应用概率统计　第二卷
第三期　1986年8月

Chinese Journal of Applied Probability
and Statistics Vol. 2. No. 3. Aug. 1986

条件密度近邻-核估计的强相合性

刘志军
（中国科学技术大学）

1. 引言　设(X, Y)为取值于$R^p\times R^q$的随机向量，在给定$X=x$的条件下Y具有条件密度函数$f(y|x)$. 它是(x, y)的Borel可测函数. 设(X_1, Y_1), $\cdots$, (X_n, Y_n)为(X, Y)的i. i. d. 观测值. 我们的目的是要利用这些观测值来估计条件密度$f(y|x)$. 赵林城给出了条件密度的两类估计. 本文主要讨论其中一类被称为近邻-核估计(NN-K)的相合性.

首先约定在本文中，$u=(u^{(1)}, \cdots, u^{(d)})\in R^d$的模$\|u\|$取为实线性空间中的$L_2$-模或$L_\infty$-模. 对固定的$x\in R^p$, 将$(X_1, Y_1)$, $\cdots$, (X_n, Y_n)按照

$$\|X_{R_1}-x\|\leqslant\|X_{R_2}-x\|\leqslant\cdots\leqslant\|X_{R_n}-x\| \tag{1}$$

的次序重新排列，约定$\|X_i-x\|=\|X_j-x\|$而$i<j$时，$\|X_i-x\|$在(1)式中排在$\|X_j-x\|$之前. 设$h=h_n>0$, $n=1, 2, \cdots$, 为一常数列. $k=k_n\leqslant n$, $n=1, 2, \cdots$, 为一正整数列. $K(v)$为R^q上的概率密度函数，令

$$f_n(y|x)=\frac{1}{kh^q}\sum_{i=1}^{k}K\left(\frac{y-Y_{Ri}}{h}\right). \tag{2}$$

则可用$f_n(y|x)$来估计$f(y|x)$. 这就是赵林城给出的条件密度的近邻-核估计. 下面是我们的主要结果.

定理　设存在常数$M>0$, $\rho>0$使得

$$h\to 0,\ \frac{k}{n}\to 0,\ \frac{kh^q}{\log n}\to\infty,\ (n\to\infty). \tag{3}$$

$$K(v)\leqslant MI(\|v\|\leqslant\rho) \tag{4}$$

I)　记$C(f(y|x))=\{(x, y): (x, y)$为$f(y|x)$的连续点$\}$则当$(x, y)\in C(f(y|x))$时有

$$f_n(y|x)\xrightarrow{\text{a.s.}}f(y|x),\ (n\to\infty). \tag{5}$$

II)　设$F(x)$为X的边缘分布，记$\log^+f(y|x)=\max\{0, \log f(y|x)\}$, 若

$$\int_A f(y|x)\log^+f^+(y|x)dy\,dF(x)<\infty \tag{6}$$

对$R^p\times R^q$的任何有界Borel子集A成立，则

$$f_n(y|x)\xrightarrow{\text{a.s.}}f(y|x),\ (n\to\infty),\ \text{a.e. } F\times L \tag{7}$$

L为R^q上的Lebesgue测度.

注意定理中未对(X, Y)的联合密度及X的边缘密度有任何假定，它们甚至可以不存在，对h, k加的条件众所周知是几乎不能减弱的.

本文1985年3月25日收到.

2．若干引理 先引述和证明下列事实．

引理1([1]) 设 X_{R_k} 为在距离$\|\cdot\|$之下相对于 x 的第 k 个近邻点．若 $\frac{k}{n}\to 0(n\to\infty)$，则

$$P(\|X_{R_k}-x\|>\varepsilon)\leqslant 2\exp\left(-\frac{1}{10}P(\|X-x\|\leqslant\varepsilon)n\right) \tag{8}$$

$X_{R_k}\xrightarrow{\text{a.s.}}x$，$(n\to\infty)$，是上式的直接推论．

引理2(Bernstein [2]) 设 $Y_1,\cdots,Y_n$ 独立，均值为零，且存在有限常数 b，使得 $P(|Y_i|\leqslant b)=1$，$i=1,2,\cdots,n$．又

$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}Y_i$．则对任何 $\varepsilon>0$ 和 $n=1,2,\cdots$，有

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}Y_i\right|\geqslant\varepsilon\right)\leqslant 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2+b\varepsilon}\right). \tag{9}$$

引理3([3]) 设 μ 为 R^d 上的 Lebesgue-Stieltjes 测度．g 为 R^d 上的 Borel 可测函数，且关于 μ 在任何有界 Borel 集上可积．那么存在 B^d 中的一个集合 A，使得 $\mu(A^c)=0$，且对任何 $x\in A$，有

$$\lim_{r\to 0}\int_{S_{x,r}}|g(y)-g(x)|d\mu(y)/\mu(S_{x,r})=0 \tag{10}$$

其中

$$S_{x,r}=\{y:\|x-y\|\leqslant r\}.$$

设 $h(x,y)$ 为定义在 R^{p+q} 上的 Borel 可测函数，μ，ν 分别为 R^p，R^q 上的 Lebesgue-Stieltjes 测度，$\{r_n\}$为一串正数，且 $\lim_{n\to\infty}r_n=0$，定义

$$\hat{h}_k(x,y)=\sup_{n\geqslant k}\int_{S_{y,r_n}}h(x,w)d\nu(w)/\nu(S_{y,r_n}),$$

$$h^*(x,y)=\sup_{0<\rho<\infty}\int_{S_{y,\rho}}h(x,w)d\nu(w)/\nu(S_{y,\rho}). \tag{11}$$

引理4．A 为 R^{p+q} 的有界 Borel 子集，若

$$\int_A h(x,y)\log^+h(x,y)d\mu(x)d\nu(y)<\infty,$$

则对任何 $r_n\to 0$ 及 k，有

$$\int_A \hat{h}_k(x,y)d\mu(x)d\nu(y)<\infty. \tag{12}$$

证明 令 $\varphi(x,y)=h(x,y)I\left(h(x,y)\geqslant\frac{t}{2}\right)(t\geqslant 0)$．因此

$$h\leqslant\varphi+\frac{t}{2},\ \hat{h}_k\leqslant\varphi_k+\frac{t}{2},\ h^*\leqslant\varphi^*+\frac{t}{2}.$$

由 $\hat{h}_k(x,y)$的定义可知其一定是一 Borel 可测函数，记 $Q=\mu\times\nu$．由 Fubini 定理及[3]，有

$$Q(\hat{\varphi}_k(x,y)>b)=\int\nu(y:\hat{\varphi}_k(x,y)>b)d\mu(x)$$

$$\leqslant\int\nu(y:\varphi^*(x,y)>b)d\mu(x)\leqslant\frac{a_q}{b}\iint\varphi(x,y)d\mu(x)d\nu(y), \tag{13}$$

其中 a_q 是一仅与维数 q 有关的常数．因此

$$Q(\hat{h}_k(x,y)>t)\leqslant Q\left((\hat{\varphi}_k(x,y)>\frac{t}{2}\right)\leqslant\frac{2a_q}{t}\int_{(h(x,y)<\frac{t}{2})}h(x,y)d\mu(x)d\nu(y).$$

于是我们有

$$\int_A \hat{h}_k(x, y) d\mu(x) d\nu(y) = \int_0^\infty Q(I(A)\hat{h}_k(x, y) > t) dt$$

$$\leqslant Q(A) + 2a_q \int_1^\infty \frac{1}{t} dt \int_{\{h(x, y) > \frac{t}{2}\}} I(A) h(x, y) d\mu(x) d\nu(y)$$

$$\leqslant Q(A) + 2a_q \int_{\{h(x, y) > \frac{1}{2}\}} I(A) h(x, y) \left[\int_1^{2h(x,y)} \frac{1}{t} dt\right] d\mu(x) d\nu(y)$$

$$= Q(A) + 2a_q \int_A h(x, y) \log^+[2h(x, y)] d\mu(x) d\nu(y) < \infty. \tag{14}$$

引理 5 设 $f(y|x)\log^+ f(y|x)$ 在 R^{p+q} 的任何有界 Borel 集上关于 $F\times L$ (X 的边缘测度和 R^q 上的 L-测度的乘积) 可积，则对任何 $\rho_1=\rho_{1n}$, $\rho_2=\rho_{2n}$, $\lim\limits_{n\to\infty}\rho_1=\lim\limits_{n\to\infty}\rho_2=0$ 有

$$R_n = \int_{S_{x,\rho_1}} \int_{S_{y,\rho_2}} |f(\nu|u) - f(y|x) d\nu dF(u) / F(S_{x,\rho_1}) \rho_2^q \to 0, \ (n\to\infty). \ \text{a.e.}(x, y) F\times L. \tag{15}$$

证明 令 $h(u, v) = |f(v|u) - f(y|x)|$. 由引理 4 所证

$$\hat{h}_k(u, y) = \sup_{n>k} \int_{S_{y,\rho_2}} |f(v|u) - f(y|x)| dv / \rho_2^q.$$

在 R^{p+q} 的任何有界 Borel 子集上 $F\times L$ 可积. 由 Fubini 定理, 存在 Borel 集 $B\subset R^q$, $L(B)=0$, 对任何 $y\in B$, $\hat{h}_k(u, y)$ 关于 u 在 R^p 的任何有界 Borel 子集上 F 可积. 因此用有理数逼近的方法我们有

$$\limsup_{n\to\infty} R_n \leqslant \limsup_{n\to\infty} \int_{S_{x,\rho_1}} \left[\sup_{n\geqslant k}\left\{\int_{S_{y,\rho_2}} |f(v|u) - f(y|x)| dv/\rho_2^q\right\}\right] dF(u) / F(S_{x,\rho_1})$$

$$= \limsup_{n\to\infty} \int_{S_{x,\rho_1}} \hat{h}_k(u, y) dF(u) / F(S_{x,\rho_1}) = \hat{h}_k(x, y), \quad \text{a.e.}(x) F.$$

由于 k 为任意正整数, 我们有

$$\hat{h}_k(x, y) = \sup_{n\geqslant k} \int_{S_{y,\rho_2}} |f(v|x) - f(y|x)| dv/\rho_2^q \to 0, \quad k\to\infty, \ \text{a.e.}(y) L$$

从而, $\limsup\limits_{n\to\infty} R_n = 0$ a. e. $x(F)$ 及 a. e. $y(L)$.

3. 定理的证明 为行文方便以下证明过程中我们总以 C 记与 n 和样本点无关的常数, 但可能与 (x, y) 有关, 即使在同一式中出现也可能取不同的值. 另记

$$\Delta = (X_1, X_2, \cdots). \ \widetilde{E}(\cdot) = E(\cdot|\Delta).$$

注意到

$$|f_n(y|x) - f(y|x)| \leqslant |f_n(y|x) - \widetilde{E} f_n(y|x)| + |\widetilde{E} f_n(y|x) - f(y|x)| \triangleq |U_1| + |U_2|. \tag{15}$$

首先考虑 U_2, 由于

$$|U_2| = \left|\frac{1}{k}\sum_{i=1}^k h^{-q} \int K\left(\frac{y-v}{h}\right) f(v|X_{R_i}) dv - f(y|x)\right|$$

$$\leqslant \frac{1}{k}\sum_{i=1}^k h^{-q} \int K\left(\frac{y-v}{h}\right) |f(v|X_{R_i}) - f(y|x)| dv$$

$$\leqslant \frac{c}{k}\sum_{i=1}^k h^{-q} \int_{S_{y,h_\rho}} |f(v|X_{R_i}) - f(y|x)| dv \triangleq cJ_n$$

再记 $\|X_{R_{k+1}} - x\| = \lambda$, $E^*(\cdot) = E(\cdot|\lambda)$, 则我们有

$$E^*J_n\leqslant c\int_{S_{x,\lambda}}\int_{S_{y,\lambda_\rho}}|f(v|u)-f(y|x)|dv\,dF(u)/F(S_{x,\lambda})h^q \tag{16}$$

此时若$(x, y)\in c(f)$则显然有

$$E^*J_n\to 0,\quad \text{a.s. } (n\to\infty).$$

若$f(y|x)$满足II)中的条件，由引理5亦可得

$$E^*J_n\to 0.\ \text{a.s. a.e.}(x, y)F\times L(n\to\infty).$$

记

$$g(u)=h^{-q}\int_{S_{y,\lambda_\rho}}|f(v|u)-f(y|x)|dv,$$

$$J_n=\frac{1}{k}\sum_{i=1}^k g(X_{R_i})$$

在给定$\|X_{R_{k+1}}-x\|=\lambda$的条件下，$J_n$与$\frac{1}{k}\sum_{i=1}^k g(V_i)$同分布，其中$V_1, \cdots, V_k$ i.i.d., V_1的分布为：

$$\tilde{F}(\cdot)=F(\cdot\cap S_{x,\lambda})/F(S_{x,\lambda}) \tag{17}$$

注意到

$$|g(V_i)-Eg(V_i)|\leqslant ch^{-q}\quad i=1, \cdots, k.$$

$$\frac{1}{k}\sum_{i=1}^k \mathrm{Var}\, g(V_i)\leqslant\frac{1}{k}\sum_{i=1}^k Eg^2(V_i)\leqslant ch^{-q} \tag{18}$$

再由引理2我们有

$$P(|J_n-E^*J_n|>\varepsilon|\lambda)=P\left(\left|\frac{1}{k}\sum_{i=1}^k g(V_i)-Eg(V_1)\right|>\varepsilon\right)\leqslant 2\exp(-ckh^q)\quad c>0. \tag{19}$$

现在来考虑U_1. 记

$$\xi_i=h^{-q}\left[K\left(\frac{y-Y_{R_i}}{h}\right)-\int K\left(\frac{y-v}{h}\right)f(v|X_{R_i})dv\right] \tag{20}$$

则$V_1=\frac{1}{k}\sum_{i=1}^k\xi_i$. 由于在$\Delta$给定的条件下$\xi_i$, $i=1, 2, \cdots, k$, 独立，且

$$|\xi_i|<ch^{-q}$$

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^k\mathrm{Var}\xi_i&\leqslant\frac{1}{k}\sum_{i=1}^k\tilde{E}\xi_i^2\\&\leqslant ch^{-q}\frac{1}{k}\sum_{i=1}^k h^{-q}\int K\left(\frac{y-v}{h}\right)f(v|X_{R_i})dv=ch^{-q}\tilde{E}f_n(y|x)\end{aligned} \tag{21}$$

记$\psi_n=\tilde{E}f_n(y|x)$，由引理2，并注意到

$|\psi_n-f(y|x)|\leqslant|U_2|\leqslant cJ_n$. 则对$\forall\varepsilon>0$, $\delta>0$，我们有

$$\begin{aligned}P(|U_1|>\varepsilon)&=EP(|U_1|>\varepsilon|\Delta)\leqslant 2E\exp\left(-\frac{ck}{\psi_n+1}\right)\\&\leqslant 2E\exp\left(-\frac{ckh^q}{\psi_n+1}\right)I(J_n>\delta)+2E\exp\left(-\frac{ckh^q}{\psi_n+1}\right)I\quad(J_n\leqslant\delta)\\&\leqslant 2P(J_n>\delta)+2\exp(-ckh^q)\end{aligned} \tag{22}$$

由(19)式，我们有

$$\begin{aligned}P(J_n>\delta)&\leqslant P\left(|J_n-E^*J_n|>\frac{\delta}{2}\right)+P\left(E^*J_n>\frac{\delta}{2}\right)\\&=EP\left(|J_n-E^*J_n|>\frac{\delta}{2}\Big|\lambda\right)+P\left(E^*J_n>\frac{\delta}{2}\right)\\&\leqslant 2\exp(-ckh^q)+P\left(E^*J_n>\frac{\delta}{2}\right)\end{aligned} \tag{23}$$

由前面关于 E^*J_n 的事实, 要使 $E^*J_n>\frac{\delta}{2}$, 必然有某常数 $\lambda_0>0$ 使得 $\lambda\geqslant\lambda_0$, 则由引理 1 有

$$P\left(E^*J_n>\frac{\delta}{2}\right)\leqslant P(\lambda\geqslant\lambda_0)\leqslant 2\exp\left(-\frac{1}{10}P(\|X-x\|\leqslant\lambda_0)n\right) \tag{24}$$

我们取 x 属于 F 的支撑集 $S(F)$. 综合(22), (23), (24), (19)我们有

$$P(|f_n(y|x)-f(y|x)|>\varepsilon)\leqslant C\exp(-ckh^q)+C\exp(-cn) \tag{25}$$

再由 Borel-Cantelli 引理, 即可得

$$f_n(y|x)\to f(y|x)\quad \text{a. s. a. e.}(x,\ y)F\times L,\ (n\to\infty).$$

致谢: 本文是在赵林城老师指导下完成的, 作者深表谢意.

参 考 文 献

[1] Cover, T. M., Heart, P. E., Nearest neighbor pattern classfication. IEEE Trans. Inform. Theory. IT 13 (1967), 21—27.

[2] Hoeffding, W. Probability inequalites for sums of bounded random variables. J. A. S. A. 58 (1963), 13—30.

[3] Wheeden, R. L., Zygmund, A., Measure and Integral, Marcel Dekker, New York, 1977.

STRONG CONSISTENCY OF THE NEAREST NEIGHBOR-KERNEL ESTIMATORS OF CONDITIONAL DENSITY FUNCTION

LIU ZHIJUN

(*University of Science and Technology of China*)

Let (X, Y), (X_1, Y_1), $\cdots$, (X_n, Y_n) be $R^p\times R^q$-valued i.i.d. random vectors, and $f(y|x)$ the conditional density function of Y, given $X=x$. Note that the existence of the density of (X,Y) is not assumed here. In this paper, we introduce the nearest neighbor-kernel estimator $f_n(y|x)$ of $f(y|x)$, and establish the strong consistency of $f_n(y|x)$ under some mild conditions.