УДК 519.7

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ БЫСТРОГО ОПРЕДЕЛЕНИЯ КОЭФФИЦИЕНТА РАЗМЫТОСТИ ЯДЕРНОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ

© **А. В. Лапко**[1, 2], **В. А. Лапко**[1, 2]

[1] *Институт вычислительного моделирования СО РАН,*
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Сибирский государственный университет науки и технологий им. академика М. Ф. Решетнева,*
660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Предлагается модификация быстрого алгоритма выбора коэффициента размытости ядерных функций в непараметрической оценке плотности вероятности типа Розенблатта — Парзена. Быстрые алгоритмы оптимизации ядерных оценок плотностей вероятностей позволяют значительно снизить временны́е затраты при выборе их коэффициентов размытости по сравнению с традиционным подходом, что особенно актуально при обработке статистических данных большого объёма. Основу предлагаемого метода составляют анализ формулы оптимального расчёта коэффициентов размытости ядерных функций и обнаруженная зависимость между нелинейным функционалом от второй производной восстанавливаемой плотности вероятности и коэффициентом контрэксцесса. Предлагаемый алгоритм выбора коэффициента размытости обеспечивает снижение ошибки аппроксимации плотности вероятности по сравнению с традиционным подходом. Полученные выводы подтверждаются результатами вычислительных экспериментов. Особое внимание уделяется зависимости этих свойств от объёма исходной информации.

Ключевые слова: ядерная оценка плотности вероятности, быстрый алгоритм оптимизации, выбор коэффициентов размытости, коэффициент контрэксцесса, симметричные плотности вероятности, вторая производная плотности вероятности.

DOI: 10.15372/AUT20200602

Введение. Непараметрические оценки плотности вероятности типа Розенблатта — Парзена $\bar{p}(x)$ используются при построении алгоритмов обработки статистических данных в условиях априорной неопределённости вида исследуемых закономерностей. Аппроксимационные свойства $\bar{p}(x)$ во многом зависят от выбора коэффициентов размытости ядерных функций, значения которых убывают с ростом объёма исходных данных. При решении этой задачи различают два подхода. Первый основан на выборе коэффициентов размытости ядерных функций из условия минимума статистической оценки среднеквадратического отклонения $\bar{p}(x)$ от плотности вероятности $p(x)$ [1–4]. Однако реализация этого подхода требует больших временны́х затрат, которые растут с увеличением объёма статистической информации, что характерно, например, при обработке данных дистанционного зондирования [5–8]. Для преодоления этой проблемы разработан быстрый алгоритм выбора коэффициентов размытости ядерных функций $\bar{p}(x)$, который использует результаты анализа его оптимального значения. Идея подхода состоит в вычислении оптимальных значений коэффициентов размытости c^* для непараметрических оценок, составляющих тестовое семейство плотностей вероятностей. Затем они обобщаются при определении оценки $\bar{c}^*$ в процедуре коллективного типа [9, 10]. В [11] предложена методика оценивания интеграла от квадрата второй производной восстанавливаемой плотности вероятности, которая является составляющей формулы расчёта оптимального коэффициента размытости.

В данной работе предлагается модификация быстрого алгоритма выбора коэффициентов размытости ядерных функций, которая основана на оценивании нелинейного функционала от второй производной восстанавливаемой плотности вероятности по значениям коэффициента контрэксцесса.

Свойства непараметрической оценки плотности вероятности. Рассмотрим непараметрическую оценку плотности вероятности $p(x)$ одномерной случайной величины

$$\bar{p}(x) = \frac{1}{nc} \sum_{i=1}^{n} \Phi\Big(\frac{x - x^i}{c}\Big), \tag{1}$$

которая восстанавливается по выборке статистически независимых наблюдений (x^i, $i = \overline{1, n}$) объёма n. Ядерные функции $\Phi(u)$ в статистике (1) удовлетворяют условиям:

$$\Phi(u) = \Phi(-u), \qquad 0 \leqslant \Phi(u) < \infty, \qquad \int \Phi(u)\, du = 1, \qquad \int u^2 \Phi(u)\, du = 1,$$

$$\int u^m \Phi(u)\, du < \infty, \qquad 0 \leqslant m < \infty.$$

Здесь и далее бесконечные пределы интегрирования опускаются. Значения коэффициентов размытости ядерных функций убывают: $c \to 0$, а $nc \to \infty$ при $n \to \infty$.

При исследовании свойств $\bar{p}(x)$ получено асимптотическое выражение средней квадратической ошибки аппроксимации

$$W(c) \sim \frac{1}{nc} \|\Phi(u)\|^2 + \frac{c^4}{4} \|p^{(2)}(x)\|^2, \tag{2}$$

где $p^{(2)}(x)$ — вторая производная $p(x)$ по x, а $\|\Phi(u)\|^2 = \int \Phi^2(u)\, du$, $\|p^{(2)}(x)\|^2 = \int (p^{(2)}(x))^2\, dx$.

Оптимальное значение коэффициента размытости ядерных функций определяется из условия минимума критерия (2):

$$c^* = \Big[\frac{\|\Phi(u)\|^2}{n\|p^{(2)}(x)\|^2}\Big]^{1/5}. \tag{3}$$

После несложных преобразований формула (3) представляется в виде

$$c^* = \beta \sigma n^{-1/5}, \tag{4}$$

где

$$\beta = \Big(\frac{\|\Phi(u)\|^2}{\sigma^5 \|p^{(2)}(x)\|^2}\Big)^{1/5}. \tag{5}$$

Составляющая выражения (5)

$$\lambda = \sigma^5 \|p^{(2)}(x)\|^2 \tag{6}$$

Таблица 1

Значения констант λ, β и коэффициента контрэксцесса δ
для семейства законов распределения одномерных случайных величин

Вид плотности вероятности	δ	λ	β
Нормальный	0,577	0,212	1,049
Логистический	0,488	0,467	0,895
Стьюдента (число степеней свободы $s = 5$)	0,337	0,73	0,819
Стьюдента ($s = 6$)	0,409	0,56	0,863
Стьюдента ($s = 7$)	0,447	0,472	0,893
Стьюдента ($s = 8$)	0,471	0,418	0,915
Стьюдента ($s = 9$)	0,488	0,382	0,932
Стьюдента ($s = 10$)	0,5	0,357	0,945
Стьюдента ($s = 15$)	0,531	0,294	0,982
Стьюдента ($s = 20$)	0,544	0,269	1
Стьюдента ($s = 25$)	0,552	0,256	1,01
Стьюдента ($s = 30$)	0,556	0,247	1,017

является константой для одномодальных плотностей вероятностей, которая определяется видом плотности и не зависит от её параметров [12, 13].

Возникает задача возможности оценивания константы λ от количественных характеристик восстанавливаемой плотности вероятности. Результаты её решения позволят повысить вычислительную эффективность методики быстрого выбора коэффициентов размытости ядерных функций при построении непараметрической оценки плотности вероятности случайных величин по статистически независимым наблюдениям.

Методика быстрого выбора коэффициентов размытости ядерных функций. Пусть законы распределения случайных величин являются симметричными и одномодальными. Соответствующие им значения λ, β и коэффициента контрэксцесса δ приведены в табл. 1.

По данным табл. 1 определена функциональная зависимость между константой λ и коэффициентом контрэксцесса δ:

$$\bar{\lambda} = -2{,}185\,\delta + 1{,}4635 \tag{7}$$

при средней относительной ошибке аппроксимации

$$\bar{\rho} = \frac{1}{N}\sum_{j=1}^{N}\frac{|\lambda^j - \bar{\lambda}^j|}{\lambda^j},$$

равной 0,0365. Здесь N — количество законов распределения случайных величин в табл. 1, а $\bar{\lambda}^j$ — значение константы λ, которая вычисляется по формуле (7) при конкретном значении коэффициента контрэксцесса δ^j.

Графическая иллюстрация зависимости (7) приведена на рис. 1.

Возникает вопрос о влиянии ошибки оценивания константы λ на аппроксимационные свойства непараметрической оценки плотности вероятности (1). Для этого проведён анализ отношения среднеквадратической ошибки аппроксимации $p(x)$ статистикой $\bar{p}(x)$ при оптимальном коэффициенте размытости c^* и его оценке $\bar{c}^*$ с учётом критерия (2). Обозначим через $(1 - \bar{\rho})\,\lambda$ результат оценивания λ (6) с помощью модели (7). Тогда

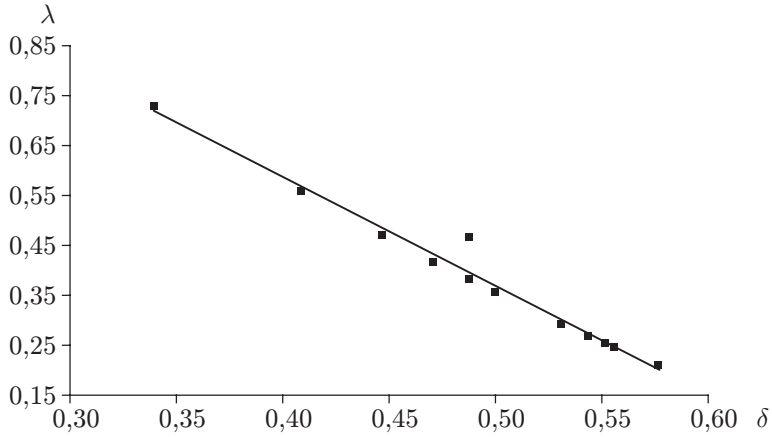

Рис. 1. Зависимость между константой λ и коэффициентом контрэксцесса (7). Точки на рисунке соответствуют данным табл. 1

$\bar{c}^* = (1-\rho)^{-1/5}\, c^*$, а отношение асимптотических выражений среднеквадратических отклонений (2) определяется значением

$$\frac{W(\bar{c}^*)}{W(c^*)} = \frac{1-(4/5)\bar{\rho}}{(1-\bar{\rho})^{4/5}}. \tag{8}$$

В этом случае при средней относительной ошибке $\bar{\rho} = 0{,}0365$ отношение (8) соответствует значению 1. Если допустить, что ошибка оценивания $\bar{\rho} = 0{,}1$, то отношение $W(\bar{c}^*)/W(c^*) = 1{,}001$, а при $\bar{\rho} = 0{,}3$ значение $W(\bar{c}^*)/W(c^*) = 1{,}01$.

Модифицированный алгоритм быстрого выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности (1) предполагает выполнение следующих действий:

1. По выборке $V = (x^i,\ i = \overline{1,n})$, которая имеется при восстановлении плотности вероятности $p(x)$ с использованием статистики (1), оценить среднеквадратическое отклонение σ случайной величины x и коэффициент контрэксцесса

$$\bar{\delta} = 1/\sqrt{\bar{\eta}},$$

где

$$\bar{\eta} = \frac{1}{n}\sum_{i=1}^{n}(x^i-\bar{x})^4 \Big/ \Big(\frac{1}{n}\sum_{i=1}^{n}(x^i-\bar{x})^2\Big)^2, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^i.$$

2. Используя модель (7) зависимости между константой λ и коэффициентом контрэксцесса δ, вычислить значения $\bar{\lambda}$ и

$$\bar{\beta} = (\|\Phi(u)\|^2/\bar{\lambda})^{1/5}$$

при выбранном виде ядерной функции.

3. При известных значениях n, $\bar{\sigma}$, $\bar{\beta}$ в соответствии с формулой (4) вычислить оценку коэффициента размытости ядерных функций статистики (1):

$$\bar{c}^* = \bar{\beta}\bar{\sigma}\, n^{-1/5}. \tag{9}$$

4. Оценить эффективность выбора коэффициента размытости ядерных функций непараметрической оценки плотности вероятности (1), используя формулу (2) при $c = \bar{c}^*$.

Анализ результатов вычислительных экспериментов. Предложенную методику рассмотрим при восстановлении плотности вероятности по выборке $V = (x^i,\ i = \overline{1,n}\,)$ для $n \in [100;500]$. Выборки V формировались с нормальной плотностью вероятности $p(x) = N(0;0{,}5)$ и логнормальным законом распределения

$$p(x) = \frac{1}{sx\sqrt{2\pi}} \exp\Big(- \frac{(\ln x - a)^2}{2s^2}\Big)$$

при $a = 0$, $s = 0{,}5$, который отсутствует в данных табл. 1.

Синтез непараметрической оценки плотности вероятности (1) осуществлялся на основе ядерной функции [14]

$$\Phi(u) = \begin{cases} 3/(4\sqrt{5}\,) - 3u^2/(20\sqrt{5}\,), & \forall\ |u| < \sqrt{5}; \\ 0, & \forall\ |u| \geqslant \sqrt{5}. \end{cases}$$

В этих условиях $\|\Phi(u)\|^2 = 3/(5\sqrt{5}\,)$.

Проводится сравнение эффективности методов быстрого выбора коэффициента размытости, рассмотренного в [9], и предлагаемого в данной работе. При конкретных значениях объёма n статистических данных вычислительный эксперимент повторялся 50 раз, полученные результаты расчётов оценок коэффициентов размытости $\bar{c}^*$ ядерных функций и соответствующих им значений среднеквадратических отклонений $\bar{p}(x)$ усреднялись, например, для нормального закона распределения $N(0;0{,}5)$ параметр $\bar{\sigma} = 0{,}494$, $\bar{\delta} = 0{,}591$ в условиях $n = 500$. При применении традиционного метода [9] с учётом всей информации табл. 1 получены расчётные значения константы $\bar{\beta} = 0{,}943$ и оценки коэффициента размытости ядерных функций $\bar{c}^* = 0{,}134$ статистики (1). При этом среднеквадратическое отклонение (2) $\bar{p}(x)$ от плотности вероятности $p(x)$ соответствует значению $W(\bar{c}^*) = 4{,}546 \cdot 10^{-3}$. В принятых условиях при исключении сведений о нормальном законе распределения из табл. 1 получим следующие результаты: $\bar{\beta} = 0{,}934$, $\bar{c}^* = 0{,}133$, $W(\bar{c}^*) = 4{,}56 \cdot 10^{-3}$.

При использовании предлагаемого метода имеем $\bar{\beta} = 1{,}099$, $\bar{c}^* = 0{,}157$, $W(\bar{c}^*) = 4{,}442 \cdot 10^{-3}$. Заметим, что в этом методе наблюдается минимальное значение $W(\bar{c}^*)$ и наиболее близкое соответствие теоретического значения $\beta = 1{,}049$ [9] расчётному $\bar{\beta}$.

При восстановлении логнормального закона распределения по выборке объёма $n = 500$ параметры $\bar{\sigma} = 0{,}599$ и $\bar{\delta} = 0{,}405$. Если использовать традиционный подход [9] (метод 1 с учётом нормального закона распределения в табл. 1), то в статистике (1) коэффициент размытости ядерных функций $\bar{c}^*(1) = 0{,}163$. В этих условиях среднеквадратическое отклонение $\bar{p}(x)$ от восстанавливаемой плотности вероятности $W(\bar{c}^*(1)) = 0{,}012$. В предлагаемом подходе (метод 2) $\bar{c}^*(2) = 0{,}149$, а $W(\bar{c}^*(2)) = 0{,}0094$.

Оценим преимущество метода 2 по сравнению с методом 1 по значению критерия

$$W_{12} = \frac{|W(\bar{c}^*(1)) - W(\bar{c}^*(2))|}{W(c^*)},$$

где $W(c^*) = 0{,}0065$ — значение среднеквадратического критерия (2) при оптимальном коэффициенте размытости для логнормального закона распределения $c^* = 0{,}103$. Эффективность предлагаемого метода определяется значением $W_{12} = 0{,}4$, так как $W(\bar{c}^*(2)) < W(\bar{c}^*(1))$.

Для логнормального закона распределения зависимости коэффициентов размытости ядерных функций $\bar{c}^*(1)$, $\bar{c}^*(2)$ в статистике (1) от объёма n статистических данных, рассчитанных по методам 1, 2, и соответствующие им значения $W(\bar{c}^*(1))$, $W(\bar{c}^*(2))$ приведены на рис. 2, 3.

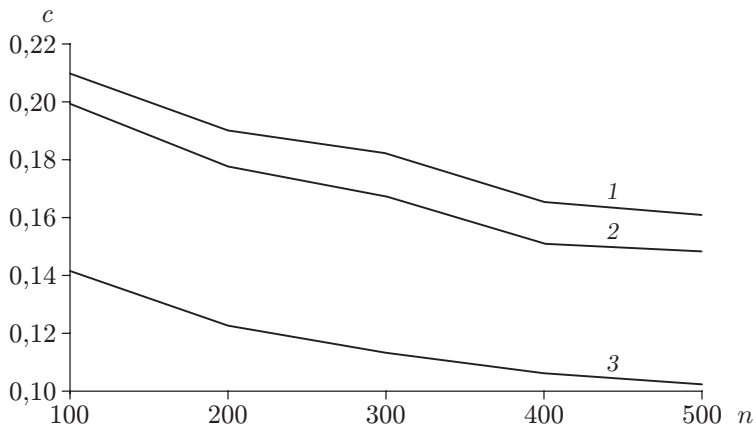

Рис. 2. Зависимости коэффициентов размытости ядерных функций в статистике (1) от объёма n статистических данных при восстановлении логнормальной плотности вероятности. Кривые *1, 2, 3* соответствуют значениям $\bar{c}^*(1)$, $\bar{c}^*(2)$, c^*

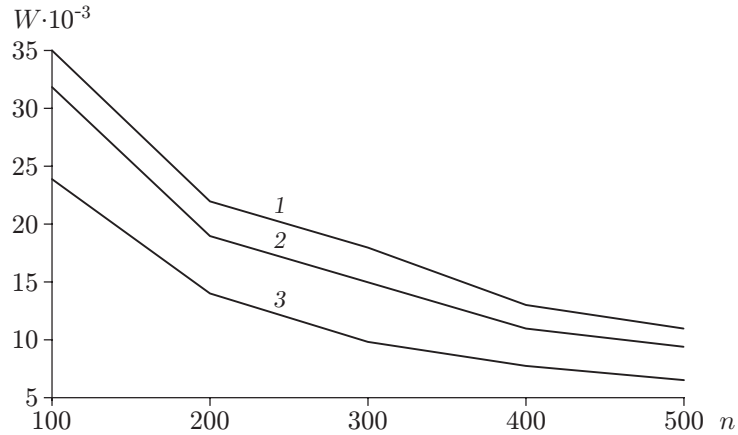

Рис. 3. Зависимости среднеквадратической ошибки аппроксимации (2) логнормальной плотности вероятности от объёма n статистических данных. Кривые *1, 2, 3* соответствуют значениям $W(\bar{c}^*(1))$, $W(\bar{c}^*(2))$, $W(c^*)$

На всём диапазоне изменения объёма статистических данных коэффициент размытости ядерных функций, рассчитанный по методу 1, превышает значение $\bar{c}^*(2)$, определённое по методу 2. Расчёт коэффициента размытости ядерных функций по формуле (9) позволяет более близко приблизиться к оптимальному значению c^*. По методу 1 коэффициент размытости ядерных функций $\bar{c}^*(1) \in [0{,}212; 0{,}163]$, по методу 2 значения $\bar{c}^*(2) \in [0{,}199; 0{,}149]$, оптимальные значения $c^* \in [0{,}142; 0{,}103]$.

Уточнение коэффициента размытости ядерных функций в статистике (1) при использовании метода (2) позволяет снизить ошибку аппроксимации плотности вероятности (рис. 3). Значения $W(\bar{c}^*(1)) \in [0{,}037; 0{,}012]$ уменьшаются с ростом объёма статистических данных $n \in [100; 500]$, а значения $W(\bar{c}^*(2)) \in [0{,}032; 0{,}0094]$.

С ростом n эффективность метода 2 по сравнению с методом 1 возрастает в 2 раза от значения критерия $W_{12} = 0{,}2$ до $W_{12} = 0{,}4$. При увеличении объёма статистических данных до $n = 5000$ отмеченная тенденция изменения критерия W_{12} сохраняется. Этим условиям соответствуют следующие результаты вычислительных экспериментов: $\bar{c}^*(1) = 0{,}104$, $W(\bar{c}^*(1)) = 1{,}85 \cdot 10^{-3}$, $\bar{c}^*(2) = 0{,}09$, $W(\bar{c}^*(2)) = 1{,}37 \cdot 10^{-3}$. Оптимальный

коэффициент размытости $c^* = 0{,}065$ вычислен по формуле (3), а соответствующий ему критерий $W(c^*) = 1{,}036 \cdot 10^{-3}$.

Полученные результаты являются достоверными, так как гипотеза H_0: $p(W(\bar{c}^*(1))) \equiv \equiv p(W(\bar{c}^*(2)))$ о тождественности законов распределения по критерию Смирнова отвергается с риском 0,05 [15].

Заключение. Быстрые алгоритмы выбора коэффициента размытости ядерных функций в непараметрической оценке плотности вероятности позволяют на несколько порядков сократить временны́е затраты. Предлагаемая методика основывается на результатах анализа формулы расчёта оптимального коэффициента размытости ядерных функций, полученной из условия минимума асимптотического выражения среднеквадратического отклонения непараметрической оценки плотности вероятности. Основная составляющая формулы расчёта оптимального коэффициента размытости является нелинейным функционалом от второй производной плотности вероятности, который определяется видом плотности вероятности и не зависит от её параметров. Она является константой для семейства одномодальных и близких к симметричным плотностям вероятностей. Зависимость обнаруженной константы от коэффициента контрэксцесса случайной величины близка к линейной. Ошибки в определении коэффициента размытости ядерных функций незначительно влияют на аппроксимационные свойства непараметрической оценки плотности вероятности. Обнаруженные закономерности подтверждаются аналитическими выводами и результатами вычислительных экспериментов при оценивании логнормального закона распределения, который характерен при анализе спектральных данных дистанционного зондирования.

Дальнейшее развитие предлагаемого подхода состоит в развитии методики быстрого выбора коэффициентов размытости ядерных функций непараметрической оценки многомерной плотности вероятности.

СПИСОК ЛИТЕРАТУРЫ

1. **Rudemo M.** Empirical choice of histogram and kernel density estimators // Scandinavian Journ. Statist. 1982. **9**, N 2. P. 65–78.
2. **Bowman A. W.** A comparative study of some kernel-based non-parametric density estimators // Journ. Statist. Comput. Simulation. 1982. **21**, Iss. 3–4. P. 313–327.
3. **Hall P.** Large-sample optimality of least squares cross-validation in density estimation // Ann. Statist. 1983. **11**, N 4. P. 1156–1174.
4. **Лапко А. В., Лапко В. А.** Анализ методов оптимизации непараметрической оценки плотности вероятности по коэффициенту размытости ядерных функций // Измерительная техника. 2017. № 6. С. 3–8.
5. **Нежевенко Е. С.** Нейросетевая классификация трудноразличимых типов растительности по гиперспектральным признакам // Автометрия. 2019. **55**, № 3. С. 62–70. DOI: 10.15372/AUT20190308.
6. **Лапко А. В., Лапко В. А., Им С. Т. и др.** Непараметрический алгоритм выделения классов, соответствующих одномодальным фрагментам плотности вероятности многомерных случайных величин // Автометрия. 2019. **55**, № 3. С. 22–30. DOI: 10.15372/AUT20190303.
7. **Лапко А. В., Лапко В. А.** Методика проверки гипотез о распределениях многомерных спектральных данных с использованием непараметрического алгоритма распознавания образов // Компьютерная оптика. 2019. **43**, № 2. С. 238–244. DOI: 10.18287/2412-6179-2019-43-2-238-244.
8. **Борзов С. М., Потатуркин О. И.** Спектрально-пространственные методы классификации гиперспектральных изображений. Обзор // Автометрия. 2018. **54**, № 6. С. 64–86. DOI: 10.15372/AUT20180607.

9. **Лапко А. В., Лапко В. А.** Быстрый алгоритм выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности // Измерительная техника. 2018. № 6. С. 16–20. DOI: 10.32446/0368-1025it-2018-6-16-20.
10. **Лапко А. В., Лапко В. А.** Быстрый алгоритм выбора коэффициентов размытости в многомерных ядерных оценках плотности вероятности // Измерительная техника. 2018. № 10. С. 19–23. DOI: 10.32446/0368-1025it.2018-10-19-23.
11. **Лапко А. В., Лапко В. А.** Оценивание интеграла от квадрата производных симметричных плотностей вероятностей одномерных случайных величин // Метрология. 2020. № 1. С. 15–27.
12. **Scott D. W.** Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons, 2015. 384 p.
13. **Лапко А. В., Лапко В. А.** Зависимость между параметрами гистограммы и ядерной оценки одномодальной плотности вероятности // Измерительная техника. 2019. № 9. С. 3–8. DOI: 10.32446/0368-1025it.2019-9-3-8.
14. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. **14**, № 1. С. 156–161.
15. **Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.** Сложные системы. М.: Высш. шк., 1977. 248 с.

Поступила в редакцию 22.04.2020
После доработки 24.08.2020
Принята к публикации 12.10.2020