

Литература

1. Ленг К. Астрофизические формулы. В 2-х частях. М.: Мир, 1978.
2. Riess A. G. e. a. Observational evidence from supernovae for an accelerating universe and a cosmological constant // Astron. J. 1998. V. 116. P. 1009—1038.
3. Perlmutter S. e. a. Measurements of Ω and Λ from 42 high-redshift supernovae // Astrophys. J. 1999. V. 517. P. 565—586.
4. Эльясберг П. Е. Измерительная информация: сколько ее нужно? Как ее обрабатывать? М.: Наука, 1983.
5. Хампель Ф. и др. Робастность в статистике. Подход на основе функций влияния. М.: Мир, 1989.
6. Fischer R. A. Theory of statistical estimation // Proc. Cambridge Phil. Soc. 1925. V. 22. N. 5. P. 700—725.
7. Математическая энциклопедия. Т. 2. М.: Советская энциклопедия, 1979. С. 656.
8. Tonry J. L. e. a. Cosmological results from high-z supernovae // Astrophys. J. 2003. V. 594. P. 1—24.
9. Branch D., Tammann G. A. Type Ia supernovae as standard candles // Annual Reviews Astronomy & Astrophysics. 1992. V. 30. P. 359—389.

10. Hicken M. e. a. Improved dark energy constraints from ~100 new CfA supernova type Ia light curves // Astrophys. J. 2009. V. 700. P. 1097—1140.

11. Шмидт Б. П. Ускоренное расширение Вселенной по наблюдениям далеких сверхновых // УФН. 2013. Т. 183. № 10. С. 1078—1089.

12. Левин С. Ф. Оптимальная интерполяционная фильтрация статистических характеристик случайных функций в детерминированной версии метода Монте—Карло и закон красного смещения. М.: АН СССР, Научный совет по комплексной проблеме «Кибернетика», 1980.

13. Р 50.2.004—2000. ГСИ. Определение характеристик математических моделей зависимостей между физическими величинами при решении измерительных задач. Основные положения.

14. Larson D. e. a. 7 year WMAP observations: power spectra and WMAP-derived parameters // Astrophys. J. Suppl. Ser. 2011. № 192.

Дата принятия 18.12.2014 г.

ОБЩИЕ ВОПРОСЫ МЕТРОЛОГИИ И ИЗМЕРИТЕЛЬНОЙ ТЕХНИКИ

519.24

Анализ зависимости аппроксимационных свойств непараметрической оценки плотности вероятности от методов дискретизации области определения

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

Предложена методика сравнения эффективности процедур дискретизации интервала значений случайной величины при оценивании плотности вероятности. В качестве критерия эффективности использовано асимптотическое выражение среднего квадратического отклонения регрессионной оценки плотности вероятности.

Ключевые слова: плотность вероятности, непараметрическая оценка, аппроксимационные свойства, количество интервалов дискретизации, правило Хайнгольда—Гаеде, правило Брукса и Каррузера, правило Старджесса, правило Фридмана и Диакониса, правило Скотта.

The comparison procedure for the effectiveness of methods for discretization of the interval of values of a random variable at probability density estimation has been suggested. As an efficiency criterion the asymptotic expression of standard deviation of regression model of probability density has been used.

Key words: probability density, nonparametric estimation, approximation properties, number of sampling intervals, Heinhold—Gaede rule, Brooks—Carruthers rule, Sturges rule, Freedman—Diaconis rule, Scott rule.

Данный материал развивает и обобщает результаты исследований по оцениванию эффективности формул дискретизации, изложенные в [1].

Пусть x^i , $i = \overline{1, n}$ — выборка из n независимых наблюдений одномерной случайной величины x с неизвестной плот-

ностью вероятности $p(x)$, которая ограничена и непрерывна со всеми производными до второго порядка включительно.

Для условий выборок большого объема n в [2] предложена и в [3, 4] развита методика синтеза непараметрической оценки плотности вероятности, основанная на сжатии исходных статистических данных.

Разобьем область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин $X^j, j = \overline{1, N}$. В качестве характеристик X^j примем частоту $\bar{P}^j$ попадания случайной величины x в j -й интервал и его центр z^j . На основе полученной информации определим массив данных $V_1 = (z^j, y^j = \bar{P}^j / (2\beta), j = \overline{1, N})$, составленный из центров z^j введенных интервалов и соответствующих им значений оценок y^j плотности вероятности. В качестве приближения по эмпирическим данным V_1 искомым плотности вероятности $p(x)$ примем статистику [3]:

$$\tilde{p}(x) = c^{-1} \sum_{j=1}^N \bar{P}^j \Phi \left[(x - z^j) / c \right], \quad (1)$$

где ядерные функции $\Phi(u)$ удовлетворяют условиям [5, 6]:

$$\begin{aligned} \Phi(u) &= \Phi(-u), \quad 0 \leq \Phi(u) < \infty; \\ \int \Phi(u) du &= 1, \quad \int u^2 \Phi(u) du = 1; \\ \int u^m \Phi(u) du &< \infty \quad \text{при } 0 \leq m < \infty. \end{aligned}$$

Здесь и далее бесконечные пределы интегрирования опускаем.

Коэффициенты размытости $c = c(N)$ ядерных функций в (1) характеризуют области их определения. Аппроксимационные свойства непараметрической оценки плотности вероятности (1) зависят от особенностей процедуры дискретизации области определения $p(x)$.

Ниже на основе результатов анализа аппроксимационных свойств непараметрической оценки плотности вероятности осуществлено сравнение и системный анализ дискретизации интервала значений одномерной случайной величины.

Выбор оптимального количества интервалов дискретизации области значений случайной величины. Представим непараметрическую оценку плотности вероятности (1) в виде

$$\bar{p}(x) = (nc)^{-1} \sum_{i=1}^N \sum_{j=1}^n h \left((x^j - z^i) / \beta \right) \Phi \left((x - z^i) / c \right), \quad (2)$$

где индикаторные функции

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

определяют принадлежность элементов выборки $V = (x^j, j = \overline{1, n})$ интервалам $(z^i \pm \beta), i = \overline{1, N}$. В отличие от (1), в ее модификации (2) присутствуют в явном виде параметры процедуры дискретизации β, N и объем n исходных статистических данных. В [7] выведено асимптотическое выражение среднего квадратического отклонения (СКО) $M \int (p(x) - \bar{p}(x))^2 dx$ непараметрической оценки $\bar{p}(x)$ (2) от восстанавливаемой плотности вероятности $p(x)$:

$$\begin{aligned} W_2(N) &= \left(\frac{(\|\Phi(u)\|^2)^4 \|p^{(2)}(x)\|^2}{2N^4} \right)^{1/5} \times \\ &\times \left(\frac{3}{2} (\Delta \|p(x)\|^2)^{4/5} + \frac{N^2}{n (\Delta \|p(x)\|^2)^{1/5}} \right). \end{aligned} \quad (3)$$

Здесь $\|p(x)\|^2 = \int p^2(x) dx$; $\|p^{(2)}(x)\|^2 = \int (p^{(2)}(x))^2 dx$; $\|\Phi(u)\|^2 = \int \Phi^2(u) du$; $p^{(2)}(x)$ — вторая производная $p(x)$ по x ; Δ — длина интервала значений случайной величины; M — математическое ожидание.

Из условия минимума асимптотического выражения СКО (3) получена процедура оптимального выбора количества интервалов дискретизации

$$N = \sqrt{\Delta \|p(x)\|^2 n}, \quad (4)$$

определяемая видом восстанавливаемой плотности вероятности, областью определения Δ и объемом n исходных статистических данных.

Анализ аппроксимационных свойств $\bar{p}(x)$. Зависимость аппроксимационных свойств непараметрической оценки плотности вероятности (2) от ряда методов дискретизации области значений одномерной случайной величины x проанализировали по данным вычислительных экспериментов [1]. В отличие от данного направления исследований, оценкой эффективности широкого круга методов дискретизации с помощью критерия (3), что позволит осуществить их объективную систематизацию.

Пусть восстанавливаемая плотность вероятности случайной величины x имеет нормальный закон распределения (НЗР) с нулевым математическим ожиданием и СКО $\sigma = 1$:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{x^2}{2} \right). \quad (5)$$

Тогда

$$\|p(x)\|^2 = \frac{1}{2\sqrt{\pi}}; \quad \|p^{(2)}(x)\|^2 = \frac{3}{8\sqrt{\pi}},$$

а случайная величина $\Delta = 6$ с вероятностью 0,997.

Для выбора количества интервалов дискретизации области значений случайной величины используют (4), а также формулы:

$$\text{Фридмана и Диакониса } N = \Delta / (2 IQR \cdot n^{-1/3}) \quad [8], \quad (6)$$

где IQR — квартильный размах, определяемый разностью между третьим и первым квартилями распределения случайной величины;

$$\text{Хайнхольда и Гаеде } N = \sqrt{n} \quad [9]; \quad (7)$$

$$\text{Брукса и Каррузера } N = 5 \lg n; \quad (8)$$

$$\text{Таушанова и Тоневой } N = 4 \lg n; \quad (9)$$

$$\text{Скотта } N = \Delta / (3,49 \sigma n^{-1/3}) \quad [10]; \quad (10)$$

$$\text{И. У. Алексеевой } N = \frac{4}{\chi} \lg \frac{n}{10}, \quad (11)$$

где $\chi = 1/\sqrt{3}$ — контрэксцесс НЗР (5);

$$\text{Старджесса } N = \log_2 n + 1. \quad (12)$$

При использовании (4) в задаче оценивания (5) коэффициент $\sqrt{\Delta \|p(x)\|^2} = 1,3$.

В качестве ядерной функции $\Phi(u)$ используем оптимальное ядро Епанечникова [6]:

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

для которого

$$\|\Phi(u)\|^2 = \frac{3}{5\sqrt{5}}.$$

Выражение (3) запишем в виде

$$W_2 = \frac{3}{10} \left(\frac{2}{15N^4} \right)^{1/5} \left(\frac{9}{2\sqrt{\pi}} + \frac{N^2}{n} \right).$$

При увеличении объема $n \in [100, 500]$ исходных статистических данных применение исследуемых методов дискретизации интервала значений случайной величины приводит к уменьшению W_2 (табл. 1). Наблюдаемое улучшение аппроксимационных свойств $\bar{p}(x)$ можно объяснить увеличением объема N массива данных V_1 , используемого при построении непараметрической оценки плотности вероятности (табл. 2), что согласуется с условиями ее асимптотической сходимости [3, 11].

Исследуемые формулы дискретизации области значений случайной величины условно можно разбить на две группы по значениям СКО W_2 аппроксимации $\bar{p}(x)$. Первую группу составляют формулы (4), (6)—(8), использование которых позволяет получить меньшую погрешность аппроксимации плотности вероятности по сравнению с (9)—(12). Применение формулы (4) при выборе количества N интервалов дискретизации более предпочтительно по сравнению с другими, так как она получена на основе минимизации асимптотического выражения СКО (3).

Таблица 1

Значения критерия W_2 качества оценивания плотности вероятности, соответствующего конкретным объемам n исходных данных и формулам дискретизации

n	Формулы дискретизации							
	Оптимальный выбор количества интервалов	Фридмана и Диакониса	Хайнхольда и Гаеде	Брукса и Каррузера	Таушанова и Тоневой	Скотта	И. У. Алексеевой	Старджесса
100	0,109	0,112	0,112	0,112	0,121	0,121	0,128	0,121
150	0,093	0,096	0,096	0,099	0,106	0,106	0,113	0,113
200	0,083	0,087	0,085	0,090	0,102	0,097	0,102	0,102
250	0,076	0,081	0,078	0,086	0,093	0,089	0,093	0,099
300	0,070	0,076	0,073	0,083	0,091	0,083	0,091	0,097
350	0,066	0,071	0,068	0,078	0,090	0,081	0,085	0,096
400	0,063	0,069	0,065	0,076	0,089	0,076	0,084	0,089
450	0,060	0,066	0,062	0,075	0,083	0,075	0,083	0,088
500	0,057	0,063	0,059	0,074	0,082	0,071	0,078	0,087

Таблица 2

Зависимость количества интервалов N , на которые разбивается область значений случайной величины, от объема n исходных данных и используемых формул дискретизации

n	Формулы дискретизации							
	Оптимальный выбор количества интервалов	Фридмана и Диакониса	Хайнхольда и Гаеде	Брукса и Каррузера	Таушанова и Тоневой	Скотта	И. У. Алексеевой	Старджесса
100	13	10	10	10	8	8	7	8
150	16	12	12	11	9	9	8	8
200	18	13	14	12	9	10	9	9
250	21	14	16	12	10	11	10	9
300	23	15	17	12	10	12	10	9
350	24	16	19	13	10	12	11	9
400	26	16	20	13	10	13	11	10
450	28	17	21	13	11	13	11	10
500	29	18	22	13	11	14	12	10

Зависимости W_2 от объема n исходных данных при использовании формул (6), (7) близки. Им свойственны сопоставимые значения количества N интервалов дискретизации области изменения случайной величины (см. табл. 2). При малых $n \leq 150$ значения W_2 , соответствующие указанным формулам дискретизации, отличаются незначительно, однако с ростом $n \geq 200$ предпочтительно применение формулы (7) при синтезе плотности вероятности с НЗР.

Формула дискретизации (8) из первой группы методов имеет преимущество над (9)—(12) во всем диапазоне изменений n и сопоставима с (10) при $n \leq 300$. Результаты ее применения сопоставимы с формулами (7), (6) при малом объеме $n = 100$.

Вторая группа методов дискретизации (9)—(12) по сравнению с первой характеризуется относительно меньшим количеством интервалов N , на которые разбивается область значений случайных величин. Применение формулы (10) при выборе количества интервалов дискретизации более предпочтительно по сравнению с (11), (12) при $n \geq 150$. Ее эффективность сопоставима с (8) при $n \geq 300$. Преимущество формулы (11) над (9), (12) проявляется при $n \geq 350$ и $n \geq 250$, соответственно. Результаты применения (12) имеют преимущества по сравнению с использованием (9), (11) при $n \geq 200$ и $n = 100$, соответственно. Эффективность (12) сопоставима с формулой (10) при $n = 100$. Применение (9) обеспечивает меньшую эффективность при оценивании плотности вероятности. Ее преимущество наблюдается лишь над формулой (11) при $n \leq 150$. Результаты, полученные на основе (9), сопоставимы с (12) при $n \leq 150$.

Результаты сравнения эффективности методов дискретизации представлены в табл. 3.

Если в элементе b_{ij} (i, j — номера строки и столбца) присутствует знак «+», то метод дискретизации m_j обладает преимуществом над методом m_i , которое на ограниченном диапазоне изменения n сопровождается соответствующим условием. Например, метод Хайнхольда и Гаеде имеет преимущество над методом Фридмана и Диакониса при $n \geq 200$. Два метода дискретизации m_i, m_j сопоставимы по эффективности, если в элементе таблицы b_{ij} расположен знак «0» и условие на объем n исходных данных (например, $n = 100$ либо $n \geq 300$). Если в элементе таблицы b_{ij} присутствует знак «—», то метод дискретизации m_i обладает преимуществом над методом m_j .

Выводы. Методика синтеза оценки плотности вероятности предполагает использование процедуры дискретизации области ее определения. Поэтому можно оценить эффективность различных формул дискретизации и осуществить на этой основе их систематизацию.

При восстановлении плотности вероятности с НЗР целесообразно использовать формулы (4), (6)—(8). По сравнению с другими методами дискретизации им свойственны сопоставимые относительно большие значения количества интервалов дискретизации. Применение формулы (4) является более предпочтительным, а результаты использования (8) менее эффективны при оценивании плотности вероятности. Формулы дискретизации (6), (7) по своей эффективности занимают промежуточное положение.

Представленные результаты исследований имеют важное значение при решении задач проверки гипотез о рас-

Таблица 3

Результаты сравнения эффективности методов дискретизации при оценивании плотности вероятности

Методы дискретизации	Оптимальный выбор количества интервалов	Фридмана и Диакониса	Хайнхольда и Гаеде	Брукса и Каррузера	Таушанова и Тоневой	Скотта	И. У. Алексеевой	Старджесса
Оптимальный выбор количества интервалов	0	—	—	—	—	—	—	—
Фридмана и Диакониса	+	0	+ $n \geq 200$	0 $n = 100$	—	—	—	—
Хайнхольда и Гаеде	+	0 $n \leq 150$	0	0 $n = 100$	—	—	—	—
Брукса и Каррузера	+	+ $n \geq 150$	+ $n \geq 150$	0	—	0 $n \geq 300$	—	—
Таушанова и Тоневой	+	+	+	+	0	—	+ $n \geq 350$	+ $n \geq 200$
Скотта	+	+	+	+ $n \leq 300$	—	0	—	0 $n = 100$
И. У. Алексеевой	+	+	+	+	+ $n \leq 150$	+	0	+ $n = 100$
Старджесса	+	+	+	+	0 $n \leq 150$	+ $n \geq 150$	+ $n \geq 250$	0

пределениях случайных величин и доверительного оценивания плотности вероятности.

Работа выполнена в рамках проектной части государственного задания Минобрнауки РФ № 2.914.2014/К.

Л и т е р а т у р а

1. Лапко А. В., Лапко В. А. Сравнение эффективности методов дискретизации интервала изменения значений случайной величины при синтезе непараметрической оценки плотности вероятности // Измерительная техника. 2014. № 3. С. 5—8.

2. Лапко А. В., Лапко В. А. Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. Т. 39. № 1. С. 54—61.

3. Лапко А. В., Лапко В. А. Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. 2012. Т. 49. № 3.1. С. 152—156.

4. Лапко А. В., Лапко В. А. Построение доверительных границ для плотности вероятности на основе ее регрессионной оценки // Метрология. 2013. № 12. С. 3—9.

5. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. V. 33. N. 3. P. 1065—1076.

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

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. Freedman D., Diaconis P. On the histogram as a density estimator: L2 theory // Probability Theory and Related Fields (Heidelberg: Springer Berlin). 1981. V. 57. N. 4. P. 453—476.

9. Heinhold I., Gaede K. Ingenieur statistic. München: Wien, Springer, 1964.

10. Scott D. On optimal and data-based histograms // Biometrika. 1979. V. 66. N. 3. P. 605—610.

11. Лапко А. В., Лапко В. А. Анализ аппроксимационных свойств регрессионной оценки плотности вероятности // Информатика и системы управления. 2014. Т. 39. № 1. С. 80—87.

Дата принятия 09.10.2014 г.

53.072(086.48):53.088:53.089.6

Перспективы совершенствования способов калибровки при использовании аппарата редукции измерений

М. В. ЗЕЛЕНКОВА, В. Л. СКРИПКА

Московский государственный технический университет им. Н. Э. Баумана,
Москва, Россия, e-mail: viz_zelen@rambler.ru

Обоснована возможность дистанционного проведения калибровочных работ на базе теории редукции измерений с использованием существующих цифровых телекоммуникационных сетей, что позволит оптимизировать алгоритм обработки измерительной информации и обеспечит повышение достоверности результатов калибровки в различных условиях ее проведения. Рассмотрен пример применения аппарата данной теории при калибровке средств измерений сложнопрофильных поверхностей при дистанционной оценке соответствия.

Ключевые слова: калибровка, редукция измерений, моделирование.

The possibility of remote calibration works based on the measurements reduction theory with use of existing digital telecommunication networks is presented. This theory usage allows to optimize the measuring information processing algorithm with accuracy of higher calibration results reliability in different calibration conditions. The example of this theory application for complex profile surfaces measuring instruments calibration at distance estimation of conformity has been considered.

Key words: calibration, measurements, reduction, modelling.

Одним из перспективных направлений совершенствования поверочно-калибровочных работ является их проведение через удаленный канал обслуживания [1—3]. В настоящей работе на базе теории редукции измерений технически обоснована возможность дистанционного проведения калибровочных работ [4, 5]. Их проведение сводится к достоверной оценке двух малоразличимых функций, полученных эмпирически с помощью нескольких средств измерений

(СИ), одно из которых точнее остальных (или считается таковым). Математически эту процедуру можно описать системой уравнений, решения которых неустойчивы, т. е. при весьма малых отклонениях в исходных данных результаты могут сильно отличаться. Однако теоретически такие задачи можно доопределить при использовании дополнительной информации для построения устойчивых приближений к точным решениям [4].