ITM Web of Conferences **23**, 00017 (2018) https://doi.org/10.1051/itmconf/20182300017
XLVIII Seminar of Applied Mathematics

The study and comparison of one-dimensional kernel estimators – a new approach. Part 1. Theory and methods

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. In this article we compare and examine the effectiveness of different kernel density estimates for some experimental data. For a given random sample $X_1, X_2, \ldots, X_n$ we present eight kernel estimators $\hat{f}_n$ of the density function f with the Gaussian kernel and with the kernel given by Epanechnikov [1] using several methods: Silverman's rule of thumb, the Sheather–Jones method, cross-validation methods, and other better-known plug-in methods [2–5]. To assess the effectiveness of the considered estimators and their similarity, we applied a distance measure for measurable and integrable functions [6]. All numerical calculations were performed for a set of experimental data recording groundwater level at a land reclamation facility (cf. [7–8]). The goal of the paper is to present a method that allows the study of local properties of the examined kernel estimators.

1 Introduction

In many natural (e.g. meteorological or climatic) and typical engineering (e.g. hydrological or soil science) problems, it is extremely important to obtain knowledge of the density function f of the probability distribution of features (X) describing the phenomenon being studied. This is also a fundamental concept in statistics. Specifying the function f gives a natural description of the distribution of X and allows one to determine the probabilities $P(X \in (a, b))$ for a<b. One approach to density estimation is parametric estimation (cf. the assessment of climate change impacts on a river runoff [9], models of atmospheric precipitation [10], river flow prediction for future climatic conditions [11] and many other papers in the fields of meteorology and hydrology). In this article we shall not be considering parametric estimates (even the commonly used generalized gamma distributions with 3 parameters) – the approach will be more nonparametric, in that less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has a probability density f, the data will be allowed to speak for themselves in determining the estimate f more than would be the case if f were constrained to fall within a given parametric family. The oldest and most widely used nonparametric density estimator is the histogram. This naive estimator can be written as

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - X_i}{h}\right),$$

$$\text{where } w(x) = \begin{cases} \frac{1}{2} & if\ |x| < 1 \\ 0 & if\ |x| \geq 1 \end{cases},$$

for a given random sample $X_1, X_2 \ldots X_n$ and a fixed parameter h.

It is easy to see that the estimate is constructed by placing a "box" of width $2h$ (called the binwidth) and height $1/(2nh)$ on each observation and summing to obtain the estimate. The generalization of this histogram estimator consists in the replacement of the function w by a kernel function K which satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1,$$

Now, by analogy with the definition of the naive estimator, for a given random sample $X_1, X_2, \ldots, X_n$ the kernel estimator $\hat{f}_n$ with kernel K is defined by

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

where h is the window width, also called the smoothing parameter or bandwidth.

In the statistical literature, the estimator of the density function of a random variable X given by (1) under more general assumptions is called the Parzen–Rosenblatt estimator or the Akaike–Parzen–Rosenblatt estimator (see [1, 12, 13]). The results presented here are an important extension of the results of the paper [8], which considered only the Gaussian kernel and specific window smoothing dependent upon the sample size and some parameter from the kernel K. We shall consider eight kernel estimators $\hat{f}_n$ of the density function f with two kernels – the Gaussian kernel and the kernel given

---

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

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 main aim is to compare the examined kernel estimators, their effectiveness and the prognostic results.

To assess the effectiveness of the considered estimates and their similarity, we applied a distance measure for measurable and integrable functions proposed by Marczewski and Steinhaus [6]. All numerical calculations were performed for a set of hydrological data recording groundwater level at a land reclamation facility (see [7, 8]).

2 Material and methods

Land reclamation studies at the foothill site of Długopole (area approximately 1.5 ha) included long-term measurements of groundwater level using properly installed piezometers (hydrogeological observation holes). Daily registered groundwater levels were averaged based on measurements from about a dozen piezometers suitably located at the research station (the experimental data are derived from the Institute of Agricultural and Forest Improvement at Wroclaw University of Environmental and Life Sciences). The data set includes groundwater level measurements based on 10-centimeter ranges of levels from 10 up to 150 cm. The experimental data aggregated in a frequency table were reproduced by repeated use of a random number generator with a given frequency structure (see [8], Table 1). Then we obtained the vector of values of groundwater level as (x_1, x_2, …, x_{366}). Based on these data, the frequency histogram of the groundwater level is drawn below.

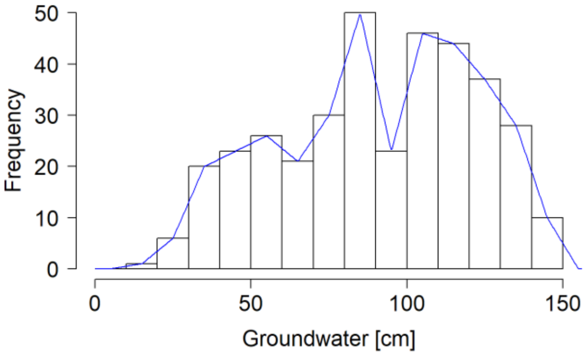

Fig. 1. Histogram of groundwater level with frequency polygon (the left axis describes non-relative frequency as number of days).

Throughout this paper we consider the two most often used kernel functions: the kernel K_g, being a function of the density of the normal distribution N(0,1) (Gaussian kernel), i.e.

$$K_g(x) = \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-x^2}{2}\right)}, \tag{2}$$

or more generally

$$K_g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-x^2}{2\sigma^2}\right)}, for\ \sigma > 0$$

(see [15]) and the Epanechnikov kernel given by

$$K_e(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}x^2\right)\mathbf{1}_{\{|x|<\sqrt{5}\}},$$

or in a simpler version

$$K_e(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{\{|x|<1\}}, \tag{3}$$

for which the optimality properties in the density estimation setting were first described by Epanechnikov (see e.g. [8]). Graphs of the kernels are shown below in Fig. 2.

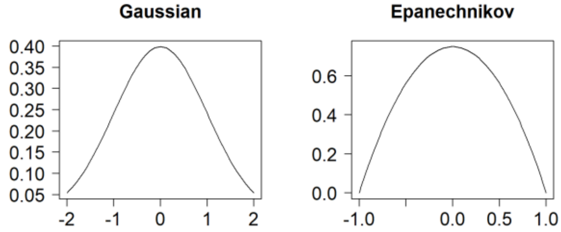

Fig. 2. The Gaussian kernel K_g (left panel) and the Epanechnikov kernel K_e (right panel).

The structure of the kernel estimator indicates that in density estimation it is important to select not only the appropriate kernel, but also the optimal bandwidth smoothing. The value chosen for the bandwidth *h* exerts a strong influence on the estimator $\hat{f}_n$ (see Figure 3). The bandwidth parameter controls the smoothness of the density estimate. An optimal solution in the problem of selection of the bandwidth *h* is to minimize the integrated squared error given by

$$\begin{aligned} ISE(h) &= \int \hat{f}^2(x)\,dx - 2E[\hat{f}(x)] + \int f^2(x)\,dx \\ &= R(\hat{f}) - 2E\{\hat{f}(x)\} + R(f) \end{aligned} \tag{4}$$

where *R(g)* denotes a measure of the roughness of a given function *g*, defined by $R(g) = \int g^2(t)\,dt$, or to minimize the mean integrated squared error

$$\begin{aligned} MISE(h) &= E\int \left(\hat{f}(x) - f(x)\right)^2 dx \\ &= \int \left\{var[\hat{f}(x)] + \left(bias[\hat{f}(x)]\right)^2\right\} dx \end{aligned} \tag{5}$$

where $bias[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$. To compute the bias term in (5), note that

$$E[\hat{f}(x)] = \frac{1}{h}\int K\left(\frac{x-t}{h}\right) f(t)dt = \int K(t) f(x - ht)dt$$

by applying a change of variable. We further analyze the expression (5), allowing *h*→ 0 and *nh*→ ∞ as *n* → ∞.

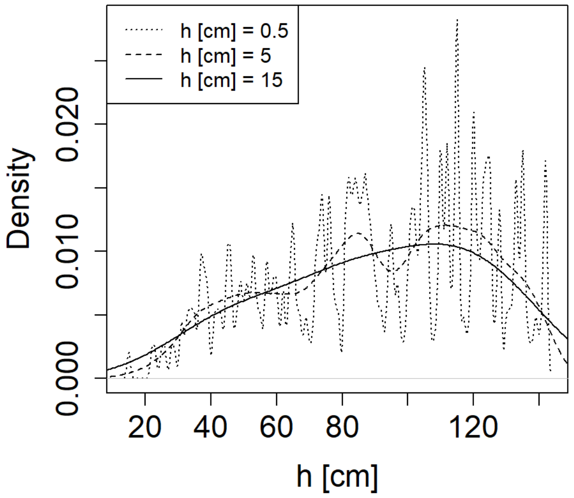

Fig. 3. Examples of bandwidth h for experimental data (groundwater levels) and the Gaussian kernel.

Note that the large bandwidth causes important features of f to be smoothed away, thereby causing bias. To understand bandwidth selection it is necessary to analyze carefully the expression MISE(h) (see e.g. [4]). In the numerical analysis of the selection of the optimal smoothing bandwidth, the AMISE is also taken into account, defined as

$$\mathrm{AMISE}\,(\mathrm{h}) = \frac{R(K)}{nh} + \frac{h^4\sigma_K^2 R(f'')}{4}, \tag{6}$$

where f must have two bounded continuous derivatives and $R(f'') < \infty$ and $\sigma_K^2 < \infty$ is the variance of the kernel K. It can be shown that the following relationship holds:

$$MISE\,(h) = AMISE\,(h) + o\left(\frac{1}{nh} + h^4\right) \tag{7}$$

where $o(h^4)$ is a quantity that converges to 0 faster than h^4 does as $h \to 0$. If $h \to 0$ and $nh \to \infty$ as $n \to \infty$ then MISE(h) $\to$ 0. The numerical considerations show that the optimal bandwidth is

$$h = \left(\frac{R(K)}{n\sigma_K^4 R(f'')}\right), \tag{8}$$

but this is a theoretical result; the quantity h essentially depends on the unknown density f.

In the statistical literature devoted to non-parametric kernel estimation of an unknown density function, many methods can be found based on various premises. Here we shall present two methods and several variants of them. One of the most interesting is the cross-validation method, which uses the data in two ways: once to calculate the kernel estimator $\hat{f}_n$ from the data, and a second time to evaluate the quality of $\hat{f}_n$ as an estimator of f. Let $\hat{f}_{-i}$ be the density estimate constructed from all of the data points except X_i, as follows:

$$\hat{f}_{-i}(x) = \frac{1}{h(n-1)}\sum_{j\neq i} K\left(\frac{x - X_k}{h}\right). \tag{9}$$

Considering the expression (4), we notice that the last term is constant, and the middle term can be estimated by $\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i)$. Now, to get a good bandwidth, it is enough to minimize

$$UCV(x) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i), \tag{10}$$

with respect to h. Use of this criterion is called unbiased cross-validation or least squares cross-validation, because choosing h to minimize UCV(h) minimizes the integrated squared error between $\hat{f}_n$ and f. Instead of the exact MISE formula given by (5), we can use a criterion based on the formula for the asymptotic MISE given by (6); this is called biased cross-validation (BCV). To get a good bandwidth we should minimize the expression given by

$$BCV(h) = \frac{R(K)}{nh} + \frac{h^4\sigma_K^4\tilde{R}(f'')}{4}, \tag{11}$$

where BCV(h) is obtained by replacing the unknown $R(f'')$ in (6) by its the estimator $\tilde{R}(f'')$ (see [14]). An alternative method of cross-validation for the smoothing of density estimates was proposed by Bowman (see [15, 16]). The smoothing parameter h is chosen to minimize the cross-validation function written as

$$CV_B(h) = \frac{1}{n}\sum_{k=1}^{n}\int\left\{I(x - X_i) - \hat{f}_{-i}(x)\right\}^2 dx, \tag{12}$$

where $I(x\text{-}X_i) = 1$ if $x\text{-}X_i \geq 0$ and $I(x\text{-}X_i) = 0$ if $x\text{-}X_i < 0$, and $\hat{f}_{-i}(x)$ is given by (9).

In addition to cross-validation methods, and with similar success, many authors have proposed so-called plug-in methods. Such a method for obtaining an optimal smoothing factor in the space L_2 of the square integrable functions was introduced by Woodroofe [17], who obtained an asymptotically optimal expression for the optimal h as a function of f and n. In a second step, he estimated the unknown functional of f (in this case, $R(f'') = E\int(f'')^2$) from the data in a nonparametric manner using a pilot bandwidth. To minimize $E\int(\hat{f}_n - f)^2$ when f is sufficiently smooth and K is a nonnegative kernel, the asymptotically optimal h has the following form:

$$h = \left(\frac{A(K)}{nR(f'')}\right),$$

$$where \quad A(K) = \int K^2(x)dx \Big/ \left(\int x^2K(x)\,dx\right)^2.$$

This formula is at the heart of the plug-in method (cf. (8)).

In our numerical analyses, we used four different plug-in approaches: Silverman's rule of thumb [3], and the methods of Sheather–Jones [18], Altman–Leger [18] and Polansky–Baker [19].

For example, Silverman's rule of thumb gives $h = (4/3n)^{1/5}\,\hat{\sigma} \approx 1.06\,\hat{\sigma} n^{-1/5}$, where $(\hat{\sigma})^2$ is the sample variance. A better solution in a fairly complex process for finding the bandwidth h is given by an approach known in the literature as the Sheather–Jones method. This is a two-stage process. At the first stage, a simple rule of thumb is used to calculate the bandwidth h_0 expressed by

$$h_0 = C_1\big(R(f''), R(f''')\big)C_2(L)h^{5/7}, \tag{13}$$

where C_1 and C_2 are functionals that depend respectively on the derivatives of f and on the kernel L.

The bandwidth h_0 is used to estimate $R(f'')$, being the only unknown in expression (14) for the optimal bandwidth given by

$$\hat{f}''(x) = \frac{1}{nh_0^3}\sum_{i=1}^{n} L''\left(\frac{x - X_k}{h_0}\right), \tag{14}$$

where L is a sufficiently differentiable kernel used to estimate f''. The estimation of $R(f'')$ in the formula (8) follows from (14). At the second stage we solve the equation

$$\left(\frac{R(K)}{n\sigma_K^4\hat{R}(f'')}\right) - h = 0, \tag{15}$$

The solution to (15) can be found using, for example, grid search or a root-finding technique such as Newton's method. This calculation method is quite complex, even using a Gaussian kernel (cf. [4]).

An alternative approach to selecting a bandwidth is to use an estimator of the asymptotically optimal bandwidth. This was proposed as a plug-in estimate by Altman and Leger (see [18]) as follows:

$$\hat{h}_{opt} = \left(0.25\hat{V}_2/\hat{B}_3\right)^{1/3} n^{-1/3}, \tag{16}$$

where $\hat{V}_2$ and $\hat{B}_3$ are estimators of the expressions V_2 and B_3 respectively, and

$$V_2 = 2\int xk(x)K(x)dx \cdot \int [f(x)]^2 W(X)dx \quad and$$

$$B_3 = 0.25\int [x^2k(x)]^2 dx \cdot \int [f'(x)]^2 f(x)W(X)\,dx$$

with $K(x) = \int_{-\infty}^{x} k(t)\,dt$ for a positive kernel k and a nonnegative weight function $W(x)$ (see [18]).

The fourth plug-in method of kernel estimation, proposed by Polansky and Baker [19], involves estimation of the asymptotically optimal bandwidth h_0 as

$$h_0 = \left(\frac{\rho(k)}{n\mu_2^2 R(f')}\right)^{1/3}, \tag{17}$$

where

$$\rho(k) = 2\int_{-\infty}^{\infty} xk(x)K(x)dx,$$

$$\mu_2(k) = \int_{-\infty}^{\infty} t^2k(t)dt \quad and$$

$$R(f') = \int_{-\infty}^{\infty} [f'(x)]^2 dx.$$

Note. The method for selecting the optimal smoothing bandwidth proposed by Altman and Leger [18] and the modified four-stage method given by Polansky and Baker [19] concern kernel estimation of the distribution function, in contrast to the methods for the density function. It turns out that the values of h that optimize global measures of the accuracy of $\hat{F}_{n,h}$ are different from those for $\hat{f}_{n,h}$. Therefore, the vast array of estimation techniques used in density estimation are not directly applicable for kernel distribution function estimates (see e.g. [19]).

To assess the efficiency of the obtained estimators, the question arises of how to compare them – that is, how to determine a distance measure between measurable and integrable functions in the same space. To compare the obtained kernel density estimates against the polygon frequency of the feature, we used the Marczewski–Steinhaus metric presented in the paper *On a certain distance of sets and the corresponding distance of functions* [6].

For non-negative and μ-integrable functions f and g we define the metric σ_μ as follows:

$$\sigma_\mu(f, g) = \frac{\int |f(x) - g(x)|\,d\mu(x)}{\int \max\left(|f(x)|, |g(x)|\right)d\mu(x)}, \tag{18}$$

Next, we use the above formula for the selected kernel estimators representing different approaches in relation to the empirical frequency polygon (see Figure 1) and we determine their effectiveness in distinct variability intervals $[x_i, x_{i+1})$ for $i=0,\ldots,14$ ($x_0 = 0$ and $x_{15} = 150$) in accordance with the experiment conducted. In this way, we obtain a distance matrix of (objects x features), where the objects are kernel estimators and the role of features is performed by relative efficiency in separate intervals. For the thus obtained distance matrix, taxonomic methods (the complete linkage method) are used to define groups (taxa) with similar behavior.

3 Discussion

It is worth noting that the use of density estimates is part of the informal investigation of the properties of a given set of data. Density estimates can give valuable indication of such features as skewness and multimodality in the data. There is a vast literature on density estimation, much of it concerned with asymptotic results (see [5]). Key features of kernel estimators include strong consistency, asymptotic unbiasedness and uniform convergence (see [1, 13]). The essence of all methods of kernel estimation is the optimal choice of the smoothing bandwidth. Our considerations in this article concern the space L_2 of square integrable functions. Among many methods of selecting the bandwidth, we can distinguish the following: the cross-validation method, various plug-in methods, the maximum smoothing principle, the bootstrap method, the projection method, the spacings method, a method based on the Greenwood statistic, etc. (see e.g. [1]). In our numerical considerations, we used two types of commonly used kernels (Gaussian and Epanechnikov) and several methods of selecting the optimal smoothing bandwidth, based on various statistical and analytical conditions.

From among a dozen obtained density estimates, we selected eight estimators for further comparative analysis and assessment of their effectiveness. Because of the breadth of the numerical material, all calculations, technical details and conclusions will be presented by the authors in a second article on applications in hydrology (Part 2).

References

1. A. Berlinet, L. Devroye, Publications de l'Institut de Statistique de l'Université de Paris, vol. XXXVIII – Fascicule **3**, 3- 59 (1994)
2. W. Feluch, J. Koronacki, Comput. Stat. Data Anal. **13**, 143-151 (1992)
3. B.W. Silverman, *Density Estimation for Statistics and Data Analysis* (London, Chapman & Hall, 1986)
4. G.H. Givens, J.A. Hoeting, *Computational Statistics* (New York, Wiley & Sons, 2005)
5. M.P. Wand, M.C. Jones, *Kernel Smoothing* (London, Chapman & Hall, 1995)
6. E. Marczewski, H. Steinhaus, Colloq Math **6**, 319-327 (1958)
7. E. Gąsiorek, A. Michalski, and A. Pływaczyk, Zeszyty Naukowe AR **192**, Wrocław (1990)
8. A. Michalski, MHWM **Vol. 4**, 1, 41-46, (2016)
9. L. Kuchar, S. Iwański, L. Jelonek, and W. Szalińska, Geografie **119(1)**, 1-25 (2014)
10. L. Kuchar, Comput. Symul. **65**, 69-75 (2004)
11. L. Kuchar, S. Iwanski, L. Jelonek, E3S Web of Conf. 17, doi: 10.1051/e3sconf/20171700046 (2017)
12. H. Akaike, Ann. Inst. Stat. Math. **6**, 127– 132 (1954)
13. L. Devroye, T.J. Wagner, Technical Report 183, Electronic Research Center the University of Texas at Austin (1976)
14. D.W. Scott, G.R. Terrell, J. Amer. Statist. Assoc. **82**, 1131-46 (1987)
15. A. W. Bowman, Biometrika **71** (1984)
16. A.W. Bowman, P. Hall, T. Prvan, Biometrika **85** (1998)
17. M. Woodroofe, An. Math. Stat. **41** (1970)
18. N. Altman, C. Léger, J. Stat. Plan, Inference **46**, 195-214 (1995)
19. A. Polansky, E.R. Baker, J. Stat. Comp. Sim. **65** (2000)