

Сравнение эффективности методов дискретизации области значений зависимых случайных величин при синтезе непараметрической оценки двумерной плотности вероятности

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

Проведено сравнение оптимального и эвристического методов дискретизации области значений двумерной случайной величины. Определены условия компетенции методов при восстановлении нормального закона распределения двух зависимых случайных величин. Рассмотрены результаты вычислительного эксперимента.

Ключевые слова: плотность вероятности, гистограмма, непараметрическая оценка, двумерная случайная величина, методы дискретизации, критерий Пирсона, правило Старджесса, правило Хайнкольда—Гаеде.

We compare the best and heuristics sampling area of two-dimensional random variable. The conditions of their competence in restoring the normal law of distribution of two independent random variables. The results of computational experiment.

Key words: probability density, histogram, nonparametric estimation, two-dimensional random variable, discretization methods, Pearson criterion, Sturges rule, Heinhold—Gaede rule.

Задача дискретизации области значений случайной величины актуальна при оценивании плотности вероятности, построении её доверительных интервалов и проверки гипотез об их распределениях. Методы дискретизации интервала значений одномерной случайной величины рассмотрены в [1—3]. Для многомерных случайных величин предложены эвристические рекомендации, основанные на пороговом значении минимального количества наблюдений, попадающих в один интервал [4]. При обработке изображений в компьютерных системах зрения используется методика построения двумерной гистограммы независимых случайных величин. Дискретизация области определения их плотности вероятности основана на применении формул Старджесса и Скотта для каждой переменной [5, 6].

В работе [7] при анализе аппроксимационных свойств непараметрической оценки плотности вероятности ядерного типа, восстанавливаемой по статистическим данным объёмом n , получена процедура выбора оптимального количества интервалов дискретизации области Δ значений одномерной случайной величины x с плотностью вероятности $p(x)$:

$$N_1^* = \sqrt{\Delta \int p^2(x) dx} n = k_1 \sqrt{n}. \quad (1)$$

Здесь и далее бесконечные пределы интегрирования опущены.

Аналитическая зависимость (1) по виду близка к формуле Гаеде и совпадает с ней при оценивании плотности вероятности случайной величины с равномерным законом распределения ($k_1 = 1$). Установлено, что коэффициент k_1 определяется только видом плотности вероятности и не зависит от значений её параметров. Полученные результаты обоб-

щены при построении непараметрической оценки двумерной плотности вероятности [8]. Пусть имеется выборка $V = (x_1^i, x_2^i, i = \overline{1, n})$ из n независимых наблюдений двумерной случайной величины с неизвестной плотностью вероятности $p(x_1, x_2)$. В качестве приближения по эмпирическим данным искомой плотности $p(x_1, x_2)$ примем статистику [8]:

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

Ядерные функции $\Phi(u_v)$ удовлетворяют следующим условиям:

$$\Phi(u_v) = \Phi(-u_v); 0 \leq \Phi(u_v) < \infty;$$

$$\int \Phi(u_v) du_v = 1; \int u_v^2 \Phi(u_v) du_v = 1,$$

где $z^i = (z_1^i, z_2^i)$ — центры двумерных интервалов дискретизации области определения плотности вероятности $p(x_1, x_2)$; $\bar{P}^i$ — частота попадания в эти интервалы случайных величин.

Коэффициенты размытости $c = c(N)$ ядерных функций в статистике (2) убывают с ростом количества N двумерных интервалов дискретизации области определения плотности вероятности.

В работе [8] получена оценка среднего квадратического отклонения (СКО) $\bar{p}(x_1, x_2)$ от $p(x_1, x_2)$ при оптимальных значениях коэффициентов размытости с ядерных функций:

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

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

$$B = \int \int \left[\sum_{v=1}^2 p_v^{(2)}(x_1, x_2) \right]^2 dx_1 dx_2.$$

Из условия минимума (3) определено оптимальное количество двумерных интервалов дискретизации

$$N^* = \left[\frac{3}{4} \Delta^2 \|p(x_1, x_2)\|^2 n \right]^{1/2} = k_2 \sqrt{n}, \quad (4)$$

где Δ — длина интервала дискретизации случайных величин x_1, x_2 в предположении $\Delta_1 = \Delta_2 = \Delta$.

Формула дискретизации (4) справедлива для зависимых и независимых случайных величин. Показано, что коэффициент k_2 для независимых случайных величин x_1, x_2 определяется только видом плотности вероятности и не зависит от значений её параметров. Например, для равномерного, линейного и нормального законов распределения коэффициент k_2 принимает значения 0,866; 1,15; 1,466. Эта закономерность сохраняется при разных значениях Δ_1, Δ_2 .

Цель настоящей работы — сравнение аппроксимационных свойств непараметрической оценки плотности вероятности при использовании оптимальной формулы дискретизации области значений зависимых двумерных случайных величин и существующих эвристических рекомендаций.

Анализ методов дискретизации области значений зависимых случайных величин. Исследуем зависимость аппроксимационных свойств непараметрической оценки плотности вероятности (2) от методов дискретизации области значений двумерной случайной величины $x = (x_1, x_2)$. Плотность вероятности двумерной случайной величины с нормальным законом распределения восстанавливается следующим образом:

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x_1-m_1)^2}{\sigma_1^2} - r \frac{2(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2} \right] \right\},$$

где r — коэффициент корреляции x_1, x_2 .

Будем считать, что значения математических ожиданий случайных величин $m_1 = m_2 = 0$, а СКО анализируемых слу-

чайных величин $\sigma_1 = \sigma_2 = 1$. Для выбора количества интервалов дискретизации области значений случайных величин x_1, x_2 , используем формулу (4) и рекомендации [4]. Синтез непараметрической оценки плотности вероятности (2) осуществляется на основе оптимальных ядерных функций [9]:

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

В принятых условиях вычислительного эксперимента составляющие выражения (3) принимают значения:

$\|\Phi(u)\|^2 = 3/(5\sqrt{5})$; $\|p(x_1, x_2)\|^2 = 0,083$; $B = 0,211$ при $r = 0,3$ и $0,133$; $2,702$ при $r = 0,8$, соответственно.

Обозначим m — минимальное количество наблюдений случайных величин x_1, x_2 , принадлежащих двумерному интервалу дискретизации области их значений. Для определения количества интервалов дискретизации $N(m)$, соответствующих значению m , используем следующую методику $D(m)$.

$$\text{Определим } \bar{\lambda}(m) \text{ из условия } \left\lfloor n \int_{\lambda(m)}^{3\sigma_1} \int_{\lambda(m)}^{3\sigma_2} p(x_1, x_2) dx_1 dx_2 \right\rfloor = m,$$

где $|\alpha|$ — целая часть числа; $\sigma_1 = \sigma_2 = 1$.

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 случайной величины x_v соответствует значению $(3 - \bar{\lambda}(m))$,

а их количество $N_v(m) = 6/(3 - \bar{\lambda}(m))$, $v = 1, 2$. Пусть

$\bar{N}(m) = \lfloor N_v(m) \rfloor$, $\bar{\bar{N}}(m) = \bar{N}(m) + 1$ — левая и правая целые границы количества интервалов дискретизации случайных величин x_1, x_2 . Для выбора их целого значения $\tilde{N}^*(m)$ воспользуемся правилом

$$\tilde{N}^*(m) = \begin{cases} \bar{N}(m), & \text{если } W_2\left(\left(\bar{N}(m)\right)^2\right) \leq W_2\left(\left(\bar{\bar{N}}(m)\right)^2\right); \\ \bar{\bar{N}}(m), & \text{если } W_2\left(\left(\bar{\bar{N}}(m)\right)^2\right) < W_2\left(\left(\bar{N}(m)\right)^2\right). \end{cases}$$

Количество двумерных интервалов дискретизации области значений случайных величин x_v , соответствующих значению m , определяется в виде $N^*(m) = \left(\tilde{N}^*(m)\right)^2$. При конкрет-

ных объёмах n статистических данных и коэффициентах корреляции r зависимость $W_2(N, n, r)$ от количества N интервалов дискретизации двумерной случайной величины имеет экстремальный характер (рисунок). Например, при $r = 0,3$ и $n = 300$; 700 ; 1100 оптимальные значения N^* принимают значения 25, 36, 49 (табл. 1). Минимальное значение $W_2(N, n, r)$ определяет оптимальное количество интервалов дискретизации $N^*(n, r)$.

Таблица 1

Зависимость СКО $W_2(N, n, r)$ от количества интервалов дискретизации N области определения двумерной плотности вероятности $p(x_1, x_2)$ с нормальным законом распределения, объема n исходных данных и коэффициента корреляции r между x_1, x_2

N	r = 0,3			r = 0,8		
	n = 300	n = 700	n = 1100	n = 300	n = 700	n = 1100
9	0,0624	0,0603	0,0597	0,1955	0,1915	0,1904
16	0,0477	0,0433	0,0422	0,1437	0,1349	0,1326
25	0,0436	0,0357	0,0335	0,1230	0,1072	0,1029
36	0,0458	0,0329	0,0295	0,1198	0,0941	0,0870
49	0,0529	0,0335	0,0283	0,1287	0,0899	0,0794
64	0,0644	0,0367	0,0291	0,1478	0,0925	0,0774
81	0,0799	0,0420	0,0317	0,1762	0,1004	0,0797
100	0,0997	0,0495	0,0358	0,2135	0,1131	0,0857
121	0,1237	0,0589	0,0413	0,2598	0,1304	0,0951
144	0,1521	0,0705	0,0482	0,3153	0,1521	0,1076
169	0,1852	0,0841	0,0566	0,3802	0,1782	0,1231
196	0,2230	0,0999	0,0663	0,4549	0,2088	0,1417

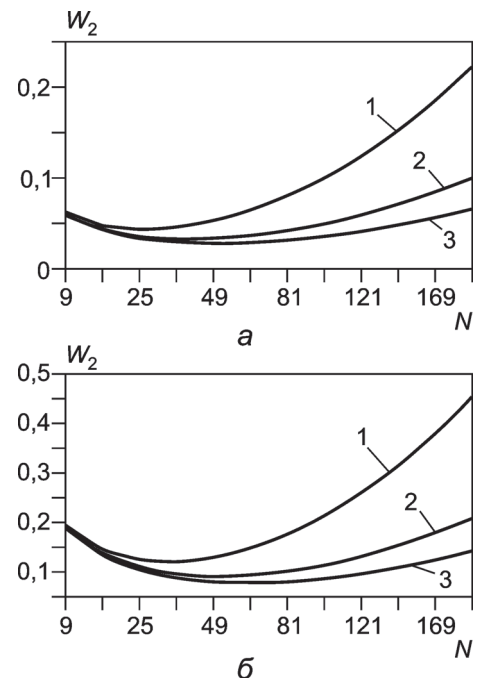
С ростом n значения $N^*(n, r)$ увеличиваются и сопровождаются снижением СКО $W_2(N^*, n, r)$ оценки плотности вероятности $\bar{p}(x_1, x_2)$ от $p(x_1, x_2)$, что согласуется с результатами аналитических исследований [8]. При конкретных значениях n, r увеличение $W_2(N, n, r)$ в условиях $N > N^*(n, r)$ объясняется уменьшением статистических данных, используемых при оценивании в статистике $\bar{p}(x_1, x_2)$ (2) вероятностей принадлежности P^i слу-

чайных величин (x_1, x_2) интервалам дискретизации $i = \overline{1, N}$. Если $N < N^*(n, r)$, то достоверность оценивания P^i повышается, но аппроксимационные свойства оценки плотности вероятности (2) снижаются с уменьшением объема промежуточных данных N в процессе её синтеза. При больших объемах n исходных статистических данных наблюдается достаточно широкий диапазон изменений N , когда значения $W_2(N, n, r)$ отклоняются незначительно (см. рисунок).

С ростом коэффициента корреляции количество оптимальных интервалов дискретизации $N^*(n, r)$ возрастает (см. рисунок, табл. 1). Данный факт объясняется уменьшением области определения плотности вероятности $p(x_1, x_2)$, что требует увеличения количества интервалов дискретизации значений каждой случайной величины x_1, x_2 . Отмеченная закономерность количественно подтверждается результатами анализа формулы оптимальной дискретизации (4). Зависимость коэффициента k_2 в принятых условиях исследования от коэффициента корреляции r остаётся постоянной для различных значений математических ожиданий и СКО случайных величин x_1, x_2 . С ростом r увеличивается значения k_2 и в соответствии с формулой (4) — оптимальное количество интервалов дискретизации $N^*(n, r)$. Зависимость k_2 от r можно представить аппроксимацией, которая характеризуется средней квадратической погрешностью 0,0012: $k_2(r) = 0,94 |r^3| + 1,466$.

Установлено, что применение формулы дискретизации (4) при выборе количества $N^*(n, r)$ двумерных интервалов дискретизации более предпочтительно, чем методов $D(m = 1)$, $D(m = 3)$, $D(m = 5)$, $D(m = 10)$ (табл. 2, 3). Данный вывод является ожидаемым, так как формула дискретизации (4) получена из условия минимума оценки СКО $\bar{p}(x_1, x_2)$ от $p(x_1, x_2)$.

При малых значениях $|r| \leq 0,5$ из сравниваемых методов дискретизации наиболее близок к оптимальному (4) метод $D(m = 1)$. Остальные методы в среднем менее эффективны, что следует из анализа табл. 1. Количество интервалов дискретизации, соответствующих этим методам, находятся левее оптимального значения $N^*(n, r)$ для зависимости $W_2(N, n, r)$ (см. рисунок и табл. 1). При $n \leq 300$ эффективность формулы дискретизации (4) сопоставима с методом $D(m = 3)$. С увеличением $|r| > 0,5$ оптимальная формула дискретизации по эффективности сопоставима с методами $D(m = 3)$ и $D(m = 5)$. Менее предпочтительно в этих условиях применять методы $D(m = 1)$, $D(m = 10)$ (см. табл. 3).



Зависимости среднего квадратического отклонения $W_2 = W_2(N, n, r)$ от количества интервалов дискретизации N области значений двумерной случайной величины с нормальным законом распределения $p(x_1, x_2)$ и коэффициентами корреляции $r = 0,3$ (а), $r = 0,8$ (б) между x_1, x_2 ; $n = 300; 700; 1100$ — кривые 1—3, соответственно

Таблица 2

Значения СКО $W_2(N^*, n, r)$ оценки плотности вероятности $\bar{p}(x_1, x_2)$, соответствующих объёму n статистических данных, коэффициенту корреляции $r = 0,3, r = 0,8$ между x_1, x_2 и сравниваемым методом дискретизации

n	$r = 0,3$					$r = 0,8$				
	Формула (4)	$D(m=1)$	$D(m=3)$	$D(m=5)$	$D(m=10)$	Формула (4)	$D(m=1)$	$D(m=3)$	$D(m=5)$	$D(m=10)$
100	0,0630	0,0713	0,0630	0,0690	0,0690	0,1740	0,2100	0,1780	0,1740	0,2100
300	0,0436	0,0458	0,0436	0,0477	0,0477	0,1200	0,1478	0,1200	0,1230	0,1440
500	0,0368	0,0368	0,0380	0,0380	0,0446	0,1016	0,1432	0,1016	0,1018	0,1120
700	0,0330	0,0335	0,0357	0,0357	0,0433	0,0900	0,1300	0,0920	0,0900	0,0940
900	0,0303	0,0303	0,0308	0,0344	0,0426	0,0832	0,1250	0,0832	0,0835	0,0900
1100	0,0283	0,0283	0,0295	0,0335	0,0335	0,0770	0,1230	0,0797	0,0770	0,0870
1300	0,0268	0,0270	0,0285	0,0330	0,0330	0,0730	0,1230	0,0740	0,0730	0,0765
1500	0,0255	0,0255	0,0278	0,0278	0,0325	0,0700	0,1100	0,0730	0,0700	0,0740

Таблица 3

Количество интервалов дискретизации области значений двумерной случайной величины с нормальным законом распределения, соответствующих объёму n статистических данных, коэффициенту корреляции $r = 0,3, r = 0,8$ между x_1, x_2 , и сравниваемым методом дискретизации

n	$r = 0,3$					$r = 0,8$				
	Формула (4)	$D(m=1)$	$D(m=3)$	$D(m=5)$	$D(m=10)$	Формула (4)	$D(m=1)$	$D(m=3)$	$D(m=5)$	$D(m=10)$
100	16	25	16	9	9	16	36	25	16	9
300	25	36	25	16	16	36	64	36	25	16
500	36	36	25	25	16	49	100	49	36	25
700	36	49	25	25	16	49	121	64	49	36
900	49	49	36	25	16	64	144	64	49	36
1100	49	49	36	25	25	64	169	81	64	36
1300	49	64	36	25	25	64	196	81	64	49
1500	64	64	36	36	25	81	196	100	64	49

Заключение. Зависимость СКО непараметрической оценки двумерной плотности вероятности от количества интервалов дискретизации имеет экстремальный характер, что обосновывает возможность нахождения его оптимального значения. Для нормального закона распределения зависимых случайных величин количество оптимальных интервалов дискретизации области их значений зависит от коэффициента корреляции, объёма исходных данных и метода дискретизации. С ростом абсолютного значения коэффициента корреляции количество оптимальных интервалов дискретизации увеличивается. При малых значениях $|r| < 0,5$ эффективность оптимальной формулы дискретизации сопоставима с методом $D(m=1)$. В условиях $|r| > 0,5$ целесообразно использовать методы дискретизации $D(m=3)$ и $D(m=5)$. Отмеченная закономерность не зависит от значений математических ожиданий и СКО двумерных случайных величин.

Предлагаемую методику можно использовать при сравнении эффективности методов дискретизации области значений двумерной случайной величины с другими законами распределения. Полученные результаты имеют важное значение при восстановлении плотности вероятности, их доверительного оценивания и проверки гипотез о распределении случайных величин с использованием критерия Пирсона.

Работа выполнена в рамках проектной части государственного задания Минобрнауки России № 2.914.2014/К и Программы СО РАН IV.35.1.

Л и т е р а т у р а

1. **Sturges H. A.** The choice of a class interval // J. American Statistical Association. 1926. V. 21. P. 65—66.

2. **Scott D. W.** Multivariate density estimation: theory, practice and visualization. N. Y.: Wiley, 1992.

3. **Hacine-Gharbi A., Ravier P., Harba R., Mohamadi T.** Low bias histogram-based estimation of mutual information for feature selection // Pattern Recognition Letters. 2012. V. 33. N. 10. P. 1302—1308.

4. **Пугачёв В. С.** Теория вероятностей и математическая статистика. Учеб. Пособие. М.: ФИЗМАТЛИТ, 2002.

5. **Minnotte M. C., Sain S. R., Scott D. W.** Multivariate Visualization by Density Estimation // Handbook of Data Visualization. 2008. Part 3. P. 389—413.

6. **Legg P. A., Rosin P. L., Marshall D., Morgan J. E.** Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule: 11th Annual Conf. in Medical Image Understanding and Analysis. 2007. P. 26—30.

7. **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.

8. **Лапко А. В., Лапко В. А.** Выбор оптимального количества интервалов дискретизации области значений двухмерной случайной величины // Измерительная техника. 2016. № 2. С. 14—17.

9. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. Т. 14. № 1. С. 156—161.

Дата принятия: 22.06.2016.