第 33 卷 第 5 期　　华侨大学学报(自然科学版)　　Vol. 33 No. 5
2012 年 9 月　　Journal of Huaqiao University (Natural Science)　　Sep. 2012

文章编号： 1000-5013(2012)05-0590-05

ND 样本最近邻密度估计的一致强相合性

刘艳，吴群英

(桂林理工大学 理学院，广西 桂林 541004)

摘要： 设 $X_1,X_2,\cdots,X_n$ 是同分布的负相依(ND)样本，具有共同的密度函数 $f(x)$，利用相应的 Bernstein 不等式，将负相关(NA)样本最近邻密度估计的一致强相合性推广到 ND 样本，得到其最近邻密度估计的一致强相合性.

关键词： 负相依序列；最近邻密度估计；一致强相合性；Bernstein 不等式

中图分类号： O 212.7　　**文献标志码：** A

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.最近邻密度估计(NN-估计)是由 Loftsgarden 和 Quesenberry[1] 于 1965 年提出的.设 $X_1,X_2,\cdots,X_n$ 是来自未知密度 f 的样本，先选定一个与 n 有关的整数 $k=k_n$，$1\leqslant k<n$，对固定的 $x\in R$，记 $a_n(x)$ 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 个.注意到，对每一个 $a>0$ 可以期望在 $X_1,X_2,\cdots,X_n$ 中大约有 $2anf(x)$ 个观察值落入区间 $[x-a,x+a]$ 中.因而值 $f(x)$ 的估计(记为 $\hat{f}_n(x)$)自然地可以通过令 $k=2an\hat{f}_n(x)$ 得到.于是定义 $\hat{f}_n(x)=k/2na_n(x)$ 为 $f(x)$ 的估计.此后，许多著名学者都讨论过它的性质．对于独立样本，Wagner[2] 证明了 NN-估计的强相合性，陈希孺[3] 证明了它的渐近正态性，而杨善朝[4] 在一定条件下给出了 NA 序列下 NN-估计的相合性.ND 相依序列是比 NA 相依序列弱的一种数列，这些负相依随机变量的概念在可靠性理论、渗透理论和多元统计分析中均有广泛的应用.因此，将独立序列或 NA 序列的一些性质推广到 ND 序列是十分必要的.文献[5-7]分别讨论了 ND 序列完全收敛性、ND 随机变量列的指数不等式等，但却很少触及 NN-估计.而 NN-估计在独立样本下已有很多优点，特别是计算简单，容易实施.因此，在 ND 样本下也应有一定的地位．本文在 ND 序列的基础上，讨论了 ND 序列的最近邻密度估计的一致强相合性.

1 定义及引理

假设总体 X 的分布密度函数为 $f(x)$，$X_1,X_2,\cdots,X_n$ 是抽自该总体的 ND 样本；$F(x)$ 是密度函数 $f(x)$ 相应的分布函数，$F_n(x):=\frac{1}{n}\sum_{i=1}^{n}I(X_i<x)$ 是样本 $X_1,X_2,\cdots,X_n$ 的经验分布函数；c 是与 n 无关的常数，在不同的地方可以表示不同的值.

定义 1[5] 称随机变量 $X_1,X_2,\cdots,X_n$，$n\geqslant 2$ 是 ND 的，若对 $\forall x_1,x_2,\cdots,x_n\in R$，都有

$$P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n)\leqslant\prod_{i=1}^{n}P(X_i\leqslant x_i),$$

及

$$P(X_1>x_1,X_2>x_2,\cdots,X_n>x_n)\leqslant\prod_{i=1}^{n}P(X_i>x_i),$$

称随机变量列 $\{X_n;n\geqslant 1\}$ 是 ND 列，如果对任意的 $n\geqslant 2$，$X_1,X_2,\cdots,X_n$ 是 ND 的.

从有限总体不放回抽样所得到的样本不是独立的，但是 ND 的.

收稿日期： 2012-01-07

通信作者： 吴群英(1961-)，女，教授，主要从事概率统计的研究. E-mail：wqy666@glite.edu.cn.

基金项目： 国家自然科学基金资助项目(11061012)；广西自然科学基金资助项目(2011GXNSFA018147)；广西研究生教育创新计划项目(2011105960202M32)

引理 1[5]　设$\{X_n;n\geqslant 1\}$是 ND 的，$\forall m\geqslant 2,A_1,A_2,\cdots,A_m$ 是集合$\{1,2,\cdots,n\}$的两两不交非空子集. 如果 $f_i(i=1,2,\cdots,m)$是对每个变元都非降(或非升)的函数，则 $f_1(X_j;j\in A_1),f_2(X_j;j\in A_2),\cdots,f_m(X_j;j\in A_m)$仍是 ND 的.

引理 2　即 Bernstein 不等式[6]. 设$\{X_i:\leqslant i\leqslant n\}$为同分布的 ND 序列，$EX_i=0,|X_i|\leqslant b$, a. s. $(i=1,2,\cdots,n)$，$t>0$ 为实数，且满足 $t\cdot\max\limits_{1\leqslant i\leqslant n} b_i\leqslant 1$，则 $\forall\varepsilon>0$，有 $P(|S_n|>\varepsilon)\leqslant 2\exp\{-t\varepsilon+t^2\sum\limits_{i=1}^{n}EX_i^2\}$，其中 $S_n=\sum\limits_{i=1}^{n}X_i$.

特别地，记 $\sigma_n^2=n^{-1}\sum\limits_{i=1}^{n}\mathrm{Var}(X_i)$，$b=\max\limits_{1\leqslant i\leqslant n}b_i$，取 $t=\varepsilon/(2\sigma_n^2+b\varepsilon)$，则 $\forall\varepsilon>0$，有

$$P(|\frac{1}{n}\sum_{i=1}^{n}X_i|>\varepsilon)\leqslant 2\exp\{-\frac{n\varepsilon}{2(2\sigma_n^2+b\varepsilon)}\}.$$

引理 3[7]　设 $F(x)$为连续的分布函数，对 $n\geqslant 3$，令 $x_{n,k}$ 满足：$F(x_{n,k})=k/n,k=1,2,\cdots,n-1$，则有

$$\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|\leqslant\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|+\frac{2}{n}.$$

引理 4　设$\{X_1,X_2,\cdots,X_n\}$为 ND 样本，$F(x)$为连续的分布函数，存在正数列$\{\tau_n\}$满足 $\tau_n\to 0$，$\frac{n\tau_n^2}{\log n}\to\infty$，则有$\sup\limits_x|F_n(x)-F(x)|=o(\tau_n)$, a. s..

证明　由条件知，$\forall\varepsilon>0$，当 n 充分大时，有$\frac{2}{n}<\frac{\varepsilon\tau_n}{2}$. 由此及引理 3 可得

$$\begin{aligned}P(\sup_x|F_n(x)-F(x)|>\varepsilon\tau_n)&\leqslant P(\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|+\frac{2}{n}>\varepsilon\tau_n)\leqslant\\&P(\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|>\frac{\varepsilon\tau_n}{2})\leqslant\\&\sum_{k=1}^{n-1}P(|F_n(x_{n,k})-F(x_{n,k})|>\frac{\varepsilon\tau_n}{2}).\end{aligned}$$

记 $\xi_i=I(X_i\leqslant x_{n,k})-EI(X_i\leqslant x_{n,k})$，则 $\xi_1,\xi_2,\cdots,\xi_n$ 仍为 ND 变量，且 $E\xi_i=0,|\xi_i|\leqslant b,(i=1,2,\cdots,n)$. 取 $t=\frac{\varepsilon\tau_n}{4}$，当 n 充分大时，t 满足引理 2 的条件，所以有

$$\begin{aligned}P(|F_n(x_{n,k})-F(x_{n,k})|>\frac{\varepsilon\tau_n}{2})&=P(|\sum_{i=1}^{n-1}\xi_i|>\frac{\varepsilon\tau_n}{2})\leqslant 2\exp\{-t\varepsilon n\tau_n/2+t^2\sum_{i=1}^{n}E\xi_i^2\}\leqslant\\&2\exp\{-t\varepsilon n\tau_n/2+t^2n\}=2\exp\{-\varepsilon^2n\tau_n^2/8+\varepsilon^2n\tau_n^2/18\}=\\&2\exp\{-\varepsilon^2n\tau_n^2/16\}\leqslant 2n^{-3}.\end{aligned}$$

因此，有

$$P(\sup_x|F_n(x)-F(x)|>\varepsilon\tau_n)\leqslant 2n^{-2}.$$

由此即得结论. 证毕.

2　主要结果

定理 1　设$\{X_n;n\geqslant 1\}$是 ND 序列，k_n 满足

$$k_n\to\infty,\qquad\frac{k_n}{n}\to 0,\qquad\frac{k_n}{\sqrt{n\log n}}\to\infty,\tag{1}$$

且 $f(x)$在 R 上一致连续，则有

$$\lim_{n\to\infty}\sup_x|\hat{f}_n(x)-f(x)|=0,\qquad\text{a. s.}.$$

证明　$\forall\varepsilon>0$，$P\{|\hat{f}_n(x)-f(x)|>\varepsilon\}=P\{\hat{f}_n(x)>f(x)+\varepsilon\}+P\{\hat{f}_n(x)<f(x)-\varepsilon\}$，当 $f(x)<\varepsilon$ 时，$f(x)-\varepsilon$ 的值为负值，由 $\hat{f}_n(x)$的非负性知，事件$\{\hat{f}_n(x)<f(x)-\varepsilon\}$是不可能事件，故其概率为零. 因此，对 $P\{\hat{f}_n(x)<f(x)-\varepsilon\}$的估计，只需考虑 $f(x)\geqslant\varepsilon$ 的情况.

记 $b_n(x)=\dfrac{k_n}{2n(f(x)+\varepsilon)}$，$c_n(x)=\dfrac{k_n}{2n(f(x)-\varepsilon/2)}$，(此时 $f(x)\geqslant\varepsilon$)，则由 $\hat{f}_n(x)$的定义可得

$$\begin{aligned}\{|\hat{f}_n(x)-f(x)|>\varepsilon\}=&\{\hat{f}_n(x)>f(x)+\varepsilon\}\cup\{\hat{f}_n(x)<f(x)-\varepsilon,f(x)\geqslant\varepsilon\}\subset\\&\{\hat{f}_n(x)>f(x)+\varepsilon\}\cup\{\hat{f}_n(x)<f(x)-\varepsilon/2,f(x)\geqslant\varepsilon\}=\\&\{a_n(x)<b_n(x)\}\cup\{a_n(x)>c_n(x),f(x)\geqslant\varepsilon\}\triangleq A_x+B_x.\end{aligned}\tag{2}$$

又由 $f(x)$一致连续且 $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1<\infty$，可知 $f(x)$是有界的，即存在 $M\geqslant 0$，使得 $f(x)<M$. 对于 $\forall\varepsilon>0$，$\exists\delta>0$，使得 $\forall x,y\in R$，当 $|x-y|<\delta$ 时，有 $|f(x)-f(y)|<\varepsilon/4$. 由条件式(1)可知

$$b_n(x)\leqslant\frac{k_n}{2n\varepsilon}\to 0,\qquad c_n(x)\leqslant\frac{k_n}{2n(\varepsilon-\varepsilon/2)}=\frac{k_n}{n\varepsilon}\to 0,\qquad n\to\infty,\tag{3}$$

关于 x 一致成立. 故当 n 充分大时，$\forall t\in(x-b_n(x),x+b_n(x))\cup(x-c_n(x),x+c_n(x))\subset(x-\delta,x+\delta)$，有

$$|f(t)-f(x)|<\varepsilon/4,\tag{4}$$

即 $f(x)-\varepsilon/4<f(t)<f(x)+\varepsilon/4$.

记 $p_n=\int_{x-b_n(x)}^{x+b_n(x)}f(t)\mathrm{d}t$，$q_n=\int_{x-c_n(x)}^{x+c_n(x)}f(t)\mathrm{d}t$，则由式(4)可得

$$p_n\leqslant(f(x)+\varepsilon/4)\times 2b_n(x)=\frac{k_n}{n}\cdot\frac{f(x)+\varepsilon/4}{f(x)+\varepsilon}\leqslant\frac{k_n}{n},\tag{5}$$

$$q_n\geqslant(f(x)-\varepsilon/4)\times 2c_n(x)=\frac{k_n}{n}\cdot\frac{f(x)-\varepsilon/4}{f(x)-\varepsilon/2}\geqslant\frac{k_n}{n}.\tag{6}$$

记 $\xi_i=I(x-b_n(x)<X_i\leqslant x+b_n(x))$，$\eta_i=I(x-c_n(x)<X_i\leqslant x+c_n(x))$，由经验分布函数的定义及式(5)可得

$$\begin{aligned}A_x=&\{\text{在}(x-b_n(x),x+b_n(x))\text{中至少有 }k_n\text{ 个 }X_i\}\subset\{\sum_{i=1}^{n}\xi_i\geqslant k_n\}=\\&\{\sum_{i=1}^{n}(I(X_i\leqslant x+b_n(x))-I(X_i\leqslant x-b_n(x)))\geqslant k_n\}=\\&\{\frac{1}{n}\sum_{i=1}^{n}[(I(X_i\leqslant x+b_n(x))-EI(X_i\leqslant x+b_n(x)))-\\&(I(X_i\leqslant x-b_n(x))-EI(X_i\leqslant x-b_n(x)))]\geqslant\frac{k_n}{n}-p_n]\}\subset\\&\{|F_n(x+b_n(x))-F(x+b_n(x))|\geqslant\frac{1}{2}(\frac{k_n}{n}-p_n)\}\cup\\&\{|F_n(x-b_n(x))-F(x-b_n(x))|\geqslant\frac{1}{2}(\frac{k_n}{n}-p_n)\}\subset\\&\{\sup_x|F_n(x)-F(x)|\geqslant\frac{1}{2}(\frac{k_n}{n}-p_n)\}\subset\\&\{\sup_x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{3\varepsilon}{8(f(x)+\varepsilon)}\}\subset\\&\{\sup_x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{3\varepsilon}{8(M+\varepsilon)}\geqslant\frac{k_n}{n}\cdot\frac{3\varepsilon}{8M}\}.\end{aligned}\tag{7}$$

同理，另一方面有

$$\begin{aligned}B_x=&\{\text{在}(x-c_n(x),x+c_n(x))\text{中最多有 }k_n\text{ 个 }X_i\}\subset\{\sum_{i=1}^{n}\eta_i\leqslant k_n\}=\\&\{\frac{1}{n}\sum_{i=1}^{n}[(I(X_i\leqslant x+c_n(x))-EI(X_i\leqslant x+c_n(x)))-\\&(I(X_i\leqslant x-c_n(x))-EI(X_i\leqslant x-c_n(x)))]\leqslant\frac{k_n}{n}-q_n\leqslant 0\}=\end{aligned}$$

$$\{|F_n(x+c_n(x))-F(x+c_n(x))|\geqslant\frac{1}{2}(q_n-\frac{k_n}{n})\}\cup$$
$$\{|F_n(x-c_n(x))-F(x-c_n(x))|\geqslant\frac{1}{2}(q_n-\frac{k_n}{n})\}\subset$$
$$\{\sup_x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8(f(x)-\varepsilon/2)}\}\subset$$
$$\{\sup_x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8f(x)}\}\subset$$
$$\{\sup_x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8M}\}.\tag{8}$$

由式(6),(7),(8)可得

$$\{|\hat{f}_n(x)-f(x)|>\varepsilon\}\subset\{\sup^x|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8M}\}\triangleq B.$$

故有
$$\sup_{x\in R}\{|\hat{f}_n(x)-f(x)|>\varepsilon\}=\bigcup_x\{|\hat{f}_n(x)-f(x)|>\varepsilon\}\subset B.$$

由引理 3,取 $x_{n,k}$ 满足 $F(x_{n,k})=k/n$, $k=1,2,\cdots,n-1$,则有

$$P\{\sup_{x\in R}|\hat{f}_n(x)-f(x)|>\varepsilon\}\leqslant P\{\sup_{x\in R}|F_n(x)-F(x)|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8M}\}\leqslant$$
$$P\{\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|+\frac{2}{n}\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8M}\}=$$
$$P\{\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{8M}-\frac{2}{n}\}.$$

因为 $k_n\to\infty$,所以当 n 充分大时,有$\frac{2}{n}<\frac{\varepsilon k_n}{16Mn}$,故有

$$P\{\sup_{x\in R}|\hat{f}_n(x)-f(x)|>\varepsilon\}\leqslant P\{\max_{1\leqslant k\leqslant n-1}|F_n(x_{n,k})-F(x_{n,k})|\geqslant\frac{k_n}{n}\cdot\frac{\varepsilon}{16M}\}\leqslant$$
$$\sum_{k=1}^{n}P\{|F_n(x_{n,k})-F(x_{n,k})|\geqslant$$
$$\frac{k_n}{n}\cdot\frac{\varepsilon}{16M}\}=\sum_{k=1}^{n}P\{|F_n(x_{n,k})-F(x_{n,k})|\geqslant\frac{ck_n}{n}\}=$$
$$\sum_{k=1}^{n}P\{\frac{1}{n}\sum_{i=1}^{n}(I(X_i\leqslant x_{n,k})-EI(x\leqslant x_{n,k}))\geqslant\frac{ck_n}{n}\}.$$

由于 $I(X_i\leqslant x_{n,k})-EI(x\leqslant x_{n,k})$关于 X_i 单调非降,所以 $\zeta_i=I(X_i\leqslant x_{n,k})-EI(x\leqslant x_{n,k})$也为 ND 序列,且满足引理 2 的条件,而 $\sigma_n^2=n^{-1}\sum_{i=1}^{n}\mathrm{Var}(\zeta_i)=n^{-1}\sum_{i=1}^{n}E\zeta_i^2\leqslant1$,且$\frac{k_n}{n}\to0$,所以由引理 2 可得

$$P\{\sup_x|\hat{f}_n(x)-f(x)|>\varepsilon\}\leqslant\sum_{k=1}^{n}2\exp\{-\frac{n(\frac{ck_n}{n})^2}{2(2\sigma_n^2+2\frac{ck_n}{n})}\}\leqslant$$
$$\sum_{k=1}^{n}2\exp\{-\frac{k_n^2}{n}\cdot\frac{c^2}{4(\sigma_n^2+\frac{ck_n}{n})}\}\leqslant\sum_{k=1}^{n}2\exp\{-\frac{k_n^2}{n}\cdot\frac{c^2}{4(1+c)}\}\leqslant\sum_{k=1}^{n}2\exp\{-\frac{ck_n^2}{n}\}.$$

由条件(1)可得,当 n 充分大时,$\frac{ck_n^2}{n}>3\log n=\log n^3$,则有

$$\sum_{n=1}^{\infty}P\{\sup_x|\hat{f}_n(x)-f(x)|>\varepsilon\}<\infty.$$

由 Borel-Cantelli 引理可知

$$P\{\sup_x|\hat{f}_n(x)-f(x)|>\varepsilon,\mathrm{i.o.}\}=0.$$

即
$$\lim_{n\to\infty}\sup_x|\hat{f}_n(x)-f(x)|=0,\qquad \mathrm{a.s.}.$$

定理证毕.

3 应用举例

设在可靠性问题中 r.v. X 的分布函数记为 $F(x)$,密度函数为 $f(x)$,生存函数和失效率函数分别定义为 $\bar{F}(x)=1-F(x)=P(X\geqslant x)$ 和 $r(x)=\dfrac{f(x)}{\bar{F}(x)}$. 又设 $X_1,X_2,\cdots,X_n$ 为从总体 $F(x)$ 中抽取的同分布 ND 样本,则生存函数和失效率函数的自然估计为

$$\bar{F}(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i\geqslant x),\qquad r_n(x)=\frac{\hat{f}_n(x)}{\bar{F}_n(x)}.$$

定理 2 设定理 1 的条件满足,则对任何满足 $F(c)<1$ 的 c,均有 $\lim\limits_{n\to\infty}\sup\limits_{x\leqslant c}|r_n(x)-r(x)|=0$, a.s..

证明 记 $\bar{F}_n(x)=1-F(x)$, $\bar{F}_n(x)=1-F_n(x)$,显然有

$$r_n(x)-r(x)=\frac{1}{\bar{F}(x)\bar{F}_n(x)}\cdot\{\bar{F}(x)[\hat{f}_n(x)-f(x)]-f(x)[\bar{F}_n(x)-\bar{F}(x)]\}.\tag{9}$$

注意到 $0<\bar{F}(c)\leqslant\bar{F}(x)\leqslant 1$, $\forall x\leqslant c$, $\sup\limits_{x}f(x)\leqslant M<\infty$. 由定理 1 和引理 4 知, $\sup\limits_{x\leqslant c}|\hat{f}_n(x)-f(x)|\to 0$, a.s., 而且 $\sup\limits_{x\leqslant c}|\bar{F}_n(x)-\bar{F}(x)|\to 0$, a.s.. 从而当 n 充分大时, 对于 $x\leqslant c$, 一致地有 $\bar{F}(c)>\bar{F}(x)-\bar{F}(c)/2>\bar{F}(c)/2>0$. 由这些事实和式(9)即得定理 2 的结论. 证毕.

参考文献:

[1] LOFTSGARDEN D O, QUENSBERRY C D. A Nonparametric estimator of a multivariate density function[J]. Ann Statist, 1965, 36(3): 1049-1051.

[2] WAGNER T J. Strong consistency of a nonparametic estimate of a density function[J]. IEEE Trans Systems Man Cybenet, 1973(3): 289-290.

[3] CHENG Xi-ru. Convergence rates for nearest neighbor density estimator[J]. Sci China Ser A, 1980(12): 1419-1428.

[4] YANG Shan-chao. Consistency of nearest neighbor density function for pairwise NA sequence[J]. Acta Math Application, 2003, 26(3): 385-395.

[5] BOZORGNIA A, PATTERSON R F, TAYLOR R L. Limit theorems for ND r.v.'s[R]. Athens: University of Georgia, 1993.

[6] WU Qun-ying, JIANG Yuan-ying. The strong consistency of estimator in linear model for negatively dependent random samples[J]. Communications in Statistics-Theory and Methods, 2011, 40(3): 467-491.

[7] 杨善朝. NA 样本最近邻密度估计的相合性[J]. 应用数报, 2003, 26(3): 385-408.

Uniform Strong Consistency of Nearest Neighbor Estimator of Density Function for Negative Dependent Samples

LIU Yan, WU Qun-ying

(College of Science, Guilin University of Technology, Guilin 541004, China)

Abstract: Suppose that $X_1,X_2,\cdots,X_n$ are negative dependent (ND) samples, with a common density function $f(x)$. By the use of the corresponding Bernstein inequality, the uniform strong consistency of nearest neighbor estimator of density function for negatively associated (NA) samples is extended to ND samples, and the latter uniform strong consistency of nearest neighbor estimator of density function is obtained.

Keywords: negative dependent sequence; nearest neighbor density estimator; Bernstein inequality

(责任编辑: 黄晓楠 英文审校: 黄心中)