УДК 519.7

БЫСТРЫЙ ВЫБОР КОЭФФИЦИЕНТОВ РАЗМЫТОСТИ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ ДВУХМЕРНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ С ЗАВИСИМЫМИ КОМПОНЕНТАМИ

© **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

Предложена методика быстрого выбора коэффициентов размытости ядерных функций в непараметрической оценке двухмерной случайной величины с зависимыми компонентами. Методика основана на результатах анализа асимптотических свойств ядерной оценки плотности вероятности Розенблатта — Парзена. Исследованы свойства быстрого алгоритма выбора коэффициентов размытости в рассматриваемой непараметрической оценке плотности вероятности.

Ключевые слова: непараметрическая оценка плотности вероятности двухмерной случайной величины, зависимые случайные величины, ядерная оценка плотности вероятности, быстрый выбор коэффициентов размытости.

DOI: 10.15372/AUT20230204

Введение. Непараметрические оценки плотности вероятности типа Розенблатта — Парзена $\bar{p}(x)$ используются при разработке алгоритмов принятия решений в условиях априорной неопределённости [1–3]. Аппроксимационные свойства $\bar{p}(x)$ определяются выбором оптимальных значений коэффициентов размытости ядерных функций. Традиционный подход к оптимизации непараметрической оценки плотности вероятности состоит в минимизации статистической оценки её среднего квадратического отклонения. Вычислительная эффективность этого подхода снижается при увеличении объёма исходных статистических данных [4–8]. Для решения этой проблемы предложены методики быстрого выбора коэффициентов размытости ядерных функций, основанные на анализе результатов исследования асимптотических свойств непараметрической оценки плотности вероятности для ряда семейств законов распределения случайных величин [9–14].

В работе [15] обоснована возможность выбора коэффициентов размытости ядерных функций в непараметрической регрессии из условия минимума оценок средних квадратических ошибок аппроксимации плотности вероятности зависимых случайных величин. Созданы условия значительного повышения вычислительной эффективности непараметрической регрессии на основе быстрых процедур выбора коэффициентов размытости в ядерной оценке плотности вероятности зависимых случайных величин.

Цель данной работы — исследование методики быстрого выбора коэффициентов размытости ядерных функций непараметрической оценки плотности вероятности двухмерной случайной величины с зависимыми компонентами.

Методика быстрого выбора коэффициентов размытости непараметрической оценки плотности вероятности зависимых случайных величин. Пусть имеется выборка $V = (x^i, y^i,\ i = \overline{1,n})$ статистически независимых наблюдений случайных величин x, y, распределённых с неизвестной плотностью вероятности $p(x, y)$.

При оценивании $p(x,y)$ используем непараметрическую статистику $\bar{p}(x,y)$ типа Розенблатта — Парзена [16, 17]:

$$\bar{p}(x,y)=\frac{1}{nc_1c_2}\sum_{i=1}^{n}\Phi\Big(\frac{x-x^i}{c_1}\Big)\Phi\Big(\frac{y-y^i}{c_2}\Big). \tag{1}$$

Ядерные функции $\Phi(u)$ в непараметрической оценке плотности вероятности (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 c_1, c_2 в непараметрической оценке плотности вероятности (1) зависят от интервала изменений случайных величин x и y. Поэтому примем $c_1=c\sigma_1$, $c_2=c\sigma_2$, где σ_1, σ_2 — средние квадратические отклонения случайных величин x, y соответственно, а c — неопределённый параметр.

Тогда асимптотическое выражение среднего квадратического отклонения $\bar{p}(x,y)$ от $p(x,y)$

$$W(c)=M\iint(\bar{p}(x,y)-p(x,y))^2\,dx\,dy$$

с учётом результатов работы [16] запишем в виде

$$\bar{W}(c)=\frac{1}{nc^2\sigma_1\sigma_2}\,\|\Phi(u_1)\|^2\,\|\Phi(u_2)\|^2+\frac{c^4}{4}\,B. \tag{2}$$

Здесь приняты следующие обозначения:

$$\|\Phi(u_1)\|^2=\int\Phi^2(u_1)\,du_1;\qquad \|\Phi(u_2)\|^2=\int\Phi^2(u_2)\,du_2;$$

$$B=\iint(\sigma_1^2p_1^{(2)}(x,y)+\sigma_2^2p_2^{(2)}(x,y))^2\,dx\,dy; \tag{3}$$

$p_1^{(2)}(x,y)$, $p_2^{(2)}(x,y)$ — вторые производные функции $p(x,y)$ по переменным x,y соответственно; M — знак математического ожидания.

Оптимальное значение параметра c коэффициентов размытости c_1, c_2 ядерных функций непараметрической оценки плотности вероятности (1)

$$c^*=(2(\|\Phi(u)\|^2)^2/(nB\sigma_1\sigma_2))^{1/6} \tag{4}$$

определим из условия минимума критерия (2) для ядерных функций $\Phi(u_v)=\Phi(u)$, $v=1,2$.

С учётом ранее принятых предположений оптимальные коэффициенты размытости c_1, c_2 ядерных функций в статистике (1) запишем как

$$c_v^*=c^*\sigma_v=\beta\sigma_vn^{-1/6},\qquad v=\overline{1,2}\,, \tag{5}$$

где

$$\beta = (2(\|\Phi(u)\|^2)^2/(B\sigma_1\sigma_2))^{1/6}. \tag{6}$$

Для вычисления оптимального параметра c^* коэффициентов размытости ядерных функций (4) необходимо определить оценку нелинейного функционала $\lambda = B\sigma_1\sigma_2$ от вторых производных плотности вероятности $p(x, y)$ по переменным x, y.

Используем семейство нормальных законов распределения зависимых случайных величин x и y:

$$p(x, y) = (2\pi\sigma_1\sigma_2\sqrt{1 - r^2})^{-1} \times$$

$$\times \exp\Big\{ - \frac{1}{2(1 - r^2)} \Big[\Big(\frac{x - m_1}{\sigma_1} \Big)^2 - 2r \, \frac{(x - m_1)(y - m_2)}{\sigma_1\sigma_2} + \Big(\frac{y - m_2}{\sigma_2} \Big)^2 \Big] \Big\}, \tag{7}$$

где (m_1, σ_1) и (m_2, σ_2) — математические ожидания и средние квадратические отклонения случайных величин x и y соответственно; неопределённые вещественные числа r подчиняются условию $-1 \leqslant r \leqslant 1$.

Нормальные распределения типа (7) часто встречаются в различных прикладных задачах. Значимость нормального закона распределения объясняется тем, что анализируемая случайная величина, являющаяся суммой большого числа независимых помех, имеет закон распределения, близкий к нормальному [18].

На этой основе методика быстрого выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности (1) состоит в выполнении следующих действий:

1. По исходной выборке $V = (x^i, y^i,\ i = \overline{1, n})$, которая имеется при восстановлении $p(x, y)$, оценить средние квадратические отклонения σ_1, σ_2 случайных величин x, y и параметра r, например, как коэффициента корреляции между ними. Обозначим через $\bar\sigma_1$, $\bar\sigma_2$ и $\bar r$ статистические оценки параметров σ_1, σ_2 и r.

2. При конкретном объёме n выборки V и выбранном виде ядерной функции $\Phi(u)$ в соответствии с формулами (4) и (7) определить оценку параметра

$$\bar c^* = (2(\|\Phi(u)\|^2)^2/(n\bar B\bar\sigma_1\bar\sigma_2))^{1/6}$$

коэффициентов размытости ядерных функций c_1, c_2 в статистике $\bar p(x, y)$ (1). Здесь значение $\bar B$ вычислим в соответствии с выражением (3) при $\sigma_1 = \bar\sigma_1$, $\sigma_2 = \bar\sigma_2$, $r = \bar r$. В качестве $\bar r$ используем статистическую оценку коэффициента корреляции.

3. Определить оценки коэффициентов размытости c_1, c_2 в ядерных функциях в непараметрической статистике (1):

$$\bar c_1^* = \bar c^*\bar\sigma_1, \qquad \bar c_2^* = \bar c^*\bar\sigma_2.$$

Анализ результатов вычислительных экспериментов. Исследуем зависимость оптимальных коэффициентов размытости c_1^*, c_2^* непараметрической оценки плотности вероятности $\bar p(x, y)$ (1) от составляющих их параметров c^*, β и λ. Проведём анализ зависимости асимптотического выражения среднего квадратического отклонения $\bar W(c^*)$ от значений параметра r плотности вероятности $p(x, y)$ (7). При организации вычислительных экспериментов положим, что для зависимых случайных величин x и y значения их средних квадратических отклонений $\sigma_1 = 1$ и $\sigma_2 = (1, 2, 3)$.

Таблица 1

Результаты анализа аппроксимационных свойств непараметрической оценки плотности вероятности $\bar{p}(x, y)$ зависимых случайных величин x, y

r	σ_2	λ	β	c^*	c_1^*	c_2^*	$\bar{W}(c^*)$
0,175	1	0,175	0,968	0,449 0,344	0,449 0,344	0,449 0,344	0,005346 0,001828
	2			0,505 0,386		0,899 0,687	0,002829 0,0009674
	3			0,54 0,413		1,348 1,031	0,002059 0,0007043
0,225	1	0,186	0,958	0,445 0,34	0,445 0,34	0,445 0,34	0,005458 0,001866
	2			0,499 0,382		0,89 0,68	0,002888 0,0009876
	3			0,534 0,409		1,335 1,021	0,002102 0,0007189
0,25	1	0,193	0,952	0,442 0,338	0,442 0,338	0,442 0,338	0,005526 0,00189
	2			0,496 0,379		0,884 0,676	0,002924 0,001
	3			0,531 0,406		1,326 1,014	0,002129 0,000728
0,275	1	0,201	0,946	0,439 0,336	0,439 0,336	0,439 0,336	0,005603 0,001916
	2			0,493 0,377		0,878 0,671	0,002965 0,001014
	3			0,527 0,403		1,317 1,007	0,002158 0,0007381
0,325	1	0,221	0,931	0,432 0,33	0,432 0,33	0,432 0,33	0,005787 0,001979
	2			0,485 0,371		0,864 0,661	0,003062 0,001047
	3			0,519 0,397		1,296 0,991	0,002229 0,0007623

Продолжение таблицы 1

r	σ_2	λ	β	c^*	c_1^*	c_2^*	$\bar{W}(c^*)$
0,525	1	0,405	0,842	0,391 0,299	0,391 0,299	0,391 0,299	0,007079 0,002421
	2			0,438 0,335		0,781 0,597	0,003746 0,001281
	3			0,469 0,359		1,172 0,896	0,002727 0,0009325
0,675	1	0,894	0,738	0,342 0,262	0,342 0,262	0,342 0,262	0,009212 0,003151
	2			0,384 0,294		0,685 0,524	0,004875 0,001667
	3			0,411 0,314		1,027 0,786	0,003549 0,001214
0,75	1	1,611	0,669	0,31 0,237	0,31 0,237	0,31 0,237	0,011 0,003834
	2			0,348 0,266		0,621 0,475	0,005932 0,002029
	3			0,373 0,285		0,931 0,712	0,004318 0,001477
0,825	1	3,701	0,582	0,27 0,207	0,27 0,207	0,27 0,207	0,015 0,005059
	2			0,303 0,232		0,54 0,413	0,007828 0,002677
	3			0,324 0,248		0,811 0,62	0,005698 0,001949
0,975	1	433,4	0,263	0,122 0,093	0,122 0,093	0,122 0,093	0,072 0,025
	2			0,137 0,105		0,244 0,187	0,038 0,013
	3			0,147 0,112		0,366 0,28	0,028 0,009535

Определим значения параметра $\bar{r} = r \pm \alpha r$ при известных $r = 0{,}25$ и $r = 0{,}75$. Значения $\alpha = 0{,}1$, $\alpha = 0{,}3$ характеризуют погрешности оценивания r. Например, при $r = 0{,}25$ и $\alpha = 0{,}1$ значения $\bar{r} = 0{,}225$, $\bar{r} = 0{,}275$ соответственно.

Верхние строки элементов табл. 1 в столбцах c^*, c_1^*, c_2^*, $\bar{W}(c^*)$ соответствуют объёму статистических данных $n = 100$, а нижние — $n = 500$.

Значения функционала λ при конкретных значениях параметра r не зависят от объёма n статистических данных и изменения $\sigma_2 \in [1; 3]$. Например, при $r = 0{,}175$ значение $\lambda = 0{,}175$ для $n \in [100; 500]$, а при $r = 0{,}75$ в принятых условиях $\lambda = 1{,}611$. С ростом коэффициента корреляции r значения функционала λ увеличиваются от 0,175 при $r = 0{,}175$ до 1,611 при $r = 0{,}75$. При $r = 0{,}975$ наблюдается значительное увеличение λ до 433,4, что может служить одним из критериев линейной зависимости между случайными величинами (см. табл. 1).

Зависимость функционала β (6) от значений параметров r и σ_2 является близкой к вышеотмеченной зависимости $\lambda = B\sigma_1\sigma_2$. Для $\sigma_2 \in [1; 3]$ и $n \in [100; 500]$ значения $\beta = 0{,}968$, $\beta = 0{,}669$ при $r = 0{,}175$, $r = 0{,}75$ соответственно. При зависимости между случайными величинами, близкой к линейной ($r = 0{,}975$), значение $\beta = 0{,}263$ для $\sigma_2 \in [1; 3]$ и $n \in [100; 500]$. Оптимальные значения параметра c^* коэффициентов размытости c_1^*, c_2^* (4) зависят от r, σ_2 и объёма статистических данных n. При фиксированных значениях r с ростом σ_2 значения параметра c^* увеличиваются. Например, при $r = 0{,}175$ и $n = 100$ значения $c^* = 0{,}449$, $c^* = 0{,}54$ в условиях $\sigma_2 = 1$, $\sigma_2 = 3$ соответственно. С увеличением r и n значения c^* уменьшаются, что особенно проявляется при $r = 0{,}975$ и $n = 500$. В этих условиях $c^* \in [0{,}093; 0{,}112]$ для $\sigma_2 \in [1; 3]$ (см. табл. 1).

Сравним аппроксимационные свойства непараметрической оценки плотности вероятности $\bar{p}(x, y)$ в зависимости от погрешности оценивания коэффициента корреляции r, используя результаты вычислительных экспериментов (см. табл. 1). Для этого вычислим отношение $\bar{W}(\bar{c}^*)/\bar{W}(c^*)$ для различных значений $\bar{r}$ и $r = 0{,}25$, $r = 0{,}75$ при $\alpha = 0{,}1$, $\alpha = 0{,}3$. Здесь значения $\bar{W}(\bar{c}^*)$, $\bar{W}(c^*)$ определим по формуле (2) при значениях $\bar{r}$ и r, где параметр r считается известным, а $\bar{r}$ — его предполагаемая оценка.

Вычислим аналитические выражения отношений $\bar{c}^*/c^*$ и $\bar{W}(\bar{c}^*)/\bar{W}(c^*)$ с учётом принятых условий исследования. Используя формулы (4), (2), имеем

$$\bar{c}^*/c^* = (B(r)/B(\bar{r}))^{1/6}, \tag{8}$$

$$\bar{W}(\bar{c}^*)/\bar{W}(c^*) = (B(\bar{r})/B(r))^{1/3}, \tag{9}$$

где $\bar{c}^*$ — значение c^* (4) при $r = \bar{r}$.

Результаты анализа выражений (8), (9) представлены в табл. 2.

При известном малом значении $r = 0{,}25$ и относительно больших погрешностей его оценивания $0{,}175 \leqslant \bar{r} \leqslant 0{,}325$ отношение $\bar{c}^*/c^* \in [1{,}017; 0{,}977]$, а соответствующие им значения $\bar{W}(\bar{c}^*)/\bar{W}(c^*) \in [0{,}967; 1{,}047]$. При малых r и больших значениях параметра погрешностей их оценивания $\alpha \in [0{,}1; 0{,}3]$ аппроксимационные свойства непараметрической оценки плотности вероятности $\bar{p}(x, y)$ зависимых случайных величин изменяются незначительно. С увеличением параметра r аппроксимационные свойства $\bar{p}(x, y)$ снижаются. Например, при известном значении $r = 0{,}75$ и $\alpha = 0{,}1$ $\bar{c}^*/c^* \in [1{,}103; 0{,}871]$, а $\bar{W}(\bar{c}^*)/\bar{W}(c^*) \in [0{,}822; 1{,}32]$. При $\alpha = 0{,}3$ эти показатели аппроксимационных свойств $\bar{p}(x, y)$ значительно снижаются (см. табл. 2).

Отсюда следует, что при малых значениях r, характерных для нелинейных зависимостей между x, y в условиях достаточно больших параметров помех α, целесообразно принимать значения $c_1 = c\sigma_1$, $c_2 = c\sigma_2$. При бóльших значениях $\bar{r} > 0{,}825$ соблюдение этого условия сопряжено со значительным ухудшением свойств $\bar{p}(x, y)$.

Таблица 2

Результаты сравнения аппроксимационных свойств $\bar{p}(x, y)$ от погрешности оценивания коэффициента корреляции

$\bar{r}$	$B(\bar{r})/B(r)$	$\bar{c}^*/c^*$	$\bar{W}(\bar{c}^*)/\bar{W}(c^*)$
0,175	0,906	1,017	0,967
0,225	0,963	1,006	0,988
0,275	1,042	0,993	1,014
0,325	1,048	0,977	1,047
0,525	0,252	1,258	0,631
0,675	0,555	1,103	0,822
0,825	2,298	0,871	1,32
0,975	269,108	0,394	6,456

Заключение. Для упрощения задачи оптимизации непараметрической оценки плотности вероятности двухмерной случайной величины с зависимыми компонентами целесообразно представлять коэффициенты размытости ядерных функций в виде произведения параметра c и среднего квадратического отклонения случайной величины. По результатам анализа асимптотических свойств исследуемой непараметрической оценки плотности вероятности оптимальное значение параметра c^* зависит от коэффициента корреляции r между зависимыми случайными величинами и объёма n исходных статистических данных. При фиксированных значениях r с ростом среднего квадратического отклонения σ_2 функции y наблюдается рост значений c^*. С увеличением r и n значения c^* уменьшаются. При малых значениях r и относительно больши́х погрешностях их оценивания аппроксимационные свойства рассматриваемой непараметрической оценки плотности вероятности изменяются незначительно. С увеличением r аппроксимационные свойства $\bar{p}(x, y)$ снижаются, что особенно характерно для больших погрешностей оценивания r.

Полученные результаты являются основой развития быстрых процедур оптимизации непараметрической регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. **Лапко А. В., Лапко В. А., Бахтина А. В.** Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о независимости переменных неоднозначных функций // Измерительная техника. 2022. № 1. С. 17–22. DOI: 10.32446/0368-1025it.2022-01-17-22.
2. **Лапко А. В., Лапко В. А., Бахтина А. В.** Исследование методики проверки гипотезы о независимости двухмерных случайных величин с использованием непараметрического классификатора // Автометрия. 2021. **57**, № 6. С. 90–100. DOI: 10.15372/AUT20210610.
3. **Зеньков И. В., Лапко А. В., Лапко В. А. и др.** Методика последовательного формирования набора компонент многомерной случайной величины с использованием непараметрического алгоритма распознавания образов // Компьютерная оптика. 2021. **45**, № 6. С. 926–933. DOI: 10.18287/2412-6179-CO-902.
4. **Rudemo M.** Empirical choice of histogram and kernel density estimators // Scandinav. Journ. Statist. 1982. **9**. P. 65–78.
5. **Bowman A. W.** A comparative study of some kernel-based non-parametric density estimators // Journ. Statist. Comput. Simulation. 1985. **21**, N 3–4. P. 313–327. DOI: 10.1080/00949658508810822.

6. **Hall P.** Large-sample optimality of least squares cross-validation in density estimation // Ann. Statist. 1983. **11**, N 4. P. 1156–1174. DOI: 10.1214/aos/1176346329.
7. **Jiang M., Provost S. B.** A hybrid bandwidth selection methodology for kernel density estimation // Journ. Statist. Comput. and Simulation. 2014. **84**, N 3. P. 614–627. DOI: 10.1080/00949655.2012.721366.
8. **Dutta S.** Cross-validation revisited // Communications in Statistics — Simulation and Computation. 2016. **45**, N 2. P. 472–490. DOI: 10.1080/03610918.2013.862275.
9. **Silverman B. W.** Density Estimation for Statistics and Data Analysis. London: Chapman & Hall, 1986. 175 p.
10. **Sheather S., Jones M.** A reliable data-based bandwidth selection method for kernel density estimation // Journ. Royal Statist. Soc. Ser. B. 1991. **53**, N 3. P. 683–690. DOI: 10.1111/j.2517-6161.1991.tb01857.x.
11. **Sheather S. J.** Density estimation // Statist. Sci. 2004. **19**, N 4. P. 588–597. DOI: 10.1214/088342304000000297.
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.** Быстрый выбор коэффициентов размытости ядерной оценки плотности вероятности независимых случайных величин // Автометрия. 2022. **58**, № 1. С. 33–39. DOI: 10.15372/AUT20220104.
14. **Лапко А. В., Лапко В. А., Бахтина А. В.** Оптимизация ядерной оценки плотности вероятности двухмерной случайной величины с независимыми составляющими // Измерительная техника. 2021. № 12. С. 17–21. DOI: 10.32446/0368-1025it.2021-12-17-21.
15. **Лапко А. В., Лапко В. А.** Нетрадиционная методика выбора коэффициентов размытости ядерных функций в непараметрической регрессии // Измерительная техника. 2022. № 2. С. 3–7. DOI: 10.32446/0368-1025it.2022-2-3-7.
16. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. **14**, № 1. С. 156–161.
17. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
18. **Пугачёв В. С.** Теория вероятностей и математическая статистика: уч. пособие. М.: Физматлит, 2002. 496 с.

Поступила в редакцию 12.05.2022
После доработки 14.06.2022
Принята к публикации 24.06.2022