



HIV 模型的统计诊断

周杰^①, 刘三阳^①, 周芳^②, WU HuLin^③

① 西安电子科技大学应用数学系, 西安 710071;

② 河南大学附属淮河医院血液肿瘤内科, 开封 475000;

③ Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA

E-mail: zhoujie@xidian.edu.cn

2011-10-12 收稿, 2012-02-17 接受

国家自然科学基金(60974082, 61075055)和中央高校基本科研业务费专项资金(K50510700007)资助

摘要 描述 HIV 的数学模型是一组非线性常微分方程, 其中包括 CD4+T 细胞再感染率和死亡率, HIV 病毒死亡率等多个重要未知参数, 准确估计这些参数有助于正确了解患者病情发展, 以采用个人化治疗方案. 针对此模型已提出多个参数估计方法, 包括基于数值解的非线性最小二乘, 截面似然及两步估计. 另一方面, 由于客观因素影响, 各观测数据对于参数估计的影响大小不同, 找出对参数估计影响大的观测值, 进一步分析其原因具有重要意义. 基于 HIV 模型的两步估计, 构造了用于影响分析的非参数 Cook 型统计量, 并给出其大样本分布. 通过模拟数据和临床数据分析发现: (1) 所构造统计量可以检测中度以上偏移. (2) 相对于内点, 边界点对 HIV 模型的参数估计影响更大. 基于以上结论, 建议对 HIV 模型进行参数估计时, 对边界点应加以格外关注. 提出的方法也可以推广到其他线性常微分方程模型的统计诊断问题中.

关键词

HIV 模型

核光滑

影响分析

Cook 距离

艾滋病是威胁人类健康的严重疾病, 从 20 世纪 80 年代就开始了相关药物和治疗方案的研究, 以 Ho 等人^[1]的研究工作为基础所提出的鸡尾酒疗法已大大延长了病人的生存时间和质量. 在这些研究中, 对 HIV 病毒动力学的深入理解发挥了关键的基础作用, 而对其所建立的数学模型也逐步从简单的线性回归模型发展到目前复杂的微分方程模型. 这些微分方程模型可以描述病毒与宿主之间不同的状态关系, 譬如针对抗病毒治疗前所建立的模型不同于进行抗病毒治疗后的模型, 采用抗逆转录酶抑制剂治疗与采用蛋白酶抑制剂治疗所对应的模型也是不同的. 如果考虑到病毒的耐药性, 模型相应就更加复杂. 总之, 由于实际问题的需要, 利用常微分方程对 HIV 病毒动力学的建模越来越深入细致, Perelson^[2]对此作了全面的综述. 考虑下面的常微分方程, 它描述进行抗病毒治疗的艾滋病患者体内 HIV 病毒数量, 淋巴细胞数量之间的动态关系

$$\begin{cases} DT_u(t) = \lambda - \rho T_u(t) - \eta(t)T_u(t)V(t), \\ DT_l(t) = \eta(t)T_u(t)V(t) - \delta T_l(t), \\ DV(t) = N\delta T_l(t) - cV(t), \end{cases} \quad (1)$$

其中 $T_u(t)$, $T_l(t)$, $V(t)$ 分别表示 t 时刻患者体内未被感染 CD4+T 细胞数量, 已被感染 CD4+T 细胞数量和 HIV 病毒数量. λ 表示体内新生 CD4+T 细胞的速率, ρ 表示正常 CD4+T 细胞的死亡率, η 表示正常 CD4+T 细胞被感染的速率, δ 表示已感染 CD4+T 细胞的死亡率, N 表示单个已感染 CD4+T 细胞在死亡前产生病毒的数量, c 表示病毒的死亡率. 此模型在临床实验中已被广泛采用, 能够很好地和实际观测数据相匹配.

上述模型中的参数对于判断患者免疫系统对病毒的反应以及病情发展具有关键性的作用, 所以参数估计问题就成为理论和应用上的一个重要问题. 对于给定的观测数据, 目前文献中主要有 3 种类型的参数估计方法, 分别是基于微分方程数值解的非线性最小二乘法^[3]; 基于惩罚样条的截面似然法^[4-6];

基于局部多项式核光滑的两步估计^[7-9].

虽然 HIV 模型具有多种参数估计方法,但相应的统计诊断方法还没有建立起来.统计诊断是现代统计学的重要分支之一,通过统计诊断,可以发现模型的设定是否存在问题,观测值中是否存在异常点、强影响点等.若发现模型和数据存在问题,则需对模型和数据进行更深入的分析,改进统计推断结果.众所周知,临床实践中的数据由于各种客观条件的影响,譬如设备更换、数据记录人员失误、病毒变异等,均可造成数据意料之外的改变,从而影响模型参数的估计值.通过统计诊断,发现 HIV 模型参数估计中的强影响点,进而分析其成因,有助于更好地了解患者的病情发展,制定有针对性的治疗方案,这对于临床实践具有重要的意义.

本文针对上述 HIV 的动力学方程,考察了基于两步估计的统计诊断方法,即经典的数据删除模型.我们得到了在数据删除前后参数估计变化的近似表达式,在此基础上,构造了 Cook 型的诊断统计量,证明了其分布收敛至卡方分布.利用构造的诊断统计量,我们发现在 HIV 模型的两步估计中,边界点相对于内点对参数估计具有更大的影响.基于此,在对 HIV 模型的参数估计中,我们建议对边界点应给予更大的关注,以得到更稳健的参数估计.本文提出的统计诊断方法适用于所有带测量误差的线性常微分方程,我们也讨论了方法可能的推广与改进.

1 HIV 模型的两步估计

这一节我们简要回顾一下 HIV 模型的两步估计法.对于模型(1),病毒数量 $V(t)$ 和 CD4+T 细胞量 $T(t)=T_U(t)+T_I(t)$ 在临床上是可以测量的,设测量数据为

$$Y^{(V)}(t_n)=V(t_n)+\varepsilon_{1n}, Y^{(T)}(t_n)=T(t_n)+\varepsilon_{2n}, n=1,2,\cdots N,$$

其中 $\varepsilon_{1n}, \varepsilon_{2n}$ 表示相互独立的测量误差,满足 $E\varepsilon_{1n}=E\varepsilon_{2n}=0$, $\text{Var}(\varepsilon_{1n})=\sigma_1^2$, $\text{Var}(\varepsilon_{2n})=\sigma_2^2$. 给定上述测量值,若无其他限制条件,则模型(1)中的参数是不可识别的.为此文献中的通常方法是假定一些参数为事先给定的经验值,而估计其他感兴趣的参数.这里我们假定已被 HIV 感染的 CD4+T 细胞 T_I 的死亡率为已知的,即 $\delta=0.5$,这里和文献中已有研究是一致的^[10].

由模型(1)

$$DT_I(t)+DT_U(t)=\lambda-\rho T_U(t)-\delta T_I(t),$$

定义 $T(t)=T_U(t)+T_I(t)$,则由上式,

$$DT(t)=\lambda-\rho[T(t)-T_I(t)]-\delta T_I(t),$$

也即

$$T_I(t)=\frac{-\lambda}{\rho-\delta}+\frac{\rho}{\rho-\delta}T+\frac{1}{\rho-\delta}DT(t),$$

把此式代入(1)中的第三式

$$DV(t)=\alpha_0+\alpha_1T(t)+\alpha_2DT(t)-cV(t), \quad (2)$$

其中

$$\alpha_0=\frac{N\delta\lambda}{\rho-\delta}, \quad \alpha_1=\frac{N\delta\rho}{\rho-\delta}, \quad \alpha_2=\frac{N\delta}{\rho-\delta},$$

注意到

$$\lambda=-\frac{\alpha_0}{\alpha_2}, \quad \rho=\frac{\alpha_1}{\alpha_2}, \quad N=\frac{\alpha_1}{\delta}-\alpha_2,$$

因此若有了 $\alpha_0, \alpha_1, \alpha_2$ 和 c 的估计,就可得到原始参数 λ, ρ, N, c 的估计,而对应的标准差可利用 δ 方法得到.

所谓两步估计,第一步首先利用非参数的方法(局部多项式核光滑或样条光滑)估计 $V(t), T(t)$ 及其导数 $DV(t), DT(t)$,然后把所得到的估计 $\hat{V}(t), \hat{DV}(t), \hat{T}(t), \hat{DT}(t)$ 代入(2)式

$$\hat{DV}(t)=\alpha_0+\alpha_1\hat{T}(t)+\alpha_2\hat{DT}(t)-c\hat{V}(t)+\Delta(t), \quad (3)$$

其中

$$\Delta(t)=\left(\hat{DV}(t)-DV(t)\right)-\alpha_1\left(\hat{T}(t)-T(t)\right)-\alpha_2\left(\hat{DT}(t)-DT(t)\right)+c\left(\hat{V}(t)-V(t)\right),$$

在本文中 $V(t), T(t)$ 采用局部线性估计,而 $DV(t), DT(t)$ 采用局部二次估计.第二步求解下述最小化问题得到参数 $\theta=(\alpha_0, \alpha_1, \alpha_2, c)$ 的估计

$$\hat{\theta}=\arg \min _{\theta} S_N, \quad (4)$$

其中

$$S_N=\sum_{n=1}^N\left(\hat{DV}\left(t_n\right)-\alpha_0-\alpha_1 \hat{T}\left(t_n\right)-\alpha_2 \hat{DT}\left(t_n\right)+c \hat{V}\left(t_n\right)\right)^2.$$

由于(3)式中 $\Delta(t)$ 的均值不为零,所以称这一步为拟最小二乘,易知

$$\hat{\theta}=\left(U^{\mathrm{T}} U\right)^{-1} U^{\mathrm{T}} Y, \quad (5)$$

其中

$$U=\left(u_1, u_2, \ldots, u_N\right)^{\mathrm{T}}=\left[\begin{array}{cccc} 1 & \hat{T}\left(t_1\right) & \hat{DT}\left(t_1\right) & -\hat{V}\left(t_1\right) \\ 1 & \hat{T}\left(t_2\right) & \hat{DT}\left(t_2\right) & -\hat{V}\left(t_2\right) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & \hat{T}\left(t_N\right) & \hat{DT}\left(t_N\right) & -\hat{V}\left(t_N\right) \end{array}\right],$$

$$Y = (y_1, y_2, \dots, y_N)^T = (\hat{D}V(t_1), \hat{D}V(t_2), \dots, \hat{D}V(t_N))^T.$$

由此所得到的参数估计是 $\hat{\theta}$ 相合估计, 其渐近分布是正态分布^[7].

2 统计诊断: 基于两步估计的数据删除模型

在第二部分两步估计方法的基础之上, 我们考虑基于数据删除模型的 HIV 模型统计诊断. 记删除第 i 个观测之后 $V(t)$, $T(t)$, $DV(t)$, $DT(t)$ 的估计分别为 $\hat{V}_i(t)$, $\hat{T}_i(t)$, $\hat{D}V_i(t)$, $\hat{D}T_i(t)$. 与(5)类似, 此时参数 θ 的估计为

$$\hat{\theta}(i) = \arg \min_{\theta} S_N(i) = (U^T(i)U(i))^{-1} U^T(i)Y(i), \quad (6)$$

其中

$$S_N(i) = \sum_{n \neq i} \left(\hat{D}V_i(t_n) - \alpha_0 - \alpha_1 \hat{T}_i(t_n) - \alpha_2 \hat{D}T_i(t_n) - c \hat{V}_i(t_n) \right)^2,$$

$U(i)$ 表示将矩阵 U 中的第 i 行删除, 同时将元素 $\hat{V}_i(t_n)$, $\hat{T}_i(t_n)$, $\hat{D}T_i(t_n)$ 分别替换为 $\hat{V}_i(t_n)$, $\hat{T}_i(t_n)$, $\hat{D}T_i(t_n)$ 后所得矩阵, 而 $Y(i) = (y_{i1}, \dots, y_{iN})$ 表示将向量 Y 中第 i 个元素删除, 同时将元素 $\hat{D}V(t_n)$ 替换为 $\hat{D}V_i(t_n)$ 后所得向量.

由附录, 数据删除前后参数估计的关系为

$$\begin{aligned} & \hat{\theta} - \hat{\theta}(i) \\ &= \left(\sum_{n=1}^N u_n u_n^T \right)^{-1} \left\{ \frac{u_i (y_i - u_i^T \hat{\theta})}{1 - Q_{ii}} + \sum_{n \neq i} \left[(e_{i1}^{(V)}(t_n) - e_{i1}^T \hat{\theta}) u_n \right] \right\} \\ &+ o_p \left(\frac{1}{Nh} \right), \end{aligned} \quad (7)$$

其中

$$Q = (Q_{ij})_{N \times N} = U(U^T U)^{-1} U^T,$$

$$e_{i1}^{(V)}(t_n) = \hat{D}V(t_n) - \hat{D}_i V(t_n),$$

$$e_{in} = (0, e_{i0}^{(T)}(t_n), e_{i1}^{(T)}(t_n), e_{i0}^{(V)}(t_n))^T = u_n - u_{in}.$$

由(7)式可得

$$\frac{Nh}{\sigma \mu_2(K)} (\hat{\theta} - \hat{\theta}(i)) \sim N(0, (E \tilde{u}_n \tilde{u}_n^T)^{-1}), \quad (8)$$

其中 K 为核估计中的核函数, h 为窗宽,

$$\mu_n(K) = \int x^n K(x) dx, \quad \tilde{u}_n = (1, T(t_n), DT(t_n), -V(t_n))^T,$$

$$\sigma^2 = \sigma_1^2 + \alpha_2^2 \sigma_2^2.$$

(7)式和(8)式的证明参见附录. 由此定义 Cook 距离为

$$D_i = \frac{N^2 h^2}{\sigma^2 \mu_2^2(K)} (\hat{\theta} - \hat{\theta}(i))^T E \tilde{u}_n \tilde{u}_n^T (\hat{\theta} - \hat{\theta}(i)) \sim \chi^2(4). \quad (9)$$

由上式可知计算 Cook 距离 D_i 须计算数据删除前后参数估计差 $(\hat{\theta} - \hat{\theta}(i))$, 而由(7)式, 计算 $(\hat{\theta} - \hat{\theta}(i))$ 须首先

计算 $e_{i0}^{(V)}(t_n)$, $e_{i1}^{(V)}(t_n)$, $e_{i0}^{(T)}(t_n)$, $e_{i1}^{(T)}(t_n)$. 由附录中引理 1, 这些量的计算并不需要对数据删除后的数据重新作核光滑, 而只需利用完全数据情形下的核光滑产生的残差即可计算, 这大大缩短了算法的运行时间. 其次, 计算 D_i 还需要给出未知参数 σ^2 和 $E \tilde{u}_n \tilde{u}_n^T$ 的估计, 这里 σ^2 利用完全数据下由核估计所得到的残差进行估计, 而

$$E \tilde{u}_n \tilde{u}_n^T \text{ 利用完全数据下的经验均值 } \frac{1}{N} \sum_{n=1}^N u_n u_n^T \text{ 估计,}$$

这些估计都是相合估计. 下面给出计算 Cook 距离 D_i 的算法:

步骤 1. 利用完全数据.

(I) 计算参数 θ 的两步估计 $\hat{\theta}$, 误差方差 σ^2 的估计

$$\sigma^2 = \hat{\sigma}_1^2 + \hat{\alpha}_2^2 \hat{\sigma}_2^2, \quad E \tilde{u}_n \tilde{u}_n^T \text{ 的估计 } \frac{1}{N} \sum_{n=1}^N u_n u_n^T;$$

(II) 对 $1 \leq n \leq N$, 计算 $\hat{V}(t_n)$, $\hat{T}(t_n)$, $\hat{D}T(t_n)$, $\hat{D}V(t_n)$;

(III) 对 $1 \leq n \leq N$, $j = 0, 1$, 计算核光滑残差 $r_{ij}^{(V)}(t_n)$, $r_{ij}^{(T)}(t_n)$ (定义可见附录中引理 1).

步骤 2. 利用步骤 1(III) 中的残差 $r_{ij}^{(V)}(t_n)$, $r_{ij}^{(T)}(t_n)$

计算 $e_{i0}^{(V)}(t_n)$, $e_{i1}^{(V)}(t_n)$, $e_{i0}^{(T)}(t_n)$, $e_{i1}^{(T)}(t_n)$.

步骤 3. 把上述统计量代入(9)计算 Cook 距离 D_i .

3 数据分析

3.1 模拟数据分析

这一节利用模拟数据研究第三节提出的 HIV 模型强影响点检测方法. 对模型(1), 其中的参数设定为 $\lambda = 7$, $\rho = 0.108$, $\eta = 9 \times 10^{-6}$, $\delta = 0.5$, $N = 5$, $c = 5$, 这些参数设定值在 HIV 感染者是常见的^[7]. 我们利用四阶 Runge-Kutta 算法计算模型在区间 $[0, 20]$ 内的数值解, 步长为 0.01, 所得到的数值解记 $T(t_n) = T_U(t_n) + T_I(t_n)$, $V(t_n)$, 测量值为

$$Y^{(V)}(t_n) = V(t_n) + \varepsilon_{1n}, \quad Y^{(T)}(t_n) = T(t_n) + \varepsilon_{2n}, \quad n = 1, 2, \dots, N,$$

其中 $\varepsilon_{1n} \sim N(0, 20^2)$, $\varepsilon_{2n} \sim N(0, 6^2)$, ε_{1n} 和 ε_{2n} 是相互独立的测量误差. 对 $Y^{(V)}(t_n)$, $Y^{(T)}(t_n)$, 利用局部线性核光滑估计 $\hat{V}(t)$ 和 $\hat{T}(t)$, 利用局部二次核光滑估计 $\hat{D}V(t)$ 和 $\hat{D}T(t)$, 以上核估计的窗宽都根据 CV 准则选取.

为检验算法的有效性, 对第 100 个观测值, 我们对病毒和 T 细胞测量值加上均值偏移 a , 即

$$Y^{(V)}(t_{100}) = V(t_{100})(1 + a) + \varepsilon_{1(100)},$$

$$Y^{(T)}(t_{100}) = T(t_{100})(1 + a) + \varepsilon_{2(100)}.$$

给定 a , 利用(9)式可计算每一个数据点的 Cook 距离. 我们定义若第 100 个数据点对应的 Cook 距离是所有数据点中 Cook 距离最大的, 则算法是有效的. 对每一个 a , 重复上述过程 M 次, 其中有效次数为 n , 则有效率定义为 $p = n/M$. 我们考虑了 a 的四种选择, 即 $a = 0.1, 0.3, 0.6, 1$, 并取 $M = 500$, 对应的有效率列于表 1 的第二行. 因为在边界点处核光滑会产生较大的均方误差, 所以我们也计算了去掉前 10 个和后 10 个边界点后的有效率, 列于表 1 的第 3 行.

由表 1 可知, 由(9)式所定义的 Cook 距离对于中等程度的偏移检测是有效的, 而对于较小的偏移效果不好; 其次, 边界点对于强影响点的检测有效率有着很大的影响, 我们认为这是由核光滑的性质所导致的. 众所周知, 局部多项式核估计在边界点处具有较大的均方误差, 所以在此基础之上建立的两步估计必然在边界点处具有较大的变异, 也即相对于内点, 边界点对于 HIV 模型的参数估计具有更大的影响. 由此对 HIV 模型进行两步估计的时候, 需要对边界点加以慎重考虑. 图 1 给出了不同偏移下 Cook 统计量, 由图可知相对于内点, Cook 统计量在边界点的

确取到较大值.

3.2 临床试验数据分析

本节我们应用上述统计诊断方法考察临床试验中艾滋病患者的实际测量数据. 此数据由罗彻斯特大学医学研究中心(URMC) 采集, 已有多文献利用此数据对艾滋病动力学方程做了研究^[8,10,11]. 此数据对两位进行抗病毒和免疫调节治疗的 HIV 感染者, 分别在前两周(以小时为单位) 0, 1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 28, 32, 40, 46, 52, 58, 64, 70, 144, 240 时刻进行测量, 然后在第 3, 4, 6, 8, 12, 14, 16, 20, 24, 28, 32, 36, 40, 44 和 48 周各测量一次. 图 2 中给出了两位感染者 HIV 病毒量, CD4+T 细胞及相应导数的变化曲线. 可以看出, 接受治疗后, 两位感染者的病毒数量呈下降趋势, 而 CD4+T 细胞数量呈递增趋势.

表 1 强影响点检测有效率

偏移	$a=0.1$	$a=0.3$	$a=0.6$	$a=1$
有效率 p_1	0.5%	11%	62%	85.5%
有效率 p_2	8.23%	60%	85%	96%

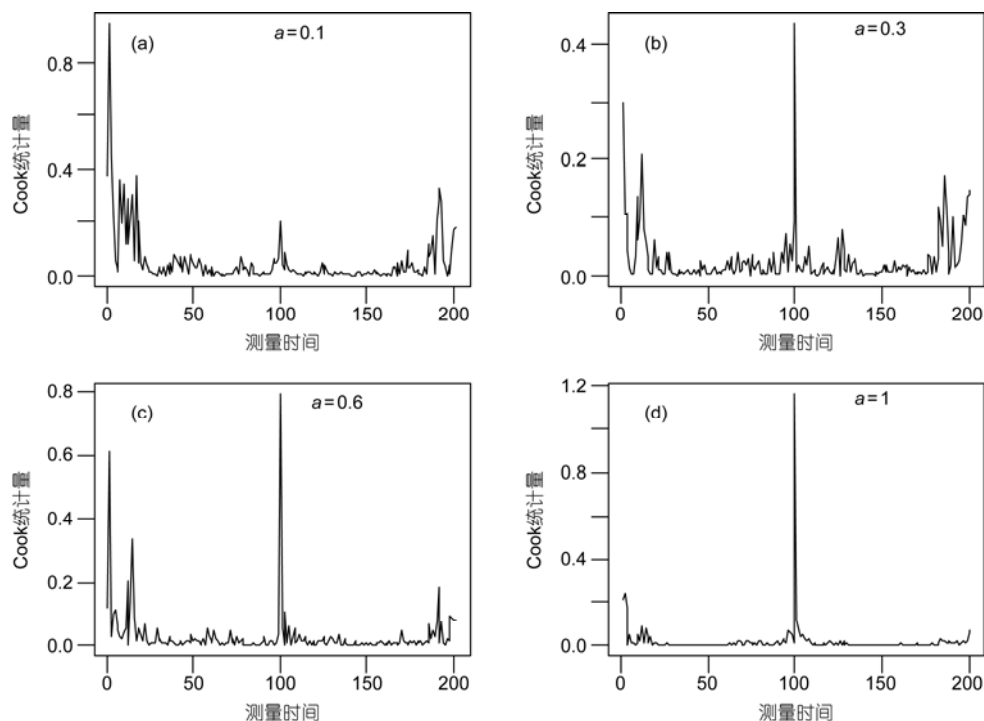


图 1 不同偏移下的 Cook 距离, 4 种情形下在边界点处 Cook 距离取值都较大

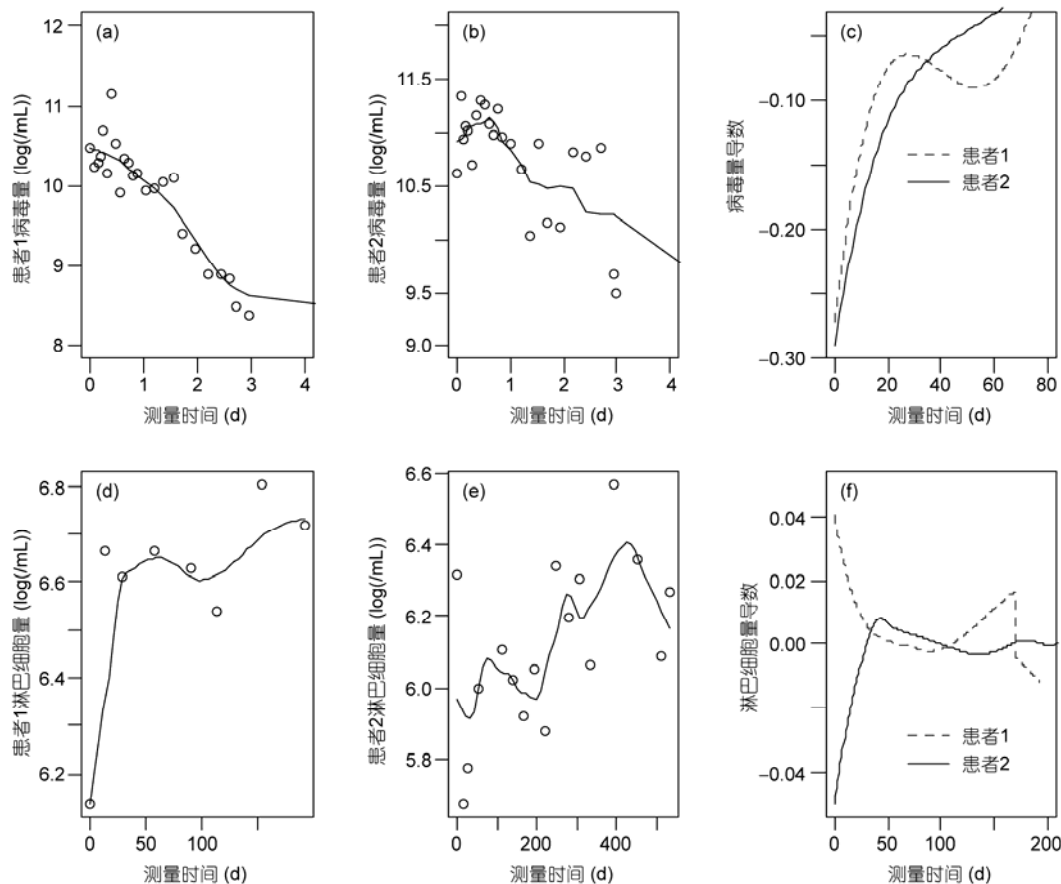


图2 两感染者 HIV 量及其导数的核估计(第1行), 两感染者淋巴细胞量及其导数的核估计(第2行)

对上述数据作对数变换, 对变换后的数据采用两步估计法, 其中在利用局部多项式核估计求曲线及其导数时, 窗宽的选择根据 CV 标准, 核函数 K 采用正态核, 参数估计的结果在表 2 的第 2 行和第 3 行中给出, 其中括号内的数字表示参数估计的标准差. 图 3 给出了两名患者对应的 Cook 统计量, 显然边界点对于参数的估计有着更大的影响.

为了评价边界点对参数估计的作用, 在对曲线及其导数进行局部多项式估计之后, 在两步估计的第 2 步, 对两个患者分别去掉前 3 个和后面 3 个观测点, 只利用剩余内点重新估计参数, 结果列于表 2 的后两行.

由表 2, 去掉边界点之后参数估计的变化虽然不是特别大, 但是和标准差相比, 这个变化还是很显著的. 所以在对 HIV 模型的两步估计中, 如何利用边界点是一个需要进一步讨论的问题, 这可以看作是两步估计相对于其他估计方法的缺陷.

表 2 HIV 感染者的参数估计

参数	λ	ρ	N	c
患者 1	7.012(2.3)	0.0546(0.03)	5.60(0.023)	0.0248(0.003)
患者 2	7.61(1.82)	-0.0613(0.041)	5.02(0.019)	0.035(0.002)
患者 1	7.036(2.41)	0.0629(0.052)	5.35(0.025)	0.029(0.004)
患者 2	7.65(1.88)	-0.0938(0.047)	5.41(0.023)	0.028(0.006)

4 结论

HIV 模型的统计推断问题在临床医学中有着重要的应用, 精确的参数估计对于制定个人化的治疗方案起着关键的作用. 本文基于 HIV 模型的两步估计, 研究了 HIV 模型参数估计的统计诊断问题, 主要包括: (1) 在 HIV 模型两步估计的基础上, 给出了用于强影响点检测的 Cook 距离表达式和渐近分布. (2) 利用所给 Cook 距离, 发现 HIV 模型两步估计的边界点对于参数估计总会产生较大的影响. 基于

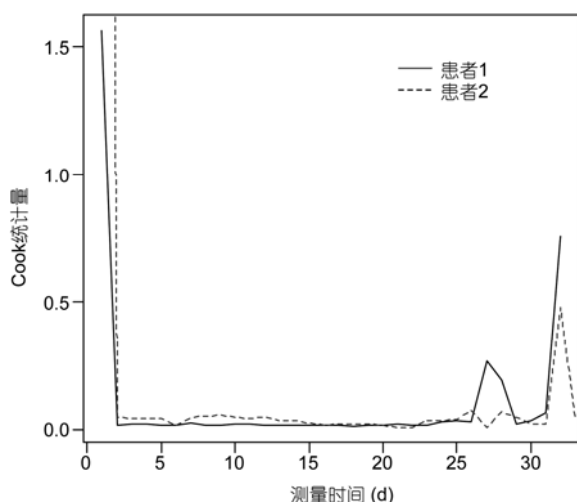


图3 两 HIV 感染者 Cook 距离序列图

上述结论, 对 HIV 模型的参数估计, 我们建议对边界点应加以格外的关注, 譬如在两步估计中的第二步,

可删除部分边界点, 以增加参数估计的稳健性.

本文提出的方法也适用于其他线性常微分方程的统计诊断问题. 非线性常微分方程在应用中更常见, 譬如对 HIV 模型, 如果只有病毒量的测量值, 则模型(1) 将只能化简为非线性的常微分方程, 而得不到类似(2) 式的线性微分方程, 在这种情形下模型的统计诊断问题还需进一步的研究.

对于 HIV 模型, 文献中还有其他的参数估计方法, 譬如非线性最小二乘法, 截面似然法等, 在这些估计方法的基础上对 HIV 模型进行统计诊断值得进一步研究. 另外, 本文考虑的是数据删除模型, 对 HIV 模型我们也可以考虑其他的统计诊断方法, 譬如局部影响分析, 这也是很有意义的. 总之, 对 HIV 模型及其相关的微分方程模型, 统计诊断在理论上和临床上都具有重要的意义, 但在文献中还没有得到应有的重视, 与此相关的大量工作还需要进一步的研究.

参考文献

- 1 Ho D D, Neumann A U, Perelson A S, et al. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 1995, 373: 123-126
- 2 Perelson A S. Mathematical analysis of HIV-1 dynamics *in vivo*. *SIAM Rev*, 1999, 41: 3-44
- 3 Bates D M, Watts D G. *Nonlinear Regression and Its Application*. New York: John Wiley & Sons, Inc, 1988
- 4 Ramsay J O, Hook G, Campbell V, et al. Parameter estimation for differential equation: A generalized smoothing approach. *J Roy Stat Soc B*, 2007, 69: 741-796
- 5 Cao J, Ramsay J O. Parameter cascades and profiling in functional data analysis. *Comput Stat*, 2007, 22: 335-351
- 6 Poyton A A, Varziri M S, McAuley K B, et al. Parameter estimation in continuous time dynamic models using principal differential analysis. *Comput Chem Eng*, 2006, 30: 698-708
- 7 Liang H, Wu H L. Parameter estimation for differential equation models using a framework of measurement error in regression models. *J Am Stat Assoc*, 2008, 103: 1570-1583
- 8 Chen J, Wu H L. Efficient local estimation for time varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *J Am Stat Assoc*, 2008, 103: 369-384
- 9 Chen J, Wu H L. Estimation of time varying parameters in deterministic dynamic models. *Statist Sin*, 2008, 18: 987-1006
- 10 Liang H, Miao H Y, Wu H L. Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. *Ann Appl Stat*, 2010, 4: 460-483
- 11 Fan J, Gijbels I. *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall, 1996
- 12 Cook R D, Weisberg S. *Residuals and Influence in Regression*. London: Chapman & Hall, 1982
- 13 韦博成, 林金官, 解锋昌. 统计诊断. 北京: 高等教育出版社, 2009
- 14 Miller K S. On the inverse of sum of matrices. *Math Mag*, 1981, 54: 67-72

Statistical diagnosis of HIV model

ZHOU Jie¹, LIU SanYang¹, ZHOU Fang² & WU HuLin³

¹Department of Mathematics, Xidian University, Xi'An 710071, China;

²Department of Oncology Affiliated Huaihe Hospital, Henan University, KaiFeng 475000, China;

³Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA

The dynamics of the human immunodeficiency virus (HIV) can be described by ordinary differential equations (ODEs), in which the unknown parameters include the infected rate and death rate of CD4+ T cell, the death rate of virus etc. Accurate estimation of these parameters plays a major role in the personalized treatment. Presently the main parameter estimation approaches include numerical solution based nonlinear OLS, generalized profiling likelihood and two stage estimation. On the other hand, influential analysis for dynamics of HIV, though important for the reliable parameter estimation, has received little attention in the literatures. Based on the two stage estimation of HIV dynamics, a nonparametric Cook type statistic is constructed in this paper to detect the influential measurements. The limit distribution of such Cook type statistic is established. By applying the proposed Cook statistic to the simulated and clinical trial data we find (1) the middle bias can be detected by the proposed Cook statistic, (2) compared to the interior measurements, the boundary measurements have more impact on the estimation of the parameters. Based on these observations, we suggest that the boundary measurements should be paid more attentions than the interior measurements for the parameters estimation of HIV dynamics. The proposed influential measurements detection procedure can be generalized to other linear ODEs with measurement error.

HIV modeling, kernel smoothing, influential analysis, Cook distance

doi: 10.1360/972011-1978

补充材料

附录

本文的以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.



附录:

本节我们给出论文中相关理论结果的证明概要. 为表述方便, 不妨设对 $V(t)$, $T(t)$, $DV(t)$, $DT(t)$ 的估计中采用相同的核函数 $K(t)$, 窗宽为 h , 并满足以下假设条件: (I) 核函数 K 关于 0 点对称且支撑为 $[-1, 1]$. (II) 测量时间 t_i 是独立同分布的随机变量, 其密度函数 $p(t) \neq 0$, 且具有连续一阶导数. (III) 窗宽 h 满足 $\lim_{N \rightarrow +\infty} \frac{1}{Nh} = 0$.

记

$$T_{1t} = \begin{pmatrix} 1 & t_1 - t \\ 1 & t_2 - t \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & t_N - t \end{pmatrix}_{N \times 2}, \quad T_{2t} = \begin{pmatrix} 1 & t_1 - t & (t_1 - t)^2 \\ 1 & t_2 - t & (t_2 - t)^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & t_N - t & (t_N - t)^2 \end{pmatrix}_{N \times 3},$$

$$K_h(t) = \frac{1}{h} K(t/h), \quad W_t = \text{diag}(K_h(t_1 - t), \dots, K_h(t_N - t)),$$

$$P^{(j)}(t) = (P_{mn}^{(j)}(t))_{N \times N} = W_t^{1/2} T_{jt} (T_{jt}^T W_t T_{jt})^{-1} T_{jt}^T W_t^{1/2},$$

则由 Fan 等^[11], $V(t)$ 和 $T(t)$ 的局部线性核估计为

$$\hat{V}(t) = \xi_1^T (T_{1t}^T W_t T_{1t})^{-1} T_{1t}^T W_t Y^{(V)} = \xi_1^T \hat{\alpha}_{1t}^{(V)}, \quad \hat{T}(t) = \xi_1^T (T_{1t}^T W_t T_{1t})^{-1} T_{1t}^T W_t Y^{(T)} = \xi_1^T \hat{\alpha}_{1t}^{(T)}, \quad (10)$$

$\hat{DV}(t)$ 和 $\hat{DT}(t)$ 的局部二次核估计为

$$\hat{DV}(t) = \xi_2^T (T_{2t}^T W_t T_{2t})^{-1} T_{2t}^T W_t Y^{(V)} = \xi_2^T \hat{\alpha}_{2t}^{(V)}, \quad \hat{DT}(t) = \xi_2^T (T_{2t}^T W_t T_{2t})^{-1} T_{2t}^T W_t Y^{(T)} = \xi_2^T \hat{\alpha}_{2t}^{(T)}, \quad (11)$$

其中 $\xi_1 = (1, 0)^T$, $\xi_2 = (0, 1, 0)^T$, $Y^{(V)}$ 和 $Y^{(T)}$ 分别表示由病毒细胞和 CD4+T 细胞的测量值所组成的 N 维列向量. 类似地, 删除第 i 个观测之后, 我们可得到相应的核估计 $\hat{D}_i V(t)$, $\hat{D}_i T(t)$, $\hat{V}_i(t)$, $\hat{T}_i(t)$.

记

$$e_{i1}^{(V)}(t) = \hat{DV}(t) - \hat{D}_i V(t), \quad e_{i1}^{(T)}(t) = \hat{DT}(t) - \hat{D}_i T(t)$$

$$e_{i0}^{(V)}(t) = \hat{V}(t) - \hat{V}_i(t), \quad e_{i0}^{(T)}(t) = \hat{T}(t) - \hat{T}_i(t), \quad e_i(t) = (0, e_{i0}^{(T)}(t), e_{i1}^{(T)}(t), e_{i0}^{(V)}(t))^T,$$

为了计算上述函数改变量及导数改变量, 需要在数据删除前后分别进行局部多项式核光滑, 这在数据量较大时是不可行的. 下面给出的引理说明我们只需利用完全数据下的局部多项式核光滑即可计算这些改变量, 而无需对删除后的数据进行核光滑.

引理 对 $j=0, 1$, 记 $\xi_n^{(j+1)}(t) = (1, t_n - t, \dots, (t_n - t)^{j+1})^T$, $r_{ij}^{(V)}(t) = Y^{(V)}(t_i) - (\xi_i^{(j+1)})^T \hat{\alpha}_{(j+1)t}^{(V)}$, $r_{ij}^{(T)}(t) = Y^{(T)}(t_i) - (\xi_i^{(j+1)})^T \hat{\alpha}_{(j+1)t}^{(T)}$, 则

$$e_{ij}^{(V)}(t) = \frac{r_{ij}^{(V)}(t) K_h(t_i - t) (t_i - t)^{j+1}}{N p(t) h^{2j+1}} + o_p \left(\frac{1}{N h^j} \right), \quad (12)$$

$$e_{ij}^{(T)}(t) = \frac{r_{ij}^{(T)}(t) K_h(t_i - t) (t_i - t)^{j+1}}{N p(t) h^{2j+1}} + o_p \left(\frac{1}{N h^j} \right), \quad (13)$$

其中 $p(t)$ 为 t_i ($i=1, 2, \dots, N$) 的密度函数, $\alpha_u^{(V)}$, $\alpha_u^{(T)}$ 由(10)和(11)式所定义.

证明: 我们只对 $e_{ij}^{(V)}(t)$ 证明 $j=1$ 时的情形, 其他情况证明是类似的. 由 Cook 等^[12]或韦博成等^[13]可知

$$\begin{aligned} e_{il}^{(V)}(t) &= \hat{D}V(t) - \hat{D}_i V(t) = \xi_2^T (\hat{\alpha}_{2t}^{(V)} - \hat{\alpha}_{2t}^{(V)}(i)) \\ &= \frac{\xi_2^T K_h(t_i - t) (T_{2t}^T W_t T_{2t})^{-1} \xi_i^{(2)}(t) r_{il}^{(V)}(t)}{1 - P_{ii}^{(2)}(t)}, \end{aligned}$$

其中 $P_{ii}^{(j)}$ 为 $P^{(j)}$ 第 i 个对角元. 由

$$\frac{1}{N} \sum_{n=1}^N (t_n - t)^l K_h(t_n - t) = \begin{cases} h^l \mu_l(K) p(t) + o_p(h^l) & l \text{ 为偶数} \\ h^{l+1} \mu_{l+1}(K) Dp(t) + o_p(h^{l+1}) & l \text{ 为奇数} \end{cases}$$

可知

$$\frac{1}{N} T_{2t}^T W_t T_{2t} = M_3 (p(t) N_3 + h Dp(t) Q_3) M_3 + o_p(M_3 I M_3),$$

其中 $M_3 = \text{diag}(1, h, h^2)$, N_3 , Q_3 , I 为 3×3 的矩阵, 元素分别为 $\mu_{i+j-2}(K)$, $\mu_{i+j-1}(K)$ 和 1, 见(8)式后 $\mu_n(K)$ 的定义. 由此

$$\begin{aligned} \xi_2^T \left(\frac{1}{N} T_{2t}^T W_t T_{2t} \right)^{-1} &= \xi_2^T M_3^{-1} (p(t) N_3 + h Dp(t) Q_3)^{-1} M_3^{-1} + o_p(h^{-1} M_3^{-1}) \\ &= \xi_2^T M_3^{-1} p^{-1}(t) \left(N_3^{-1} - h \frac{Dp(t)}{p(t)} N_3^{-1} Q_3 N_3^{-1} \right) M_3^{-1} + o_p(h^{-1} M_3^{-1}) \\ &= \xi_2^T p^{-1}(t) h^{-1} \left(N_3^{-1} - h \frac{Dp(t)}{p(t)} N_3^{-1} Q_3 N_3^{-1} \right) M_3^{-1} + o_p(h^{-1} M_3^{-1}), \end{aligned} \quad (14)$$

由(14)及 $P_{ii}^{(2)}(t)$ 的定义

$$\begin{aligned} P_{ii}^{(2)}(t) &= K_h(t_i - t) (\xi_i^{(2)}(t))^T (T_{2t}^T W_t T_{2t})^{-1} \xi_i^{(2)}(t) \\ \hat{\theta}(i) &= \hat{\theta} - \frac{A^{-1} u_i (y_i - \hat{y}_i)}{1 - Q_{ii}} - A^{-1} \sum_{n \neq i} [(e_{il}^{(V)} - e_i^T(t_n) \hat{\theta}) u_n] + o_p\left(\frac{1}{Nh}\right), \end{aligned}$$

其中 C 是仅依赖于核函数 K 的常数. 由此

$$\begin{aligned} e_{il}^{(V)}(t) &= \frac{1}{N} r_{il}^{(V)}(t) K_h(t_i - t) \xi_2^T \left(\frac{1}{N} T_{2t}^T W_t T_{2t} \right)^{-1} \xi_i^{(2)}(t) \\ &= \frac{r_{il}^{(V)}(t) K_h(t_i - t) (t_i - t)^2}{N p(t) h^3} + o_p\left(\frac{1}{Nh}\right), \end{aligned}$$

此即 $j=1$ 时的(12)式, 证毕.

下面的(16)式用到矩阵和的求逆公式^[14], 为了方便阅读, 这里给出相关的结果.

矩阵和的求逆公式 设 A 和 $A+B$ 为非奇异矩阵, B 的秩 $r>0$. 设 $B = B_1 + \dots + B_r$, 其中对 $1 \leq i \leq r$, B_i 的秩都为 1, 且对 $1 \leq k \leq r$, 有 $C_{k+1} = A + B_1 + \dots + B_k$ 是非奇异矩阵. 设 $C_1 = A$, 则 $C_{k+1}^{-1} = C_k^{-1} - \frac{1}{1 + g_k} C_k^{-1} B_k C_k^{-1}$, 其中 $g_k = \text{trace}\{C_k^{-1} B_k\}$. 特别地, $(A+B)^{-1} = C_r^{-1} - \frac{1}{1 + g_r} C_r^{-1} B_r C_r^{-1}$.

(7)式的证明: 定义 $\sigma^2 = \sigma_1^2 + \alpha_2^2 \sigma_2^2$, $B_i = -u_i u_i^T$, $C_{in} = -u_n e_i^T(t_n)$, $D_{in} = e_i(t_n) e_i^T(t_n)$, 于是由 $\hat{\theta}(i)$ 的定义

$$\begin{aligned} \hat{\theta}(i) &= \left(U^T(i) U(i) \right)^{-1} U^T(i) Y(i) = \left(A + B_i + \sum_{n \neq i} C_{in} + \sum_{n \neq i} C_{in}^T + \sum_{n \neq i} D_{in} \right)^{-1} \\ &\quad \times \left(\sum_{n=1}^N u_n y_n - u_i y_i - \sum_{n \neq i} e_{il}^{(V)}(t_n) u_n - \sum_{n \neq i} e_i(t_n) y_n + \sum_{n \neq i} e_{il}^{(V)}(t_n) e_i(t_n) \right), \end{aligned} \quad (15)$$

由矩阵和的求逆公式

$$\begin{aligned} (U^T(i)U(i))^{-1} &= \left(A + B_i + \sum_{n \neq i} (C_{in} + C_{in}^T) + \sum_{n \neq i} e_i(t_n) e_i^T(t_n) \right)^{-1} \\ &= A^{-1} - \frac{A^{-1} B_i A^{-1}}{1 - Q_{ii}} + \sum_{n \neq i} \frac{A^{-1} (u_n e_{in}^T + e_{in}^T u_n) A^{-1}}{1 - P_{nn}^{(e)}} + o_p \left(\frac{1}{N^2 h} \right), \end{aligned} \quad (16)$$

这里 Q_{ii} 为矩阵 Q 的第 i 个对角元, 而矩阵 Q 在(7)式中定义, $P_{nn}^{(e)} = e_i^T(t_n) A^{-1} u_n$. 把(16)式代入(15)式有

$$\hat{\theta}(i) = \hat{\theta} - \frac{A^{-1} u_i (y_i - \hat{y}_i)}{1 - Q_{ii}} - A^{-1} \sum_{n \neq i} \left[(e_{i1}^{(V)} - e_i^T(t_n) \hat{\theta}) u_n \right] + o_p \left(\frac{1}{Nh} \right),$$

其中 $\hat{y}_i = u_n^T \hat{\theta}$. 由此

$$\hat{\theta} - \hat{\theta}(i) = \left(\sum_{n=1}^N u_n u_n^T \right)^{-1} \left\{ \frac{u_i (y_i - \hat{y}_i)}{1 - Q_{ii}} + \sum_{n \neq i} \left[(e_{i1}^{(V)}(t_n) - e_i^T(t_n) \hat{\theta}) u_n \right] \right\} + o_p \left(\frac{1}{Nh} \right) \quad (17)$$

(8)式的证明: 定义矩阵 $\tilde{U} = (\tilde{u}_1, \dots, \tilde{u}_N)^T$ 为将矩阵 U 中元素 $\hat{V}(t_n)$, $\hat{T}(t_n)$, $\hat{D}T(t_n)$ 分别替换为 $V(t_n)$, $T(t_n)$, $DT(t_n)$ 后所得矩阵, $\tilde{U}(i)$ 表示将 $\tilde{U}$ 的第 i 行删除后所得矩阵. 考虑(17)式

$$\begin{aligned} \sum_{n \neq i} \left[(e_{i1}^{(V)}(t_n) - e_i^T(t_n) \hat{\theta}) u_n \right] &= \sum_{n \neq i} \left[(e_{i1}^{(V)}(t_n) - \alpha_2 e_{i1}^{(T)}(t_n) \hat{\theta}) u_n \right] + o_p \left(\frac{1}{Nh} \right) \\ &= \tilde{U}^T(i) \tau_i + o_p \left(\frac{1}{Nh} \right), \end{aligned} \quad (18)$$

其中 τ_i 为 $N-1$ 维列向量, 其第 n 个元素定义为

$$\tau_i(t_n) = e_{i1}^{(V)}(t_n) - \alpha_2 e_{i1}^{(T)}(t_n).$$

由引理可知

$$\begin{aligned} \tau_i &= \frac{1}{Nh^3} \begin{pmatrix} \gamma_i(t_1) \xi_i^T (I - P^{(2)}(t_1)) W_{t_1}^{1/2} (Y^{(V)} - \alpha_2 Y^{(T)}) \\ \dots \\ \gamma_i(t_N) \xi_i^T (I - P^{(2)}(t_N)) W_{t_N}^{1/2} (Y^{(V)} - \alpha_2 Y^{(T)}) \end{pmatrix}_{(N-1) \times 1} \\ &= \frac{1}{Nh^3} \text{diag} \{ \gamma_i(t_1), \dots, \gamma_i(t_N) \} I_{N-1} \otimes \xi_i^T F_i (X(t) + \varepsilon) = I + II, \end{aligned} \quad (19)$$

其中

$$\gamma_i(t) = \frac{(t_i - t)^2 K_h^{1/2}(t_i - t)}{p(t)}, \quad F_i = \begin{pmatrix} (I - P^{(2)}(t_1)) W_{t_1}^{1/2} \\ \dots \\ (I - P^{(2)}(t_N)) W_{t_N}^{1/2} \end{pmatrix}_{N(N-1) \times N},$$

$X(t) = (V(t_1) - \alpha_2 T(t_1), \dots, V(t_N) - \alpha_2 T(t_N))^T$, $\varepsilon = (\varepsilon_{11} - \alpha_2 \varepsilon_{21}, \dots, \varepsilon_{1N} - \alpha_2 \varepsilon_{2N})^T$, ξ_i 表示第 i 个元素为 1 其余元素为 0 的 N 维列向量, I_N 表示 $N \times N$ 的单位阵. 对于(19)式的第一项,

$$I = \frac{1}{Nh^3} \text{diag} \{ \gamma_i(t_1), \dots, \gamma_i(t_N) \} I \otimes \xi_i^T F_i X(t) = o_p \left(\frac{1}{Nh} \right). \quad (20)$$

对于(19)式的第二项, 给定观测时间 $t_1, \dots, t_N$,

$$\text{Var}\{II \mid t_1, \dots, t_N\} = \frac{\sigma^2}{N^2 h^6} \Sigma,$$

其中

$$\sigma^2 = \sigma_1^2 + \alpha_2^2 \sigma_2^2,$$

$$\Sigma = (\sigma_{n_1 n_2}^2)_{(N-1) \times (N-1)} = \text{diag}(\gamma_i(t_1), \dots, \gamma_i(t_N)) I \otimes \xi_i^T F_i F_i^T I \otimes \xi_i \text{diag}(\gamma_i(t_1), \dots, \gamma_i(t_N)),$$

化简可得

$$\begin{aligned} \sigma_{n_1 n_2}^2 &= \gamma_i(t_{n_1}) \gamma_i(t_{n_2}) \xi_i^T \left(I_N - P^{(2)}(t_{n_1}) \right) W_{t_{n_1}}^{1/2} W_{t_{n_2}}^{1/2} \left(I_N - P^{(2)}(t_{n_2}) \right) \xi_i \\ &= \gamma_i(t_{n_1}) \gamma_i(t_{n_2}) K_h^{1/2}(t_i - t_{n_1}) K_h^{1/2}(t_i - t_{n_2}) + o_p(1), \end{aligned}$$

从而

$$\begin{aligned} \frac{1}{N^2} \tilde{U}^T(i) \Sigma \tilde{U}(i) &= \frac{1}{N} \sum_{n_1 \neq i} K_h(t_i - t_{n_1}) (t_i - t_{n_1})^2 \frac{\tilde{u}_{n_1}}{p(t_{n_1})} \\ &\times \frac{1}{N} \sum_{n_2 \neq i} K_h(t_i - t_{n_2}) (t_i - t_{n_2})^2 \frac{\tilde{u}_{n_2}^T}{p(t_{n_2})} \rightarrow h^4 \mu_2^2(K) \tilde{u}_i \tilde{u}_i^T, \end{aligned} \quad (21)$$

由(17)–(21)

$$\begin{aligned} \text{Var}\{\hat{\theta} - \hat{\theta}(i)\} &= \frac{\sigma^2}{N^2 h^6} E(\tilde{U}^T \tilde{U})^{-1} \tilde{U}^T \Sigma \tilde{U} (\tilde{U}^T \tilde{U})^{-1} \\ &= \frac{\sigma^2}{N^2 h^6} E \left(\frac{1}{N} \tilde{U}^T \tilde{U} \right)^{-1} \left(\frac{1}{N^2} \tilde{U}^T \Sigma \tilde{U} \right) \left(\frac{1}{N} \tilde{U}^T \tilde{U} \right)^{-1} \\ &= \frac{\sigma^2 \mu_2^2(K)}{N^2 h^2} (E \tilde{u}_n \tilde{u}_n^T)^{-1}, \end{aligned}$$

由中心极限定理得

$$\frac{Nh}{\sigma \mu_2(K)} (\hat{\theta} - \hat{\theta}(i))^d \sim N(0, (E \tilde{u}_n \tilde{u}_n^T)^{-1}),$$

此即(8)式, 证毕.