

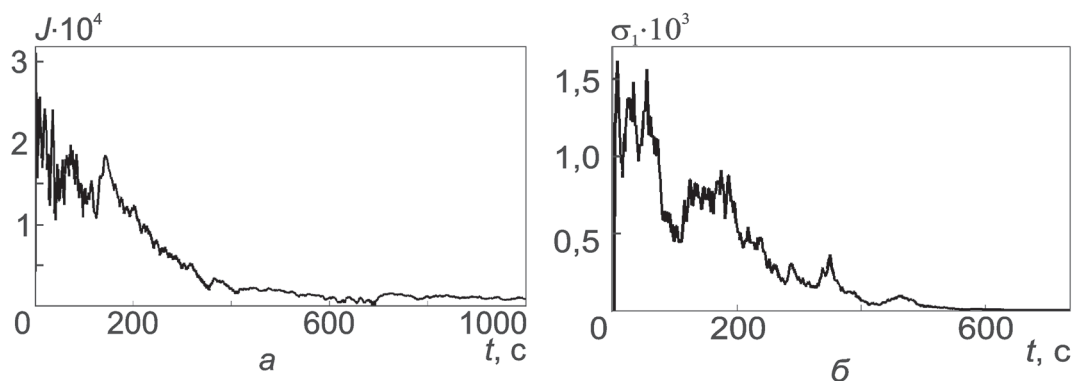


Рис. 3. Графики эволюции критерия $J(a)$ и параметра настройки $\sigma_1(b)$

Литература

1. Девятисильный А. С., Числов К. А. Нейроподобная инерциально-спутниковая векторная гравиметрическая система // Измерительная техника. 2013. № 10. С. 3—8.
2. Андреев В. Д. Теория инерциальной навигации. Автономные системы. М.: Наука, 1966.
3. Андреев В. Д. Теория инерциальной навигации. Корректируемые системы. М.: Наука, 1967.

4. Калман Р., Фалб П., Арбиб М. Очерки по математической теории систем / Пер. с англ., под ред. Э. З. Цыпкина. М.: Мир, 1971.

5. Девятисильный А. С., Числов К. А. Интегрированная инерциально-спутниковая навигационная система, корректируемая по наблюдениям одной звезды // Измерительная техника. 2013. № 4. С. 3—5.

6. Лескова Н. Л. В стране лилипутов // В мире науки. 2014. № 6. С. 48—53.

Дата принятия 16.01.2015 г.

519.24

Выбор оптимального количества интервалов дискретизации области значений двумерной случайной величины

А. В. ЛАПКО^{1, 2}, В. А. ЛАПКО^{1, 2}

¹ Институт вычислительного моделирования СО РАН, Красноярск, Россия

² Сибирский государственный аэрокосмический университет им. акад. М.Ф. Решетнева, Красноярск, Россия, e-mail: lapko@icm.krasn.ru

Исследованы асимптотические свойства непараметрической оценки двумерной плотности вероятности, синтез которой предполагает декомпозицию статистических данных. Определена зависимость количества двумерных интервалов дискретизации от объема исходной информации.

Ключевые слова: плотность вероятности, непараметрическая оценка, аппроксимационные свойства, количество интервалов дискретизации, критерий Пирсона.

The asymptotic properties of nonparametric estimation of two-dimensional probability density have been studied. The synthesis of this estimation presupposes decomposition of statistical data. The dependence is determined of the number of two-dimensional sampling intervals on the volume of the original data.

Key words: probability density, nonparametric estimation, approximation properties, number of sampling intervals, Pearson criterion.

Выбор количества интервалов дискретизации области значений случайных величин очень важен в задачах определения плотности вероятности, её доверительного оценивания и проверке статистических гипотез с использованием критерия Пирсона. Для этого разработаны традиционные

методы, характеризующиеся широким разбросом числа одномерных интервалов дискретизации [1—4]. В [5] при анализе аппроксимационных свойств непараметрической оценки плотности вероятности ядерного типа, восстанавливаемой по статистическим данным объема n , получена процедура

выбора оптимального количества интервалов дискретизации N_1^* области Δ значений случайной величины x с плотностью вероятности $p(x)$:

$$N_1^* = \left(n \Delta \int_{-\infty}^{+\infty} p(x)^2 dx \right)^{1/2} = \bar{k}_1 \sqrt{n},$$

где $\bar{k}_1$ — коэффициент.

Данная аналитическая зависимость по виду близка к формуле Гаеде [3] и совпадает с ней при оценивании плотности вероятности случайной величины с равномерным законом распределения. Установлено, что $\bar{k}_1$ определяется только видом плотности вероятности и не зависит от её параметров.

Получим аналитическую зависимость оптимального количества двумерных интервалов дискретизации от объёма исходных статистических данных и характеристик анализируемой плотности вероятности, исследовав асимптотические свойства непараметрической оценки двумерной плотности вероятности.

Синтез непараметрической оценки плотности вероятности. Пусть дана выборка $V = (x^j, j=1, \bar{n})$ из n независимых наблюдений двумерной случайной величины $x=(x_1, x_2)$ с неизвестной плотностью вероятности $p(x_1, x_2)$. Предположим, что $p(x_1, x_2)$ можно разложить в ряд Тейлора по всем аргументам в каждой точке $x=(x_1, x_2)$. Для упрощения аналитических преобразований будем считать, что интервалы Δ_v изменения аргументов x_v равны, т. е. $\Delta_v = \Delta$, $v = 1, 2$. Разобьём область изменения каждого аргумента на $\bar{N}$ непересекающихся интервалов длиной 2β таких, что $\Delta = 2\beta\bar{N}$. По исходным данным V сформируем множества случайных величин X^i , $i = \overline{1, \bar{N}}$; $N = \bar{N}^2$. В качестве характеристик X^i примем частоту $\bar{P}^i$ попадания случайной величины x в i -й двумерный интервал и его центр $z^i = (z_1^i, z_2^i)$. На основе полученной информации составим статистическую выборку $V_1 = (z^i, y^i = \bar{P}^i / (2\beta)^2, i = \overline{1, \bar{N}})$. Объём N полученной выборки может быть значительно меньше объёма n исходных статистических данных.

В качестве приближения по эмпирическим данным искомым плотностью $p(x_1, x_2)$ примем статистику [6, 7]:

$$\bar{p}(x_1, x_2) = \frac{1}{c^2} \sum_{i=1}^N \bar{P}^i \prod_{v=1}^2 \Phi\left(\frac{x_v - z_v^i}{c}\right), \quad (1)$$

где c — коэффициент размытости; $\Phi(u_v)$ — ядерные функции, удовлетворяющие условиям [8, 9]:

$$\begin{aligned} \Phi(u_v) &= \Phi(-u_v); \quad 0 \leq \Phi(u_v) < \infty; \\ \int \Phi(u_v) du_v &= 1; \quad \int u_v^2 \Phi(u_v) du_v = 1. \end{aligned} \quad (2)$$

Здесь и далее бесконечные пределы интегрирования опущены.

Коэффициенты размытости $c = c(N)$ ядерных функций в статистике (1) убывают с ростом количества N двумерных

интервалов дискретизации. Их оптимальное значение, минимизирующее асимптотическое выражение среднеквадратического отклонения (СКО) аппроксимации оценки $\bar{p}(x_1, x_2)$ от $p(x_1, x_2)$, можно определить как [6, 7]:

$$\bar{c} = \left(2 \|p(x_1, x_2)\|^2 \left(2\beta \|\Phi(u)\|^2 \right)^2 / B \right)^{1/6}, \quad (3)$$

где $\|\Phi(u)\|^2 = \int \Phi^2(u) du$; $\|p(x_1, x_2)\|^2 = \int \int p^2(x_1, x_2) dx_1 dx_2$;

$B = \int \int \left(\sum_{v=1}^2 p_v^{(2)}(x_1, x_2) \right)^2 dx_1 dx_2$; $p_v^{(2)}(x_1, x_2)$ — вторая производная

плотности вероятности $p(x_1, x_2)$ по аргументу x_v .

Асимптотические свойства непараметрической оценки плотности вероятности. Представим непараметрическую оценку плотности вероятности (1) в виде

$$\bar{p}(x_1, x_2) = \frac{1}{nc^2} \sum_{i=1}^N \sum_{j=1}^n \prod_{v=1}^2 h\left(\frac{x_v^j - z_v^i}{\beta}\right) \prod_{v=1}^2 \Phi\left(\frac{x_v - z_v^i}{c}\right), \quad (4)$$

где индикаторные функции

$$h\left(\frac{x_v^j - z_v^i}{\beta}\right) = \begin{cases} 1, & \text{если } |x_v^j - z_v^i| \leq \beta; \\ 0, & \text{если } |x_v^j - z_v^i| > \beta \end{cases}$$

определяют принадлежность элементов выборки $V = (x^j, j = \overline{1, \bar{n}})$ двумерным интервалам $(z_v^i \pm \beta, v = 1, 2)$, $i = \overline{1, \bar{N}}$.

При анализе асимптотических свойств статистики (4) будем использовать технологию преобразований, предложенную в [9]. В отличие от (1) в модификации (4) в явном виде присутствуют параметры β , N процедуры дискретизации и объём n исходных статистических данных. Поэтому проанализировав (4), можно получить асимптотическое выражение СКО $W_2(c, N)$, зависящее от коэффициента размытости ядерных функций оценки плотности (4) и количества интервалов дискретизации N . Подставив в $W_2(c, N)$ оптимальное значение $\bar{c}$ из (3), получим $W_2(N)$. Минимизация $W_2(N)$ по параметру N позволит найти аналитическую зависимость количества N интервалов дискретизации от объёма n исходных статистических данных.

Следуя предложенной схеме, вычислим

$$\begin{aligned} M[\bar{p}(x_1, x_2)] &= \frac{1}{nc^2} \sum_{i=1}^N \sum_{j=1}^n \int \int \int \int h\left(\frac{x_1^j - z_1^i}{\beta}\right) h\left(\frac{x_2^j - z_2^i}{\beta}\right) \times \\ &\times p(x_1^j, x_2^j) dx_1^j dx_2^j \Phi\left(\frac{x_1 - z_1^i}{c}\right) \Phi\left(\frac{x_2 - z_2^i}{c}\right) p_2(z_1^i, z_2^i) dz_1^i dz_2^i = \\ &= \frac{N}{c^2} \int \int \int \int h\left(\frac{y_1 - t_1}{\beta}\right) h\left(\frac{y_2 - t_2}{\beta}\right) p(y_1, y_2) dy_1 dy_2 \times \\ &\times \Phi\left(\frac{x_1 - t_1}{c}\right) \Phi\left(\frac{x_2 - t_2}{c}\right) p_2(t_1, t_2) dt_1 dt_2, \end{aligned} \quad (5)$$

где M — математическое ожидание; $p_2(t_1, t_2) = (4\beta^2 N)^{-1}$ — плотность вероятности распределения центров интервалов дискретизации.

Заменим в (5) переменные $y_1 = \beta u_1 + t_1$; $y_2 = \beta u_2 + t_2$; $t_1 = x_1 - c w_1$; $t_2 = x_2 - c w_2$. Разложим функцию $p(x_1 - c w_1 + \beta u_1, x_2 - c w_2 + \beta u_2)$ в ряд Тейлора в точке $x = (x_1, x_2)$.

Тогда (5) запишем в виде

$$M[\bar{p}(x_1, x_2)] \approx \frac{1}{4} \int \int \int \{ p(x_1, x_2) + \sum_{v=1}^2 [(\beta u_v - c w_v) p_v^{(1)}(x_1, x_2) + \frac{1}{2} (\beta u_v - c w_v)^2 p_v^{(2)}(x_1, x_2)] \} \times \\ \times \prod_{v=1}^2 [h(u_v) \Phi(w_v)] du_1 du_2 dw_1 dw_2.$$

С учётом свойств ядерных $\Phi(w_v)$ из (2) и индикаторных функций $h(u_v)$, $v = 1, 2$ запишем асимптотическое выражение

$$M[\bar{p}(x_1, x_2)] \sim p(x_1, x_2) + \frac{\beta^2/3 + c^2}{2} \sum_{v=1}^2 p_v^{(2)}(x_1, x_2). \quad (6)$$

Определим асимптотические свойства СКО:

$$M[\bar{p}(x_1, x_2) - p(x_1, x_2)]^2 = \\ = M[\bar{p}^2(x_1, x_2)] - 2p(x_1, x_2)M[\bar{p}(x_1, x_2)] + p^2(x_1, x_2). \quad (7)$$

Применив использованные ранее преобразования, сформируем асимптотическое выражение

$$M[\bar{p}^2(x_1, x_2)] \sim p^2(x_1, x_2) + \frac{4\beta^2}{c^2} p^2(x_1, x_2) (\|\Phi(u)\|^2)^2 + \\ + \frac{N}{nc^2} p(x_1, x_2) (\|\Phi(u)\|^2)^2 + \left(\frac{\beta^2}{3} + c^2 \right) p(x_1, x_2) \times \\ \times \sum_{v=1}^2 p_v^{(2)}(x_1, x_2) + \frac{c^4}{4} \left[\sum_{v=1}^2 p_v^{(2)}(x_1, x_2) \right]^2. \quad (8)$$

С учётом (6) при значительных объёмах статистических данных запишем второе слагаемое выражения (7) в виде

$$-2p(x_1, x_2)M[\bar{p}(x_1, x_2)] \sim \\ \sim -2p^2(x_1, x_2) - \left(\frac{\beta^2}{3} + c^2 \right) p(x_1, x_2) \sum_{v=1}^2 p_v^{(2)}(x_1, x_2). \quad (9)$$

Тогда, подставив результаты (8), (9) в (7), получим асимптотическое выражение для СКО:

$$M[\bar{p}(x_1, x_2) - p(x_1, x_2)]^2 \sim \frac{4\beta^2}{c^2} p^2(x_1, x_2) (\|\Phi(u)\|^2)^2 + \\ + \frac{N}{nc^2} p(x_1, x_2) (\|\Phi(u)\|^2)^2 + \frac{c^4}{4} \left[\sum_{v=1}^2 p_v^{(2)}(x_1, x_2) \right]^2. \quad (10)$$

Из анализа (10) следует, что при выполнении условий $\frac{\beta}{c} \rightarrow 0$; $\beta \rightarrow 0$; $c \rightarrow 0$; $\frac{N}{nc^2} \rightarrow 0$ при $n \rightarrow \infty$ статистика (4) обладает свойством сходимости в СКО, а с учётом асимптотической несмещённости (6) является состоятельной.

Анализ аппроксимационных свойств непараметрической оценки плотности вероятности. Определим оптимальное количество интервалов дискретизации области значений двумерной случайной величины $x = (x_1, x_2)$ из условия минимума асимптотического выражения СКО $\bar{p}(x_1, x_2)$ от $p = p(x_1, x_2)$, полученного при интегрировании критерия (10):

$$M \int \int [\bar{p}(x_1, x_2) - p(x_1, x_2)]^2 dx_1 dx_2 \sim \\ \sim \frac{4\beta^2}{c^2} \|p(x_1, x_2)\|^2 (\|\Phi(u)\|^2)^2 + \frac{N}{nc^2} (\|\Phi(u)\|^2)^2 + \frac{c^4}{4} B. \quad (11)$$

Подставим оптимальное значение коэффициента размытости (3) ядерных функций непараметрической оценки плотности вероятности (1) в (11) и с учётом принятой методики дискретизации заменим β на $\Delta / (2\sqrt{N})$.

В результате получим

$$W_2(N) = \left(\frac{1}{4} B^2 (\|\Phi(u)\|^2)^8 (\|p(x_1, x_2)\|^2)^4 \Delta^2 \right)^{1/6} \times \\ \times \left(\frac{3\Delta}{2N^{4/6}} + \frac{N^{8/6}}{n\Delta \|p(x_1, x_2)\|^2} \right). \quad (12)$$

Тогда оптимальное количество N^* двумерных интервалов дискретизации, которое минимизирует асимптотическую погрешность (12) аппроксимации плотности вероятности, определим из выражения

$$N^* = \left(3/4 \Delta^2 \|p(x_1, x_2)\|^2 n \right)^{1/2}. \quad (13)$$

Количество $\bar{N}^*$ интервалов дискретизации каждой компоненты x_v , $v = 1, 2$ соответствует значению $\bar{N}^* = \sqrt{N^*}$. Коэффициент

$\bar{k}_2 = \left(3/4 \Delta^2 \|p(x_1, x_2)\|^2 \right)^{1/2}$ определяется видом плотности вероятности и не зависит от её параметров. Например, для независимых случайных x_1, x_2 с равномерным, линейным и нормальным законами распределения коэффициент $\bar{k}_2$ принимает значения соответственно 0,866; 1,15; 1,46.

Проанализируем зависимость N^* от объёма n исходных статистических данных для ряда двумерных плотностей вероятности $p(x_1, x_2) = p(x_1) p(x_2)$ с указанными законами распределения. При малых n количество интервалов N^* дискретизации, полученных в соответствии с (13) для сравниваемых законов распределения, сопоставимы (рисунок).

Увеличение объёма статистических данных сопровождается ростом количества двумерных интервалов дискретизации. Для нормального закона распределения количество N^* интервалов дискретизации принимает большее значение по сравнению с линейным и равномерным. Данный факт объясняется более сложным видом нормального закона распределения, что приводит к неравномерному характеру распределения статистических данных и вызвано усложнением условий сходимости непараметрической оценки плотности вероятности.

При $n < 250$ количество n_2 значений случайных величин, приходящихся на двумерный интервал дискретизации, составляет 5—10 наблюдений для нормального закона распределения. Для линейного и равномерного законов $n_2 \in 6...13$, $n_2 \in 8...18$ наблюдениям, соответственно. Полученные результаты согласуются с основными выводами аналитических исследований [6, 7]. Например, в случае двумерного нормального закона распределения относительно малые значения n_2 ухудшают условия оценивания вероятности принадлежности значений случайной величины некоторым интервалам дискретизации в статистике (1). Однако высокие аппроксимационные свойства непараметрической оценки плотности вероятности поддерживаются увеличением количества N^* интервалов дискретизации.

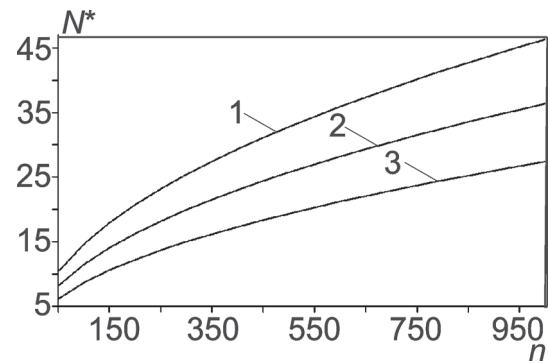
Дискретизация для многомерной случайной величины $x = (x_v, v = \overline{1, k})$ сопровождается значительными аналитическими преобразованиями. Однако при анализе предложенных формул дискретизации областей определения одномерной и двумерной плотности вероятности не исключена возможность их обобщения на многомерный случай:

$$N_k^* = \left[\alpha(k) \Delta^k \|p(x_1, x_2, \dots, x_k)\|^2 n \right]^{1/2}; \quad \bar{N}_1^* = k \sqrt{N_k^*}.$$

Коэффициент $\alpha(k) \leq 1$ и уменьшается с ростом размерности k случайной величины, причём $\alpha(1)=1$; $\alpha(2)=3/4$. Одним из перспективных направлений выбора $\alpha(k)$ является использование результатов [6, 7].

Заключение. Исследование асимптотических свойств непараметрической оценки двумерной плотности вероятности ядерного типа, синтез которой предполагает декомпозицию исходных статистических данных, позволяет сформировать критерий выбора оптимального количества двумерных интервалов дискретизации. Таким образом, можно определить зависимость количества интервалов декомпозиции области значений двумерной случайной величины от объёма статистических данных и характеристик восстанавливаемой плотности вероятности. Полученные результаты, позволяющие вывести формулу дискретизации области значений многомерной случайной величины, особенно важны в задачах проверки гипотез о распределениях двумерных случайных величин с использованием критерия Пирсона.

Работа выполнена в рамках базовой части государственного задания Минобрнауки России (СибГАУ № Б121/14) и



Зависимость количества интервалов N^* дискретизации области значений двумерной случайной величины от объёма n исходных статистических данных:

1, 2, 3 — нормальный, линейный и равномерный законы распределения, соответственно

Программы СО РАН IV.35.1 «Теоретические основы и технологии создания и применения интегрированных информационно-вычислительных систем для решения задач поддержки принятия решений».

Л и т е р а т у р а

1. **Sturges H. A.** The choice of a class interval // J. Am. Statist. Association. 1926. P. 65—66.
2. **Freedman D., Diaconis P.** On the histogram as a density estimator: L2 theory // Probab. Theory and Rel. Fields. 1981. V. 57. N. 2. P. 453—476.
3. **Heinhold I., Gaede K.** Ingenieur statistic. München, Wien: Springer Verlag, 1964.
4. **Scott D.** On optimal and data-based histograms // Biometrika. 1979. V. 66. N. 3. P. 605—610.
5. **Лапко А. В., Лапко В. А.** Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности // Измерительная техника. 2013. № 7. С. 24—27.
6. **Лапко А. В., Лапко В. А.** Регрессионная оценка многомерной плотности вероятности и её свойства // Автометрия. 2014. Т. 50. № 2. С. 50—56.
7. **Lapko A. V., Lapko V. A.** Regression Estimate of the Multidimensional Probability Density and Its Properties // Opt. Instrum. Data Processing, 2014. V. 50. N. 2. P. 148—153.
8. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. V. 33. N. 3. P. 1065—1076.
9. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. Т. 14. № 1. С. 156—161.

Дата принятия 12.03.2015 г.