УДК 519.24

Е.Г. КРАСНОГИР

О РАСХОЖДЕНИИ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ С ИСТИННОЙ ПЛОТНОСТЬЮ, МИНИМАЛЬНОМ ПО РАССТОЯНИЮ ХЕЛЛИНГЕРА

We propose and analyze special error criterion using for optimal bandwidth choice in kernel density estimation.

Одним из наиболее известных и часто используемых подходов при непараметрическом оценивании плотности вероятностей является ядерное оценивание - способ найти функцию плотности, не прибегая к параметрической модели. Такое оценивание предполагает меньшие ограничения, чем в параметрическом случае. При наличии соответствующего количества данных ядерные оценки часто выявляют особенности, которые незаметны при использовании других методов.

Пусть $X_1, \ldots, X_n$ - случайная выборка объема n из распределения с неизвестной плотностью $f(x)$. Ядерная оценка этой функции задается формулой

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

где функция K - это ядро, а h - параметр размытости. Традиционно предполагается [1], что K является симметричной плотностью распределения вероятностей, т. е.

$$\int_{-\infty}^{+\infty} K(x)dx = 1,\ K(x) \geq 0,\ \int_{-\infty}^{+\infty} xK(x)dx = 0.$$

Относительно h предполагается, что $h \to 0$, а $nh \to \infty$ при $n \to \infty$.

Хотя непараметрическое ядерное оценивание сегодня является стандартным способом анализа различных данных, до сих пор не разрешен вопрос, как достичь качественного оценивания и какой из параметров взять в качестве оптимального. Известно [1], что выбор ядра существенно не влияет на качество оценки. Основной проблемой в ядерном оценивании является выбор параметра размытости, который рассматривается как наиболее важный фактор, определяющий характерные черты ядерной оценки и ее вид. В последние годы было опубликовано множество исследований, посвященных оптимальному выбору параметра, основывающихся на различных критериях расхождения между непараметрической оценкой исследуемой функции и самой функцией.

В данной работе мы рассмотрим величину отклонения непараметрической оценки плотности распределения от истинной функции плотности, а затем на ее основе найдем оптимальное значение A.

Расхождение между непараметрической оценкой плотности и ее истинным значением будем измерять при помощи расстояния Хеллингера **[2]**

$$D(f_h(x), p) = E\int\left(f_h^{1/p}(x) - f^{1/p}(x)\right)^p dx,\ p \in Z,\ p > 1.$$

Введем следующие обозначения:

$$\nu_2(K) = \int K^2(u)du\ ,\ \mu_r(K) = \int u^r K(u)du.$$

Теорема. *Если* $\mu_3(K) = 0$, *существуют и ограничены производные функции* $f(x)$ *до четвертого порядка включительно,* $\to 0$, *а* $nh \to \infty$ *при* $n \to \infty$, *то*

$$D(f_h(x), 2) = \frac{h^4\mu_2^2(K)}{16}\int\frac{\left(f''(x)\right)^2}{f(x)}dx + \frac{\nu_2(K)}{4nh} + O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right),$$

$$D(f_h(x), p) = O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right),\ p > 2,\ p \in Z.$$

Доказательство. Математическое ожидание и дисперсию функции $f_h(x)$ по формуле Тейлора можно записать в виде

$$E(f_h(x)) = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + \frac{h^4}{24}f^{(4)}(x)\mu_4(K) + o\left(h^4\right),$$

$$\mathrm{Var}(f_h(x)) = \frac{1}{nh}f(x)\nu_2(K) + O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right).$$

Представим теперь $f_h(x)$ в виде $f_h(x) = E(f_h(x)) + \xi\sqrt{\mathrm{Var}(f_h(x))}$, где ξ - случайная величина с нулевым средним и дисперсией 1. Отсюда запишем

$$f_h(x) = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + \frac{h^4}{24}f^{(4)}(x)\mu_4(K) +$$

$$+\xi\left(\frac{1}{nh}f(x)\nu_2(K)+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right)\right)^{1/2}=$$

$$=f(x)\left(1+\frac{h^2}{2f(x)}f''(x)\mu_2(K)+\frac{h^4}{24f(x)}f^{(4)}(x)\mu_4(K)+\right.$$

$$\left.+\xi\left(\frac{1}{nh}\frac{\nu_2(K)}{f(x)}+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right)\right)^{1/2}\right).$$

Далее, по формуле Тейлора

$$f_h^{\alpha}(x)=f^{\alpha}(x)\left(1+\frac{\alpha h^2}{2f(x)}f''(x)\mu_2(K)+\frac{\alpha h^4}{24f(x)}f^{(4)}(x)\mu_4(K)+\right.$$

$$+\alpha\xi\left(\frac{1}{nh}\frac{\nu_2(K)}{f(x)}\right)^{1/2}+\frac{\alpha(\alpha-1)h^4}{8f^2(x)}\left(f''(x)\mu_2(K)\right)^2+$$

$$\left.+\xi^2\frac{\alpha(\alpha-1)}{2nh}\frac{\nu_2(K)}{f(x)}\right)+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right).$$

Тогда

$$E(f_h^{\alpha}(x))=f^{\alpha}(x)\left(1+\frac{\alpha h^2}{2f(x)}f''(x)\mu_2(K)+\frac{\alpha h^4}{24f(x)}f^{(4)}(x)\mu_4(K)+\right.$$

$$\left.+\frac{\alpha(\alpha-1)h^4}{8f^2(x)}\left(f''(x)\mu_2(K)\right)^2+\frac{\alpha(\alpha-1)}{2nh}\frac{\nu_2(K)}{f(x)}\right)+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right).$$

Домножим последнее выражение на $f^{1-\alpha}(x)$ и проинтегрируем:

$$\int E\left(f_h^{\alpha}(x)f^{1-\alpha}(x)\right)dx=\left(1+\frac{\alpha h^2\mu_2(K)}{2}\int f''(x)\,dx+\right.$$

$$+\frac{\alpha h^4\mu_4(K)}{24}\int f^{(4)}(x)\,dx+\frac{\alpha(\alpha-1)h^4\mu_2^2(K)}{8}\int\frac{\left(f''(x)\right)^2}{f(x)}dx+$$

$$\left.+\frac{\alpha(\alpha-1)\nu_2(K)}{2nh}\right)+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right).$$

По формуле бинома Ньютона

$$D(f_h(x),p)=E\int\left(f_h^{1/p}(x)-f^{1/p}(x)\right)^p dx=$$

$$=\sum_{m=0}^{p}C_p^m(-1)^m\int E\left(f_h^{1-m/p}(x)f^{m/p}(x)\right)dx=\sum_{m=0}^{p}C_p^m(-1)^m+$$

$$+\left(\frac{h^2\mu_2(K)}{2}\int f''(x)\,dx+\frac{h^4\mu_4(K)}{24}\int f^{(4)}(x)\,dx\right)\sum_{m=0}^{p-1}C_p^m(-1)^m\left(1-\frac{m}{p}\right)+$$

$$+\left(\frac{h^4\mu_2^2(K)}{8}\int\frac{\left(f''(x)\right)^2}{f(x)}dx+\frac{\nu_2(K)}{2nh}\right)\sum_{m=1}^{p-1}C_p^m(-1)^{m+1}\frac{m}{p}\left(1-\frac{m}{p}\right)+$$

$$+O\left(\frac{1}{n}\right)+O\left(\frac{h^2}{n}\right).$$

Первое слагаемое в этом выражении равно 0 для $p > 1$. Рассмотрим второе слагаемое:

$$\sum_{m=0}^{p-1} C_p^m (-1)^m \left(1-\frac{m}{p}\right) = \sum_{m=0}^{p-1} (-1)^m \frac{(p-1)!}{m!(p-m-1)!} = \sum_{m=0}^{p-1} C_{p-1}^m (-1)^m.$$

Это выражение равно 0 при $p \geq 2$. Третье слагаемое:

$$\sum_{m=1}^{p-1} C_p^m (-1)^{m+1} \frac{m}{p}\left(1-\frac{m}{p}\right) = \frac{p-1}{p}\sum_{m=1}^{p-1} (-1)^{m+1} \frac{(p-2)!}{(m-1)!(p-m-1)!} =$$

$$= \frac{p-1}{p}\sum_{m=0}^{p-2} C_{p-2}^m (-1)^m.$$

Равенство этого выражения нулю выполняется при $p \geq 3$.

Таким образом, при $p \geq 3$ выполняется

$$D(f_h(x), p) = O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right).$$

При $p = 2$ третье слагаемое $\frac{p-1}{p}\sum_{m=0}^{p-2} C_{p-2}^m (-1)^m = \frac{1}{2}$. Отсюда

$$D(f_h(x), 2) = \frac{h^4 \mu_2^2(K)}{16}\int \frac{\left(f''(x)\right)^2}{f(x)}dx + \frac{\nu_2(K)}{4nh} + O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right).$$

Теорема доказана.

Следствие. Для $p = 2$ при выполнении условий теоремы минимальная величина расстояния Хеллингера достигается, если

$$h = \left(\frac{\nu_2(K)}{\mu_2^2(K)\int_{-\infty}^{+\infty} \frac{(f''(x))^2}{f(x)}dx}\right)^{1/5} n^{-1/5}.$$

В заключение заметим, что в случае выполнения условий теоремы полученное в работе [3] выражение

$$D(f_h(x), 2) = \frac{1}{4}E\int \left(f_h(x) - f(x)\right)^2 dx =$$

$$= \frac{h^4 \mu_2^2(K)}{16}\int \left(f''(x)\right)^2 dx + \frac{\nu_2(K)}{4nh} + O\left(\frac{1}{n}\right) + O\left(\frac{h^2}{n}\right)$$

справедливо только при наличии дополнительного условия

$$\int \left(f''(x)\right)^2 dx = \int \frac{\left(f''(x)\right)^2}{f(x)}dx.$$

Таким образом, мы получили простую явную формулу для нахождения оптимального параметра размытости, которая может быть использована в задаче ядерного непараметрического оценивания неизвестной плотности распределения вероятностей наряду с другими известными формулами и алгоритмами. При практическом применении основным недостатком этой формулы является присутствие в ней оцениваемой функции, что требует дополнительных предположений о виде плотности распределения.

1. Hardle W., Muller M., Sperlich S., Werwatz A. Non- and Semiparametric Modelling. Heidelberg, 2004.

2. Basu A., Harris I. R., Basu S. // Handbook of Statistics. 1997. Vol. 15. P. 21.

3. Mugdadi Abdel-Razzaq // J. of Applied Statistical Science. 2004. Vol. 13. № 3. P. 231.

Поступила в редакцию 12.10.05.

Евгений Григорьевич Красногир - аспирант кафедры теории вероятностей и математической статистики. Научный руководитель - доктор физико-математических наук, профессор ГА. Медведев.