

Непараметрическая оценка плотности вероятности парzenовского типа с неявно заданной формой ядерной функции

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 непараметрическая оценка плотности вероятности, имеющая повышенные аппроксимационные свойства по сравнению с традиционной статистикой типа Розенблатта—Парзена. Исследована зависимость её свойств от вида ядерных функций и формул дискретизации интервала значений случайной величины.

Ключевые слова: плотность вероятности, непараметрическая оценка Розенблатта—Парзена, аппроксимационные свойства, ядерная функция, правило Старджесса, формула Хайнхольда и Гаеде.

Proposed nonparametric probability density estimate having improved approximation properties compared to the traditional statistics such as Rosenblatt—Parzen. The dependence of its properties on the type of kernel functions and formulas sampling interval of values of the random variable.

Key words: probability density, nonparametric estimation Rosenblatt—Parzen, approximation properties, kernel function, Sturges' rule, Heinhold—Gaede rule.

Пусть для оценивания плотности вероятности по выборке x^i , $i=1, n$ статистически независимых наблюдений случайной величины x используется её непараметрическая оценка [1]:

$$\bar{p}(x) = \frac{1}{nc} \sum_{i=1}^n \Phi\left(\frac{x-x^i}{c}\right), \quad (1)$$

где n — объём статистических данных.

В статистике (1) последовательности коэффициентов размытости $c = c(n)$ функции $\Phi(u)$ такие, что $c(n) \rightarrow 0$ при $n \rightarrow \infty$ и каждая ядерная функция $\Phi(u)$ удовлетворяет условиям:

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

$$\int \Phi(u) du = 1; \quad \int u^2 \Phi(u) du = 1;$$

$$\int u^m \Phi(u) du < \infty; \quad 0 \leq m < \infty.$$

Здесь и далее бесконечные пределы интегрирования опущены.

В [2] изложена теория непараметрического оценивания плотности вероятности. Рассмотрены оценки плотности вероятности, приведены результаты исследований их асимптотических свойств и оптимизации. Непараметрическая оценка плотности вероятности (1) обобщена для многомерного случая [3]. Установлены условия её асимптотической несмещённости и состоятельности. Для условий выборок большого объёма разработаны непараметрические оценки плотности вероятности, основанных на принципах декомпозиции и процедуре сжатия исходных статистических данных [4, 5].

При ограниченных объёмах n статистических данных непараметрическая оценка плотности вероятности типа (1) характеризуется повышенной дисперсией.

В данной работе предложена и исследована модификация непараметрической оценки плотности вероятности, синтез которой основан на использовании сглаживающего оператора.

Синтез непараметрической оценки плотности вероятности. В качестве приближения по эмпирическим данным одномерной плотности вероятности $p(x)$ примем статистику

$$\bar{p}_1(x) = \frac{1}{2\beta} \int_{x-\beta}^{x+\beta} \bar{p}(u) du, \quad (2)$$

где $\bar{p}(u)$ — непараметрическая оценка плотности вероятности (1); β — параметр, определяющий пределы интегрирования.

Подставив в (2) выражение (1), получим

$$\bar{p}_1(x) = \frac{1}{2\beta nc} \sum_{i=1}^n \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^i}{c}\right) du. \quad (3)$$

Статистика (3) относится к классу непараметрических оценок плотности вероятности с неявно заданной ядерной функцией

$$\frac{1}{2\beta c} \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^i}{c}\right) du. \quad (4)$$

Выбор значений β , с при оптимизации непараметрической оценки плотности вероятности (3) позволяет изменять не только область определения ядерной функции, но и её вид. Например, при ядерной функции

$$\frac{1}{c} \Phi\left(\frac{u-x^i}{c}\right) = \begin{cases} 1/(2c), & \text{если } |u-x^i| \leq c; \\ 0, & \text{если } |u-x^i| > c \end{cases} \quad (5)$$

неявно заданное ядро (4) при $c > \beta$ запишется в виде

$$\begin{cases} 0, & \text{если } |x-x^i| \geq c+\beta; \\ \frac{c+\beta-|x-x^i|}{4\beta c}, & \text{если } c-\beta < |x-x^i| < c+\beta; \\ 1/(2c), & \text{если } |x-x^i| < c-\beta. \end{cases} \quad (6)$$

Соотношения (6) определяют ядерную функцию (4) в виде трапеции при $c > \beta$ и треугольника, если $c = \beta$ (рис. 1). Нетрудно убедиться, что ядерная функция (4) обладает свойством нормированности.

Определим в интервале интегрирования функцию ядерного типа $h(u)$. Тогда модифицированную непараметрическую оценку плотности вероятности можно представить в виде

$$\bar{p}_2(x) = \frac{1}{nc} \sum_{i=1}^n \int_{\Omega(u)} \Phi\left(\frac{x-\beta u-x^i}{c}\right) h(u) du, \quad (7)$$

где $\Omega(u)$ — область интегрирования произведения ядерных функций по переменной u .

Если $\beta = 0$, то статистика (7) преобразуется в традиционную оценку плотности вероятности Розенблатта—Парзена

$$\bar{p}_2(x) = \frac{1}{nc} \sum_{i=1}^n \Phi\left(\frac{u-x^i}{c}\right) \int_{\Omega(u)} h(u) du,$$

где $\int_{\Omega(u)} h(u) du = 1$.

Пусть функция $h(u)$ является ступенчатой:

$$h(u) = \begin{cases} 1/2, & \text{если } |u| \leq 1; \\ 0, & \text{если } |u| > 1. \end{cases}$$

Тогда оценка (7) принимает вид

$$\bar{p}_2(x) = \frac{1}{2nc} \sum_{i=1}^n \int_{-1}^1 \Phi\left(\frac{x-\beta u-x^i}{c}\right) du. \quad (8)$$

Проведя замену переменных $x - \beta u = z$ и пределов интегрирования, из выражения (8) нетрудно найти непараметрическую оценку плотности вероятности $\bar{p}_1(x)$.

Статистики (3), (7) оптимизируют путём последовательного выбора параметров c, β . С учётом методики синтеза исследуемых непараметрических оценок плотностей вероятности $\bar{p}_1(x), \bar{p}_2(x), \bar{p}(x)$ коэффициенты размытости с ядерных функций равны. Поэтому для выбора параметра c используем условие максимума функции правдоподобия [6]:

$$L(c) = \prod_{j=1}^n \bar{p}(x^j),$$

где

$$\bar{p}(x^j) = \frac{1}{(n-1)c} \sum_{\substack{i=1 \\ i \neq j}}^n \Phi\left(\frac{x^j-x^i}{c}\right).$$

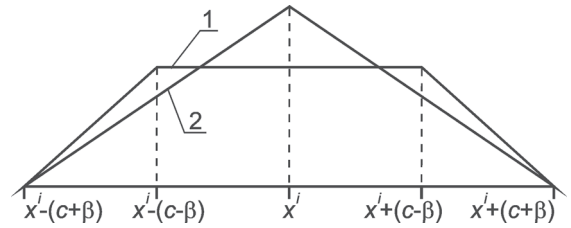


Рис. 1. Вид ядерной функции (4) в условиях (5) при $c > \beta$ (1), $c = \beta$ (2)

Параметр β выберем в соответствии с выражением $\beta = \Delta/(2N)$, где Δ — область значений случайной величины x , а N — количество интервалов её дискретизации, определяемых по формулам Старджесса [7]

$$N = \log_2 n + 1, \quad (9)$$

или Хайнхольда и Гаеде [8]

$$N = \sqrt{n}. \quad (10)$$

Асимптотические свойства непараметрической оценки плотности вероятности $\bar{p}_1(x)$. Исследуем асимптотические свойства одномерной непараметрической оценки плотности вероятности (3) в случае, когда искомая плотность вероятности $p(x)$ разлагается в ряд Тейлора в каждой точке x . При выполнении преобразований используем методику, предложенную в [3].

Для определения условий асимптотической несмещённости непараметрической оценки $\bar{p}_1(x)$ вычислим её математическое ожидание

$$\begin{aligned} M[\bar{p}_1(x)] &= \frac{1}{2\beta nc} \sum_{i=1}^n \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^i}{c}\right) du p(x^i) dx^i = \\ &= \frac{1}{2\beta c} \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-t}{c}\right) du p(t) dt, \end{aligned}$$

где наблюдения $x^i, i = \overline{1, n}$ характеризуются одной и той же плотностью вероятности $p(t)$.

Проведём последовательные замены переменных (сначала $u = x + z\beta$, а затем $(x + z\beta - t)/c = u_1$ и получим

$$\begin{aligned} M[\bar{p}_1(x)] &= \frac{1}{2c} \int_{-1}^1 \int_{-1}^1 \Phi\left(\frac{x+z\beta-t}{c}\right) p(t) dt dz = \\ &= \frac{1}{2} \int_{-1}^1 \int_{-1}^1 \Phi(u_1) p(x+z\beta-cu_1) du_1 dz. \end{aligned} \quad (11)$$

Разложим функцию $p(x + z\beta - cu_1)$ в ряд Тейлора в точке x и проинтегрируем правую часть (11) с учётом свойств ядерной функции. В результате получим асимптотическое выражение

$$M[\bar{p}_1(x)] \sim p(x) + \frac{1}{2} \left(\frac{\beta^2}{3} + c^2 \right) p^{(2)}(x), \quad (12)$$

где $p^{(2)}(x)$ — вторая производная плотности вероятности $p(x)$.

Вычислим усреднённую квадратическую ошибку аппроксимации:

$$M[\bar{p}_1(x) - p(x)]^2 = M[\bar{p}_1^2(x)] - 2p(x)M[\bar{p}_1(x)] + p^2(x). \quad (13)$$

Определим первое слагаемое в правой части (13):

$$\begin{aligned} M[\bar{p}_1^2(x)] &= \frac{1}{4n^2c^2\beta^2} \left\{ \sum_{i=1}^n \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^i}{c}\right) du \right\}^2 p(x^i) dx^i + \\ &+ \sum_{i=1}^n \sum_{j=1, j \neq i}^n \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^i}{c}\right) du \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-x^j}{c}\right) du p(x^i) p(x^j) dx^i dx^j = \\ &= \frac{1}{4n^2c^2\beta^2} \left\{ n \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-t}{c}\right) du \right\}^2 p(t) dt + \\ &+ n(n-1) \left\{ \int_{x-\beta}^{x+\beta} \Phi\left(\frac{u-t}{c}\right) du p(t) dt \right\}^2. \end{aligned}$$

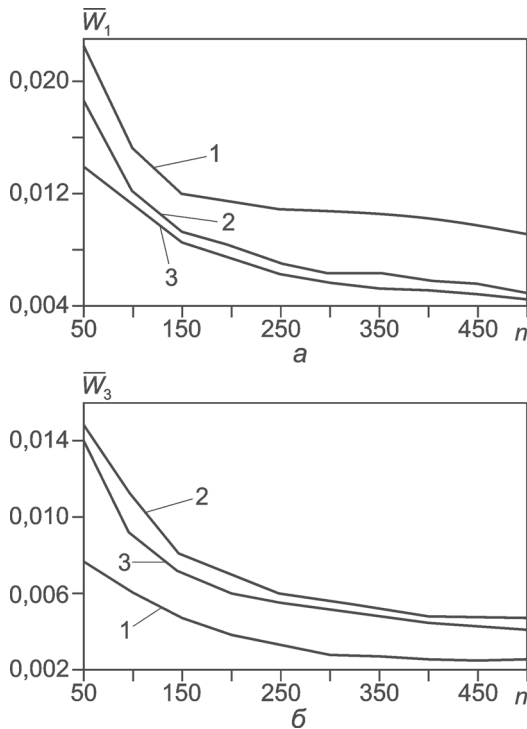


Рис. 2. Зависимость смещения $\bar{W}_1$ (а) и дисперсии $\bar{W}_3$ (б) непараметрических оценок плотности вероятности $\bar{p}(x)$, $\bar{p}_1(x)$ от объема n статистической выборки значений случайной величины с равномерным законом распределения:

1 — оценка плотности $\bar{p}_1(x)$ при ядерной функции (6) и формуле дискретизации (9); 2, 3 — статистика $\bar{p}(x)$ при использовании ядерных функций вида (5), (16)

Итак, после преобразований по приведённой выше методике запишем:

$$\begin{aligned} M[\bar{p}_1^2(x)] &\sim p^2(x) + \frac{p(x)}{nc} \int \Phi^2(u) du + \left(\frac{\beta^2}{3} + c^2 \right) p^{(2)}(x) p(x) + \\ &+ \frac{1}{4} [p^{(2)}(x)]^2 \left(\frac{\beta^2}{3} + c^2 \right)^2. \end{aligned} \quad (14)$$

При получении данного результата слагаемые малости $O(c/n)$, $O[\beta^2/(nc)]$ и выше не учитывались.

Подставив в (13) выражения (12), (14), получим

$$M[\bar{p}_1(x) - p(x)]^2 = \frac{p(x)}{nc} \int \Phi^2(u) du + \frac{1}{4} [p^{(2)}(x)]^2 \left(\frac{\beta^2}{3} + c^2 \right)^2. \quad (15)$$

Из анализа (12), (15) следует, что при выполнении условий $\beta(n) \rightarrow 0$, $c(n) \rightarrow 0$, $nc(n) \rightarrow \infty$ для $n \rightarrow \infty$ непараметрическая оценка плотности вероятности $\bar{p}_1(x)$ имеет свойства асимптотической несмещённости и сходимости в среднеквадратическом, а следовательно, является состоятельной оценкой $p(x)$.

Анализ результатов вычислительных экспериментов.

Сравним аппроксимационные свойства традиционной непараметрической оценки плотности вероятности (1) и её модификации (3) при восстановлении $p(x)$ с равномерным и нормальным законами распределения случайной величины.

При синтезе статистики $\bar{p}(x)$ использовали функцию (5) и оптимальное ядро В. А. Епанечникова [3]:

$$\Phi(u) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3u^2}{20\sqrt{5}}, & \text{если } |u| < \sqrt{5}; \\ 0, & \text{если } |u| \geq \sqrt{5}. \end{cases} \quad (16)$$

Формирование непараметрической оценки плотности вероятности (3) осуществляли на основе ядерной функции (6).

Оптимизацию $\bar{p}_1(x)$ проводили в соответствии с приведённой выше методикой. В качестве показателя эффективности сравниваемых статистик использовали оценки их смещения $\bar{W}_1$, среднеквадратического отклонения $\bar{W}_2$ и дисперсии $\bar{W}_3$. Вычислительные эксперименты при определении значений $\bar{W}_j$, $j=1, 3$ в конкретных условиях осуществляли 50 раз с последующим усреднением. При этом объём контрольной выборки, элементы которой формировались в соответствии с плотностью вероятности $p(x)$ был принят равным 1000.

Рассмотрим результаты исследований при восстановлении плотности вероятности случайной величины с равномерным законом распределения.

С ростом объёма n исходных статистических данных значения $\bar{W}_j$, $j=1, 3$ снижаются для исследуемых статистик, что согласуется с результатами анализа их асимптотических свойств (рис. 2). Предложенная оценка плотности вероятности $\bar{p}_1(x)$ по сравнению $\bar{p}(x)$ имеет в 1,5—2 раза меньшую дисперсию и среднеквадратическое отклонение в условиях, когда при их синтезе используют ядерные функции типа (5), (6)

и формулу дискретизации (9). При этом смещение $\bar{W}_1[\bar{p}_1(x)]$ статистики $\bar{p}_1(x)$ от $p(x)$ существеннее, чем $\bar{W}_1[\bar{p}(x)]$ для оценки плотности вероятности $\bar{p}(x)$ (см. рис. 2).

Если при построении $\bar{p}(x)$ применяют ядерную функцию В. А. Епанечникова (16), то наблюдается снижение на 10—15 % значений оценок её показателей $\bar{W}_j, j = \overline{1, 3}$. Однако отмеченные выше закономерности в основном сохраняются.

Применение формулы дискретизации Хайнхольда и Гаеде (10) позволяет снизить смещение $\bar{W}_1[\bar{p}_1(x), 10]$ непараметрической оценки плотности вероятности $\bar{p}_1(x)$ до 30 % по сравнению с результатами $\bar{W}_1[\bar{p}_1(x), 9]$ использования правила Старджесса (9). При этом дисперсия $\bar{W}_3[\bar{p}_1(x), 10]$ статистики $\bar{p}_1(x)$ увеличивается на 20 % от $\bar{W}_3[\bar{p}_1(x), 9]$, что особенно проявляется при больших значениях n . Однако в данных условиях дисперсия и среднеквадратическое отклонение непараметрической оценки $\bar{p}_1(x)$ меньше, чем для традиционной статистики $\bar{p}(x)$.

При оценивании нормального закона распределения аппроксимационные свойства статистик $\bar{p}(x), \bar{p}_1(x)$ существенно снижаются. Данный факт объясняется неравномерностью распределения значений случайной величины, что приводит к уменьшению оптимальных значений коэффициентов размытости и увеличению дисперсии статистик $\bar{p}(x), \bar{p}_1(x)$. Отмеченная закономерность особо проявляется при использовании процедуры дискретизации Хайнхольда и Гаеде. Однако преимущество непараметрической оценки плотности вероятности $\bar{p}_1(x)$ над $\bar{p}(x)$ сохраняется при использовании ядерных функций (5), (16) и формул дискретизации (9), (10).

Заключение. Использование сглаживающего оператора при синтезе непараметрической оценки плотности вероятности типа Розенблатта—Парзена позволяет получить её модификации с повышенными аппроксимационными свойствами. Предложенные модификации в отличие от традиционной непараметрической оценки плотности вероятности отличаются неявно заданной формой ядерной функции. Эта особенность позволяет при их оптимизации изменять не только область определения ядерной функции, но и её вид.

Оптимизация исследуемых статистик происходит путём последовательного выбора коэффициента размытости ядерной функции из условия максимума функции правдоподобия и параметра сглаживающего оператора с использованием формул дискретизации интервала значений случайной величины. По сравнению с непараметрической оценкой плотности вероятности Розенблатта—Парзена исследуемая модификация характеризуется большим значением смещения, но меньшей дисперсией и среднеквадратическим отклонением. Отмеченные свойства сохраняются при значительно отличающихся условиях вычислительных экспериментов, таких как вид восстанавливаемой плотности вероятности и ядерной функции, правила дискретизации интервала значений случайной величины.

Полученные результаты создают основу построения непараметрических оценок многомерных плотностей вероятности с неявно заданной ядерной функцией.

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

Л и т е р а т у р а

1. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. V. I. 33. N. 3. P. 1065—1076.
2. **Деврой Л., Дьерфи Л.** Непараметрическое оценивание плотности. (L_1 — подход). М.: Мир, 1988.
3. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. Т. 14. № 1. С. 156—161.
4. **Лапко А. В., Лапко В. А.** Анализ свойств смеси непараметрических оценок плотности вероятности многомерной случайной величины // Вестник СибГАУ. 2010. № 2. С. 32—35.
5. **Lapko A. V., Lapko V. A.** Regression Estimate of the Multidimensional Probability Density and Its Properties // Optoelectronics, Instrumentation and Data Processing. 2014. V. 50. N. 2. P. 148—153.
6. **Duin R. P. W.** On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions // IEE Trans. Comp. 1976. V. C-25. P. 1175—1179.
7. **Sturges H. A.** The choice of a class interval // J. American Statistical Assoc. 1926. P. 65—66.
8. **Heinhold I., Gaede K.** Ingenieur statistic. Munchen: Wien, Springer Verlag, 1964.

Дата принятия 06.08.2015 г.