Тр. V международной конференции "Актуальные проблемы электронного приборостроения" АПЭП-2000. Новосибирск, 2000. - Т. 6. - С. 17-20.

# О НАХОЖДЕНИИ ПАРАМЕТРА РАЗМЫТОСТИ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК ФУНКЦИИ ПЛОТНОСТИ[1]

Б.Ю. Лемешко, С.Н. Постовалов, А.В. Французов

Новосибирский государственный технический университет
Новосибирск, Россия. E-mail: headrd@ fpm.ami.nstu.ru

**Аннотация.** Предложен новый метод оценивания параметра размытости ядерных оценок плотности, основанный на близости ядерной оценки функции распределения к эмпирической функции распределения. Методами статистического моделирования проведено сравнение полученных оценок с “оптимальными”.

**Постановка задачи.** Для непараметрического оценивания функции плотности используются оценки Розенблата-Парзена [1], которые имеют вид

$$p_n(x) = \frac{1}{n\lambda_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{\lambda_n}\right) \qquad (1)$$

где $x_i$, $i = 1, ..., n$ – выборка наблюдений одномерной непрерывной случайной величины, $\lambda_n$ – параметр размытости, а $\varphi(u)$ – колоколообразная (ядерная) функция, удовлетворяющая следующим условиям регулярности:

$$\varphi(u) = \varphi(-u);\ 0 \le \varphi(u) \le \infty;\ \int \varphi(u)du = 1;\ \int u^2\varphi(u)du = 1; \qquad (2)$$
$$\int u^m\varphi(u)du < \infty;\ 0 \le m < \infty.$$

Тогда ядерная оценка функции распределения будет иметь вид:

$$P_n(x) = \frac{1}{n\lambda_n}\sum_{i=1}^{n}\Phi\left(\frac{x-x_i}{\lambda_n}\right), \qquad (3)$$

где $\Phi(u) = \int_{-\infty}^{u}\varphi(x)dx$.

Из условий регулярности (2) следует, что в качестве ядерной функции $\varphi(u)$ можно рассматривать функцию плотности любого симметричного распределения с конечными начальными моментами, например, нормального распределения. В [2] рассматривается квадратичная ядерная функция, обладающая наилучшими свойствами при минимизации среднеквадратичной ошибки аппроксимации:

$$\varphi_1(u) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3}{20\sqrt{5}}u^2, & \text{если } |u| \le \sqrt{5}; \\ 0, & \text{если } |u| > \sqrt{5}. \end{cases} \qquad (4)$$

На вид непараметрических оценок функции плотности существенное влияние оказывает параметр размытости $\lambda_n$, и, таким образом, существует проблема нахождения оптимального значения параметра.

[1] Работа поддержана Российским фондом фундаментальных исследований (проект № 00-01-00913)

**Методы решения.** Если исходить из условий минимума среднеквадратической ошибки аппроксимации

$$J = M\left\{\int [f(x) - p_n(x)]^2 dx\right\},$$

где $f(x)$ - функция плотности распределения выборочных наблюдений, то оптимальная оценка параметра размытости будет иметь вид [2]:

$$\lambda^* = \left[\frac{\|\varphi\|^2}{n\|f''\|^2}\right]^{1/5} \qquad (5)$$

Так, например, если наблюдается нормальный закон распределения, то оптимальное значение $\lambda^*$ равно $\left[\frac{8\sqrt{\pi}}{5n\sqrt{5}}\right]^{1/5}$, а для экспоненциального $\lambda^* = \left[\frac{6}{5n\sqrt{5}}\right]^{1/5}$.

В выражении (5) присутствует вторая производная функции плотности наблюдаемой случайной величины, которая как раз и не известна. При $n \to \infty$ параметр размытости $\lambda_n^* \to n^{-1/5}$, поэтому иногда рекомендуют в качестве параметра размытости $\tilde{\lambda}'_n = n^{-1/5}$.

Несмотря на привлекательные асимптотические свойства, оценка (5) обладает следующим недостатком. При конечных объемах выборки в случае ограниченности области определения наблюдаемой случайной величины возможно значительное отличие ядерной оценки функции распределения от эмпирической функции распределения на границах области (на “хвостах” распределения). На рис. 1 приведен следующий пример. Для смоделированной в соответствии с экспоненциальным распределением выборки из 100 наблюдений построена ядерная оценка функции распределения с квадратичным ядром и параметром размытости $\lambda_n^* = 0.31511$, вычисленным по формуле (5). Цифрой "1" обозначена истинная функция распределения, цифрой "2" – эмпирическая и цифрой "3" – ядерная оценка.

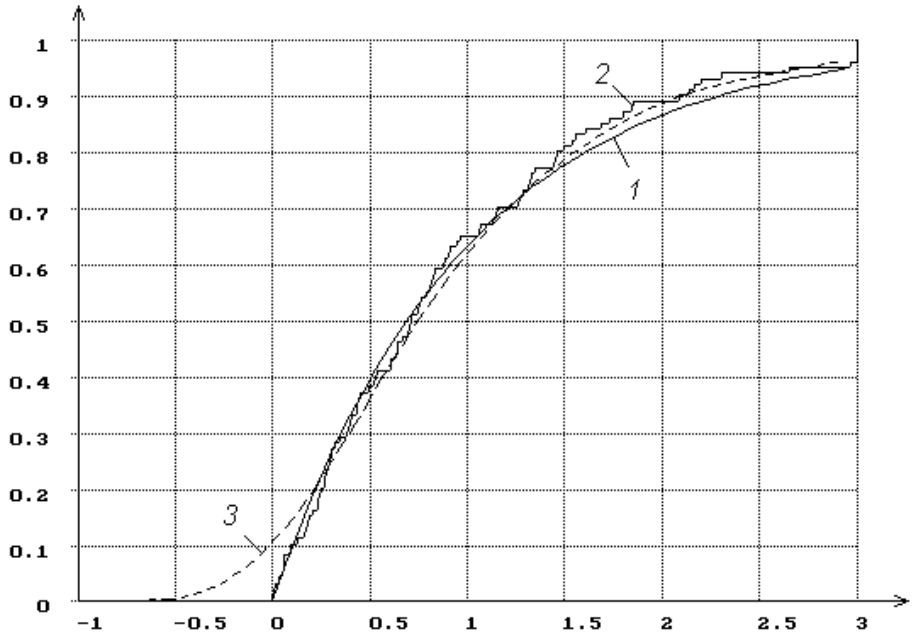


Рис. 1

На этом примере хорошо видно насколько сильно отличается ядерная оценка от истинного распределения. С ростом объема выборки это отличие уменьшается, но достаточно медленно. Например, для того чтобы ядерная оценка не заходила за ноль более чем на $\varepsilon$, необходимо, чтобы в выборке было около $(\sqrt{5}/\varepsilon)^5$ наблюдений.

Поэтому нами предложен другой метод оценивания параметра размытости ядерной оценки плотности, в основу которого положены следующие требования:

1. *Минимальное отклонение оценки от эмпирического распределения:* функция распределения, построенная по непараметрической оценке функции плотности, не должна выходить за допустимый интервал, определяемый критерием Колмогорова, т.е.

$$d(\lambda) = \sup_x \left| \int P_n(x) - F_n(x) \right| < \frac{c_\alpha}{\sqrt{n}} \qquad (6)$$

где $c_\alpha$ – квантиль распределения Колмогорова порядка $\alpha$. В случае истинности выражения (6), гипотеза о согласии *ядерной оценки с наблюдаемой выборкой* по критерию Колмогорова не отвергается с уровнем значимости $\alpha$.

2. *Максимальная гладкость:* оценка функции плотности должна иметь минимальное число максимумов.

$$\hat{\lambda} = \min_{\lambda < \lambda_{c_\alpha}} m(\lambda), \qquad (7)$$

где $m(\lambda)$ - число максимумов оценки функции плотности, а $\lambda_{c_\alpha}$ такое, что $d(\lambda_{c_\alpha}) = \frac{c_\alpha}{\sqrt{n}}$.

Алгоритм работает следующим образом. Выбирается начальное приближение и вычисляется допустимая область (6) для параметра размытости. Если начальное приближение лежит вне допустимой области, то ищется такое значение, которое принадлежит допустимой области. Далее в этой области ищется значение параметра размытости, при котором оценка функции плотности будет иметь минимальное число максимумов. Найденное значение параметра считается оптимальным в смысле предложенного алгоритма.

Одной из проблем применения алгоритма является выбор $c_\alpha$. Ниже приведена таблица, показывающая зависимость параметра размытости от $c_\alpha$ при различных объемах выборок случайных величин, распределенных по нормальному закону.

Таблица

**Значения параметра размытости в зависимости от n**

| $n$ | $\lambda_{c_\alpha}$ | | | $\lambda^*$ | $\lambda'$ |
|---|---|---|---|---|---|
| | $c_\alpha = 1.0$ | $c_\alpha = 1.3$ | $c_\alpha = 1.8$ | | |
| 10 | 0.643256 | 0.641016 | 0.639496 | 0.661671 | 0.630957 |
| 30 | 0.526134 | 0.521254 | 0.527014 | 0.531151 | 0.506496 |
| 50 | 0.479283 | 0.476743 | 0.479643 | 0.479566 | 0.457305 |
| 70 | 0.460167 | 0.454804 | 0.450064 | 0.448355 | 0.427544 |
| 100 | 0.434696 | 0.432111 | 0.424089 | 0.417486 | 0.398107 |
| 150 | 0.421704 | 0.417003 | 0.425492 | 0.384967 | 0.367098 |

Другой проблемой алгоритма является многоэкстремальность функции максимумов $m(\lambda)$, пример которой приведен на рис. 2.

**Полученные результаты.** Данный алгоритм позволяет получать более точные оценки функции плотности для односторонних законов распределений (например, для экспоненциального).

На рис. 3 и 4 показаны различные оценки параметра размытости в зависимости от объема выборки случайных величин. На рис. 3 – для нормального закона, на рис. 4 – для экспоненциального. Как видно из графиков, для нормального закона порядок сходимости функции $\lambda_n$ к нулю, полученной по предложенному алгоритму, несколько меньше, чем для оп-

тимальных оценок (5). Для экспоненциального закона наоборот – порядок сходимости функции $\lambda_n$ к нулю, полученной с использованием предложенного алгоритма, оказывается выше, чем для оптимальных оценок (5), что обусловлено более сильным влиянием ограничения (6) на выбор параметра размытости.

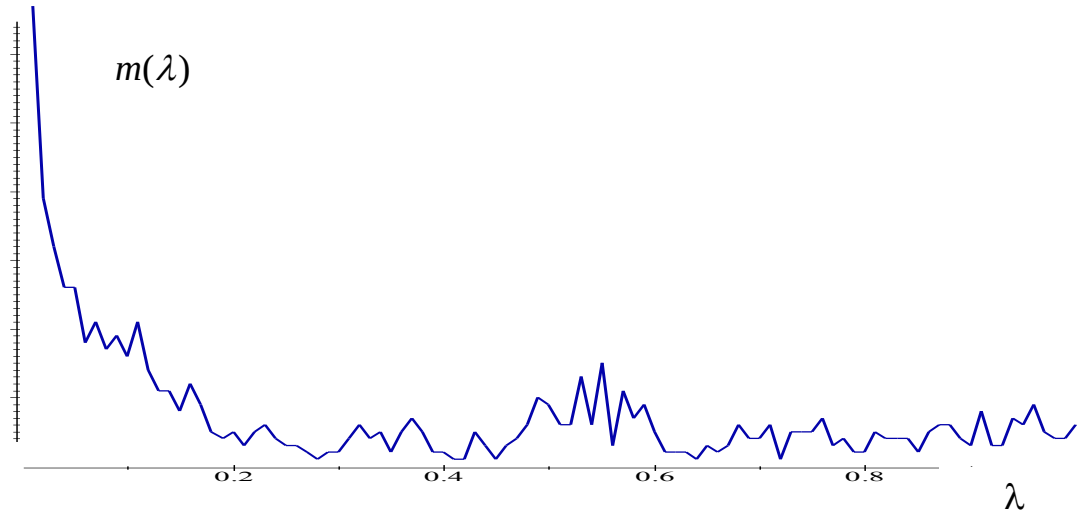


Рис. 2

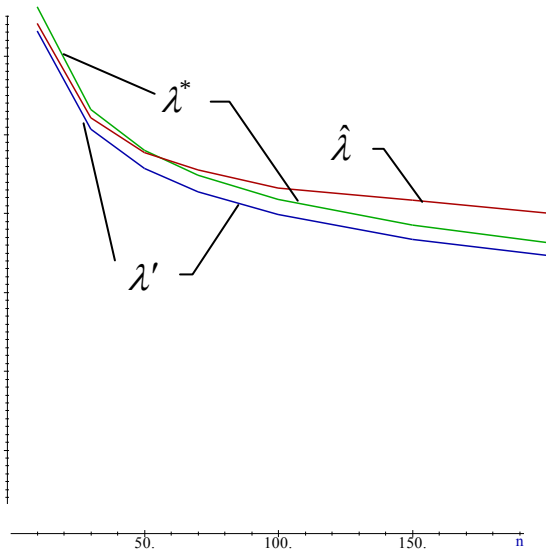


Рис. 3.

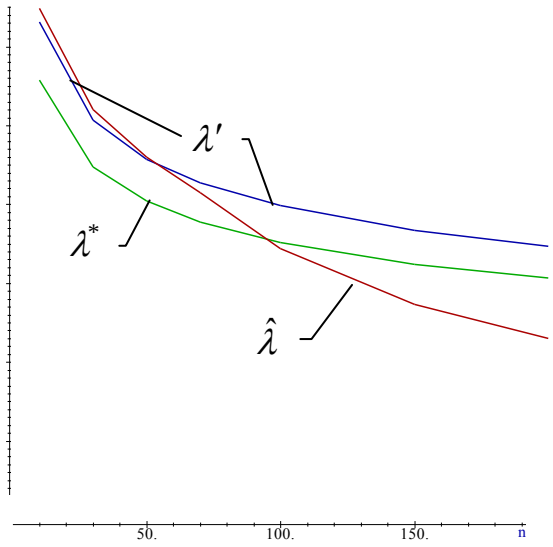


Рис. 4.

Таким образом, предложенный алгоритм вычисления параметра размытости непараметрической оценки плотности распределения непосредственно по наблюдаемой выборке позволяет получать достаточно гладкие оценки функции плотности. Соответствующие оценки интегральной функции распределения оказываются предпочтительными в тех случаях, когда область определения наблюдаемой случайной величины ограничена и функция плотности на границе положительна.

ЛИТЕРАТУРА

1. Parzen E. On the estimation of probability density function and the mode // Ann. Math. Stat., 1962. – Vol. 33. – P.1065-1076.
2. Епаненчиков В.А. Непараметрическая оценка многомерной плотности вероятности. Теория вероятностей и ее применения, 1969. – Т.14. – № 1. – С. 156-161.