THEORY AND METHODS OF SIGNAL PROCESSING

UDC 519.22/.25 (045)

A. A. Zelenkov,
A. V. Molchanov

A NON-PARAMETRICAL ESTIMATION OF THE DISTRIBUTION DENSITY FUNCTION

Institute of Information and Diagnostic Systems, National Aviation University, Kyiv, Ukraine
E-mail: molchanovlesha@ukr.net

Abstract—*Nonparametric recurrent algorithm of probability density function estimation for problem solving of complex dynamic systems reliability is considered.*

Index Terms—Probability density function; nonparametric estimation; kernel coefficient; recurrent algorithm; coordinates array; truncated function; normalizing multiplier.

I. INTRODUCTION

Statistical data defined during state control of the dynamic system in its operational process in real-life environment are single source of information about system. As a rule, the basic problem in this case is data processing to get the density of distribution of obtained random values.

Two basic ways are possible. The first one uses the parallel algorithm in which data processing is carried out after long stage their accumulation (classical problem).

The second way provides the series algorithm (recurrent) to estimate the distribution law. In this case each element of entering information is used at once to specify the obtained estimation of distribution law. It is important if the results of data processing are used in the predictive control of dynamic systems when delay of entering information is undesirable.

II. PROBLEM STATEMENT

A non-parametrical estimation of the random value distribution density, as a rule, is written in the form:

$$f^*(x) = \frac{1}{n}\sum_{i=1}^{n} k\left[\frac{x_i - x}{c}\right],$$

where $k(u)$ is the weight function (kernel), satisfying to the following conditions:

$$k(u) > 0, \quad k(-u) = k(u),$$

$$\int k(u)du = 1, \quad k(u) \to 0 \text{ for } |u| \to \infty.$$

Introduction of the kernel allows smoothing the experimental data in some environs of each experimental point. Coefficient c is the smoothing parameter and depends on experimental data volume.

The normal kernel was used to obtain the value $f^*(x)$ according to the expression:

$$f^*(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}c}\exp\left[\frac{(x_i - \bar{x})^2}{2c^2}\right].$$

In the problems of operational control and prediction of the on-board control system reliability it is necessary to use each element of the entering information at once [1]. That's why it is necessary to use the series algorithms. In this case the Parzen's algorithm may be used:

$$f^*(x) = \frac{1}{n}\sum_{i=1}^{n} k\left[\frac{x_i - \bar{x}}{h}\right],$$

where h is kernel coefficient. It defines the interval in which the kernel is not equal to zero. Optimal value with respect to minimum of approximation error is defined as $h = n^{-0,2}$ [2], [3].

III. PROBLEM SOLUTION

The distribution density is represented by step curve, restricting the row of equaled area rectan-gulars. The number of dividing intervals Q is defined by the given approximation accuracy and the interval coordinates are calculated by the recurrent way:

$$\int_{S_{q-1}}^{S_q} f(x)dx = F(S_q) - F(S_{q-1}) = \frac{1}{Q}, \quad q = \overline{1,Q}.$$

In the memory of data processing system it is necessary to store the coordinates of the intervals $\{S_q\}$ and their quantity. If we obtain the next value x_{n+1} it is necessary recalculate a new array of coordinates $\{S_q^{(n+1)}\}$ corresponding to the estimation $f_{n+1}^*(x)$.

The algorithm (Fig. 1) begins the processing from the source data input (block 1):

$$\left[n,\ Q,\ h_n,\ \{S_q^{(n)}\},\ x_{n+1}\right].$$

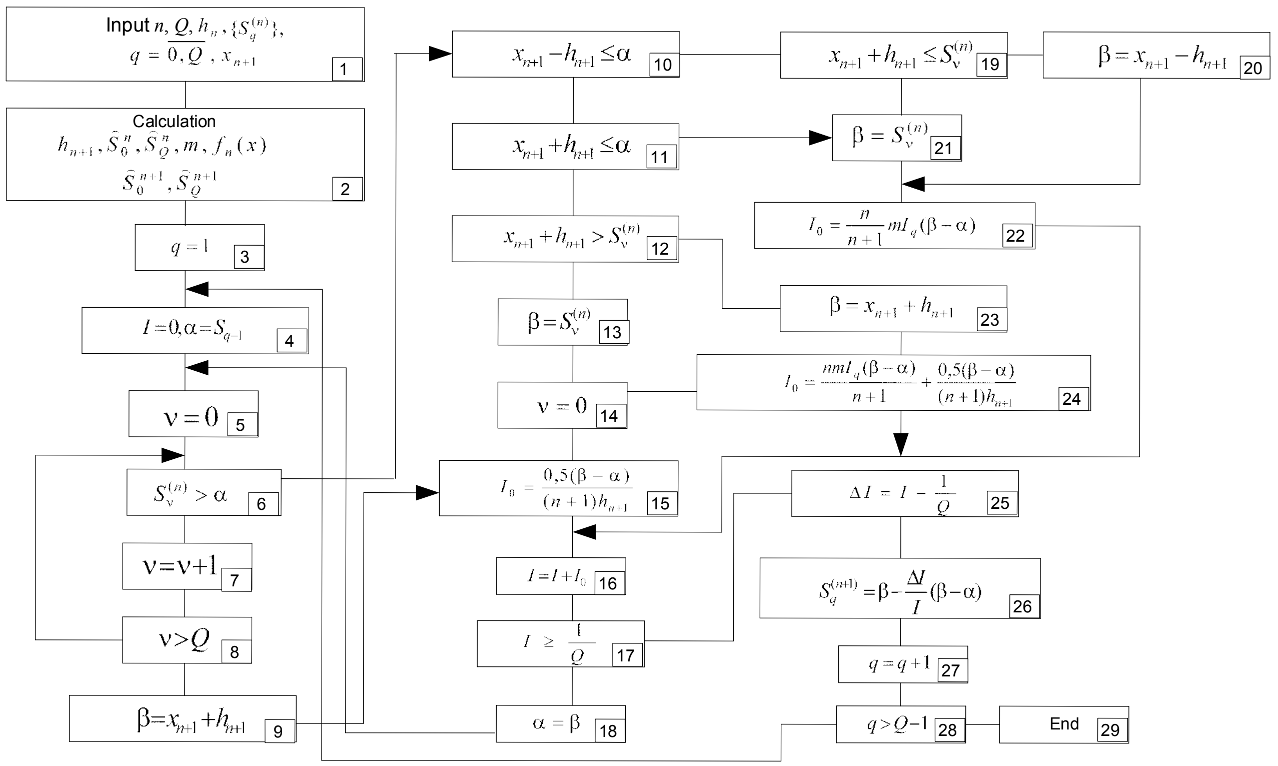

Fig. 1. Blok diagram of the considered algorithm

Then the new coefficient h_{n+1} is calculated as the greatest value from (block 2):

$$\max\begin{cases} h_{n+1}=h_n n^{0,2}(n+1)^{-0,2}, \\ h_{n+1}=0,5[S_0^{(n)}-h_n-x_{n+1}], \quad x_{n+1}<S_0^{(n)}, \\ h_{n+1}=0,5[x_{n+1}-S_0^{(n)}-h_n], \quad x_{n+1}>S_0^{(n)}. \end{cases}$$

The estimation $f_{n+1}^*(x)$ is defined as:

$$f_{n+1}^*(x)=(n+1)^{-1}\sum_{i=1}^{n}k_i^{(n+1)}(x)+(n+1)^{-1}k_{n+1}^{(n+1)}(x),$$

where

$$k_i^{(n+1)}=\frac{1}{h_{n+1}}\begin{cases}0,5 & |x-x_i|\le h_{n+1}, \\ 0 & |x-x_i|>h_{n+1}.\end{cases}$$

This expression may be written in the form:

$$f_{n+1}^*(x)=n(n+1)^{-1}\hat{f}_n(x)+(n+1)^{-1}k_{n+1}^{(n+1)}(x).$$

The limits of area $\{\hat{S}_0^{(n)},\hat{S}_Q^{(n)}\}$ for $\hat{f}_n(x)$ are found as

$$\hat{S}_0^{(n)}=S_0^{(n)}+(h_n-h_{n-1}),$$

$$\hat{S}_Q^{(n)}=S_Q^{(n)}-(h_n-h_{n-1}),$$

and written into the respective memory cell $\{S\}$.

The truncated area of $\hat{f}_n(x)$ must satisfy to the normalization condition

$$\int_{\hat{S}_0^{(n)}}^{\hat{S}_Q^{(n)}}\hat{f}_n(x)dx=1.$$

Since all rectangulars have the same area equaled Q^{-1} then the ordinate of $f_n(x)$ will be equaled

$$f_q(x)=I_q=Q^{-1}\left[S_q^{(n)}-S_{q-1}^{(n)}\right]^{-1}.$$

Applying truncation and normalization operations we obtain

$$\hat{f}_n(x)=\hat{I}_q=mI_q, \quad \hat{S}_0^{(n)}\le x\le\hat{S}_Q^{(n)},$$

where m is normalizing multiplier, taking into account the error of the step function, Fig. 2, *a*. Coefficient m is defined by the following way.

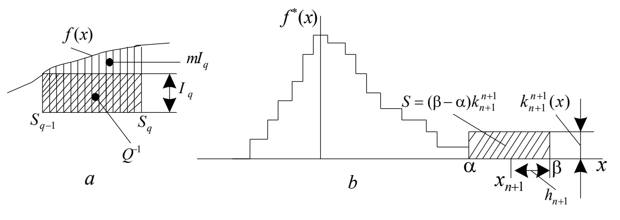

Fig. 2. Case $f(x)=0, \quad k_{n+1}^{n+1}\ne 0$

The area under approximated curve $f_n(x)$ equals 1, and truncated area equals:

$$f_1(x)(h_n - h_{n+1}) + f_Q(x)(h_n - h_{n-1}) = \frac{h_n - h_{n-1}}{(S_1 - S_0)Q} + \frac{h_n - h_{n-1}}{(S_Q - S_{Q-1})Q},$$

so that the area under the truncated function is

$$1 - Q^{-1}[(S_1 - S_0)^{-1} + (S_Q - S_{Q-1})^{-1}](h_n - h_{n-1}).$$

It is evident that the normalizing multiplier m is defined as the ratio of areas (block 2):

$$m = \left[1 - Q^{-1}[(S_1 - S_0)^{-1} + (S_Q - S_{Q-1})^{-1}](h_n - h_{n-1})\right]^{-1}.$$

Next we may define the needed array $\{S_q^{(n+1)}\}$. At first we calculate the initial $S_0^{(n+1)}$ and finite $S_Q^{(n+1)}$ quantities (block 2):

$$S_0^{(n+1)} = \inf\left[x_{n+1} - h_{n+1}, S_0^{(n)}\right];$$

$$S_Q^{(n+1)} = \sup\left[x_{n+1} + h_{n+1}, S_Q^{(n)}\right],$$

where $x_{n+1} - h_{n+1}$ and $x_{n+1} + h_{n+1}$ define the limits in which $k_{n+1} \neq 0$. Then we calculate in consecutive order the remaining components of array according to the expression:

$$\int_{S_{q-1}^{(n)}}^{S_q^{(n)}} f(x)dx = Q^{-1}, \quad q = \overline{1, Q-1},$$

where

$$f_{n+1}(x) = (n+1)^{-1}\sum_{i=1}^{n} k_i^{(n+1)}(x) + (n+1)^{-1}k_{n+1}^{(n+1)}(x) = n(n+1)^{-1}\hat{f}_n(x) + (n+1)^{-1}k_{n+1}^{(n+1)}(x) = n(n+1)^{-1}mI_q + (n+1)^{-1}k_{n+1}^{(n+1)}(x),$$

$$k_{n+1}^{(n+1)} = \frac{1}{h_{n+1}}\begin{cases} 0{,}5 & |x - x_{n+1}| \le h_{n+1}, \\ 0 & |x - x_{n+1}| > h_{n+1}. \end{cases}$$

The first and the second terms take into account the stored and the obtained information respectively. Since the estimation of $f(x)$ is the step function then the integral is defined by multiplication of ordinate by the difference of abscissas.

After calculation of a new value x_{n+1} it is necessary to estimate the location of kernel $k_{n+1}^{n+1}(x)$ with respect to the stored estimation $f_n(x)$ in order to use the basic algorithm of calculation:

$$f_{n+1}(x) = n(n+1)^{-1}mI_q + (n+1)^{-1}k_{n+1}^{(n+1)}(x).$$

For the given number q the low limit of the interval is found by equality $\alpha = S_{q-1}^{(n)}$ (blocks 8–11) and high limit β is assumed as $S_q^{(n)}$ ($S_q^{(n)} > \alpha$). It is possible two cases. In the first case no value $S_\nu^{(n)}$ from array $\{S_q^{(n)}\}$, for which $S_\nu^{(n)} > \alpha$ (block 9). It means that $\hat{f}_n(x) = 0$ (Fig. 2, *b*) and only the kernel k_{n+1}^{n+1} is not equal to zero. Then $\beta = x_{n+1} + h_{n+1}$ (block 12) and $k_{n+1}^{n+1} = 0{,}5h_{n+1}^{-1}$.

It is evident that the area limited by the function $k_{n+1}^{n+1}(x)$ equals (taking into account the weight $(n+1)^{-1}$):

$$I_0 = 0{,}5(\beta - \alpha)(n+1)^{-1}h_{n+1}^{-1},$$

realized by block 24.

In the second case (the basic case) the array $\{S_q^{(n)}\}$ contains the value $S_\nu^{(n)}$, which is greater than $\alpha (S_\nu^{(n)} > \alpha)$ and it is necessary to estimate the location of kernel with respect to the interval $[\alpha, S_\nu^{(n)}]$, Fig. 3. Then a low limit of the function existence area is compared with the given value $\alpha : x_{n+1} - h_{n+1} \le \alpha$ (block 10). If this condition does not fulfilled, then the function definition area k_{n+1}^{n+1} is located to the right of α and its beginning may be located on the interval $[\alpha, S_\nu^{(n)}]$. This possibility $(x_{n+1} - h_{n+1} < S_\nu^{(n)})$ is checked by the respective block (19).

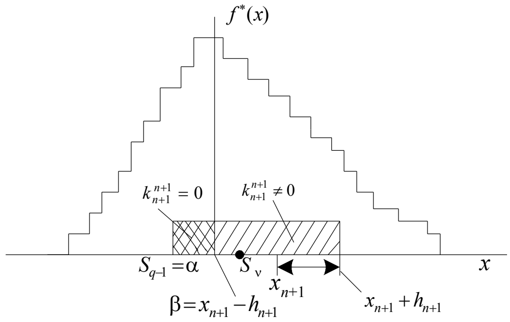

Fig. 3. Basic case $S_\nu^{(n)} > \alpha$

If the decision is positive, then $\beta = x_{n+1} - h_{n+1}$. On the contrary we have $\beta = S_\nu^{(n)}$ (blocks 20, 21). In both cases it means that $k_{n+1}^{n+1}(x) \equiv 0$ and the weighted area under the curve is defined (block 22) as

$$I_0 = n(n+1)^{-1}mA_q(\beta - \alpha). \qquad (*)$$

If the condition $x_{n+1} - h_{n+1} \le \alpha$ is fulfilled it means that the beginning of function existence area is located to the left of α and we have to define the location of its end. If $x_{n+1} + h_{n+1} \le \alpha$ is fulfilled (block 11) then the function $k_{n+1}^{n+1}(x)$ is taken into account on the previous steps and, consequently, $k_{n+1}^{n+1}(x) \equiv 0$, $\beta = S_\nu^{(n)}$ (block 16) and as before we have the same result ($*$).

If the last condition does not fulfilled then finite point $x_{n+1} + h_{n+1}$ of the function definition area k_{n+1}^{n+1} (x) may belong to interval $[\alpha, S_\nu^{(n)}]$. The block 12 checks it. Depending on results of its functioning it is assumed: $\beta = x_{n+1} + h_{n+1}$ (block 23) or $\beta = S_\nu^{(n)}$ (block 13). The variant $\beta = x_{n+1} + h_{n+1}$ corresponds to the case when both terms in the formula ($*$) are not equal to zero. The block 24 shows it. Two cases are possible in variant $\beta = S_\nu^{(n)}$. For the condition $\nu = 0$ (block 14) we have $f(x) \equiv 0$ and the control is passed to the block 15, Fig.4, *a*. Otherwise the block 24 operates.

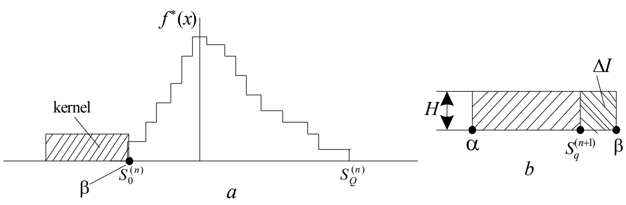

Fig. 4. *a* is the case $\nu = 0$ and $f(x) \equiv 0$;
b is the calculation of $S_q^{(n+1)}$

Thus, the operation of each algorithm branch is finished by calculation of the area increment I_0 obtained on the considered interval (blocks 22, 24, 15). The block 16 summarizes increments obtained on each step q_i, after that the content of cell I (the stored value I_0) is compared with the value Q^{-1} (block 17) according to the equation

$$\int_{S_{q-1}^{(n)}}^{S_q^{(n)}} f(x)dx = Q^{-1}, \quad q = \overline{1, Q-1}.$$

It is evident that Q is the number of intervals of the estimation $f(x)$ and Q^{-1} is the area restricted by this curve between any unknown values $\left(S_{q-1}^{(n+1)}, S_q^{(n+1)}\right)$. If the content of cell I does not exceed the value Q^{-1}, then the value of interval high limit is assumed as a new low limit, that is $\alpha = \beta$ (block 27) and the calculations repeat. If the content I exceeds the value Q^{-1} the next value $S_q^{(n+1)}$ is calculated. In this case the increment ΔI is defined as $\Delta I = I - Q^{-1}$ (block 25), that is the value ΔI is exceeding of the given area, Fig. 4, *b*. Then $\Delta I = H\left(\beta - S_q^{(n+1)}\right)$ and $S_q^{(n+1)} = \beta - \Delta I H^{-1}$. Taking into account that $I = H(\beta - \alpha)$ and $H = I(\beta - \alpha)^{-1}$ we may obtain (block 26):

$$S_q^{(n+1)} = \beta - \Delta I(\beta - \alpha) I^{-1}.$$

The obtained value $S_q^{(n+1)}$ is written into q^{th} cell of the array $\{S_{n+1}(Q)\}$. After that the content of the counter increases per unit (q + 1) (block 27). If the value (q + 1) does not exceed the value (Q – 1) then the algorithm passes on to definition of the next coordinate $S_{q+1}^{(n+1)}$. On reaching the value $q = Q - 1$ (block 28) the array $\{S_{n+1}(Q)\}$ is completely organized and the work of the algorithm is finished. The next value x_{n+2} is waited.

IV. Conclusion

In the modern conditions the data processing system to control by reliability must provide immediate use of each element of the incoming information to obtain and get more specific needed statistical estimations. These estimations may be used to solve problems of complex dynamic systems reliability.

References

[1] A. A. Zelenkov and V. M. Sineglazov, *On-board automatic control systems. Estimation of accuracy by flight test results*. Kyiv: NAU, 2009. 264 p. (in Russian).

[2] E. Parzen, "On estimation of a probability density function and mode". *Ann. Math. Stat.*, vol. 33, no. 3, 1962. pp. 1065–1076.

[3] V. A. Epanechnikov, "Nonparametric estimation of the multidimensional probability density function". *Probability theory and its application*. vol. XIV, issue 1, 1969. pp. 156–161 (in Russian).

Received August 11, 2015

Zelenkov Alexander. Candidate of Engineering. Professor.
Computerized Electrical Systems and Technologies Department, National Aviation University, Kyiv, Ukraine.
Education: Kyiv Civil Aviation Engineers Institute, Kyiv, Ukraine (1968).
Research area: Estimation of the accuracy and reliability of on-board automatic control systems.
Publication: 230.

Molchanov Alexey. Assistant.
Computerized Electrical Systems and Technologies Department, National Aviation University, Kyiv, Ukraine.
Education: National Aviation University, Kyiv, Ukraine (2009).
Research area: Estimation of the accuracy and reliability of on-board automatic control systems.
Publication: 5.
E-mail: molchanovlesha@ukr.net

О. А. Зеленков, О. В. Молчанов. Непараметрична оцінка функції густини розподілу

Розглянуто непараметричний рекурентний алгоритм оцінки функції густини розподілу визначальних параметрів для вирішення проблем точності і надійності складних динамічних систем.

Ключові слова: функція густини розподілу; непараметрична оцінка; коефіцієнт ядра; рекурентний алгоритм; масив координат; усічена функція; нормалізуючий множник.

Зеленков Олександр Аврамович. Кандидат технічних наук. Професор.
Кафедра комп'ютеризованих електротехнічних систем та технологій, Національний авіаційний університет, Київ, Україна.
Освіта: Київський інститут інженерів цивільної авіації, Київ, Україна (1968).
Напрям наукової діяльності: Оцінка точності і надійності бортових автоматичних систем управління.
Кількість публікацій: 230.

Молчанов Олексій Володимирович. Асистент.
Кафедра комп'ютеризованих електротехнічних систем та технологій, Національний авіаційний університет, Київ, Україна.
Освіта: Національний авіаційний університет, Київ, Україна (2009).
Напрям наукової діяльності: Оцінка точності і надійності бортових автоматичних систем управління.
Кількість публікацій: 5.
E-mail: molchanovlesha@ukr.net

А. А. Зеленков, А. В. Молчанов. Непараметрическая оценка функции плотности распределения

Рассмотрен непараметрический рекуррентный алгоритм оценки функции плотности распределения определяющих параметров для решения проблем точности и надежности сложных динамических систем.

Ключевые слова: функция плотности распределения; непараметрическая оценка; коэффициент ядра; рекуррентный алгоритм; массив координат; усеченная функция; нормализующий множитель.

Зеленков Александр Аврамович. Кандидат технических наук. Профессор.
Кафедра компьютеризированных электротехнических систем и технологий, Национальный авиационный университет, Киев, Украина.
Образование: Киевский институт инженеров гражданской авиации, Киев, Украина (1968).
Направление научной деятельности: Оценка точности и надежности бортовых автоматических систем управления.
Количество публикаций: 230.

Молчанов Алексей Владимирович. Ассистент.
Кафедра компьютеризированных электротехнических систем и технологий, Национальный авиационный университет, Киев, Украина.
Образование: Национальный авиационный университет, Киев, Украина (2009).
Направление научной деятельности: Оценка точности и надежности бортовых автоматических систем управления.
Количество публикаций: 5.
E-mail: molchanovlesha@ukr.net