УДК 519.7

© 2016 г. **А.В. Лапко**, д-р техн. наук,
В.А. Лапко, д-р техн. наук
(Институт вычислительного моделирования СО РАН, Красноярск,
Сибирский государственный аэрокосмический университет им.
академика М.Ф. Решетнева, Красноярск)

АНАЛИЗ ЭФФЕКТИВНОСТИ МЕТОДОВ ДИСКРЕТИЗАЦИИ ОБЛАСТИ ЗНАЧЕНИЙ ДВУМЕРНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПРИ СИНТЕЗЕ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ*

Проводится сравнение оптимального и эвристических методов дискретизации области значений двумерной случайной величины. Определяются условиях их компетенции при восстановлении нормального закона распределения двух независимых случайных величин. Эффективность методики подтверждается результатами вычислительных экспериментов.
Ключевые слова: плотность вероятности, гистограмма, регрессионная оценка плотности вероятности, двумерная случайная величина, методы дискретизации, критерий Пирсона, правило Старджесса, правило Хайнкольда – Гаеде.

DOI: 10.22250/isu.2016.49.78-85

Введение

Выбор количества интервалов дискретизации области значений случайной величины является одной из важных задач математической статистики. Результаты ее решения используются при оценивании плотности вероятности и проверке статистических гипотез о распределениях случайных величин с использованием критерия Пирсона. Начиная с работы Старджесса [1], предложен ряд известных методов дискретизации интервала значений одномерной случайной величины [2, 3].

В работе [4] при анализе аппроксимационных свойств непараметрической оценки плотности вероятности ядерного типа, восстанавливаемой по статистическим данным объема n, получена процедура выбора оптимального количества интервалов дискретизации

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

* Работа выполнена в рамках базовой части государственного задания Министерства образования и науки РФ высшим учебным заведениям на 2014 – 2016 гг. (СибГАУ № Б121/14) и Программы СО РАН IV.35.1.

области Δ значений одномерной случайной величины x с плотностью вероятности $p(x)$. 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 коэффициент $\bar{k}_1$ определяется только видом плотности вероятности и не зависит от ее параметров.

Для многомерной случайной величины используются эвристические рекомендации для выбора количества интервалов дискретизации области их значений. В работе [5] предлагается количество многомерных интервалов дискретизации выбирать из условия, чтобы минимальное число наблюдений случайной величины, попадающих в один интервал, соответствовало десяти. Существуют другие мнения – это число должно быть равно 3,5 [6]. Впервые оптимальное количество интервалов дискретизации двумерной случайной величины обосновано в работе [7].

Сравним аппроксимационные свойства двумерной непараметрической оценки плотности вероятности при использовании эвристических рекомендаций и оптимальной формулы дискретизации области значений случайной величины.

Выбор оптимального количества интервалов дискретизации области значений двумерной случайной величины

Пусть дана выборка $V = \left(x^i,\, i = \overline{1,\, n}\right)$ из n независимых наблюдений двумерной случайной величины $x = (x_1, x_2)$ с неизвестной плотностью вероятности $p(x_1, x_2)$.

Для упрощения аналитических преобразований будем считать, что интервалы Δ_v изменения аргументов x_v равные, т.е. $\Delta_v = \Delta, v = 1, 2$. Разобьем область изменения каждого аргумента на $\overline{N}$ непересекающихся интервалов длиной 2β таких, что $\Delta = 2\beta\overline{N}$. В этих условиях по исходным данным V сформируем множества случайных величин $X^i, i = \overline{1,\, N}$, $N = \overline{N}^2$. В качестве характеристик X^i примем частоту $\overline{P}^i$ попадания случайной величины x в i-й двумерный интервал и его центр $z^i = \left(z_1^i, z_2^i\right)$. На основе полученной информации составим статистическую выборку $V_1 = \left(z^i,\, y^i = \overline{P}^i/(2\beta)^2,\, i = \overline{1,\, N}\right)$.

В качестве приближения по эмпирическим данным искомой плотности $p(x_1, x_2)$ примем статистику [8]

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

где ядерные функции $\Phi(u_v)$ удовлетворяют условиям [9, 10],

$$\Phi(u_v) = \Phi(-u_v), \quad 0 \le \Phi(u_v) < \infty, \quad \int\Phi(u_v)du_v = 1, \quad \int u_v^2\,\Phi(u_v)du_v = 1. \qquad (2)$$

Здесь и далее бесконечные пределы интегрирования опускаются.

Коэффициенты размытости $c = c(N)$ ядерных функций в статистике (1)

убывают с ростом количества N двумерных интервалов дискретизации области определения плотности вероятности. Их оптимальное значение, минимизирующее асимптотическое выражение среднеквадратической ошибки аппроксимации $p(x_1, x_2)$ ее оценкой $\bar{p}(x_1, x_2)$, определяется формулой [8]:

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

здесь $\|\Phi(u)\|^2 = \int \Phi^2(u)du$, $\|p(x_1, x_2)\|^2 = \int\int p^2(x_1, x_2)dx_1 dx_2$, $B = \int\int \left(\sum_{v=1}^{2} p_v^{(2)}(x_1, x_2) \right)^2 dx_1 dx_2$;

$p_v^{(2)}(x_1, x_2)$ – вторая производная плотности вероятности $p(x_1, x_2)$ по аргументу x_v.

Представим непараметрическую оценку плотности вероятности (1) в виде

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

где индикаторные функции

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

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

В отличие от (1) в ее модификации (4) в явном виде присутствуют параметры β, N процедуры дискретизации и объем n выборки исходных статистических данных. Поэтому в результате анализа статистики (4) может быть получено соответствующее ей асимптотическое выражение среднеквадратического отклонения $W_2(c, N)$ [7]

$$M\left(\bar{p}(x_1, x_2) - p(x_1, x_2)\right)^2 \sim \frac{4\beta^2}{c^2} p^2(x_1, x_2) \left(\|\Phi(u)\|^2 \right)^2 + \\ + \frac{N}{nc^2} p(x_1, x_2) \left(\|\Phi(u)\|^2 \right)^2 + \frac{c^4}{4} \left(\sum_{v=1}^{2} p_v^{(2)}(x_1, x_2) \right)^2, \qquad (5)$$

зависящее от коэффициента размытости ядерных функций оценки плотности (4) и количества интервалов N дискретизации области значений двумерной случайной величины. При доказательстве утверждения (5) использовалась технология преобразований В.А. Епанечникова [10], развитая в работах [11 – 13].

Подставляя в $W_2(c, N)$ оптимальное значение $\bar{c}$ (3), получим выражение

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

Минимизация $W_2(N)$ по параметру N позволяет получить аналитическую зависимость количества N интервалов дискретизации от объема n исходных статистических данных

$$N^* = \sqrt{\frac{3}{4}\Delta^2 \left\| p(x_1, x_2) \right\|^2 n} \ . \qquad (7)$$

Количество N_v^* интервалов дискретизации каждой компоненты x_v, $v = 1, 2$ соответствует значению $N_v^* = \sqrt{N^*}$. Нетрудно показать, что значения коэффициента $\bar{k}_2 = \left(3/4\,\Delta^2 \left\| p(x_1, x_2) \right\|^2\right)^{1/2}$ определяются видом плотности вероятности и не зависят от ее параметров. Например, для независимых случайных x_1, x_2 с равномерным, линейным и нормальным законами распределения коэффициент $\bar{k}_2$ принимает значения 0,866; 1,15; 1,46.

Сравнение эффективности методов дискретизации области значений двумерной случайной величины

Исследуем зависимость аппроксимационных свойств регрессионной оценки плотности вероятности (1) от методов декомпозиции области значений двумерной случайной величины $x = (x_1, x_2)$ и объема n исходных статистических данных $V = \left(x^i, i = \overline{1, n}\right)$.

Будем восстанавливать плотность вероятности двумерной случайной величины с нормальным законом распределения $p(x_1, x_2) = p(x_1)\,p(x_2)$, где

$$p(x_v) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_v^2}{2}\right), \ v = 1, 2 \, .$$

Для выбора количества интервалов дискретизации области значений двумерной случайной величины используется выражение (7) и рекомендации, предложенные в работах [5, 6].

При использовании формулы (7) значение коэффициента k_2=1,46.

Синтез регрессионной оценки плотности вероятности (1) осуществлялся на основе ядерных функций В.А. Епанечникова [10]:

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

В этих условиях составляющие выражения (6) принимают значения:

$$\left\| \Phi(u) \right\|^2 = \frac{3}{5\sqrt{5}}; \ \left\| p(x_1, x_2) \right\|^2 = \frac{1}{4\pi}; \ B = \frac{1}{2\pi} \, .$$

Тогда среднеквадратический критерий качества аппроксимации (6) плотности вероятности определяется в виде

$$W_2(n,\,N)=\frac{(3/2)^{5/3}}{25\pi}\left(\frac{9}{N^{4/6}}+\frac{2\pi N^{8/6}}{3n}\right).$$

Обозначим через m минимальное количество наблюдений случайной величины x, попадающих в двумерный интервал дискретизации области ее значений. Для определения количества интервалов дискретизации $N(m)$, соответствующих значению m, используем следующую методику $D(m)$.

Из условия

$$\left|n\left[\int_{\lambda(m)}^{3\sigma_1}\int_{\lambda(m)}^{3\sigma_2} p(x_1,\,x_2)\,dx_1\,dx_2\right]\right|=m \qquad (8)$$

определим значение $\lambda(m)$. Здесь символом $|\alpha|$ обозначена целая часть числа α, а $\sigma_1=\sigma_2=1$. При реализации условия (8) используется процесс вычислительного эксперимента.

Пусть $\overline{\lambda}(m)$ – значение $\lambda(m)$, удовлетворяющее условию (8). При заданных параметрах $p(x_1,\,x_2)$ интервал наблюдений каждой случайной величины x_1, x_2 с вероятностью 0,997 ограничен значением 6. Тогда длина интервала дискретизации x_v равна $(3-\overline{\lambda}(m))$, а их количество $N_v(m)=6/\left(3-\overline{\lambda}(m)\right)$, $v=1,\,2$.

Пусть $\overline{N}(m)$ и $\overline{\overline{N}}(m)=\overline{N}(m)+1$ – левая и правая целые границы количества интервалов дискретизации анализируемых случайных величин. Для решения проблемы выбора его целого значения $\widetilde{N}^*(m)$ воспользуемся правилом

$$\widetilde{N}^*(m)=\begin{cases}\overline{N}(m),\ \textit{если}\ W_2\left(\left(\overline{N}(m)\right)^2\right)\le W_2\left(\left(\overline{\overline{N}}(m)\right)^2,\right)\\ \overline{\overline{N}}(m),\ \textit{если}\ W_2\left(\left(\overline{\overline{N}}(m)\right)^2\right)< W_2\left(\left(\overline{N}(m)\right)^2\right).\end{cases}$$

В этом случае количество двумерных интервалов дискретизации области значений случайных величин x_v, $v=1,\,2$, соответствующих значению m, определяется в виде $N^*(m)=\left(\widetilde{N}^*(m)\right)^2$.

При фиксированном объеме n исходных статистических данных зависимость $W_2(n,\,N)$ от количества N элементов дискретизации интервала значений двумерной случайной величины x имеет экстремальный характер (рис. 1). Отсюда следует, что в конкретных условиях синтеза $\overline{p}(x_1,\,x_2)$ существует количество интервалов дискретизации, при котором $W_2(n,\,N)$ достигает минимального значения. Так, при n=100, 200, 300, 500, 1000 значения $\overline{N}$ равняются 16, 25, 36, 36, 49.

Причем с ростом n значения N^* увеличиваются и сопровождаются снижением среднеквадратического отклонения $W_2\left(n,\,N^*\right)$, что согласуется с результатами аналитических исследований [8]. Увеличение $W_2(n,\,N)$ при $N>N^*$ объясняется уменьшением статистических данных, используемых при оценивании ве-

роятностей принадлежности случайных величин принятым интервалам дискретизации. В условиях $N < N^*$ достоверность оценивания вероятности принадлежности случайных величин интервалам дискретизации увеличивается, но объем N данных при синтезе оценки плотности вероятности (1) уменьшается. Поэтому ее аппроксимационные свойства снижаются. При больших значениях n наблюдается достаточно широкий диапазон изменения N, при котором $W_2(n, N)$ изменяется незначительно.

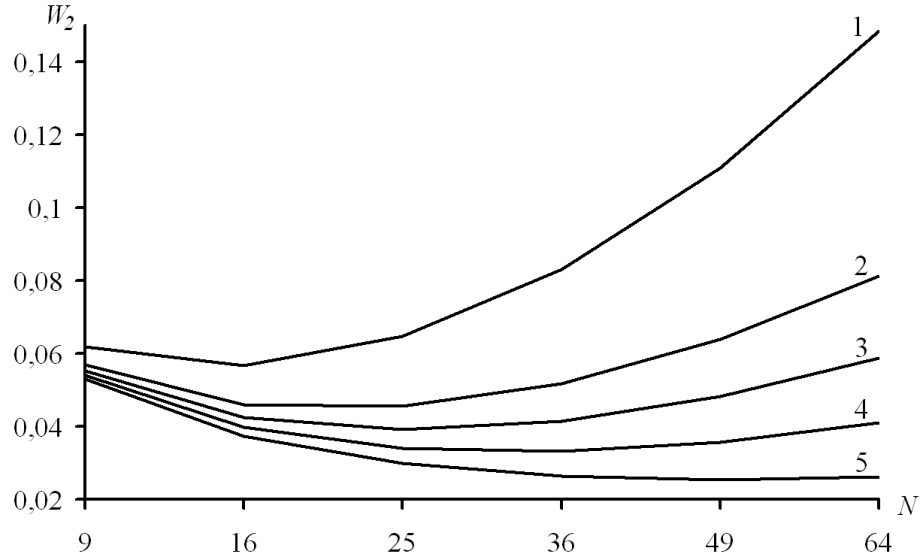

Рис. 1. Зависимость среднеквадратического отклонения $W_2 = W_2(n, N)$ от количества N интервалов дискретизации области значений двумерной случайной величины с нормальным законом распределения $p(x_1, x_2) = p(x_1)\,p(x_2)$. Кривые 1, 2, 3, 4, 5 соответствуют значениям n=100, 200, 300, 500, 1100.

Значения среднеквадратического отклонения $W_2\left(n, N^*(m)\right)$ регрессионной оценки плотности вероятности (1) от нормального закона распределения при исследовании сравниваемых методов дискретизации двумерной случайной величины приведены в табл. 1.

Таблица 1

n	Методы дискретизации				
	Формула (1)	$D(m = 1)$	$D(m = 3)$	$D(m = 5)$	$D(m = 10)$
100	0,056605	0,056605	0,061869	0,061869	0,061869
300	0,041427	0,039116	0,042517	0,042517	0,055327
500	0,03312	0,034007	0,039699	0,039699	0,039699
700	0,030247	0,031817	0,038491	0,038491	0,038491
900	0,027263	0,027582	0,030601	0,037821	0,037821
1100	0,025364	0,026323	0,029827	0,037394	0,037394
1300	0,024399	0,025451	0,029291	0,029291	0,037098
1500	0,023023	0,024812	0,028898	0,028898	0,036881

Количество интервалов дискретизации области значений двумерной случайной величины с нормальным законом распределения $p(x_1, x_2) = p(x_1)\,p(x_2)$, соответствующих конкретному объему n исходных данных и сравниваемым методам дискретизации (см. табл. 2).

Установлено, что применение формулы (7) при выборе количества двумерных интервалов дискретизации является более предпочтительным по сравнению с другими. Это следует из сравнения информации табл. 1, 2. Данный вывод ожидаем, так как формула дискретизации (7) является оптимальной в смысле минимума асимптотического выражения среднеквадратического отклонения регрессионной оценки плотности вероятности (1).

Таблица 2

n	Методы дискретизации				
	Формула (1)	$D(m = 1)$	$D(m = 3)$	$D(m = 5)$	$D(m = 10)$
100	16	16	9	9	9
300	36	25	16	16	9
500	36	25	16	16	16
700	49	25	16	16	16
900	49	36	25	16	16
1100	49	36	25	16	16
1300	64	36	25	25	16
1500	64	36	25	25	16

Значения критериев эффективности $W_2(n, N(m))$ методов дискретизации $D(m)$ при m=3; 5; 10 в условиях малых выборок ($n = 100$) одинаковы. Им соответствуют значения $N^*(m = 3) = N^*(m = 5) = N^*(m = 10) = 9$ (табл. 2). При этом соотношения между длинами $\Delta(m)$ интервалов дискретизации являются ожидаемыми

$$(\Delta(m = 3) = 2{,}05) < (\Delta(m = 5) = 2{,}24) < (\Delta(m = 10) = 2{,}53).$$

Следуя приведенной методике расчета количества интервалов дискретизации и выбора их целого значения, получим приведенный выше результат. Если $n \in (300 - 700)$ и $n \in (1300 - 1500)$, то эффективность применения методов дискретизации в условиях m=3; 5 сопоставимы. Для этих условий $N^*(m = 3) = N^*(m = 5) = 16$. При $n \in (900 - 1100)$ применение метода дискретизации $D(m = 3)$ более предпочтительно по сравнению с $D(m = 5)$.

Эффективность метода дискретизации $D(m = 5)$ имеет преимущество над $D(m = 10)$ при $n \in (1300 - 1500)$, их результативность сопоставима в условиях $n \in (500 - 1100)$.

Полученные результаты являются общими и не зависят от параметров нормального закона распределения двумерной случайной величины.

Заключение

Аппроксимационные свойства регрессионной оценки плотности вероятности двумерной случайной величины зависят от объема исходной информации и особенностей процедуры дискретизации области значений случайной величины. Зависимость среднеквадратического отклонения регрессионной оценки плотности вероятности от количества интервалов дискретизации имеет экстремальный характер, что обосновывает возможность нахождения его оптимального значения.

Полученные выводы согласуются с результатами исследований асимптотических свойств регрессионной оценки плотности вероятности и создают количественную основу сравнения эффективности методов дискретизации области ее определения на двумерные интервалы.

При восстановлении плотности вероятности двух независимых случайных величин с нормальными законами распределения целесообразно использовать формулу дискретизации (7) и методы, основанные на значениях $m = 1; 3$. Методы дискретизации, использующие значения $m = 5; 10$, являются неэффективными особенно при больших объемах исходных статистических данных. Эти выводы не зависят от параметров нормального закона распределения.

Полученные результаты имеют важное значение при восстановлении плотности вероятности, их доверительного оценивания и проверки гипотез о распределениях случайных величин с использованием критерия Пирсона.

ЛИТЕРАТУРА

1. *Sturges H.A.* The choice of a class interval // J. American Statistical Association. – 1926. – Vol. 21. – P.65-66.
2. *Шторм Р.* Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970.
3. *Heinhold I., Gaede K.* Ingeniur statistic. – München: Wien: Springler Verlag, 1964.
4. *Lapko A.V., Lapko* V.A. Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density // Measurement Techniques. – 2013. – Vol. 56, No. 7. – C.763-767.
5. *Пугачев В.С.* Теория вероятностей и математической статистики. – М.: Наука, 1979.
6. *Cochran W. G.* Some methods of strengthening the common $\chi 2$ tests // Biometrics. – 1954. – Vol. 10. – P.417-451.
7. *Лапко А.В., Лапко В.А.* Выбор оптимального количества интервалов дискретизации области значений двухмерной случайной величины // Измерительная техника. – 2016. – №2. – С.3-8.
8. *Лапко А.В., Лапко В.А.* Регрессионная оценка многомерной плотности вероятности и ее свойства // Автометрия. – 2014. – Т. 50, №2. – С.50-56.
9. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. – 1962. – Vol. 33, N 3. – P.1065-1076.
10. *Епанечников В.А.* Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. – 1969. – Т. 14, №1. – С.156-161.
11. *Лапко А.В., Лапко В.А.* Анализ свойств непараметрических оценок смеси плотностей вероятности при различных условиях распределения статистических данных // Информатика и системы управления. – 2013. – №1 (35). – С.119-126.
12. *Лапко А. В., Лапко В.А.* Свойства непараметрической оценки плотности вероятности многомерных случайных величин в условиях больших выборок // Информатика и системы управления. – 2012. – №2 (32). – С.121-126.
13. *Лапко А.В., Лапко В.А.* Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. – 2011. –№1 (43). – С.12-15.

E-mail:
Лапко Александр Васильевич – lapko@icm.krasn.ru;
Лапко Василий Александрович – lapko@icm.krasn.ru.