Journal of Physics: Conference Series

PAPER • OPEN ACCESS

About specialities of numerical estimation of smoothing parameter of probability density functions of random sequences in Parzen-Rosenblatt approximation

To cite this article: Sergey Porshnev and Alexander Koposov 2018 *J. Phys.: Conf. Ser.* **1053** 012093

View the article online for updates and enhancements.

Related content

- Radiative power and x-ray spectrum numerical estimations for wire array Z-pinches
O G Olkhovskaya, M M Basko, P V Sasorov et al.
- Numerical estimation of various influence factors on a multipoint hydrostatic leveling system
R V Tsvetkov, V V Yepin and A P Shestakov
- Numerical estimation of deformation energy of selected bulk oilseeds in compression loading
C Demirel, A Kabutey, D Herak et al.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 213.142.35.54 on 19/07/2019 at 23:46

About specialities of numerical estimation of smoothing parameter of probability density functions of random sequences in Parzen-Rosenblatt approximation

Sergey Porshnev and Alexander Koposov[1]

Engineering School of Information Technologies, Telecommunications and Control Systems, Ural Federal University named after the first President of Russia B.N. Yeltsin, street Mira 32, Ekaterinburg, 620000, Russia

[1]Email: alexkopas@gmail.com

Abstract. The methods of nonparametric statistics are very useful in data analysis. One of the most popular methods is called Parzen-Rosenblatt approximation. This method turns out to be effective, for example, in a problem of estimation of longevity of pipelines or in the analysis of the statistical characteristics of traffic flows. This paper discusses the recommendations for application of a method, which was performed by Parzen and Rosenblatt, in a problem of recovering a probability density function from a sample of random data with a bounded scattering region. It was shown that there are some difficulties during calculation of information functional. This paper gives an explanation of causes which lead to a nonmonotonicity of an information functional and which are based on a finite precision of computer calculations. It was proved a choice of initial value of smoothing parameter for different kernel types and was proposed an algorithm for finding a maximal value of information functional.

1. Preface

The main problem of mathematical statistics is recovering the distribution function from a sample of random data obtained as a result of some experiments [1]. This problem is of great practical importance, for example, when solving the problems of strength reliability of elements and objects of oil and gas equipment [2]. The problem has the following statement: using experimental sample values of a random variable $X_i, i=\overline{1,N}$ from general population to find out distribution function $F(y)=\Pr\{X \le y\}$ which is connected with probability density function $f(y)$ by the following integral statement:

$$F(y)=\int_{-\infty}^{y} f(\xi)d\xi, \tag{1}$$

There are two main approaches to solve this problem: parametric and nonparametric.

The parametric approach implies the possession of the information about type of distribution function depending on a certain parameters set. Based on this information and given sample of data it is possible to estimate parameters values which ensure maximal similarity of theoretical distribution function $F(y)$ and empirical distribution function

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N}\Theta(y - x_i), \tag{2}$$

Where Heaviside function

$$\Theta(y - x_i) = \begin{cases} 1, \text{ if } y - x_i \geq 0, \\ 0, \text{ if } y - x_i < 0, \end{cases} \tag{3}$$

in accordance with the chosen measure of proximity, depending, generally speaking, on distribution type [3].

The existence of solution of this problem is ensured by central theorem of mathematical statistics

$$\Pr\left\{\lim_{N\to\infty}\sup\left|F_N(y) - F(y)\right| = 0\right\} = 1. \tag{4}$$

The functions represented in table 1 are used as kernel functions.

Table 1. Kernel functions.

Kernel	Formula
Normal	$k(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$
Laplace	$k(t) = \frac{1}{2} e^{-\lvert t\rvert}$
Fisher	$k(t) = \frac{1}{2\pi}\left(\frac{\sin\left(\frac{t}{2}\right)}{\frac{t}{2}}\right), \left\lvert\frac{t}{2}\right\rvert \leq \pi$
Cauchy	$k(t) = \frac{1}{\pi}\left(\frac{1}{1+t^2}\right)$
Logistic	$k(t) = \frac{e^{-t}}{\left(1+e^{-t}\right)^2}$
Epanechnikov	$k(t) = \frac{3\cdot\left(1-\frac{t^2}{5}\right)}{4\sqrt{5}}, \quad \lvert t\rvert \leq \sqrt{5}$
Uniform	$k(t) = \frac{1}{2}, \quad \lvert t\rvert \leq 1$
Triangle	$k(t) = 1 - \lvert t\rvert, \quad \lvert t\rvert \leq 1$
Square	$k(t) = \frac{3\cdot\left(1-t^2\right)}{4}, \quad \lvert t\rvert \leq 1$

The nonparametric statistics are based on an approach that makes it possible to obtain adaptive estimates of empirical distributions in the form of some functionals independent of the form of the unknown a priori distribution [4]. To recovering the unknown distribution function in nonparametric statistics, a number of methods and algorithms are known [4]: histogram method, nearest neighbor method, Rosenblatt-Parzen method, decomposition for basis functions and others. For example, it was shown in [2] that the Rosenblatt-Parzen approximation proves to be very effective in the problem of

estimating the longevity of oil and gas pipelines based on an analysis of the accumulated statistical information.

Following [4], the discussed method of recovering the probability density function of the experimental sample is based on the assumption that the probability density function is estimated locally at each point x_i using elements of the training sample from some neighborhood of x_i. The total distribution function is the sum of local functions.

$$F(y) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{y - x_i}{h}\right), \tag{5}$$

where $K(t)$ is a kernel function that meets the following conditions:

a) $K(t)$ - monotonically non-decreasing function, $\forall t\, K(t) \in [0,1]$

b) $K(t) = 1 - K(-t)$

c) $h_N \to 0$ if $N \to \infty$;

$h-$ smoothing parameter that determines the smoothness of the resulting estimates

Accordingly, the probability density function is calculated by the formula

$$f(y) = \frac{1}{N \cdot h}\sum_{i=1}^{N} k\left(\frac{y - x_i}{h}\right), \tag{6}$$

where $k(y) = \frac{d}{dy}K(y).$

These estimates were proposed by Rosenblatt [5] and explored by Parzen [6].

In this method the quality of the approximation depends on the type of kernel function and on the value of the smoothing parameter h [4] (figure 1).

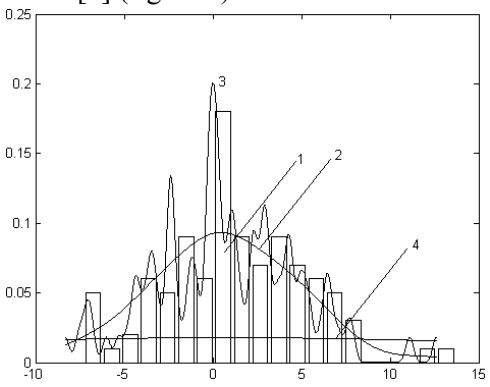

Figure 1. Probability density function of a random sequence $x_i, i = \overline{1,100}$ generated in accordance with the normal distribution law $N(1,4)$: 1 – histogram of random sequence; 2 – normal kernel $h = h^*_{opt}$; 3 – normal kernel $h < h^*_{opt}$; 4 – normal kernel $h > h^*_{opt}$.

The optimal values of the kernel function and smoothing parameter are found from the condition that the functional reaches its maximum value

$$J = \int \ln k(t) \cdot f(t) dt \tag{7}$$

In [2] it is recommended to find the optimal value h^*_{opt} for each of the kernel functions presented in table 1 and choose kernel with the most function value $\varphi(h^*_{opt})$

$$h_m^{\ *} = \arg\max\left\{\frac{1}{N}\sum_{i=1}^{N}\ln\left[\frac{1}{(N-1)\cdot h_m}\sum_{j\neq i}^{N-1}k_m\left(\frac{x_i - x_j}{h_m}\right)\right]\right\}, \tag{8}$$

$$\varphi(h_m) = \frac{1}{N}\sum_{i=1}^{N}\ln\left[\frac{1}{(N-1)\cdot h_m}\sum_{j\neq i}^{N-1}k_m\left(\frac{x_i - x_j}{h_m}\right)\right] \tag{9}$$

It can be seen from (6) that searching the optimal value of the smoothing parameter for each of the basis functions is equal to searching the solution of a complex nonlinear equation

$$\frac{\partial\varphi(h)}{\partial h} = \frac{\partial}{\partial h}\sum_{i=1}^{N}\ln\left[\frac{1}{(N-1)\cdot h}\sum_{j\neq i}^{N-1}k_m\left(\frac{x_i - x_j}{h}\right)\right] = 0, \tag{10}$$

which can only be found numerically.

The accuracy of finding the solution of this equation directly depends on the successful choice of the initial approximation, for which there are no general rules. In [2] it is proposed to search the maximum of a functional $\varphi(h_m)$ on the basis of an analysis of the values of the function calculated on the interval $\left[h_{m\min}, h_{m\max}\right]$. It turns out [2, p. 45] that the function $\varphi(h_m)$ for small values of h_m in a number of cases turns out to be non-smooth, which makes it difficult to find its maximum. At the same time, there is no explanation for the detected peculiarity of the function.

This paper discusses the reasons for the nonmonotonicity of the function $\varphi(h_m)$ and the choice of the search interval for the extremum of the function.

2. Analysis of the reasons for the nonmonotonicity of the function $\varphi(h_m)$ for kernels with an unlimited range of the function argument

Consider the results of calculating in the MATLAB the values $\varphi(h_m)$ of a random sequence $x_i, i = \overline{1,100}$ generated in accordance with the normal distribution law $N(1,4)$ for kernel No. 1 (figure 2).

It is clear from figure 2 that the function at points $h \in \left[\ 1.1643\cdot10^{-3}; 1.1645\cdot10^{-3}\right]$, $h \in \left[\ 2.6196\cdot10^{-3}; 2.6230\cdot10^{-3}\right]$ has discontinuities of the first kind and of the second kind. For other values of the smoothing parameter the function is continuous. Analysis of the values of the function for the values of the smoothing parameter located in this ranges has showed that at the points of discontinuities of the second kind, the values of the function are equal to $-\infty$. It is clear from (6) that the necessary condition for the occurrence of the situation is that the argument of the logarithm in (6) is equal to zero:

$$\frac{1}{(N-1)\cdot h}\sum_{j\neq i}^{N-1}k_1\left(\frac{x_i - x_j}{h}\right) = 0. \tag{11}$$

This result is possible when the condition $k_1\left(\frac{x_i - x_j}{h}\right) \equiv 0$ is true simultaneously for all possible combinations $i = \overline{1,100},\ j = \overline{1,100}, i \neq j$. In numerical calculations it is sufficient to perform it so that all the calculated values of the function $k_1\left(\frac{x_i - x_j}{h}\right)$ are less than the machine zero 10^{-323}. It is possible to avoid the fulfillment of this condition and to eliminate discontinuities of the second kind, using the following formula for calculating the values of the function:

$$\varphi(h) = -\ln(N-1) - \ln h + \sum_{i=1}^{N} \ln\left[\sum_{j\neq i}^{N-1} k_m\left(\frac{x_i - x_j}{h}\right)\right], \tag{12}$$

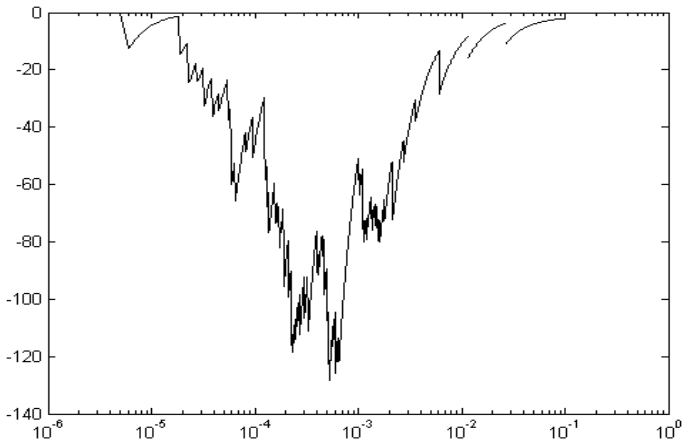

Figure 2. Function $\varphi(h_m)$, kernel No. 1, random $x_i, i = \overline{1,100}$ generated in accordance with the normal distribution law $N(1,4)$.

This formula is equivalent to (6). The calculation results of the function $\varphi(h_m)$ in accordance with (8) are shown in figure 3, from which it can be seen that with this method of calculation it is possible to eliminate discontinuities of the second kind.

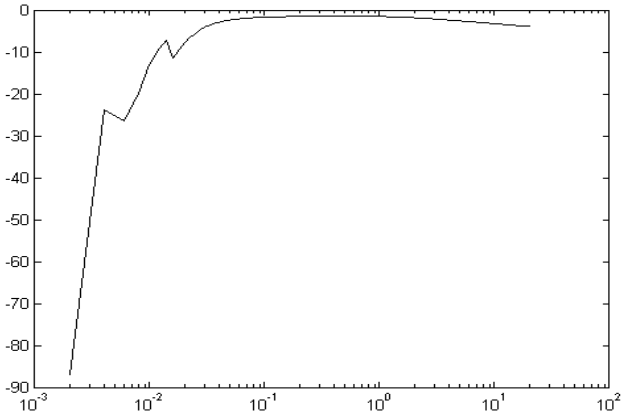

Figure 3. Function $\varphi(h_m)$, calculated in accordance with (8) kernel No. 1, random $x_i, i = \overline{1,100}$ generated in accordance with the normal distribution law $N(1,4)$.

Moreover it is possible to obtain an estimate of the minimum possible value of the smoothing parameter for kernel No. 1.

$$\exp\left(-\frac{(x_i - x_j)^2}{2h^2}\right) = 10^{-(x_i - x_j)^2/2h^2} > 10^{-323} \tag{13}$$

Where do we find

$$h_{1\min} \geq \left(\frac{\min\left((x_i - x_j)^2/2\right)}{323}\right)^{1/2} \tag{14}$$

Repeating similar arguments we find for kernels No. 2, 4, 5, respectively

$$h_{2\min} \geq \frac{\min\left(\left|x_i - x_j\right|\right)}{323 - \lg 2}, \tag{15}$$

$$h_{4\min} \geq \frac{\min\left(x_i - x_j\right)^2}{10^{323}/\pi - 1} \tag{16}$$

The minimum possible value of a variable $h_{5\min}$ can be found numerically as a solution to the inequality

$$\lg\left(1 + e^{-\min\left(\left|x_i - x_j\right|\right)/h}\right) - \frac{\min\left(\left|x_i - x_j\right|\right)}{h} \leq 323. \tag{17}$$

Note the algorithm for searching the maximum value of $\varphi(h_m)$ can be changed. To do this, note that for $h_1 \geq \max\left(\left(x_i - x_j\right)^2/2\right)^{1/2}$ for a normal kernel, for $h_{2,4,5} \geq \max\left(\left|x_i - x_j\right|\right)$ for the Laplace, Cauchy, and logistic kernel, the values of the arguments of the kernel functions belong to the interval $[-1,1]$ and the values of the kernel functions belong to the intervals $\left[\frac{1}{\sqrt{2\pi}}e^{-1}, \frac{1}{\sqrt{2\pi}}\right]$, $\left[\frac{1}{2}e^{-1}, \frac{1}{2}\right]$, $\left[\frac{1}{2\pi}, \frac{1}{\pi}\right]$, $\left[\frac{e^{-1}}{\left(1+e^{-1}\right)^2}, \frac{1}{4}\right]$ respectively. Thus, for the indicated values of the smoothing parameter, there are no computational problems associated with the precision of numbers in the computer. Therefore, it is possible with some finite step to start calculations for a particular kernel from the corresponding value $h_{1,2,4,5}$ and then move towards smaller values of the variable h until the maximum value of the function $\varphi(h_m)$. The viability of using this algorithm in practice is confirmed by the plot presented in figure 4

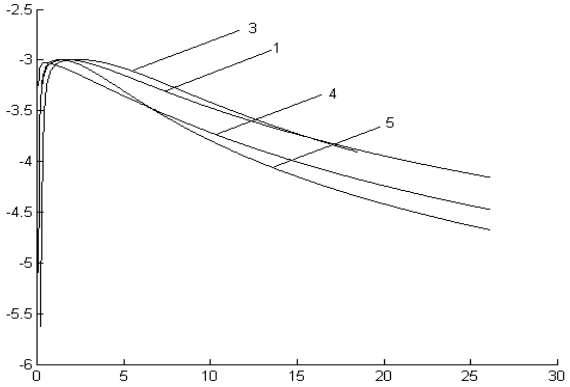

Figure 4. Function $\varphi(h_m)$, kernels No. 1,2,4,5, calculated with the proposed algorithm.

3. Properties of the function $\varphi(h_m)$ for kernels with a limited range of argument

Kernels No. 3, 6-9 have bounded scattering regions: $\left|\frac{t}{2}\right| \leq \pi$, $|t| \leq \sqrt{5}$, $|t| \leq 1$, for that reason

$$h_1 \geq \frac{\max\left(x_i - x_j\right)}{2\pi},\ h_2 \geq \frac{\max\left(x_i - x_j\right)}{\sqrt{5}},\ h_{6,7,8,9} \geq \max\left(x_i - x_j\right). \tag{18}$$

The results of calculating the values of the function $\varphi(h_m)$ for a random sequence $x_i, i = \overline{1,100}$ generated in accordance with the normal distribution law, for kernels No. 3, 6-9 are shown in figure 5

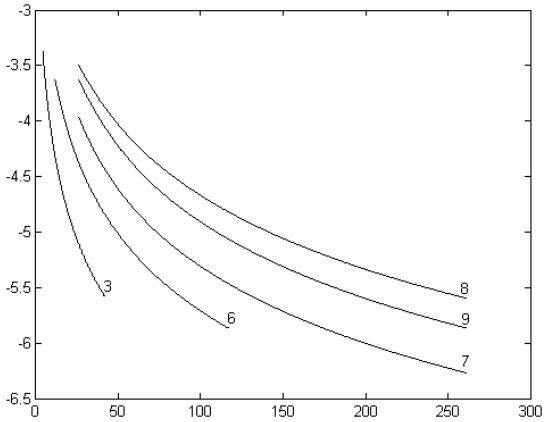

Figure 5. Function $\varphi(h_m)$, kernels No. 3,6-9, calculated with the proposed algorithm.

It is clear From figure 5 that for each of the used kernels the function $\varphi(h_m)$ turns out to be monotonically decreasing, therefore, to estimate the optimal value of the smoothing parameter it is sufficient to calculate the value of the function at points calculated in accordance with (9).

4. Conclusions

The results of the study allowed:

1. Explain the cause, which leads to a nonmonotonicity of the function $\varphi(h_m)$ due to the finite precision of computer arithmetic.
2. For kernel functions with an unbounded range of values justify the choice of the initial value and propose an algorithm for finding the maximum value of the function $\varphi(h_m)$.
3. For kernel functions with bounded range of values justify the choice of the optimal value of the parameter h, which ensures the condition (6).

References

[1] Kramer G 1975 *Mathematical Methods In Statistics* (Moscow: Mir) 648
[2] Syzrancev V N 2008 *Strength Reliability Estimation Based On Nonparametric Statistics Methods* (Novosibirsk: Nauka) 218
[3] Porshnev S V, Ovechkina E V and Kaplan V E 2006 *Theory And Algorithms For Approximation Of Empirical Dependences And Distributions* (Ekaterinburg: UrO RAN) 166
[4] Simahin V A 2011 *Robust Nonparametric Estimation: Adaptive Estimation Of Weighted Maximal Likelihood In Contexts Of Statistical Prior Indetermination* (Saarbrucken, Germany: LAP LAMBERT Academic Publishing GmbH & Co. KG) 292
[5] Rozenblatt M 1956 Remarks On Some Nonparametric Estimates Of Density Function *Ann. Math. Statist* **27** 832–835
[6] Parzen E 1962 On Estimation Of Probability Density Function And Mode *Ann. Math. Statist* **33** 1065-1076