Записки научных
семинаров ПОМИ
Том 396, 2011 г.

В. Н. Солев

ОЦЕНКА ПЛОТНОСТИ ПО КОСВЕННЫМ НАБЛЮДЕНИЯМ

§1. Введение

Пусть τ – случайное разбиение вещественной прямой, точнее, такой конечный или счетный набор интервалов:

$$\tau = \{[a_j, b_j),\ j \in J \subset \mathbb{Z}\}, \tag{1}$$

что (с вероятностью 1)

$$\mathbb{R} = \bigcup_{j \in J} [a_j, b_j), \quad \text{и} \quad [a_j, b_j) \cap [a_i, b_i) = \varnothing, \quad \text{если } i \neq j. \tag{2}$$

Для простоты мы будем предполагать, что

$$0 \in [a_0, b_0) \text{ и } \ a_{j+1} = b_j.$$

Для точки x мы обозначим $\Delta(x) = [L(x), R(x))$ интервал разбиения τ, содержащий точку x. Пусть X – случайная величина с плотностью f. Обозначим $\Delta(X) = [L(X), R(X))$ интервал разбиения τ, содержащий X. Мы будем предполагать, что X и τ независимы. Предположим, что вместо X мы наблюдаем $\Delta(X)$. Цель настоящей работы – построение оценки $\widehat{f}_n$ неизвестной плотности f распределения X по косвенным наблюдениям.

Именно, пусть $\Delta_1, \ldots, \Delta_n$ – независимые копии случайного интервала $\Delta(X)$. Мы хотим построить оценку $\widehat{f}_n$ по наблюдениям $\Delta_1, \ldots, \Delta_n$. Здесь

$$\Delta_j = [L_j, R_j)\,.$$

Эта проблема исследовалась в известной работе [4] и в большом числе работ других исследователей. В настоящей заметке мы придерживаемся подхода, изложенного в статьях [2] и [3].

Ключевые слова: цензурированные наблюдения, непараметрическая оценка, случайное разбиение.

Работа поддержана грантами РФФИ 11-01-00577, РФФИ-ННИО 09-01-91331, НШ-4472.2010.1.

§2. Плотность случайного вектора $[L(x), R(x))$

Мы предположим, что распределение случайного вектора (a_j, b_j) имеет плотность $p_j(u,v)$. Ясно, что,

$$p_j(u,v) = p_j(u,v)\mathbf{1}_{(u,\infty)}(v),$$

где $\mathbf{1}_A(x)$ – индикаторная функция множества A. Пусть x – фиксированная точка вещественной прямой $\mathbb{R}$. Поскольку

$$\mathbf{P}\left\{x \in [a_j, b_j)\right\} = \iint\limits_{u<v} \mathbf{1}_{[u,v)}(x)\, p_j(u,v)\, du\, dv,$$

то

$$\iint\limits_{u<v} \sum_{j\in J} \left\{\mathbf{1}_{[u,v)}(x)\, p_j(u,v)\right\}\, du\, dv = 1.$$

Поэтому при фиксированном x почти всюду

$$\sum_{j\in J} \left\{\mathbf{1}_{[u,v)}(x)\, p_j(u,v)\right\} < \infty. \tag{3}$$

Пусть $\psi(u,v)$ – неотрицательная функция. Сосчитаем величину $\mathbf{E}\,\psi\,(L(x), R(x))$. Получаем

$$\begin{aligned}\mathbf{E}\,\psi\,(L(x), R(x)) &= \mathbf{E}\Big[\sum_{j\in J} \psi\,(a_j, b_j)\,\mathbf{1}_{[a_j,b_j)}(x)\Big] \\ &= \sum_{j\in J} \mathbf{E}\ \psi\,(a_j, b_j)\,\mathbf{1}_{[a_j,b_j)}(x).\end{aligned}$$

Ясно, что

$$\mathbf{E}\,\psi\,(a_j, b_j)\,\mathbf{1}_{[a_j,b_j)}(x) = \iint\limits_{u<v} \psi(u,v)\mathbf{1}_{[u,v)}(x)\, p_k(u,v)\, du\, dv.$$

Поэтому,

$$\mathbf{E}\,\psi\,(L(x), R(x)) = \iint\limits_{u<v} \psi(u,v)\Big\{\sum_{j\in J} p_j(u,v)\mathbf{1}_{[u,v)}(x)\Big\}\, du\, dv. \tag{4}$$

Таким образом, так как (см. (3)) для фиксированного x и почти всех $u \leqslant x < v$ величина

$$\sum_{j\in J} p_j(u,v) < \infty, \tag{5}$$

то мы получаем, что случайный вектор $[L(x), R(x))$ имеет плотность

$$p_x(u,v) = \sum_{j\in\, J} \{p_j(u,v)\mathbf{1}_{[u,v)}(x)\} \tag{6}$$

Очевидно, что

$$p_x(u,v) = p\,(u,v)\mathbf{1}_{[u,v)}(x),$$

где функция $p\,(u,v)$ не зависит от x. Функция $p\,(u,v)$ называется базовой плотностью распределения τ. Далее будет предполагаться, что $p\,(u,v) = 0$ при $u \geqslant v$. Заметим, что функция $p\,(u,v)$, вообще говоря, не является плотностью, однако, при любом x

$$\iint\limits_{u\,\leqslant x\,<v} p\,(u,v)\,du\,dv = 1. \tag{7}$$

Приведем простой пример такой базовой плотности

$$p\,(u,v) = e^{-(v-u)}. \tag{8}$$

Фактически мы имеем дело со случайным процессом $\Delta(x) = [L(x), R(x))$, $x \in \mathbb{R}$, таким, что для любого x

$$x \in [L(x), R(x))\,, \tag{9}$$

и для любых $a \leqslant x < y < b$ события

$$\{\Delta(x) \supset [a,b)\} \quad \text{и} \quad \{\Delta(y) \supset [a,b)\} \quad \text{совпадают.} \tag{10}$$

При этом случайный вектор $(L(x), R(x))$ имеет плотность

$$p_x(u,v) = p\,(u,v)\mathbf{1}_{[u,v)}(x).$$

§3. Плотность случайного вектора $[L(X), R(X))$

Теперь предположим, что случайная величина X и разбиение τ независимы. Пусть $F(x)$ – функция распределения случайной величины X, а $f(x)$ – ее плотность. Для неотрицательной функции $\psi(u,v)$ сосчитаем $\mathbf{E}\,\psi\,(L(X), R(X))$. Поскольку X и τ независимы, получаем

$$\mathbf{E}\,\psi\,(L(X), R(X)) = \int \{\mathbf{E}\,\psi\,(L(x), R(x))\}\; f(x)\,dx.$$

Так как

$$\mathbf{E}\,\psi\,(L(x), R(x)) = \iint\limits_{u<v} \psi\,(u,v)\,p\,(u,v)\mathbf{1}_{[u,v)}(x)\,du\,dv,$$

и

$$\int \mathbf{1}_{[u,v)}(x)\, f(x)\, dx = F(v) - F(u),$$

мы выводим, что

$$\mathbf{E}\, \psi\left(L(X), R(X)\right) = \iint\limits_{u<v} \psi\left(u,v\right) p\left(u,v\right) \left(F(v) - F(u)\right)\, du\, dv. \qquad (11)$$

Таким образом, для плотности $k(u,v)$ случайного вектора $[L(X), R(X))$ получаем соотношение

$$k(u,v) = p\left(u,v\right) \left(F(v) - F(u)\right) = p_*(u,v)\, \frac{F(v) - F(u)}{v-u}, \qquad (12)$$

где

$$p_*(u,v) = p\left(u,v\right)(v-u). \qquad (13)$$

§4. Точечная оценка плотности

Пусть $X, X_1, \ldots, X_n, \ldots$ – независимые одинаково распределенные случайные величины с общей плотностью f и $\tau, \tau_1, \ldots, \tau_n, \ldots$ – независимые между собой и независимые от $X_1, \ldots, X_n, \ldots$ случайные разбиения с общей базовой плотностью $p\left(u,v\right)$. Предположим, что случайные величины $X_1, \ldots, X_n, \ldots$ недоступны наблюдению. Вместо них мы наблюдаем случайные вектора $W_j = (L_j(X_j), R_j(X_j))$. Здесь мы обозначаем $[L_j(x), R_j(x))$ интервал разбиения τ_j, содержащий x. Проблема состоит в оценивании неизвестной функции f по наблюдениям $W_1, \ldots, W_n$, когда базовая плотность $p\left(u,v\right)$ известна.

Мы будем предполагать, что функция f имеет компактный носитель:

$$f(x) = 0 \ \text{ при } \ |x| \geqslant R. \qquad (14)$$

Для точки $x \in \mathbb{R}$ и $\varepsilon > 0$ обозначим

$$B(x;\varepsilon) = \left\{(u,v) :\ x - \varepsilon \leqslant u \leqslant x - \varepsilon/2,\ x + \varepsilon/2 \leqslant v \leqslant x + \varepsilon\right\}. \qquad (15)$$

Наши предположения относительно базовой плотности $p\left(u,v\right)$ касаются ее поведения вблизи диагонали $v = u$: при всех x и достаточно малых $\varepsilon > 0$

$$\iint\limits_{B(x;\varepsilon)} p\left(u,v\right) du\, dv \geqslant C\varepsilon^2. \qquad (16)$$

Заметим, что условие (16) выполнено, если базовая плотность отделена от нуля в окрестности диагонали $v = u$: при некоторых δ, $C_1 > 0$

$$p\,(u,v) \geqslant C_1 \quad \text{при} \quad 0 < v - u \leqslant \delta.$$

Обозначим

$$A(x;\varepsilon) = \{x - \varepsilon \leqslant L(X) \leqslant x - \varepsilon/2, x + \varepsilon/2 \leqslant R(X) \leqslant x + \varepsilon\}\,. \tag{17}$$

Очевидно,

$$\mathbf{P}\,\{A(x;\varepsilon)\} = \iint\limits_{B(x;\varepsilon)} p_*(u,v)\,\frac{F(v) - F(u)}{v - u}\,du\,dv.$$

и для малых ε и гладкой функции f

$$\mathbf{P}\,\{A(x;\varepsilon)\} \approx f(x) \iint\limits_{B(x;\varepsilon)} p_*(u,v)\,du\,dv.$$

Последнее соотношение подсказывает использовать для оценивания значения функции f в точке x отношение

$$\widehat{f}_{n;\,\varepsilon}(x) = \frac{\mathbf{P}_n\,\{A(x;\varepsilon)\}}{\mu(x;\varepsilon)},$$

где $\mathbf{P}_n\,\{A(x;\varepsilon)\}$ – эмпирическая версия величины $\mathbf{P}\,\{A(x;\varepsilon)\}$,

$$\mathbf{P}_n\,\{A(x;\varepsilon)\} = \frac{1}{n}\,\#\,\{j:\,(L_j(X_j), R_j(X_j)) \in B(x;\varepsilon)\}\,, \tag{18}$$

и

$$\mu(x;\varepsilon) = \iint\limits_{B(x;\varepsilon)} p_*(u,v)\,du\,dv. \tag{19}$$

Здесь $\#\,\{A\}$ – число элементов A.

Пусть далее

$$f_\varepsilon(x) = \frac{\mathbf{P}\,\{A(x;\varepsilon)\}}{\mu(x;\varepsilon)} = \frac{1}{\mu(x;\varepsilon)} \iint\limits_{B(x;\varepsilon)} p_*(u,v)\,\frac{F(v) - F(u)}{v - u}\,du\,dv. \tag{20}$$

Так что $\mathbf{E}\,\widehat{f}_n(x) = f_\varepsilon(x)$.

§5. Большие уклонения точечной оценки

В этом пункте мы дадим оценку сверху для вероятностей

$$\begin{aligned}\mathbf{P}\left\{\left|\widehat{f}_n(x)-f_\varepsilon(x)\right|>y\right\}&=\\&=\mathbf{P}\left\{\left|\mathbf{P}_n\left\{A(x;\varepsilon)\right\}-\mathbf{P}\left\{A(x;\varepsilon)\right\}\right|>y\mu(x;\varepsilon)\right\}.\end{aligned}\tag{21}$$

Приведем удобную для наших целей формулировку одного результата Массара, (см. [1], теорема 2). Нас будет интересовать оценка сверху для вероятности

$$\mathbf{P}\big\{\big|S-np\big|>nz\big\}.\tag{22}$$

где случайная величина S имеет биномиальное распределение $\mathscr{B}(n;p)$. Массар установил, что при $q=1-p$

$$\mathbf{P}\left\{S-np>nz\right\}\leqslant\exp\left\{-\frac{nz^2}{2\,(p+z/3)\,(q-z/3)}\right\},\quad 0\leqslant z\leqslant q\,,\tag{23}$$

и

$$\mathbf{P}\left\{S-np<-nz\right\}\leqslant\exp\left\{-\frac{nz^2}{2\,(p-z/3)\,(q+z/3)}\right\},\quad 0\leqslant z\leqslant p\,.\tag{24}$$

Нас будет интересовать случай, когда $p=p_n\to 0$ при $n\to\infty$. В этом случае несколько огрубляя оценку Массара и учитывая то обстоятельство, что

$$\begin{aligned}\mathbf{P}\left\{S-np>nz\right\}&=0\quad\text{при}\quad z>q\quad\text{и}\\ \mathbf{P}\left\{S-np<-nz\right\}&=0\quad\text{при } z>p,\end{aligned}\tag{25}$$

получаем, что

$$\mathbf{P}\left\{|S-np|>nz\right\}\leqslant\begin{cases}\exp\left\{-\frac{3nz}{8}\right\} & \text{при } z>p\,;\\ 2\exp\left\{-\frac{3nz^2}{8p}\right\} & \text{при } z\leqslant p\,.\end{cases}\tag{26}$$

Полагая $z=y\mu(x;\varepsilon)$, $p=\mathbf{P}\left\{A(x;\varepsilon)\right\}=f_\varepsilon(x)\mu(x;\varepsilon)$, получаем :

$$\begin{aligned}&\mathbf{P}\left\{\left|\widehat{f}_n(x)-f_\varepsilon(x)\right|>y\right\}\\ &\qquad\leqslant\begin{cases}\exp\left\{-\frac{3ny\mu(x;\varepsilon)}{8}\right\} & \text{при}\quad y>f_\varepsilon(x)\,;\\ 2\exp\left\{-\frac{3ny^2\mu(x;\varepsilon)}{8f_\varepsilon(x)}\right\} & \text{при}\quad y\leqslant f_\varepsilon(x).\end{cases}\end{aligned}\tag{27}$$

5.1. Оценка величины E $\left|\widehat{f}_n(x) - f_\varepsilon(x)\right|$. В этом пункте мы возьмем $\varepsilon = n^{-1/3}$. Заметим, что при условии (16) справедливо неравенство

$$\mu(x;\varepsilon) \geqslant C\varepsilon = Cn^{-1/3}. \tag{28}$$

Далее,

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| = \int_0^\infty \mathbf{P}\left\{\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| > y\right\} dy.$$

Обращаясь к (28), получаем

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \leqslant \int_0^\infty \exp\left\{-\frac{3ny\mu(x;\varepsilon)}{8}\right\} dy + 2\int_0^\infty \exp\left\{-\frac{3ny^2\mu(x;\varepsilon)}{8f_\varepsilon(x)}\right\} dy.$$

Поэтому, если $n\mu(x;\varepsilon) \to \infty$ при $n \to \infty$, то при достаточно большом n (например, при $\sqrt{n\mu(x;\varepsilon)} > 3/8$) получаем

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \leqslant K\frac{1}{\sqrt{n\mu(x;\varepsilon)}} \tag{29}$$

с константой

$$K = 2\pi\sqrt{f_\varepsilon(x)} + 1\,.$$

Таким образом, при условии (16) и $\varepsilon = n^{-1/3}$ при достаточно большом n

$$\mathbf{E}\left|\widehat{f}_n(x) - f_\varepsilon(x)\right| \leqslant K_1 n^{-1/3}, \quad K_1 = K/\sqrt{C}. \tag{30}$$

Следующая лемма является почти очевидным следствием неравенства (32). Поэтому она приводится без доказательства.

Лемма 5.1. *Предположим, что функция f имеет носитель, содержащийся в интервале $[-R, R]$, и удовлетворяет условию*

$$|f(x) - f(y)| \leqslant L|x - y|. \tag{31}$$

Предположим также, что $\varepsilon = n^{-1/3}$ и выполнено условие (16). *Тогда*

$$\mathbf{E}\left|\widehat{f}_n(x) - f(x)\right| \leqslant Mn^{-1/3}, \tag{32}$$

где константа M зависит лишь от L, R и C.

Мы сохраним обозначение $\widehat{f}_{n;\,\varepsilon}(x)$ для кусочно-постоянной функции

$$\begin{aligned} \widehat{f}_{n;\,\varepsilon}(x) &= \sum_{k\in\mathbb{Z}} \frac{\mathbf{P}_n\left\{A(x_k;\varepsilon)\right\}}{\mu(x_k;\varepsilon)} \mathbf{1}_{[x_k-\varepsilon,\,x_k+\varepsilon)}(x) \\ &= \sum_{k\in\mathbb{Z}} \widehat{f}_{n;\,\varepsilon}(x_k) \mathbf{1}_{[x_k-\varepsilon,\,x_k+\varepsilon)}(x), \end{aligned} \tag{33}$$

при $x_k = 2k\varepsilon$. Ясно, что

$$\widehat{f}_{n;\,\varepsilon}(x_k) = \frac{\mathbf{P}_n\left\{A(x_k;\varepsilon)\right\}}{\mu(x_k;\varepsilon)}. \tag{34}$$

Далее, пусть $\widetilde{f}_{\varepsilon}(x)$ – неслучайная кусочно-постоянной функция, определенная соотношением

$$\widetilde{f}_{\varepsilon}(x) = \sum_{k\in\mathbb{Z}} f(x_k) \mathbf{1}_{[x_k-\varepsilon,\,x_k+\varepsilon)}(x). \tag{35}$$

Теорема 5.2. *В условиях леммы* 5.1

$$\mathbf{E}\left\|\widehat{f}_n - f\right\|_1 \leqslant M_* n^{-1/3}, \tag{36}$$

где константа M_* *зависит лишь от* L, R *и* C.

Доказательство. Утверждение теоремы непосредственно следует из неравенства (32), неравенства

$$\left\|\widetilde{f}_{\varepsilon} - f\right\|_1 \leqslant 2L\varepsilon(2R+1),$$

и соотношения

$$\mathbf{E}\left\|\widehat{f}_n - \widetilde{f}_{\varepsilon}\right\|_1 = \sum_{k\in\mathbb{Z}} \int\limits_{x_k-\varepsilon}^{x_k+\varepsilon} \mathbf{E}\left|\widehat{f}_n(x_k) - \widetilde{f}(x_k)\right| dx.$$

Литература

1. P. Massart, *The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality.* — Ann. Probab. **18** (1990), 1269–1283.
2. C. Huber-Carol, F. Vonta, *Frailty models for arbitrarily censored and truncated data.* — Lifetime Data Analysis **10** (2004), 369–388.
3. C. Huber-Carol, V. Solev, F. Vonta, *Estmation of density for arbitrarily censored and truncated data.* — In: M. S. Nikulin, D. Commenges, C. Huber-Carol (Eds.) Probability, Statistics, and Modelling in Public Healt, Springer (Kluwer Acad. Publ.), New York (2006), pp. 246–265.
4. B. W. Turnbull, *The empirical distribution function with arbitrary grouped, censored and truncated data.* — J. Royal Statist. Soc. **38** (1976), 290–295.

Solev V. N. Estumation of density on indirect observation.

In this paper it is investigated the accuracy of the estimating of the unknown density in the L_1-space on indirect observation. We suggest a simple nonparametric estimator $\widehat{f}_n$ for unknown density f and under some appropriate conditions prove the consistency of this estimator.

С.-Петербургское отделение
Математического института
им. В. А. Стеклова РАН, Фонтанка 27,
Санкт-Петербург 191023, Россия
E-mail: vnsolev@gmail.com

Поступило 23 ноября 2011 г.