Vol. 35 (2015)
No. 3

数 学 杂 志
J. of Math. (PRC)

END 样本最近邻密度估计的强相合速度

兰冲锋 [1] , 吴群英 [2]

(1. 阜阳师范学院经济与管理学院, 安徽 阜阳 236037)

(2. 桂林理工大学理学院, 广西 桂林 541004)

摘 要: 本文研究了扩展负相依（END）样本最近邻密度估计的强相合性问题. 利用 END 序列的 Bernstein 型不等式和截尾的方法, 获得了 END 样本最近邻密度估计的强相合速度, 推广了 NA 样本和 ND 样本最近邻密度估计的相应结果.

关键词: END 序列; 最近邻密度估计; 强相合速度

MR(2010) **主题分类号**: 62G07 **中图分类号**: O212.7

文献标识码: A **文章编号**: 0255-7797(2015)03-0665-07

1 引言

设总体 X 的分布函数为 $F(x)$, 其对应的密度函数为 $f(x)$, $X_1, X_2, \cdots, X_n$ 是抽自该总体的 END 样本, 而 $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x)$ 是样本 $X_1, X_2, \cdots, X_n$ 的经验分布函数. 那么, 显然有 X_i $(i = 1, 2, \cdots, n)$ 与 X 是同分布的, 且 $X_1, X_2, \cdots, X_n$ 是 END 随机变量. 设 $\{k_n; n \geq 1\}$ 为给定的正整数列, 满足 $1 \leq k_n < n$. 令 $a_n(x)$ 为最小的正数 a 使得 $[x-a, x+a]$ 中至少包含 $X_1, X_2, \cdots, X_n$ 中的 k_n 个, 则 X 的密度函数 $f(x)$ 的最近邻密度估计为

$$f_n(x) = \frac{k_n}{2na_n(x)}.$$

概率密度估计和非参数非线性回归是非参数估计中两大问题. 而最近邻密度估计 (NN-估计) 是由 Loftsgarden 等 [1] 于 1965 年提出的, 它是一种比较常用的非参数概率密度估计的方法. 由于它的广泛应用, 此后很多著名学者都研究过它的收敛性质. 在独立样本情形, 文献 [1–5] 等对最近邻密度估计 $f_n(x)$ 的强、弱相合性和一致强、弱相合性以及收敛速度等做了比较深入的研究, 也都得出了比较好的结论. 在相依样本情形, 柴根象 [6] 对 φ 混合样本讨论了 $f_n(x)$ 的相合性和一致强相合性及其收敛速度; 杨善朝 [7] 就 NA 样本下最近邻密度估计的相合性及收敛速度做了深入的研究; 而文献 [8–11] 则将最近邻密度估计的相合性以及收敛速度推广到了 ND 序列.

为了得出本文的主要结论, 下面我们先给出 END 序列的定义:

定义 [12] 称随机变量 $\{X_n; n \geq 1\}$ 是 END（Extended Negatively Dependent）的, 若存在常数 $M > 0$, 使得

$$P(\bigcap_{i=1}^{n}(X_i \leq x_i)) \leq M\prod_{i=1}^{n} P(X_i \leq x_i),$$

*收稿日期: 2013-08-11 接收日期: 2013-12-30

基金项目: 国家自然科学基金资助 (11061012); 安徽省自然科学研究项目基金资助 (KJ2013Z265; 2014KJ006).

作者简介: 兰冲锋 (1981–), 男, 安徽阜阳, 讲师, 主要研究方向: 概率极限理论.

$$P(\bigcap_{i=1}^{n}(X_i > x_i)) \le M\prod_{i=1}^{n}P(X_i > x_i),$$

对每个 $n = 1, 2, \cdots$ 和所有的 $x_1, x_2, \cdots, x_n \in R$ 都成立.

END 序列的概念是刘 [12,13] 在研究相依重尾随机变量的偏差中首先提出来的. 文献 [12] 中例 4.1 表明, END 序列不仅反应了负相依结构, 而且在某种程度上体现了正相依结构, 它是一种非常广泛的相依随机变量序列. 比如, 当定义中的 $M = 1$ 时, END 序列就是 ND 序列. 显然, END 序列包含了独立序列, 而文献 [14] 举例说明了 NA 序列一定是 ND 序列, 但 ND 序列不一定是 NA 序列, 而 ND 序列又是 END 序列, 反之则不成立. 这说明了 END 序列是比独立序列、NA 序列和 ND 序列更弱的、更广泛的一种随机变量序列. 因此, 对 END 序列的研究在理论和实际应用中都是非常有意义的, 而将独立序列或 NA 序列的一些性质推广到 END 序列也是很有必要的. 沈 [15] 研究了 END 序列的概率不等式及其应用, 而对于 END 样本下的最近邻密度的估计问题, 则还未见文献报道. 基于此, 本文主要研究 END 样本最近邻密度估计的强相合速度问题, 在更弱的条件下, 得到了与 NA 序列相同的结论, 从而推广了文献 [7] 的结果.

本文用“$\ll$”表示通常的大“O”.

2 引理

为了证明本文的主要结论, 本节建立一个 END 序列的 Bernstein 型不等式, 并给出一些相关的引理.

引理 1 [12] 若随机变量 $\{X_n; n \ge 1\}$ 是 END 的, 则

(1) $\{g_i(X_i); i = 1, 2, \cdots\}$ 仍是 END 的, 其中 $g_i(\cdot), i = 1, 2, \cdots$ 均为单调递增或单调递减的函数;

(2) 对任意的 $n = 1, 2, \cdots$, 存在常数 $M > 0$, 使得

$$E\left(\prod_{i=1}^{n}X_i^+\right) \le M\prod_{i=1}^{n}EX_i^+.$$

由引理 1 立即可得下面的引理 2.

引理 2 设随机变量 $\{X_n; n \ge 1\}$ 是 END 序列, $t_1, t_2, \cdots, t_n$ 都是非正或者都是非负的实数, 对任意的 $n = 1, 2, \cdots$, 存在常数 $M > 0$ 使得

$$E\left(\exp\left(\sum_{i=1}^{n}t_iX_i\right)\right) \le M\prod_{i=1}^{n}E(\exp(t_iX_i)).$$

特别的, 对任意的 $t \in R$, 都有

$$E\left(\exp\left(\sum_{i=1}^{n}tX_i\right)\right) \le M\prod_{i=1}^{n}E(\exp(tX_i)).$$

引理 3 (Bernstein 不等式) 设随机变量 $\{X_n; n \ge 1\}$ 是 END 序列, $EX_i = 0$, $|X_i| \le b_i$ a.s. $(i = 1, 2, \cdots, n)$, $t > 0$ 为实数, 且满足 $t\max\limits_{1\le i\le n}b_i \le 1$, 则 $\forall\varepsilon > 0$, 有

$$P\left(\left|\sum_{i=1}^{n}X_i\right| > \varepsilon\right) \le 2M\exp\left\{-t\varepsilon + t^2\sum_{i=1}^{n}EX_i^2\right\}.$$

证 因 $Y_n \triangleq \sum\limits_{k=0}^{n} \frac{(tX_i)^k}{k!} \to e^{tX_i}$, $n \to \infty$, 由 $|tX_i| \le 1$ a.s. 得 $|Y_n| \le e$ a.s.. 所以由 Lebesgue 控制收敛定理得

$$\begin{aligned} E(e^{tX_i}) &= E\left(\lim_{n\to\infty}\sum_{k=0}^{n}\frac{(tX_i)^k}{k!}\right) = \lim_{n\to\infty}\sum_{k=0}^{n}\frac{E(tX_i)^k}{k!} = \sum_{k=0}^{\infty}\frac{E(tX_i)^k}{k!} \\ &\le 1 + E(tX_i)^2\sum_{k=2}^{\infty}\frac{1}{k!} \le 1 + t^2EX_i^2 \le e^{t^2EX_i^2}. \end{aligned}$$

由 Markov 不等式, 结合引理 2, 对 $\forall\varepsilon > 0$, $t > 0$, 对任意的 $n = 1, 2, \cdots$, 存在常数 $M > 0$ 使得

$$\begin{aligned} P\left(\sum_{i=1}^{n}X_i > \varepsilon\right) &= P\left(e^{t\sum\limits_{i=1}^{n}X_i} > e^{t\varepsilon}\right) \le e^{-t\varepsilon}Ee^{t\sum\limits_{i=1}^{n}X_i} \le e^{-t\varepsilon}M\prod_{i=1}^{n}Ee^{tX_i} \\ &\le Me^{-t\varepsilon}\prod_{i=1}^{n}e^{t^2EX_i^2} = M\exp\left\{-t\varepsilon + t^2\sum_{i=1}^{n}EX_i^2\right\}. \end{aligned}$$

同理, 以 $-X_i$ 代替上式中的 X_i 得

$$P\left(\sum_{i=1}^{n}(-X_i) > \varepsilon\right) = P\left(\sum_{i=1}^{n}X_i < -\varepsilon\right) \le M\exp\left\{-t\varepsilon + t^2\sum_{i=1}^{n}EX_i^2\right\}.$$

故

$$\begin{aligned} P\left(\left|\sum_{i=1}^{n}X_i\right| > \varepsilon\right) &= P\left(\sum_{i=1}^{n}X_i > \varepsilon\right) + P\left(\sum_{i=1}^{n}X_i < -\varepsilon\right) \\ &\le 2M\exp\left\{-t\varepsilon + t^2\sum_{i=1}^{n}EX_i^2\right\}. \end{aligned}$$

证毕.

引理 4 [16] (推广的 Borel-Cantelli 引理):

(i) 若 $\sum\limits_{n=1}^{\infty}P(A_n) < \infty$, 则 $P(A_n, \text{i.o.}) = 0$;

(ii) 若 $P(A_kA_m) \le P(A_k)P(A_m), k \ne m$, 且 $\sum\limits_{n=1}^{\infty}P(A_n) = \infty$, 则 $P(A_n, \text{i.o.}) = 1$.

3 主要结果及证明

定理 1 设随机变量 $\{X_n; n \ge 1\}$ 是 END 序列, 存在正数列 $\{q_n; n \ge 1\}$, 使得 k_n 和 q_n 满足

$$k_n \to \infty,\ q_n \to 0,\ \frac{k_n}{nq_n} \to 0,\ \frac{k_nq_n}{(n\log n)^{1/2}} \to \infty,$$

且 $f(x)$ 在 R 上满足 Lipschitz 条件, $f(x) > 0$, 则当 $n \to \infty$ 时, 有

$$|f_n(x) - f(x)| = o(q_n) \quad \text{a.s..}$$

推论 1 设随机变量 $\{X_n; n \geq 1\}$ 是 END 序列, $f(x)$ 在 R 上满足 Lipschitz 条件, 且 $f(x) > 0$, 若取

$$k_n = n^{3/4}(\log n)^{1/4}, \quad q_n = n^{-1/4}(\log n)^{1/4}\log\log n,$$

容易验证 k_n 和 q_n 满足定理 1 的条件, 则当 $n \to \infty$ 时, 有

$$|f_n(x) - f(x)| = o(n^{-1/4}(\log n)^{1/4}\log\log n) \text{ a.s..}$$

注 1 定理 1 在更弱的条件下, 获得了与 NA 序列下相同的结论.

注 2 由推论 1 可知, $f_n(x)$ 的强相合收敛速度几乎为 $n^{-1/4}$, 这一结论与 NA 序列下是相同的, 但是与独立情形的 $n^{-1/3}$ 还有一些差距.

定理 1 的证明 $\forall \varepsilon > 0$, 有

$$P(|f_n(x) - f(x)| > \varepsilon q_n) = P(f_n(x) > f(x) + \varepsilon q_n) + P(f_n(x) < f(x) - \varepsilon q_n).$$

当 $f(x) \leq \varepsilon q_n$ 时, $P(f_n(x) < f(x) - \varepsilon q_n) = 0$, 故估计 $P(f_n(x) < f(x) - \varepsilon q_n)$ 时只需考虑 $f(x) > \varepsilon q_n$ 的情况. 令

$$b_n(x) = \frac{k_n}{2n(f(x) + \varepsilon q_n)}, \ \ c_n(x) = \frac{k_n}{2n(f(x) - \varepsilon q_n/2)} \ \ (\text{此时要求} f(x) > \varepsilon q_n),$$

则由 $f_n(x)$ 的定义, 有

$$\begin{aligned}
A_x &\triangleq \{|f_n(x) - f(x)| > \varepsilon q_n\} \\
&= \{f_n(x) > f(x) + \varepsilon q_n\} \bigcup \{f_n(x) < f(x) - \varepsilon q_n\} \\
&\subset \{f_n(x) > f(x) + \varepsilon q_n\} \bigcup \{f_n(x) < f(x) - \varepsilon q_n/2\} \\
&= \{a_n(x) < b_n(x)\} \bigcup \{a_n(x) > c_n(x)\} \\
&\subset \left\{F_n(x + b_n(x)) - F_n(x - b_n(x)) \geq \frac{k_n}{n}\right\} \bigcup \left\{F_n(x + c_n(x)) - F_n(x - c_n(x)) \leq \frac{k_n}{n}\right\} \\
&\triangleq B_x + C_x, \qquad (3.1)
\end{aligned}$$

其中 $F_n(\cdot)$ 表示样本的经验分布函数.

由微分中值定理, 存在 $\theta_1 \in (x - b_n(x), x + b_n(x))$ 和 $\theta_2 \in (x - c_n(x), x + c_n(x))$, 使

$$F(x + b_n(x)) - F(x - b_n(x)) = 2b_n(x)f(\theta_1), \qquad (3.2)$$

$$F(x + c_n(x)) - F(x - c_n(x)) = 2c_n(x)f(\theta_2). \qquad (3.3)$$

故由 (3.1) 式中的 B_x 和 (3.2) 式可得

$$\begin{aligned}
&F_n(x + b_n(x)) - F_n(x - b_n(x)) - F(x + b_n(x)) + F(x - b_n(x)) \\
\geq\ & \frac{k_n}{n} - 2b_n(x)f(\theta_1) = \frac{k_n}{n}\frac{f(x) - f(\theta_1) + \varepsilon q_n}{f(x) + \varepsilon q_n}. \qquad (3.4)
\end{aligned}$$

由 (3.1) 式中的 C_x 和 (3.3) 式可得

$$\begin{aligned} & F_n(x+c_n(x))-F_n(x-c_n(x))-F(x+c_n(x))+F(x-c_n(x)) \\ \leq\ & \frac{k_n}{n}-2c_n(x)f(\theta_2)=\frac{k_n}{n}\frac{f(x)-f(\theta_2)-\varepsilon q_n/2}{f(x)-\varepsilon q_n/2}. \end{aligned} \tag{3.5}$$

从以上证明过程可以看出 (3.4) 和 (3.5) 式分别是 (3.1) 式中的 B_x 和 C_x, 因此它们是成立的. 由 $f(x)$ 在 R 上满足 Lipschitz 条件和 $f(x)>0$, 以及 $q_n\to 0$, $\frac{k_n}{nq_n}\to 0$ 知, 当 $n\to\infty$ 时, 存在一个常数 L, 使得

$$|f(x)-f(\theta_1)|\leq L|x-\theta_1|\leq Lb_n(x)=\frac{Lk_n}{2n(f(x)+\varepsilon q_n)}<\varepsilon q_n/2, \tag{3.6}$$

$$|f(x)-f(\theta_2)|\leq Lc_n(x)<\varepsilon q_n/4. \tag{3.7}$$

显然, 密度函数 $f(x)$ 是有界的, 不妨记 $M=\sup\limits_x f(x)<\infty$, 结合 (3.6) 式, 有

$$\frac{k_n}{n}\frac{f(x)-f(\theta_1)+\varepsilon q_n}{f(x)+\varepsilon q_n}\geq\frac{k_n}{n}\frac{-\varepsilon q_n/2+\varepsilon q_n}{f(x)+\varepsilon q_n}\geq\frac{k_nq_n}{n}\frac{\varepsilon}{4M}. \tag{3.8}$$

由 (3.7) 式同理可得

$$\frac{k_n}{n}\frac{f(x)-f(\theta_2)-\varepsilon q_n/2}{f(x)-\varepsilon q_n/2}\leq-\frac{k_nq_n}{n}\frac{\varepsilon}{4M}. \tag{3.9}$$

令 $\frac{\varepsilon}{8M}=u$, 由 (3.4) 和 (3.8) 式, 得

$$\begin{aligned} & B_x=\left\{F_n(x+b_n(x))-F_n(x-b_n(x))\geq\frac{k_n}{n}\right\} \\ =\ & \left\{F_n(x+b_n(x))-F_n(x-b_n(x))-F(x+b_n(x))+F(x-b_n(x))\geq\frac{k_n}{n}-2b_n(x)f(\theta_1)\right\} \\ \subset\ & \left\{F_n(x+b_n(x))-F_n(x-b_n(x))-F(x+b_n(x))+F(x-b_n(x))\geq\frac{2k_nq_n}{n}u\right\} \\ \subset\ & \left\{|F_n(x+b_n(x))-F(x+b_n(x))|\geq\frac{k_nq_n}{n}u\right\} \\ & \bigcup\left\{|F_n(x-b_n(x))-F(x-b_n(x))|\geq\frac{k_nq_n}{n}u\right\} \\ \triangleq\ & D_{1x}\bigcup D_{2x}. \end{aligned} \tag{3.10}$$

由 (3.5) 和 (3.9) 式, 得

$$\begin{aligned} & C_x=\left\{F_n(x+c_n(x))-F_n(x-c_n(x))\leq\frac{k_n}{n}\right\} \\ =\ & \left\{F_n(x+c_n(x))-F_n(x-c_n(x))-F(x+c_n(x))+F(x-c_n(x))\leq\frac{k_n}{n}-2c_n(x)f(\theta_2)\right\} \\ \subset\ & \left\{F_n(x+c_n(x))-F_n(x-c_n(x))-F(x+c_n(x))+F(x-c_n(x))\leq-\frac{2k_nq_n}{n}u\right\} \end{aligned}$$

$$\begin{aligned} \subset & \left\{|F_n(x+c_n(x))-F(x+c_n(x))| \geq \frac{k_nq_n}{n}u\right\} \\ & \bigcup\left\{|F_n(x-c_n(x))-F(x-c_n(x))| \geq \frac{k_nq_n}{n}u\right\} \\ \triangleq & D_{3x}\bigcup D_{4x}. \end{aligned} \tag{3.11}$$

由 (3.1)、(3.10) 和 (3.11) 式, 易得

$$A_x \subset D_{1x}\cup D_{2x}\cup D_{3x}\cup D_{4x}. \tag{3.12}$$

令 $X_i^c = I(X_i < x+b_n(x)) - EI(X_i < x+b_n(x))$, 则由引理 1 知, $X_1^c, X_2^c, \cdots, X_n^c$ 仍为 END 序列, 且 $EX_i^c = 0$, $|X_i^c| \leq 2$. 取 $t = \frac{k_nq_nu}{2n}$, 则当 $n\to\infty$ 时, 由条件 $q_n \to 0$ 以及 $\frac{k_n}{nq_n} \to 0$ 可知 $2t = \frac{k_nq_nu}{n} \to 0$, 故满足引理 3 的条件, 由引理 1 和条件 $\frac{k_nq_n}{(n\log n)^{1/2}} \to \infty$ 知

$$\begin{aligned} P(D_{1x}) &= P\left(\left|\sum_{i=1}^n X_i^c\right| \geq k_nq_nu\right) \leq 2M\exp\left\{-tk_nq_nu + t^2\sum_{i=1}^n X_i^{c2}\right\} \\ &\ll \exp\left\{-tk_nq_nu + t^2n\right\} \\ &\leq \exp\left\{-\frac{k_nq_nu}{2n}k_nq_nu + n\left(\frac{k_nq_nu}{2n}\right)^2\right\} \\ &= \exp\left\{-\frac{(k_nq_nu)^2}{4n}\right\} \leq n^{-2}. \end{aligned} \tag{3.13}$$

同理, 我们可以得到

$$P(D_{jx}) \leq n^{-2},\ \ j = 2,3,4. \tag{3.14}$$

于是, 由 (3.12)、(3.13) 和 (3.14) 式, 当 n 充分大时, 有

$$\sum_{n=1}^\infty P(|f_n(x)-f(x)| > \varepsilon q_n) = \sum_{n=1}^\infty P(A_x) \leq 4\sum_{n=1}^\infty n^{-2} < \infty.$$

由引理 4 可知

$$|f_n(x)-f(x)| \leq \varepsilon q_n \quad \text{a.s.},$$

此即

$$|f_n(x)-f(x)| = o(q_n) \quad \text{a.s.}.$$

从而定理 1 得证.

参 考 文 献

[1] Loftsgarden D O, Quesenberry C D. A nonparametric estimate of a multivariate density function [J]. Ann. Statist., 1965, 36: 1049–1051.

[2] Wagner T J. Stronger consistency of a nonparametric estimate of a density function [J]. IEEE Trans. Systems Man. Cybernet, 1973, 3: 289–290.

[3] 陈希孺. 最近邻密度估计的收敛速度 [J]. 中国科学, 1981, 12: 1419–1428.

[4] Devroye L P, Wagner T J. The strong uniform consistency of nearest neighbor density estimates [J]. Ann. Statist., 1977, 5: 536–540.

[5] Chen X. The rate of uniformly consistency of nearest neighbor density estimator [J]. J. Math. Reasearch Exposition, 1983, 3(1): 61–68.

[6] 柴根象. 平稳序列最近邻密度估计的相合性 [J]. 数学学报, 1989, 32(3): 423–432.

[7] 杨善朝. NA 样本最近邻密度估计的相合性 [J]. 应用数学学报, 2003, 26(3): 385–395.

[8] 倪展, 吴群英, 施生塔. ND 序列下最近邻密度估计的强相合速度 [J]. 山东大学学报 (理学版), 2012, 47(12): 6–9.

[9] 刘艳, 吴群英. ND 样本最近邻密度估计的一致强相合性 [J]. 华侨大学学报 (自然科学版), 2012, 33(5): 590–594.

[10] 刘永辉, 吴群英. ND 样本最近邻密度估计的相合性 [J]. 吉林大学学报 (理学版), 2012, 50(6): 1141–1145.

[11] 施生塔, 吴群英, 倪展. ND 样本下最近邻密度估计的一致强相合速度 [J]. 桂林理工大学学报, 2012, 32(4): 631–634.

[12] Liu L. Precise large deviations for dependent random variables with heavy tails[J]. Statis. Prob. Letters, 2009, 79(9): 1290–1298.

[13] Liu L. Necessary and sufficient conditions for moderate deviations of dependent random variables with heavy tails[J]. Science China Mathematics, 2010, 53(6): 1421–1434.

[14] Wu Q. Complete convergence for negatively dependent sequences of random variables[J]. Journal of Inequalities and Applications, 2010, Article ID 507293.

[15] Shen A. Probability inequalities for END sequence and their applications[J]. Journal of Inequalities and Applications, 2011, 1: 1–12.

[16] 吴群英. 混合序列的概率极限理论 [M]. 北京: 科学出版社, 2006.

RATE OF STRONG CONSISTENCY OF NEAREST NEIGHBOR ESTIMATOR OF DENSITY FUNCTION FOR END SAMPLES

LAN Chong-feng[1] , WU Qun-ying[2]

(*1. School of Economics and Management, Fuyang Teachers College, Fuyang 236037, China*)

(*2. College of Science, Guilin University of Technology, Guilin 541004, China*)

Abstract: In this paper, we discuss the strong consistency of nearest neighbor estimator of density function for END samples. By applying Bernstein type inequality and truncation methods, the rate of strong consistency of nearest neighbor estimator of density function for END samples is obtained. Our results extend the corresponding ones of nearest neighbor estimator of density function for NA samples and ND samples.

Keywords: extended negatively dependent sequences; nearest neighbor estimator of density function; rate of strong consistency

2010 MR Subject Classification: 62G07