ITM Web of Conferences **23**, 00018 (2018) https://doi.org/10.1051/itmconf/20182300018
XLVIII Seminar of Applied Mathematics

The study and comparison of one-dimensional kernel estimators – a new approach. Part 2. A hydrology case study

Maciej Karczewski[1], and *Andrzej* Michalski[1,*]

[1] Department of Mathematics, Wroclaw University of Environmental and Life Sciences
Grunwaldzka 53, 50-357 Wroclaw, Poland

Abstract. The main purpose of this article is to present the numerical consequences of selected methods of kernel estimation, using the example of empirical data from a hydrological experiment [1, 2]. In the construction of kernel estimators we used two types of kernels – Gaussian and Epanechnikov – and several methods of selecting the optimal smoothing bandwidth (see Part 1), based on various statistical and analytical conditions [3–6]. Further analysis of the properties of kernel estimators is limited to eight characteristic estimators. To assess the effectiveness of the considered estimates and their similarity, we applied the distance measure of Marczewski and Steinhaus [7]. Theoretical and numerical considerations enable the development of an algorithm for the selection of locally most effective kernel estimators.

1 Introduction

The results presented in this paper are an essential extension of the results of the paper [2], which considered only the Gaussian kernel *K* and specific window smoothing dependent upon the sample size *n* and some parameter σ from the kernel *K*. There the estimator $\hat{f}_n$ of the unknown function *f* was expressed as

$$\hat{f}_n(x) = \frac{1}{\sqrt{2\pi n}\sigma}\sum_{i=1}^{n} e^{\left(\frac{-(X_i-x)^2 n}{2\sigma^2}\right)},$$

Here we consider over a dozen kernel estimators $\hat{f}_n$ of the density function *f* for two kernels – the Gaussian kernel and the kernel given by Epanechnikov – using several different methods: Silverman's rule of thumb, the Sheather–Jones method, cross-validation methods and other selected plug-in methods. The goal of this paper is to present a method that allows the study of local properties of the examined kernel estimators.

2 Numerical results and discussion

Based on the experimental data (groundwater levels) presented in Figure 1 (Part 1; for details see [2]), 17 kernel estimators were numerically determined with the optimal smoothing bandwidth according to the methods described in section 2 (Part 1). For further analysis, we selected eight characteristic kernel estimators: Silverman's [5], Sheather & Jones [8], unbiased cross-validation, biased cross-validation [9], Altman & Leger [8], Bowman with Gaussian kernel, Bowman with Epanechnikov kernel [10, 11], and Polansky & Baker [12] (see Figure 1).

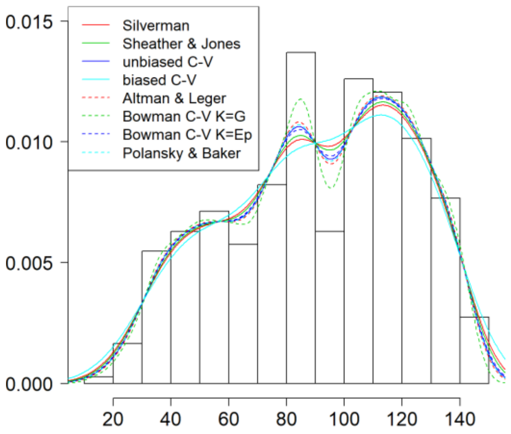


Fig. 1. Selected kernel estimators of the density function against the frequency histogram of groundwater levels.

Figure 1 shows the very similar behavior of individual estimators against the background of the source frequency histogram, and it appears that the best bimodality for these hydrological data is produced by Bowman's cross-validation estimator with the Gaussian kernel. To get a more accurate assessment of the differences between estimators compared locally (for each interval $[x_i , x_{i+1})$), some adequate measure of distance (similarity measure) for these estimated density functions of the continuous random variable is required. To compare the obtained kernel density estimates ($\hat{f}_i$) against the frequency polygon ($\hat{f}_o$) for the considered feature, we used the Marczewski–Steinhaus metric, as follows:

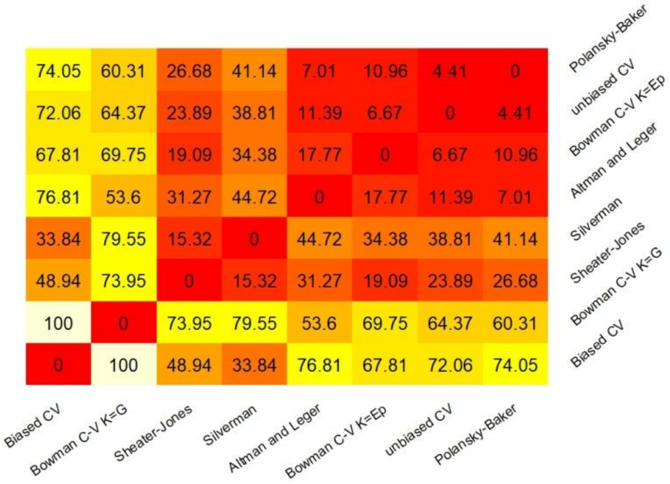
[*] Corresponding author: apm.mich@gmail.com

$$\sigma_\mu(\hat{f}_i,\hat{f}_0)=\frac{\int|\hat{f}(x)-\hat{f}_0(x)|d\mu(x)}{\max\left(|\hat{f}(x)|,|\hat{f}_0(x)|)dx\right)},$$

We use the above formula for the selected kernel estimators representing different approaches in relation to the empirical frequency polygon (see Figure 1 in Part 1 and in Part 2) and we determine their effectiveness in distinct variability intervals [x_i , x_{i+1}) for i = 0,…,14 (x_0 = 5 and x_{15} = 155) in accordance with the experiment conducted. As a result, we obtain a distance matrix of size 8 x 15 (objects x characteristics), where the objects are the kernel estimators and the role of features (characteristics) is performed by relative efficiency in separate intervals. In Table 1 we present the obtained numerical results: the last row contains the minimum values for individual ranges, and the last column shows values of the metric in the entire range of variability (i.e. [5, 155] for individual estimates). The minimum values in the whole matrix are marked in bold font.

Analysis of the obtained distance matrix allows one to identify the best of the studied estimators (digits in bold): mainly Bowman's cross-validation estimator with the Gaussian kernel, in the variability range from 0 to 115 cm, next the Sheather–Jones estimator in the interval (115; 125), Silverman's estimator in the interval (115; 135), the unbiased cross-validation estimator in the interval (135; 145), the Polansky–Baker estimator in the interval (135; 145) and Bowman's cross-validation estimator with the Epanechnikov kernel in the interval (135; 155). Hence, it is difficult to indicate a single estimator having the best properties in all intervals.

Next, for the thus obtained distance matrix, to define groups (taxa) with similar behavior, taxonomic methods were used. In our case the complete linkage method was applied. Figure 2 below presents numerical results for the distance matrix of the examined estimators, and Figure 3 shows the corresponding dendrogram – the smaller the distance measure, the greater the degree of similarity between the test functions.



	Biased CV	Bowman C-V K=G	Sheater-Jones	Silverman	Altman and Leger	Bowman C-V K=Ep	unbiased CV	Polansky-Baker
Polansky-Baker	74.05	60.31	26.68	41.14	7.01	10.96	4.41	0
unbiased CV	72.06	64.37	23.89	38.81	11.39	6.67	0	4.41
Bowman C-V K=Ep	67.81	69.75	19.09	34.38	17.77	0	6.67	10.96
Altman and Leger	76.81	53.6	31.27	44.72	0	17.77	11.39	7.01
Silverman	33.84	79.55	15.32	0	44.72	34.38	38.81	41.14
Sheater-Jones	48.94	73.95	0	15.32	31.27	19.09	23.89	26.68
Bowman C-V K=G	100	0	73.95	79.55	53.6	69.75	64.37	60.31
Biased CV	0	100	48.94	33.84	76.81	67.81	72.06	74.05

Fig. 2. Distance matrix of kernel estimators (values of *d* in %).

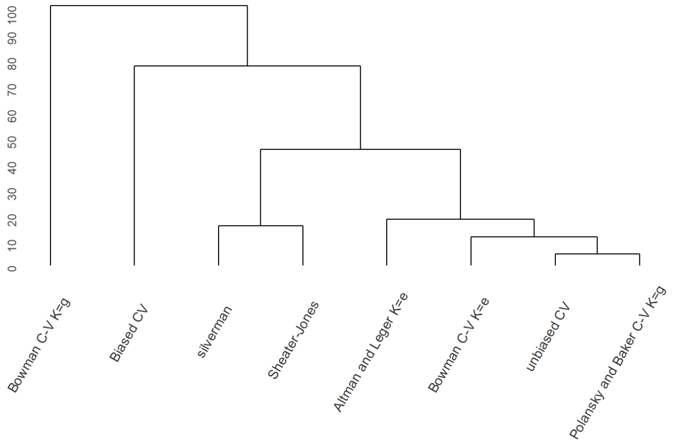


Fig. 3. Dendrogram of similarity of kernel estimators (similarity s = 100–d).

Table 1. Matrix (8 x 15) of distances according to the Marczewski–Steinhaus measure.

estimator	5-15	15-25	25-35	35-45	45-55	55-65	65-75	75-85	85-95	95-105	105-115	115-125	125-135	135-145	145-155	5-155
Silverman	0.606	0.293	0.112	0.086	0.043	0.080	0.121	0.129	0.162	0.133	0.081	**0.014**	**0.013**	0.046	0.236	0.094
Sheather & Jones	0.550	0.252	0.105	0.074	0.037	0.074	0.113	0.119	0.154	0.126	0.072	**0.014**	0.015	0.038	0.163	0.085
unbiased C-V	0.438	0.176	0.093	0.053	0.027	0.064	0.097	0.100	0.135	0.109	0.056	0.019	0.019	**0.033**	0.086	0.071
biased C-V	0.730	0.391	0.130	0.121	0.061	0.096	0.137	0.151	0.176	0.146	0.105	0.044	0.040	0.068	0.385	0.121
Altman & Leger	0.386	0.141	0.087	0.045	0.023	0.060	0.089	0.091	0.125	0.100	0.050	0.023	0.021	0.035	0.126	0.066
Bowman C-V K=G.	**0.191**	**0.052**	**0.077**	**0.019**	**0.016**	**0.051**	**0.060**	**0.060**	**0.069**	**0.053**	**0.028**	0.039	0.024	0.066	0.404	**0.051**
Bowman C-V K=E.	0.473	0.199	0.096	0.059	0.030	0.067	0.102	0.106	0.142	0.115	0.061	0.017	0.018	**0.033**	**0.079**	0.075
Polansky & Baker	0.417	0.162	0.090	0.050	0.025	0.062	0.094	0.096	0.131	0.106	0.054	0.021	0.020	**0.033**	0.099	0.069

Looking closely at the dendrogram of similarity of the kernel estimators and making cuts at the level d = 20%, we obtain the following division of estimators into taxa, i.e. into groups of objects with similar properties. Figures 4–7 show the obtained groups (taxa) of kernel estimators with similar behavior in the ranges of variability of the examined feature (groundwater level) defined by the experimenter. This analysis allows us to further reduce the set consisting of a large number of different kernel estimators obtained on the basis of different analytical and statistical concepts.

Note. All numerical calculations were performed using the authors' own original procedures on the R platform and using packages from the literature: [13–15].

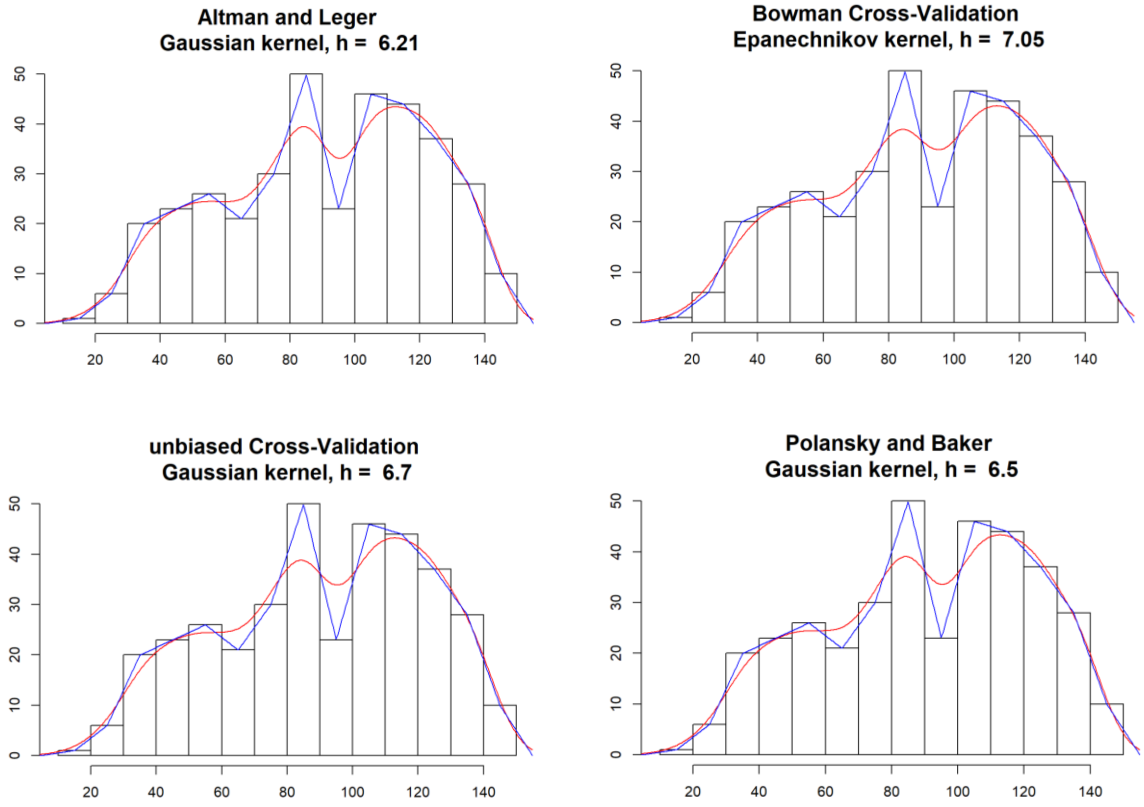


Fig. 4. Taxon I with similarity measure s = 80% (or equivalently distance d = 20%).

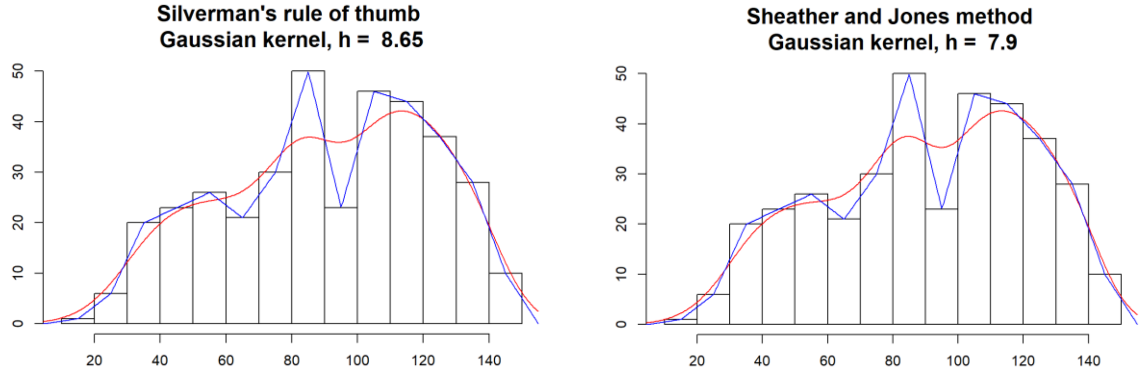


Fig. 5. Taxon II with similarity measure s = 80% (or equivalently distance d = 20%).

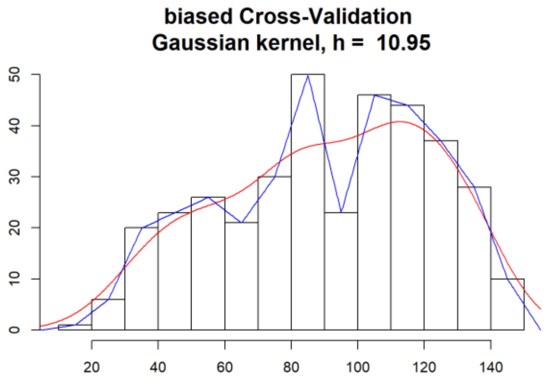


Fig. 6. Taxon III with similarity measure *s* = 20% (*d* = 80%).

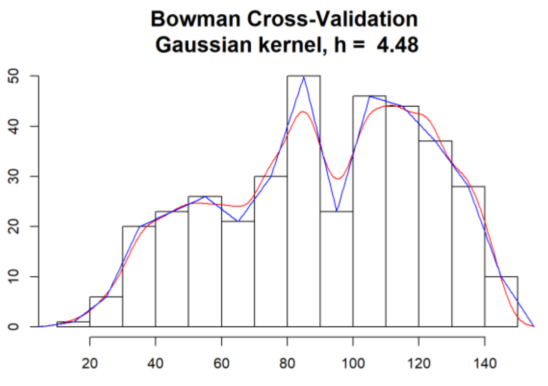


Fig. 7. Taxon IV with similarity measure *s* = 5% (*d* = 95%).

3 Conclusions

In the statistical literature there can be found a wealth of different approaches to obtaining the best estimate of the unknown density function of a continuous type random variable by nonparametric kernel estimation methods. Knowledge of the unknown density function that best reflects the probabilistic data structure is invaluable in the process of predicting values of the studied phenomenon. Using the example of hydrological data, the authors have suggested a way of classifying kernel estimators, chosen from those most often used in practice. Based on our calculations, we have shown that none of the considered estimators have optimal properties on the entire region. This is a known phenomenon in mathematical statistics related to the study of the admissibility of statistical decision rules. Each of the assessed estimators has good local properties, but their behavior is strictly dependent on empirical data. For example, Bowman et al. (1998) performed a simulation study comparing this method with the plug-in method of Altman and Leger. Better results are obtained, in general, with cross-validation (cf. [11]). Plug-in methods apply a pilot bandwidth to estimate one or more important features of the density function *f*. The bandwidth for estimating *f* itself is then estimated at a second stage using a criterion that depends on the estimated features. The best plug-in methods have proven to be very effective in diverse applications and are more popular than cross-validation approaches (see [4, 16]). However, other authors offer arguments against the uncritical rejection of cross-validation approaches.

Our results and the considerations of many authors give us the incentive to look for a solution that will allow us to use the best behavior of the tested estimators in a given area of variability, by suggesting, for example, an estimator that would be a convex linear combination of selected estimators. In 1989 Devroye introduced and developed the very interesting concept of the double kernel method for density estimation [17], and its usefulness has been demonstrated in extensive simulation studies [16]. In the double kernel method, we take two different kernels *K* and *L* whose characteristic functions do not coincide on any open neighborhood of the origin.

Only comprehensive knowledge of the efficiency and properties of the kernel density estimators for the one-dimensional case will allow us to consider cases of two-dimensional or three-dimensional random variables with greater awareness. The problem of stochastic modeling of hydrological or meteorological data using methods of multivariate density function estimation is more difficult and complex [18].

References

1. E. Gąsiorek, A. Michalski, and A. Pływaczyk, Zeszyty Naukowe AR **192** (1990)
2. A. Michalski, MHWM **Vol. 4**, 1, 41-46 (2016)
3. W. Feluch, J. Koronacki, Comput. Stat. Data Anal **13**, 143-151 (1992)
4. G.H. Givens, J.A. Hoeting, *Computational Statistics* (New York, Wiley & Sons, 2005)
5. B.W. Silverman, *Density Estimation for Statistics and Data Analysis* (London, Chapman & Hall, 1986)
6. M.P. Wand, M.C. Jones, *Kernel Smoothing* (London, Chapman & Hall, 1995)
7. E. Marczewski, H. Steinhaus, Colloq Math **6**, 319-327 (1958)
8. N. Altman, C. Léger, J. Stat. Plan, Inference **46**, 195-214 (1995)
9. D.W. Scott, G.R. Terrell, J. Amer. Statist. Assoc. **82**, 1131-46 (1987)
10. A.W. Bowman, Biometrika **71** (1984)
11. A.W. Bowman, P. Hall, T. Prvan, Biometrika **85** (1998)
12. A. Polansky, E.R. Baker, J. Stat. Comp. Sim **65** (2000)
13. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018)
URL https://www.R-project.org/ (2018)

14. A. Quintela-del-Rio, G. Estevez-Perez, J. Stat. Softw. **50(8)**, 1-21 (2012)
15. A.C. Guidoum. *kedd: Kernel estimator and bandwidth selection for density and its derivatives*, R package version 1.0.3.

    http://CRAN.R-project.org/package=kedd (2015)
16. A. Berlinet, L. Devroye, Publications de l'Institut de Statistique de l'Université de Paris, vol. XXXVIII – Fascicule **3**, 3- 59 (1994)
17. L. Devroye, Annales de l'I H. P. **25**, 533-580 (1989)
18. D.W. Scott, *Multivariate density estimation. Theory, Practice and Visualization* (J. Wiley & Sons, Inc., 1992)