A. V. Lapko
Institute of Computational Modelling, Russian Academy of Sciences, Siberian Branch, Russia, Krasnoyarsk

V. A. Lapko
Siberian State Aerospace University named after academician M. F. Reshetnev, Russia, Krasnoyarsk

THE ANALYSIS OF NONPARAMETRIC MIXTURE PROPERTIES WITH A PROBABLITY DENSITY OF A MULTIDIMENSIONAL RANDOM VARIABLE

The asymptotic properties of a mixture with nonparametric estimations of probability density with a multidimensional random variable are researched in this article. They are compared with the properties of the traditional Rosenblatt–Parzen type nonparametric probability density estimation, depending on the quantity of the composed mixture and dimension of the random variable.

Keywords: mixture of probability densities, nonparametric estimation, large samples, asymptotic properties.

The application of nonparametric statistics methods based on the estimations of Rosenblatt–Parzen type probability density [1; 2] is a rapidly developing modelling method of priori uncertainty systems. However, when the research conditions of the system are complicated, there appear methodical and computing difficulties in traditional nonparametric algorithms and models; this can be clearly observed during the processing of statistical data in great amounts.

The perspective "detour" direction of the arisen problems consists in the application of decomposition principles of training samples according to their size, and the application of the parallel calculation technology.

The purpose of this work is to prove the effective usage of decomposition principles when processing large-scale arrays of statistical data, on the basis of the asymptotic properties' analysis for a nonparametric estimation of probability density mixture.

Let sample $V = \left(x^i, i = \overline{1, n}\right)$ from n independent observations of $k-$ dimensional random variable $x = \left(x_v, v = \overline{1, k}\right)$ be with a probability density $p(x)$. The type $p(x)$ is a priori unknown.

Let's divide sample V into T observation groups $V_j = \left(x^i, i \in I_j\right)$, $j = \overline{1, T}$. Multiple observation numbers x in the group with number j shall be identified as I_j. While: $\bigcup_{j=1}^{T} I_j = I = \left(i = \overline{1, n}\right)$. The quantity $n_j = \left|I_j\right|$ of units in samples $V_j = \left(x^i, i \in I_j\right)$ is equal and equals $\overline{n} = \frac{n}{T}$.

At each sample V_j let us construct a nonparametric estimation of probability density with a multidimensional random variable x [1]:

$$\overline{p}_j(x) = \frac{1}{\overline{n} \prod_{v=1}^{k} c_v} \sum_{i \in I_j} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right), \quad j = \overline{1, T}. \qquad (1)$$

In statistics (1), the nuclear function $\Phi(u_v)$ is satisfied to conditions of normalization, positivity, and symmetry. The parameters of nuclear $c_v = c_v(\overline{n})$ functions decrease with the increase of $\overline{n}$.

Let the intervals of component x_v value change for vector x be identical. In these conditions it is reasonable to assume that the values of coefficients c_v in nonparametric estimations of probability densities $\overline{p}_j(x)$, $j = \overline{1, T}$ are identical and equal to c. Then estimation (1) will look as:

$$\overline{p}_j(x) = \frac{1}{\overline{n}\, c^k} \sum_{i \in I_j} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c}\right), \quad j = \overline{1, T}. \qquad (2)$$

As for magnifying $p(x)$ with statistical sample V we shall use a mixture of nonparametric estimations of a probability density type:

$$\overline{\overline{p}}(x) = \frac{1}{T} \sum_{j=1}^{T} \overline{p}_j(x). \qquad (3)$$

Statistics (3) allows the usage of parallel calculation technology while estimating the probability density in conditions of large samples.

The asymptotic properties $\overline{\overline{p}}(x)$ are defined by the following statement.

The theorem. Let $p(x)$ and its first two derivatives from each component x_v, $v = \overline{1, k}$ be limited and continuous; the nuclear functions satisfy $\Phi(u_v)$ conditions:

$$\Phi(u_v) = \Phi(-u_v), \quad 0 \le \Phi(u_v) < \infty,$$
$$\int \Phi(u_v)\, du_v = 1, \quad \int u_v^2\, \Phi(u_v)\, du_v = 1,$$
$$\int u_v^m\, \Phi(u_v)\, du_v < \infty, \quad 0 \le m < \infty; \; v = \overline{1, k},$$

of sequence $c = c(\overline{n})$ for blur coefficient in nuclear functions are such, that at $\overline{n} \to \infty$ values, $c \to 0$ and $\overline{n}\, c^k \to \infty$.

Then at finite values T the nonparametric estimation (3) of the probability density $p(x)$ has a property of asymptotic unbiasedness and competence.

Hereinafter infinite limits of integration are omitted.

The proof:

1. By definition:

$$M\left(\overline{\overline{p}}(x)\right)=\frac{1}{T}\sum_{j=1}^{T}M\left(\overline{p}_j(x)\right)=\frac{1}{T}\sum_{j=1}^{T}\frac{1}{\overline{n}c^k}\sum_{i\in I_j}\int\ldots$$

$$\ldots\int\prod_{v=1}^{k}\Phi\left(\frac{x_v-x_v^i}{c}\right)p(x_1^i,\ldots,x_k^i)\,dx_1^i\ldots dx_k^i=$$

$$=\frac{1}{c^k}\int\ldots\int\prod_{v=1}^{k}\Phi\left(\frac{x_v-t_v}{c}\right)p(t_1,\ldots,t_k)\,dt_1\ldots dt_k=$$

$$=\int\ldots\int\prod_{v=1}^{k}\Phi(u_v)\,p(x_1-cu_1,\ldots,x_k-cu_k)\,du_1\ldots du_k\ ,$$

where M – is a mathematical expectations sign. When performing the conversion, it is considered that statistical sample units V_j, $j=\overline{1,T}$ are values of the same random variable t with a density probability of $p(t_1,\ldots,t_k)$.

Let's spread out $p(x_1-cu_1,\ldots,x_k-cu_k)$ in the Taylor row at point $x=x_1,\ldots,x_k$ and being limited by the first two terms of the series, we get:

$$W_1=M\left(\overline{\overline{p}}(x)-p(x)\right)\sim\frac{c^2}{2}\sum_{v=1}^{k}p_v^{(2)}(x),\qquad(4)$$

where $p_v^{(2)}(x)$ – is the second derivative of the probability density $p(x)$ at component x_v.

From here, in condition that $c\to 0$ at $\overline{n}\to\infty$, appears the property of the asymptotic unbiasedness for a mixture of nonparametric probability density estimations (3).

2. For convergence proof of $\overline{\overline{p}}(x)$ in square mean we shall consider the following expression:

$$M\int\ldots\int\left(p(x)-\overline{\overline{p}}(x)\right)^2dx_1\ldots dx_k=$$

$$=M\int\ldots\int\left[\frac{1}{T}\sum_{j=1}^{T}\left(p(x)-\overline{p}_j(x)\right)\right]^2dx_1\ldots dx_k=$$

$$=\frac{1}{T^2}M\left[\sum_{j=1}^{T}\int\ldots\int\left(p(x)-\overline{p}_j(x)\right)^2dx_1\ldots dx_k+\right.$$

$$\left.+\sum_{j=1}^{T}\sum_{\substack{t=1\\t\neq j}}^{T}\int\ldots\int\left(p(x)-\overline{p}_j(x)\right)\left(p(x)-\overline{p}_t(x)\right)dx_1\ldots dx_k\right].\qquad(5)$$

Let's find the asymptotic component expression for the second part of expression (5):

$$M\int\ldots\int\left(p(x)-\overline{p}_j(x)\right)\left(p(x)-\overline{p}_t(x)\right)dx_1\ldots dx_k=$$

$$=\int\ldots\int p^2(x)\,dx_1\ldots dx_k-M\int\ldots\int\overline{p}_t(x)p(x)\,dx_1\ldots dx_k-$$

$$-M\int\ldots\int\overline{p}_j(x)p(x)\,dx_1\ldots dx_k+$$

$$+M\int\ldots\int\overline{p}_j(x)\overline{p}_t(x)\,dx_1\ldots dx_k.\qquad(6)$$

Let's transform its last part:

$$M\int\ldots\int\overline{p}_j(x)\overline{p}_t(x)\,dx_1\ldots dx_k=$$

$$=\int\ldots\int M\left(\overline{p}_j(x)\right)M\left(\overline{p}_t(x)\right)dx_1\ldots dx_k,$$

which, with great enough volumes of statistical data considering expression (4) is presented as:

$$\int\ldots\int\left(p(x)+\frac{c^2}{2}\sum_{v=1}^{k}p_v^{(2)}(x)\right)^2dx_1\ldots dx_k\ .\qquad(7)$$

Notice that the asymptotic statistics expression of type:

$$M\int\ldots\int\overline{p}_t(x)p(x)\,dx_1\ldots dx_k$$

corresponds to:

$$\int\ldots\int\left(p(x)+\frac{c^2}{2}\sum_{v=1}^{k}p_v^{(2)}(x)\right)p(x)\,dx_1\ldots dx_k\ .\qquad(8)$$

Substituting expression (7), (8) in (6), after a series of simple conversions will give:

$$M\int\ldots\int\left(p(x)-\overline{p}_j(x)\right)\left(p(x)-\overline{p}_t(x)\right)dx_1\ldots dx_k\sim$$

$$\sim\frac{c^4}{4}\int\ldots\int\left(\sum_{v=1}^{k}p_v^{(2)}(x)\right)^2dx_1\ldots dx_k=\frac{c^4}{4}B\ .\qquad(9)$$

In V. A. Epanechnikov's research [2] – an asymptotic expression for the purpose of square deviation in nonparametric probability density estimation $p(x)$, composing the first part of expression (5), is received:

$$M\int\ldots\int\left(p(x)-\overline{p}_j(x)\right)^2dx_1\ldots dx_k\sim$$

$$\sim\frac{\prod_{v=1}^{k}\int\Phi^2(u_v)\,du_v}{\overline{n}\,c^k}+\frac{c^4}{4}B.\qquad(10)$$

Accounting (9) and (10), expression (5) with enough $\overline{n}$ values is represented as:

$$M\int\ldots\int\left(p(x)-\overline{\overline{p}}(x)\right)^2dx_1\ldots dx_k\sim$$

$$\sim\frac{\prod_{v=1}^{k}\int\Phi^2(u_v)\,du_v}{T\,\overline{n}\,c^k}+\frac{c^4}{4}B.\qquad(11)$$

It is not difficult to notice that in conditions $c\to 0$ at $\overline{n}c^k\to\infty$ the estimation $\overline{n}\to\infty$ of probability density mixture (3) converges in square mean to $p(x)$; considering the property of its asymptotic unbiasedness is well-founded.

At $T=1$ the received result (11) coincides with Epanechnikov's theorem [2], which confirms the correctness of the fulfilled conversions.

The analysis of approximating properties of statistics $\overline{\overline{p}}(x)$. For the efficiency analysis of a nonparametric estimation of probability densities mixture (3) and the Rosenblatt–Parzen estimations of a probability density:

$$\tilde{p}(x)=\frac{1}{n\,c^k}\sum_{i=1}^{n}\prod_{v=1}^{k}\Phi\left(\frac{x_v-x_v^i}{c}\right)$$

let's consider the ratio of asymptotic expressions, corresponding to deviation squares for the best coefficients of blur values in nuclear functions.

Let's define the minimum value W_2 of expression (11) with optimal coefficient c^* values of blur nonparametric estimations $\overline{p}_j(x)$ composing the probability densities mixture. In the accepted assumption value:

$$c^* = \left(\frac{k \prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{\overline{n} B} \right)^{\frac{1}{(k+4)}}.$$

Then:

$$W_2 = \left[\left(\frac{\prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{\overline{n}} \right)^4 B^k \right]^{\frac{1}{(k+4)}} \frac{4 + T k}{4 T k^{\frac{k}{(k+4)}}}. \quad (12)$$

If $k = 1$, then W_2 – is coincides with the minimal asymptotic expression of square deviation for the mixture of nonparametric probability densities estimations, obtained in study [3].

At $T = 1$ and $\overline{n} = n$ expression (12) corresponds to the minimal asymptotic expression W_2' for a deviation square of the probability density Rosenblatt–Parzen type estimation [2].

After simple conversions we get:

$$R_2 = \frac{W_2}{W_2'} = \frac{4 + T k}{(4 + k) T^{\frac{k}{(k+4)}}}.$$

By analogy we shall calculate the ratio for the minimal values of the main dispersing composing statistics $\overline{\overline{p}}(x)$ and $\tilde{p}(x)$:

$$W_3 = \frac{1}{T k^{\frac{k}{(k+4)}}} \left[\left(\frac{\prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{\overline{n}} \right)^4 B^k \right]^{\frac{1}{(k+4)}},$$

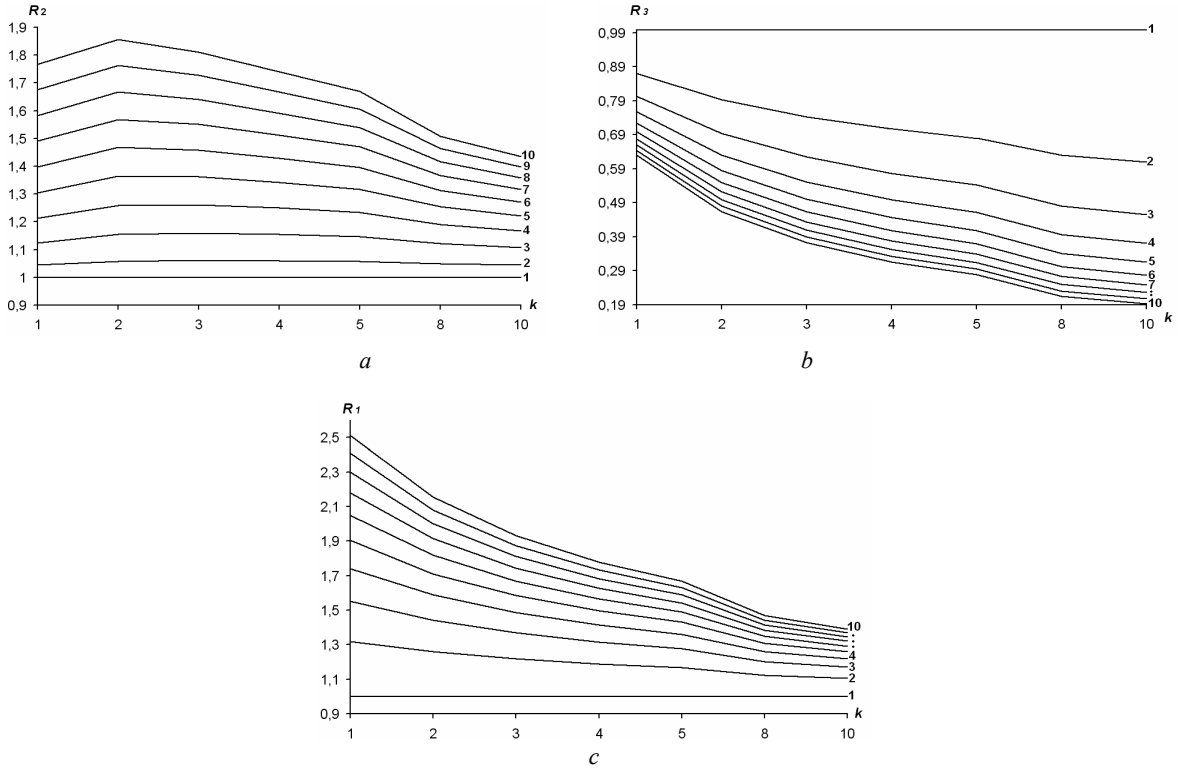
$$W_3' = \frac{1}{k^{\frac{k}{(k+4)}}} \left[\left(\frac{\prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{n} \right)^4 B^k \right]^{\frac{k}{(k+4)}}.$$

Their ratio looks as:

$$R_3 = \frac{W_3}{W_3'} = \frac{1}{T^{\frac{k}{(k+4)}}}.$$

It is easy to be convinced, that the ratio of asymptotic expressions offset: W_1, W_1' for the estimated probability density $\overline{\overline{p}}(x)$ and $\tilde{p}(x)$ at optimal blur coefficients for nuclear functions, is equal to:

$$R_1 = \frac{W_1}{W_1'} = T^{\frac{2}{(k+4)}}.$$



Dependences of ratios R_2 (*a*), R_3 (*b*), R_1 (*c*) from the dimension of random variable k and $x = \left(x_v,\, v = \overline{1, k} \right)$ quantity T = 1–10 (curves 1, …, 10), composing the nonparametric estimations mixture of probability density $\overline{\overline{p}}(x)$ (3)

With growth of component quantity T of the nonparametric estimations mixture of probability density, there is an increase in ratio values $R_2 > 1$ (figure, *a*), $R_1 > 1$ (figure, *c*). The noticed deterioration of approximating mixture properties $\overline{\overline{p}}(x)$ in comparison to traditional nonparametric estimation of density probability $\tilde{p}(x)$ (12), points to the decrease in sample sizes used during the estimation of compositions $\overline{\overline{p}}(x)$. This is a special feature of minor dimensions k of random variables. When complicating the estimating probability density with efficiency k, the growth of nonparametric estimations $\overline{\overline{p}}(x)$ also decreases $\tilde{p}(x)$. Criteria corresponding to them W_2, W_2' and W_1, W_1' become commensurable; this is evident in the decreasing of ratio R_2 and R_1 values.

The offered mixture $\overline{\overline{p}}(x)$ of probability density estimations has a lesser dispersion in comparison to the nonparametric estimation $\tilde{p}(x)$, which is identified by its structure, since statistics synthesis $\overline{\overline{p}}(x)$ is carried out on the basis of an averaging operator (figure, *b*). With a quantity increase in T composing the mixture of nonparametric estimations $\overline{\overline{p}}(x)$, the density probability and dimension k of random dimensions increases.

On the basis of the asymptotic properties analysis for nonparametric estimations mixtures of probability density with a multidimensional random variable, the decomposition possibility for initial statistical data under a synthesis of nonparametric statistics in large samples conditions is justified. The researched statistics, in comparison to the traditional Rosenblatt – Parzen nonparametric evaluation, has a considerably smaller dispersion and allows using parallel calculating technologies.

References

1. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. Vol. 33. P. 1065–1076.

2. Epanechnikov V. A. Nonparametric estimation of a many-dimensional probability density // Teoriya veroyatnosti i ee primeneniya, 1969. Vol. 14. № 1. P. 156–161.

3. Lapko V. A., Varochkin S. S., Egorochkin I. A. Development and research of a nonparametric estimation of the probability density grounded on a principle of decomposition of learning sample on its size // Vestnik SibSAU. 2009. Vol. 1 (22). P. 45–49.

D. V. Lichargin
Siberian Federal University, Russia, Krasnoyarsk

GENERATION OF THE STATE TREE BASED ON GENERATIVE GRAMMAR OVER TREES OF STRINGS

In the article the principle of state trees generation is considered based on the generative grammars over trees of strings in such objects as the sentences of natural languages, as well as two and tree dimensional images. The image of the object as a forest is considered; including the trees of object different layouts for the purpose of complex system modeling.

Keywords: natural language generation, generative grammars, semantics.

The problem of natural language sentences generation is one of the key issues in the field of computer science and formal grammar theories. The issue of meaningful speech generation applies to the area of semantics and computer science [1–7]. The states tree generation issue is studied well enough in computer science and in system analysis. In respect to the question of meaningful phrases tree generation the problem is first of all connected to the method of sentence generation by means of Chomsky's generative grammars. Generative grammars are successfully applied in software such as electronic translation systems, expert systems, systems of orthography checking, etc.

The flash point of the article is the analysis prospects for using generative grammars not over strings, but over trees of strings. In this respect it is possible to solve the task of generating grammatically and semantically meaningful speech more effectively and increasing the efficiency of different images analysis and synthesis aspects.

The importance of the issue on effective generating language meaningful constructions and two or three dimensional images is generally understood and is connected with the demands of linguistic and other software.

The purpose of this research is to apply generative grammars on the necessity basis over trees as means of meaningful speech generation connected with greater heterogeneous context.

The novelty of the work is in the application of generative grammars not over strings but over trees of strings.