KYBERNETIKA — VOLUME 39 (2003), NUMBER 3, PAGES 341 - 346

NONPARAMETRIC ESTIMATIONS OF NON–NEGATIVE RANDOM VARIABLES DISTRIBUTIONS[1]

FRANTIŠEK VÁVRA, PAVEL NOVÝ, HANA MAŠKOVÁ, MICHALA KOTLÍKOVÁ AND DAVID ZMRHAL

The problem of estimation of distribution functions or fractiles of non-negative random variables often occurs in the tasks of risk evaluation. There are many parametric models, however sometimes we need to know also some information about the shape and the type of the distribution. Unfortunately, classical approaches based on kernel approximations with a symmetric kernel do not give any guarantee of non-negativity for the low number of observations. In this note a heuristic approach, based on the assumption that non-negative distributions can be also approximated by means of kernels which are defined only on the positive real numbers, is discussed.

Keywords: distribution function, kernel approximation, non-negative random variable
AMS Subject Classification: 62G07

1. INTRODUCTION

The problem of estimation of distribution functions or fractiles of non-negative random variables often occurs in the tasks of risk evaluation. For example an estimation of period between events, duration of power equipment outage, claims and others. There are many parametric models, however sometimes we need to know also some information about the shape and the type of the distribution. This information can be used sometimes as a starting point, another time as a final result. At present classical processes based on kernel approximations with a symmetric kernel do not give any guarantee of non-negativity for the low number of observations. It means that an estimation can have a part of its definition domain even on the negative part of R_1 (where R_1 denotes one-dimensional space of real numbers). Therefore we bring forward to discussion one possible approach based on heuristic that non-negative distributions can be also approximated by means of kernels which are defined only on the positive part of R_1. In this note we give rather an impulse to discussion than a collection of our knowledge.

[1]Presented at the Workshop "Perspectives in Modern Statistical Inference II" held in Brno on August 14–17, 2002.

2. ASSUMPTIONS

Let $K(x)$ be some function with the following features:

1. $K(x) = 0 \quad \forall\, x \leq 0$,
2. $K(x) \geq 0 \quad \forall x > 0$,
3. $K(x)$ is increasing and differentiable for $\forall\, x \geq 0$,
4. $\lim_{x\to\infty} K(x) = 1$,
5. Let $\int_{x=0}^{\infty}(1 - K(x))\,\mathrm{d}x = m$ exists,
6. Let $2\int_{x=0}^{\infty} x(1 - K(x))\,\mathrm{d}x = m_2$ exists and let $\kappa(x)$ be its derivation.

Let us have n independent observations $x_1, \ldots, x_n$ of a non-negative random variable X with the distribution function $F(x)$ and the density $f(x)$. Further let $a > 0$ be a real number (we permit dependence upon n and upon the observations $x_1, \ldots, x_n$). Then:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{n(x - x_i) + am}{a}\right) \quad \text{is the distribution function} \tag{1}$$

and

$$f_n(x) = \frac{1}{a}\sum_{i=1}^{n} \kappa\left(\frac{n(x - x_i) + am}{a}\right) \quad \text{is the corresponding density.} \tag{2}$$

Our work aims at some connections between these estimations and the original distribution $F(x)$ and the density $f(x)$.

As an example of kernels mentioned above, the following one can be used:

$$K(x) = (1 - e^{-x})^r \text{ for } x > 0 \qquad \text{and} \qquad \kappa(x) = r(1 - e^{-x})^{r-1}e^{-x} \text{ for } x > 0$$
$$= 0 \text{ otherwise} \qquad\qquad\qquad\qquad = 0 \text{ otherwise}$$

where $r \geq 1$ and

$$m = \sum_{i=0}^{r-1} \frac{1}{i+1}; \quad m_2 = -\sum_{i=0}^{r-1} I_i \quad \text{where} \quad I_i = \frac{i}{i+1} I_{i-1} - \frac{i}{(i+1)^2}; \; I_0 = -1. \tag{3}$$

3. FEATURES

Let us denote:

$$E\{F_n(x)\} = \frac{1}{n}\sum_{i=1}^{n} \int_0^{\infty} K\left(\frac{n(x - x_i)}{a} + m\right) f(x_i)\,\mathrm{d}x_i,$$

where x_i is the ith observation of a random variable with non-negative definition domain and with the density $f(x)$. Of course, we suppose that individual observations are independent and identically distributed. Then we get:

$$E\{F_n(x)\} = K\left(m + \frac{n}{a}\right) \cdot F\left(x + \frac{a}{n}(m - \xi)\right), \tag{4}$$

where $F(x)$ is the distribution function of the observed values and $\xi \in \langle 0, m + \frac{n}{a}x\rangle$.

If the limit $\lim_{n\to\infty} \frac{n}{a} = \infty$ is satisfied for any observed data, the estimation of the distribution function (1) will be asymptotically unbiased, i.e. $E\{F_n(x)\} \to F(x)$. Moreover the following trivial inequality:

$$0 \leq E\{F_n(x)\} \leq K\left(m + \frac{n}{a}\right) \cdot F\left(x + m\frac{a}{n}\right)$$

is satisfied from one side. Further we denote:

$$E_n\{x\} = \int_0^\infty (1 - F_n(x))\,\mathrm{d}x$$

the mean value of a random variable x, which holds the distribution (1). After short computation we get: $E_n\{x\} = \frac{1}{n}\sum_{i=1}^n x_i$. Therefore the estimation of the distribution function (1) has the same mean value as the sampling average of observations.

For $E_n\{x^2\} = 2\int_0^\infty x(1 - F_n(x))\,\mathrm{d}x$ we can infer:

$$E_n\{x^2\} = \frac{1}{n}\sum_{i=1}^n x_i^2 + 2m_2\left(\frac{a}{n}\right)^2$$

using quite simple rearrangement and for the variance computed for the estimation of the distribution function (1) (and thereby, in our case, also for the estimation of the density):

$$\sigma_n^2\{x\} = \frac{1}{n}\sum_{i=1}^n \left(x_i - E_n\{x\}\right)^2 + 2m_2\left(\frac{a}{n}\right)^2. \tag{5}$$

If we require the sample variance $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - E_n\{x\})^2$ to be equal to the variance (5), we will get the formula for determining the parameter a:

$$\frac{1}{n}\sum_{i=1}^n \left(x_i - E_n\{x\}\right)^2 + 2m_2\left(\frac{a}{n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^n \left(x_i - E_n\{x\}\right)^2 = s_n^2 \quad \text{and} \quad s_n = \sqrt{s_n^2}.$$

We get:

$$a = \sqrt{\frac{1}{2m_2}}\, s_n \sqrt{n}. \tag{6}$$

Comparing with the classical parameter of smoothing h used for non-parametric estimations [2] of densities (2), we get:

$$h = \frac{a}{n} = \sqrt{\frac{1}{2m_2}}\sqrt{\frac{s_n^2}{n}}.$$

We can look at asymptotically unbiased estimation also in another way. We can easily see that the following holds true:

$$\int_0^\infty \Big(F_n(x) - F(x)\Big)\,\mathrm{d}x = \int_0^\infty \Big((F_n(x) - 1) + (1 - F(x))\Big)\,\mathrm{d}x = E\{x\} - E_n\{x\}.$$

Of course, if $E\{x\}$ exists. Also, if both mean value and variance exist for an observed variable, it holds: $\mathrm{Prob}\,(\lim_{n\to\infty} E_n\{x\} = E\{x\}) = 1$ (see the strong law of large numbers [5]). With the assumptions mentioned above it holds:

$$\mathrm{Prob}\Big(\lim_{n\to\infty} \int_0^\infty (F_n(x) - F(x))\,\mathrm{d}x = 0\Big) = 1. \tag{7}$$

Using analogous method, it is possible to prove the validity of the formula:

$$\mathrm{Prob}\Big(\lim_{n\to\infty} \int_0^\infty x(F_n(x) - F(x))\,\mathrm{d}x = 0\Big) = 1. \tag{8}$$

Validity of the formula (7) is independent of the choice of the parameter a, validity of the formula (8) is contingent on the selection of (6). Because it holds:

$$\frac{n}{a} = \sqrt{\frac{2m_2 n}{s_n^2}}$$

for the selection of (6), the condition of asymptotically unbiased estimation (1) is satisfied. However it is the statement, which is inaccurate: s_n^2 is a random variable, which, in a sense of probability, tends to $\sigma^2\{x\}$. Thus, again and more at large, the probability that the estimation of the distribution function tends to be unbiased will be 1.

4. EXPERIMENTS AND SIMULATION

We have verified the features resulting from the previous theory using the kernels (3) by data simulation with the following blending distribution $0.8R(5;20) + 0.2R(45;50)$. This distribution is one of possible analogy to distributions of outage duration caused by power equipment failures (high probability of short-time outages caused by minor failures and low probability of long-time outages caused by major failures). The proposed model considerably simplifies the real situation but it preserves the fact that outage times are separated by fairly "large interval". Bounds of such interval are unfortunately difficult to determine. In the following Figures 1 and 2 there is a simulation for 30 observed values, where the parameter of the kernel function $r = 2$.

Figure 3 represents the influence of the kernel choice. All parameters are the same as in the previous simulation apart from the parameter of the kernel function $r = 20$.

The smoothing and accuracy process of the estimation improves with the increasing number of observed values. This can be seen in Figure 4.

(Received November 26, 2002.)

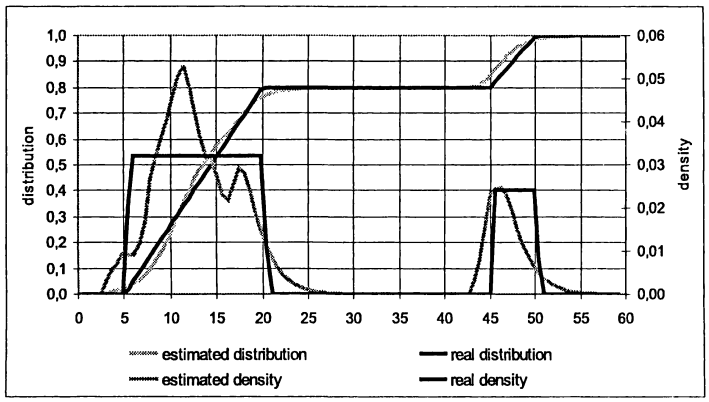

Fig. 1. Simulation for 30 observed values, $r = 2$, "relatively successful".

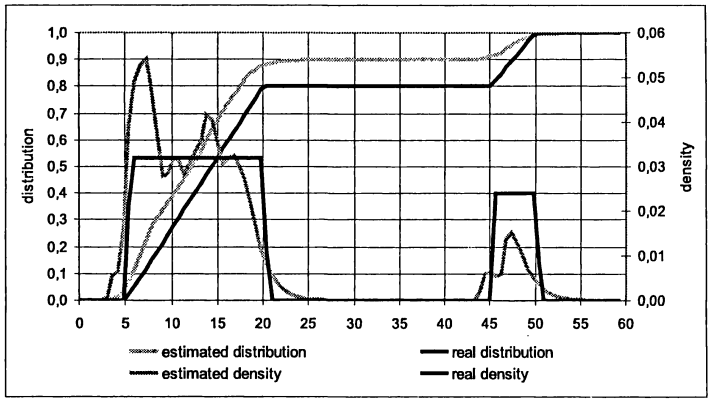

Fig. 2. Simulation for 30 observed values, $r = 2$, "relatively unsuccessful".

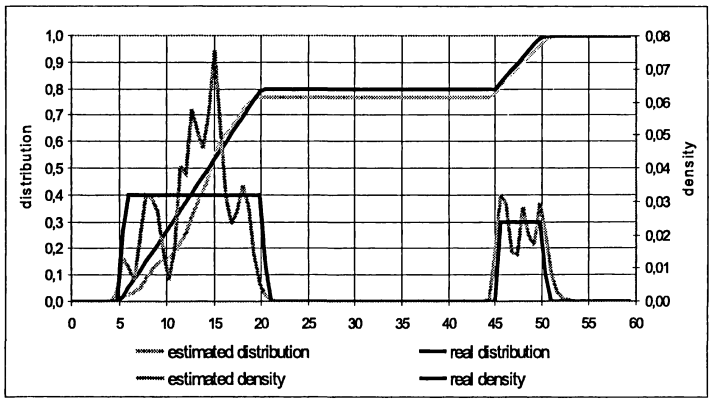

Fig. 3. Simulation for 30 observed values, $r = 20$.

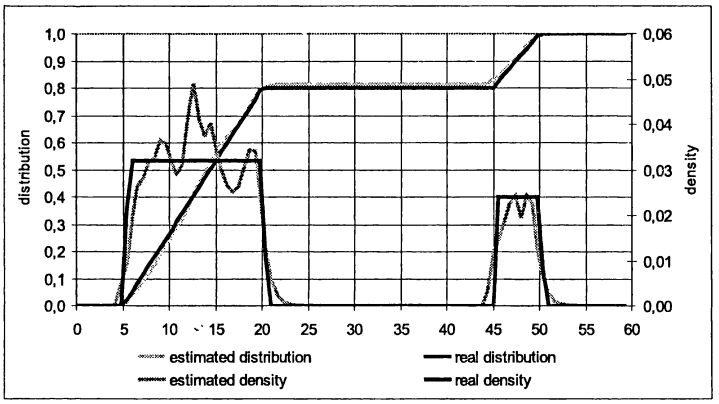

Fig. 4. Simulation for 200 observed values, $r = 2$.

REFERENCES

[1] H. German: Learning about risk: some lessons from insurance. European Finance Review *2* (1999), 113–124.

[2] L. Devroye and L. Györfi: Nonparametric Density Estimation the L_1-view. Wiley, New York 1985.

[3] H. Albrecher: Dependent Risks and Ruin Probabilities in Insurance. Interim Report, IIASA, IR 98 072.

[4] C. R. Rao: Linear Statistical Inference and its Applications (Czech translation). Academia, Praha 1978.

[5] A. Rényi: Theory of Probability (Czech translation). Academia, Praha 1972.

Doc. Ing. František Vávra, CSc., Ing. Pavel Nový, Ph.D., Ing. Hana Mašková, Ing. Michala Kotlíková and Ing. David Zmrhal, Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň. Czech Republic.

e-mails: vavra,maskova,kotlikova@kiv.zcu.cz