

条件中位数 L_1 -模最近邻估计的逐点相合性*

洪 圣 岩

(安徽大学数学系, 合肥 230039)

设 $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ 为取值于 $R^q \times R^1$ 上的独立同分布随机向量序列. Y 对 X 的条件中位数 $\theta(x)$ 定义为在给定 $X=x$ 时 Y 的条件分布函数的中位数, 出于统计稳健性的考虑, 关于 $\theta(x)$ 的估计问题近年来受到了许多学者的重视. 如文献[1]—[5], 但是较回归函数 $m(x) = E\{Y|X=x\}$ 来说, 目前对 $\theta(x)$ 的估计问题的研究还很不充分, 其中之一就是估计量的构造问题. 本文将定义 $\theta(x)$ 的 L_1 -模最近邻估计, 同时在很弱的条件下证明了它的逐点相合性. 注意到我们这里定义的估计包括了文献[1]—[3]所考虑的最近邻估计.

对固定的 $x \in R^q$, 将 $X_1, \dots, X_n$ 按下述方式重排

$$\|X_{R_1} - x\| \leq \|X_{R_2} - x\| \leq \dots \leq \|X_{R_n} - x\|$$

(按小下标在前的方式消结). 设 $\{V_{ni}, 1 \leq i \leq n\}$ 为一列给定的非负常数, $\sum_{i=1}^n V_{ni}$ 令

$$W_{niR_i} = V_{ni} \quad 1 \leq i \leq n$$

那么 $\theta(x)$ 的 L_1 -模最近邻估计 $\theta_n(x)$ 定义为下述极小值问题的解:

$$\sum_{i=1}^n W_{ni} |Y_i - \theta_n(x)| = \inf \left\{ \sum_{i=1}^n W_{ni} |Y_i - \theta|, \theta \in R^1 \right\} \quad (1)$$

特别地, 若取 $V_{ni} = k_n^{-1}, 1 \leq i \leq k_n; V_{ni} = 0, k_n < i \leq n$ (其中 $k_n \leq n$ 为适当的正整数), 则易见(1)式的解为

$$\theta_n(x) = \{Y_{R_1}, \dots, Y_{R_{k_n}}\} \text{ 的中位数}$$

此即文献[1]—[3]中所考虑的最近邻估计.

以下以 F 记 X 的边缘分布函数, 为明确计, 本文均假定 $x \in X$ 的支撑集. 本文的主要结果为

定理. 设对 a. e. $x(F)$ 条件中位数 $\theta(x)$ 唯一. 又存在一列正整数 $\{k=k_n, n \geq 1\}$ 使得当 $n \rightarrow \infty$ 时

$$(i) \quad k \rightarrow \infty, k/n \rightarrow 0; (ii) \quad k \max_{1 \leq i \leq k} V_{ni} = o(1), \sum_{i=1}^k V_{ni} = o(1)$$

则对 a. e. $x(F)$ $\theta_n(x)$ 是 $\theta(x)$ 的弱相合估计. 若 k 还满足, (iii) $k/\log n \rightarrow \infty$.

则对 a. e. $x(F)$ $\theta_n(x)$ 是 $\theta(x)$ 的强相合估计.

* 1990年6月11日收到, 本工作得到国家自然科学基金 18901001 的资助.

定理的证明.

令

$$W_{nR_i}' = \begin{cases} U_{n_i} & 1 \leq i \leq k \\ 0 & k < i \leq n \end{cases}$$

对 $1 \leq i \leq n$, $\theta \in R'$, 记

$$\Phi_{n_i}(\theta) = W_{n_i}'(|Y_i - \theta(x)| - |Y_i - \theta|); \quad R_{n_i}(\theta) = \Phi_{n_i}(\theta) - E^* \Phi_{n_i}(\theta)$$

这里及以下用 E^* 和 P^* 分别表示给定 $X_1, \dots, X_n$ 时所取的期望和概率. 又 $\Phi_{n_i}(\theta)$ 的期望存在有限是因为 $|\Phi_{n_i}(\theta)| \leq |\theta - \theta(x)| < \infty$.

我们只证 $\theta_n(x)$ 的弱相合性. 由证明过程和条件(iii)即得 $\theta_n(x)$ 的强相合性.

任意取定 $\varepsilon > 0$. 首先我们证对任意 $\delta \in (0, \varepsilon)$

$$P\left\{\sup_{|\theta - \theta(x)| \leq \varepsilon} \left| \sum_{i=1}^n R_{n_i}(\theta) \right| \geq \delta\right\} \rightarrow 0 \quad (2)$$

为此, 将区间 $\{\theta: |\theta - \theta(x)| \leq \varepsilon\}$ 等分成 $m = \lceil 8\varepsilon/\delta \rceil$ 个区间 $H_1, \dots, H_m$, 每个区间长度 $\leq \delta/4$. 以 b_j 记在 H_j 中选定的一点. 注意到当 $\theta \in H_j$ 时

$$\begin{aligned} |\Phi_{n_i}(\theta) - \Phi_{n_i}(b_j)| &\leq W_{n_i}'|\theta - b_j| \leq \frac{\delta}{4} W_{n_i}'; \text{ 而 } \sum_{i=1}^n W_{n_i}' \leq 1, \text{ 易见} \\ \sup_{|\theta - \theta(x)| \leq \varepsilon} \left| \sum_{i=1}^n R_{n_i}(\theta) \right| &= \max_{1 \leq j \leq m} \sup_{\theta \in H_j} \left| \sum_{i=1}^n R_{n_i}(\theta) \right| \\ &\leq \max_{1 \leq j \leq m} \left| \sum_{i=1}^n R_{n_i}(b_j) \right| + \max_{1 \leq j \leq m} \sup_{\theta \in H_j} \left| \sum_{i=1}^n (R_{n_i}(\theta) - R_{n_i}(b_j)) \right| \\ &\leq \max_{1 \leq j \leq m} \left| \sum_{i=1}^n R_{n_i}(b_j) \right| + \frac{\delta}{2} \end{aligned} \quad (3)$$

注意到对某 $0 < C_1 < \infty$

$$\begin{aligned} |\Phi_{n_i}(b_j)| &\leq C_1 k^{-1} |b_j - \theta(x)| \leq C_1 \varepsilon k^{-1} \\ \sum_{i=1}^n E^* \Phi_{n_i}^2(b_j) &\leq \sum_{i=1}^n V_{n_i}^2(b_j - \theta(x))^2 \leq C_1 \varepsilon^2 k^{-1} \end{aligned}$$

由(3)式和 Bernstein 不等式 得

$$\begin{aligned} P\left\{\sup_{|\theta - \theta(x)| \leq \varepsilon} \left| \sum_{i=1}^n R_{n_i}(\theta) \right| \geq \delta\right\} &= EP^*\left\{\sup_{|\theta - \theta(x)| \leq \varepsilon} \left| \sum_{i=1}^n R_{n_i}(\theta) \right| \geq \delta\right\} \\ &\leq EP^*\left\{\max_{1 \leq j \leq m} \left| \sum_{i=1}^n R_{n_i}(b_j) \right| \geq \delta/2\right\} \leq 2m \exp\{-2^{-1} \delta^2/2(C_1 \varepsilon^2 k^{-1} + C_1 \varepsilon k^{-1})\} \leq C_2 e^{-C_2 \varepsilon^2} \end{aligned}$$

其中 $0 < C_2 < \infty$. 因而由条件(i)即得(2)式. 现在记

$$\begin{aligned} p^\pm(t) &= E\{(|Y - \theta(x)| - |Y - (\theta(x) \pm \varepsilon)|) | X = t\} \\ h^\pm(t) &= |p^\pm(t) - p^\pm(x)| \end{aligned}$$

那么有

$$h^\pm(t) \leq |p^\pm(t)| + |p^\pm(x)| \leq 2\varepsilon \quad (4)$$

$$\begin{aligned} \sum_{i=1}^n E^* \Phi_{n_i}(\theta(x) \pm \varepsilon) &= \sum_{i=1}^n V_{n_i}(E\{(|Y_{R_i} - \theta(x)| - |Y_{R_i} - (\theta(x) \pm \varepsilon)|) | X_{R_i}\}) \\ &\leq C_1 k^{-1} \sum_{i=1}^n h^\pm(X_{R_i}) + p^\pm(x) \triangleq C_1 I_n^\pm + p^\pm(x) \end{aligned} \quad (5)$$

以 $S_{\rho,x}$ 记中心为 x , 半径为 ρ 的闭球. 那么在给定 $\rho_n = \|X_{R_{n+1}} - x\|$ 的条件下, $I_n^\pm$ 与 $k^{-1} \sum_{i=1}^k h^\pm(v_i)$ 同分布, 其中 $V_1, \dots, V_k$ 相互独立, 同分布于

$$\tilde{F}(\cdot) = F(\cdot \cap S_{\rho_n,x}) / F(S_{\rho_n,x})$$

记在此分布下所取的概率和期望分别为 $\tilde{P}$ 和 $\tilde{E}$ 对任意 $\delta > 0$, 由 Lebesgue 密度定理 (见 [6]) 知存在 $\eta > 0$, 使当 $\rho_n \leq \eta$ 时有

$$\tilde{E}h^\pm(V_1) = \int_{S_{\rho_n,x}} |p^\pm(t) - p^\pm(x)| F(dt) / F(S_{\rho_n,x}) \leq \delta. \quad \text{对 a. e. } x(F)$$

从而再由 (4) 式知

$$\tilde{E}(h^\pm(V_1))^2 \leq 2\epsilon\delta$$

改由 (4) 式和 Bernstein 不等式得当 $\rho_n \leq \eta$ 时

$$\begin{aligned} \tilde{P}\{I_n^\pm \geq 2\delta\} &= \tilde{P}\left\{\sum_{i=1}^k h^\pm(V_i) \geq 2\delta k\right\} \leq \tilde{P}\left\{\sum_{i=1}^k (h^\pm(V_i) - \tilde{E}(h^\pm(V_1))) \geq \delta k\right\} \\ &\leq 2\exp\{-\delta^2 k^2 / (2(k \tilde{E}(h^\pm(V_1))^2 + 2\epsilon\delta k))\} \leq 2e^{-C_3 k} \end{aligned} \quad (6)$$

其中 $0 \leq C_3 < \infty$. 再由条件 (i) 知 (同 [7] 引理 3 之证)

$$\rho_n = \|X_{R_{n+1}} - x\| \xrightarrow{P} 0$$

由此及 (6) 式即得当 $n \rightarrow \infty$ 时

$$P\{I_n^\pm \geq 2\delta\} \leq P\{\rho_n \geq \eta\} + \tilde{E}\tilde{P}\{I_n^\pm \geq 2\delta, \rho_n \geq \eta\} \rightarrow 0$$

因而再由 (5) 式知

$$\sum_{i=1}^n E^* \Phi_{n,i}(\theta(x) \pm \epsilon) \leq p^\pm(x) + o_p(1) \quad (7)$$

其中当 $n \rightarrow \infty$ 时 $o_p(1)$ 以概率趋于零. 综合 (2)、(7) 式和条件 (ii) 得对任意给定的 $\epsilon > 0$,

$$\begin{aligned} \sum_{i=1}^n W_{n,i} |Y_i - \theta(x)| - \inf_{|\theta - \theta(x)| \leq \epsilon} \sum_{i=1}^n W_{n,i} |Y_i - \theta| &= \sup_{|\theta - \theta(x)| \leq \epsilon} \sum_{i=1}^n W_{n,i} (|Y_i - \theta(x)| - |Y_i - \theta|) \\ &\leq \epsilon \sum_{i \geq k} V_{n,i} + \sup_{|\theta - \theta(x)| \leq \epsilon} \sum_{i=1}^n (\Phi_{n,i}(\theta) - E^* \Phi_{n,i}(\theta)) \sup_{|\theta - \theta(x)| \leq \epsilon} \sum_{i=1}^n E^* \Phi_{n,i}(\theta) \\ &\leq o(1) + \sup_{|\theta - \theta(x)| \leq \epsilon} \left| \sum_{i=1}^n R_{n,i}(\theta) \right| + \sup_{\theta = \theta(x) \pm \epsilon} \sum_{i=1}^n E^* \Phi_{n,i}(\theta) \leq \max p^\pm(x) + o_p(1) \end{aligned} \quad (8)$$

另一方面, 易算得

$$\begin{aligned} p^+(x) &= -2 \int_{\theta(x)}^{\theta(x)+\epsilon} P\{\theta(x) < Y < y | X = x\} dy + \epsilon(1 - 2P\{Y \leq \theta(x) | X = x\}) \\ p^-(x) &= -2 \int_{\theta(x)-\epsilon}^{\theta(x)} P\{y \leq Y < \theta(x) | X = x\} dy + \epsilon(1 - 2P\{Y \geq \theta(x) | X = x\}) \end{aligned}$$

因 $\theta(x)$ 为给定 $X=x$ 时 Y 的条件中位数, 由上述表达式易见 $p^\pm(x) \leq 0$. 若 $p^\pm(x) = 0$, 则必有

$$\int_{\theta(x)}^{\theta(x)+\epsilon} P\{\theta(x) < Y < y | X = x\} dy = 0 \quad (9)$$

$$P\{Y \leq \theta(x) | X = x\} = \frac{1}{2} \quad (10)$$

由(9)式及中值定理知存在 $y_\varepsilon \in (\theta(x), \theta(x) + \varepsilon)$ 使 $P\{\theta(x) < Y < y_\varepsilon | X = x\} = 0$, 由此和(10)式易证 y_ε 也是给定 $X = x$ 时 Y 的条件中位数, 这与题设矛盾. 因而必有 $p^+(x) < 0$. 同理可得 $p^-(x) < 0$. 故

$$\max p^+(x) = b < 0 \quad (11)$$

再注意到 $\sum_{i=1}^n W_{n,i} |y_i - \theta|$ 是 θ 的凸函数, 由(8)、(11)式得 $\sum_{i=1}^n W_{n,i} |Y_i - \theta(x)| - \inf_{|\theta - \theta(x)| \geq \varepsilon} \sum_{i=1}^n W_{n,i} |Y_i - \theta| \leq b + o_p(1)$, 由此和(11)式又得

$$P\{|\theta_n(x) - \theta(x)| \leq \varepsilon\} \rightarrow 1 \quad (n \rightarrow \infty)$$

即 $\theta_n(x)$ 是 $\theta(x)$ 的弱相合估计. 定理得证.

关于 $\theta_n(x)$ 的更进一步的渐近性质, 我们将另文逐一研究.

参 考 文 献

- [1] 郑忠国 条件中位数的最近邻估计和它的 *Bootstrap* 统计量的渐近性质. 中国科学 1984, 1074—1088
- [2] 郑忠国 条件中位数的最近邻估计的渐近性质. 数学学报
- [3] 陈希孺、赵林城 中位数回归的近邻估计的最佳收敛速度. 工程数学学报 1984
- [4] Truong, Y. K. *Asymptotic properties of kernel estimators based on local medians*. Ann. Statist. 17(1989), 606—617
- [5] Härdle, W. and Luckhaus. S. *Uniform consistency of a class of regression function estimators*. Ann. Statist. 12(1984), 612—623
- [6] Wheeden, R. L. and Iygmund, A. "Measure and Integral". Marcel Dekker, NEW York 1977
- [7] 赵林城、苏淳 非参数回归函数最近邻估计的强收敛速度. 数学学报 29(1986), 63—69

Pointwise Consistency of L_1 -Norm Nearest Neighbour Estimator of the Conditional Median

Hong Shengyan
(Anhui University)

Abstract

Let the population (X, Y) be an $R^d \times R^1$ -valued random variable. Based on a sequence of random samples drawn from this population, we construct L_1 -norm nearest neighbor estimator of the conditional median of Y on X which is defined as the median of the conditional distribution function of Y for given X . We obtain the pointwise consistency of this estimator under mild conditions.