constructing with maximum smoothing parameter and kernel function indicated by automatic method.

Automatic procedure of kernel method selection in all cases give us bigger values of smoothing parameters (compering to parameters from acceptable intervals) but

kernel function suggested by this method is kernel of order two and more. Whats more, kernel function of order more than two is used in the constructing kernel density estimator in theses samples when the left boundary of acceptable interval is small.

Values of left end in proposed new interval of values of smoothing parameter are in all cases bigger, so the kernel function indicated in automatic choice may have properties of low level of smoothing in kernel density estimation. It should be used always with bigger values of smoothing parameters.

It may mean that regarded methods of choosing smoothing parameter cannot be treated as opposed methods. In these cases there is a need to consider also the form of kernel function in kernel density estimation.

References

[1] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London 1986.

[2] M. P. Wand, M. C. Jones, Kernel Smoothing, Chapman and Hall, London 1995.

[3] S. Efromovitch, Nonparametric Curve Estimation: Methods, Theory and Applications, Springer-Verlag, New York 1999.

[4] L. Györfi, M. Kohler, A. Krzyak and H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer-Verlag, New York 2002.

[5] A. Baszczyńska, *Empirical and kernel estimation of the ROC curve*, Acta Universitatis Lodziensis. Folia Oeconomica **311** (2015), 49–55.

[6] A. Baszczyńska, *Kernel estimation of cumulative distribution function for random variable with bounded support*, Statistics in Transition new series **17**, no. 3 (2016), 541–556.

[7] S. J. Sheather, J. S. Marron, *Kernel quantile estimators*, Journal of the American Statistical Association **85** (1990), 410–416.

[8] M. Jones, *Estimating densities, quantiles, quantile densities and density quantiles*, Annals of the Institute of Statistical Mathematics **44**, no. 4 (1992), 721–727.

[9] P. Kulczycki, A. L. Dawidowicz, *Kernel estimator of quantile*, Universitatis Iagellonicae Acta Mathematica **XXXVII** (1999), 325–336.

[10] Q. Li, J. S. Racine, Nonparametric Econometrics. Theory and practice, Princeton University Press, Princeton and Oxford 2007.

[11] A. Baszczyńska, *On statistical kernel testing for equality of two probability density functions*, Models and Methods for Analysing and Forecasting Economic Processes, Cracow University of Economics Press (2014), 191–200.

[12] M. Rosenblatt, *Remarks on some nonparametric estimation of a density function*, Annals of Mathematical Statistics **27**, no. 3 (1956), 832–837.

[13] E. Parzen, *On estimation of a probability density function and mode*, Annals of Mathematical Statistics **33**, no. 3 (1962), 1065–1076.

[14] S. Chiu, *Bandwidth selection for kernel density estimation*, The Annals of Statistics **19**, no. 4 (1991), 1883–1905.

[15] W. Härdle, Smoothing Techniques with Implementation in S, Springer Series in Statistics, Springer-Verlag, Berlin Heidelberg 1991.
[16] R. Cao, A. Cuevas and W.G. Manteiga, *A comparative study of several smoothing methods in density estimation*, Computational Statistics and Data Analysis **17** (1994), 153–176.
[17] S. Chiu, *A comparative review of bandwidth selection for kernel density estimation*, Statistica Sinica **6** (1996), 129–145.
[18] M. Jones, D. Signorini, *A comparison of higher-order bias kernel density estimators*, Journal of the American Statistical Association **92**, no. 439 (1997), 1063–1073.
[19] I. Horová, J. Koláček and J. Zelinka, Kernel Smoothing in Matlab. Theory and Practice of Kernel Smoothing, World Scientific, New Jersey 2012.
[20] A. Baszczyńska, *Testing significance of peaks in kernel density estimator by SiZer map*, Proceedings of the 8th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena (2014), 9–17.

Department of Statistical Methods
University of Łódź
Rewolucji 1905 41, PL-90-214 Łódź
Poland
E-mail: aleksandra.baszczynska@uni.lodz.pl

Presented by Julian Ławrynowicz at the Session of the Mathematical-Physical Commission of the Łódź Society of Sciences and Arts on June 22, 2017.

WARTOŚCI PARAMETRU WYGŁADZANIA W AUTOMATYCZNEJ PROCEDURZE I W PRZEDZIAŁACH AKCEPTOWALNYCH W JĄDROWEJ ESTYMACJI FUNKCJI GĘSTOŚCI

S t r e s z c z e n i e

Automatyczna procedura określania parametrów metody jądra pozwala na jednoczesny wybór dwóch parametrów metody: funkcji jądra i parametru wygładzania. To podejście upraszcza procedurę wyboru parametrów, a jednocześnie zapewnia dobre właściwości estymatorów jądra. Drugą procedurą, która jest w pracy, jest akceptowalny odstęp wartości parametrów wygładzania, co pozwala na bardziej uogólnione podejście do wyboru parametru wygładzania w szacowaniu jądra. W artykule przedstawiono wyniki analizy wartości parametrów wygładzania, ustalonych w procedurze automatycznej oraz procedury akceptowalnego odstępu wartości parametrów wygładzania w oszacowaniu funkcji gęstści. Porównanie tych wartości odbywa się w oparciu o wyniki stosowania metod symulacji. Na podstawie badań symulacyjnych proponuje się i przeanalizuje nowe odstępy wartości parametrów wygładzania.

Słowa kluczowe: estymacja jądrowa funkcji gęstości, parametr wygładzania, funkcja jądra, wybór automatyczny