带跳单指标模型的半参数跳点检测估计

赵彦勇¹, 林金官^{1*}, 杜秀丽²

1. 南京审计大学统计科学与大数据研究院, 南京 211815;

2. 南京师范大学数学科学学院, 南京 210023

E-mail: yyzhao@nau.edu.cn, jglin@nau.edu.cn, duxiuli@njnu.edu.cn

收稿日期: 2018-01-27; 接受日期: 2018-05-09; 网络出版日期: 2019-04-22; * 通信作者

国家自然科学基金 (批准号: 11701286 和 11571073)、国家社会科学基金 (批准号: 17CTJ016)、江苏省自然科学基金 (批准号: BK20171073)、全国统计科学研究项目 (批准号: 2017LY71)、江苏省高等学校自然科学基金 (批准号: 17KJB110006) 和江苏省高校哲学社会科学研究基金 (2017SJB0350) 资助项目

摘要 单指标模型是统计学中常用的维数约减模型. 在实际应用中, 连接函数可能有奇异点, 包括某些未知位置上有跳点和某些相关过程的结构变点. 检测这些奇异点对于系数估计和了解结构改变非常重要. 本文基于精细最小平均条件方差估计和函数二阶导数的零穿越性质, 提出一个跳点检测方法, 然后利用检测出的跳点给出参数向量和连接函数的半参数跳点检测估计量, 并讨论程序参数的选择. 在较弱的假设条件下, 本文建立跳点检测程序和所提估计量的大样本性质. 数值模拟和实例分析验证了所提方法在有限样本下的表现.

关键词 带跳单指标模型 零穿越性质 半参数跳点检测估计量 年薪数据

MSC (2010) 主题分类 62G08, 62H12

1 引言

近年来, 单指标模型受到统计研究者的青睐, 被广泛应用于经济学、统计学和医学等领域 (参见文献 [1, 2]), 其表达式为

$$Y = g(\theta_0^T X) + \varepsilon, \quad E(\varepsilon | X) = 0, \quad (1.1)$$

其中, $g(\cdot)$ 是未知连接函数, 有有界支撑 $U = [a, b]$; θ_0 是 p 维参数向量; X 是 p 维解释变量; ε 是独立同分布的随机误差项. 为了模型的可识别性, 假定 $\|\theta_0\| = 1$ ($\|\cdot\|$ 为 Euclid 范数), 且 θ_0 的第一个元素是正的. 若 $g(\cdot)$ 光滑且事先给定, 则模型 (1.1) 退化为广义线性模型. 因此, 单指标模型的结构更加灵活. 与纯非参数模型相比, 单指标模型通过一维变量 $\theta^T X$ 刻画 Y 的信息, 这使得它有更好的解释能力, 并避免了非参数模型的“维数灾难”问题. 文献中常见的方差函数 $E(\varepsilon^2 | X)$ 的假设有两种: 一是方差函数是纯非参的, 二是方差函数有像均值一样的降维结构. 广义线性模型即是最常见的有降维结

英文引用格式: Zhao Y Y, Lin J G, Du X L. Semiparametric jump-detection-based estimation for single-index models with jumps (in Chinese). Sci Sin Math, 2019, 49: 1021–1040, doi: 10.1360/N012018-00019

构的此类模型. 在半参数一般框架下, 文献 [3] 指出, 当中心方差子空间和中心均值子空间有相同维数时, 第二种假设成立. 关于此假设下模型 (1.1) 的推断可参见文献 [4–6]. 本文假设方差函数有像均值一样的降维结构, 则给定 $\theta^T X$ 下, Y 和 X 条件独立:

$$Y \perp\!\!\!\perp X \mid \theta^T X. \quad (1.2)$$

单指标模型均值的估计已被广泛研究. 例如, 文献 [1, 7] 提出了平均导数方法估计其指标系数和连接函数, 并证明了相应估计量的大样本性质. 文献 [2, 8] 提出了半参数最小二乘估计方法. 该方法采用参数模型的最小二乘估计思想, 在指标系数给定下, 用去除一点的 N-W (Nadaraya-Watson) 估计方法估计未知连接函数 $g(\cdot)$, 所得估计是 β 的函数, 再用非线性最小二乘方法最小化残差平方和, 得到 β 的估计. 文献 [9] 提出了最小平均方差估计方法同时估计连接函数和指标系数. 该方法在没有欠光滑连接函数估计量的情形下使参数的估计达到更快的收敛速度. 文献 [10] 提出了精细最小平均条件方差估计 (refined minimum average conditional variance estimation, rMAVE) 估计单指标模型, 在更弱的条件下研究了它们的渐近性质. 文献 [11] 介绍了估计方程方法研究单指标模型. 文献 [12] 利用多项式样条方法研究单指标模型, 证明了相应估计量的渐近性质. 文献 [13] 建议用非凹惩罚最小二乘方法估计指标系数和连接函数. 与其他变量选择和估计方法相比, 该方法能同时估计参数并选择变量, 证明了当参数维数固定时, 参数估计量具有 Oracle 性质. 文献 [14] 利用局部线性光滑和最小二乘方法研究了固定效应面板数据单指标模型的指标系数和连接函数的估计, 并证明了估计量的渐近正态性. 文献 [15] 在面板数据下提出了估计方程方法来研究未知连接函数和指标系数的估计, 并证明了其渐近分布. 最近, 对单指标模型的指标系数存在变点的研究也取得了一些成果. 文献 [16] 在原假设下提出了一个检验统计量, 以检验指标系数是否存在变点, 并给出了其渐近分布. 文献 [17] 研究了指标系数变点的估计问题, 证明了变点的收敛速率和渐近分布, 并给出了其渐近有效置信区间.

上述文献的结果是在假定未知连接函数 $g(\cdot)$ 光滑情形下得到的. 然而, 在许多领域的研究中, 经常遇到未知函数存在跳点的情形, 如工程学、经济学、气象学、生物医学和流行病学. 众所周知, 在估计回归模型时, 忽略任何一个重要事项可能会导致严重的偏差, 进而显著降低估计效率. 因此, 当单指标模型的连接函数存在跳时, 需要发展一种合适的估计方法能适用这种情形. 但是, 据我们所知, 目前还未有此类研究文献出现. 本文结合局部二次光滑技术和函数二阶导数的零穿越性质提出一个跳检测方法检测连接函数中的跳, 基于检测出的跳, 提出一个跳检测估计量, 提高了模型的估计效率. 理论分析显示, 新方法估计稳定, 并被证明在连续区间和跳邻域内都具有相合性. Monte Carlo 模拟和实例分析说明了本文提出的估计方法不仅在系数函数的连续区间而且在跳点的邻域内都有很好的表现.

本文余下内容结构如下: 第 2 节介绍估计方法; 第 3 节给出估计的渐近性质; 第 4 节报告提出的方法在有限样下的数值模拟; 第 5 节用本文提出的方法处理南京师范大学的教师年薪数据; 第 6 节给出所有定理的证明; 第 7 节是本文的简要总结和一些讨论.

2 估计方法

假设连接函数 $g(\cdot)$ 有如下的表达式:

$$g(v) = f(v) + \sum_{k=1}^q d_k I(v > s_k), \quad (2.1)$$

其中, $f(\cdot)$ 是一元光滑函数; $I(u)$ 是示性函数当 $u = \text{“真”}$ 时取值 1, 否则取值 0; q 表示 $g(\cdot)$ 中跳点的个数, d_k 和 s_k ($k = 1, \dots, q$) 分别表示跳点的幅度和位置. 在 (2.1) 中, $f(\cdot)$ 和 (q, d_k, s_k) 未知. 称满足 (2.1) 的模型 (1.1) 为带跳单指标模型.

2.1 精细最小平均条件方差估计

考虑简单回归模型 (1.1), 参数 θ 是

$$\min_{\theta} [E\{Y - E(Y | \theta^T X)\}^2] \quad (2.2)$$

的解. 对任意 θ , 在给定 $\theta^T X$ 下, 模型 (1.1) 的条件方差为

$$\sigma_{\theta}^2(\theta^T X) = E\{[Y - E(Y | \theta^T X)]^2 | \theta^T X\}. \quad (2.3)$$

进而,

$$E\{Y - E(Y | \theta^T X)\}^2 = E\{\sigma_{\theta}^2(\theta^T X)\}.$$

所以, 最小化 (2.2) 等价于最小化

$$E\{\sigma_{\theta}^2(\theta^T X)\}, \quad \text{满足 } \theta^T \theta = 1. \quad (2.4)$$

假设 $\{(X_i, Y_i), i = 1, \dots, n\}$ 是来自总体 (X, Y) 的一组样本. 令 $g(v) = E(Y | \theta^T X = v)$, 则对任意给定的 X_j , $E(Y_i | \theta^T X_i)$ 在 X_j 点处的局部二次展式为

$$E(Y_i | \theta^T X_i) \approx a + b\theta^T X_{ij} + \frac{c}{2}(\theta^T X_{ij})^2, \quad (2.5)$$

其中, $a = g(\theta^T X_j)$, $b = g'(\theta^T X_j)$, $c = g''(\theta^T X_j)$, $X_{ij} = X_i - X_j$. 因此, 根据局部线性光滑估计思想, 我们可用近似式

$$\sum_{i=1}^n \{Y_i - E(Y_i | \theta^T X_i)\}^2 w_{ij} \approx \sum_{i=1}^n \left[Y_i - \left\{ a + b\theta^T X_{ij} + \frac{c}{2}(\theta^T X_{ij})^2 \right\} \right]^2 w_{ij} \quad (2.6)$$

估计 $\sigma_{\theta}^2(\theta^T X_j)$, 其中, $w_{ij} = K_{h_n}(\theta^T X_{ij}) / \sum_{l=1}^n K_{h_n}(\theta^T X_{lj})$, $K_{h_n}(\cdot) = h_n^{-1} K(\cdot/h_n)$, h_n 是需要选择的带宽, $K(\cdot)$ 是一个对称密度函数. 那么, $\sigma_{\theta}^2(\cdot)$ 在 $\theta^T X_j$ 处估计量刚好是 (2.6) 的最小值, 即

$$\hat{\sigma}_{\theta}^2(\theta^T X_j) = \min_{a, b, c} \sum_{i=1}^n \left[Y_i - \left\{ a + b\theta^T X_{ij} + \frac{c}{2}(\theta^T X_{ij})^2 \right\} \right]^2 w_{ij}. \quad (2.7)$$

因此, θ 的估计可通过最小化下式获得:

$$\min_{\theta: \theta^T \theta = 1} \left\{ \sum_{j=1}^n \hat{\sigma}_{\theta}^2(\theta^T X_j) \right\} = \min_{\substack{\theta: \theta^T \theta = 1 \\ a_j, b_j, c_j, j=1, \dots, n}} \left(\sum_{j=1}^n \sum_{i=1}^n \left[Y_i - \left(a_j + b_j \theta^T X_{ij} + \frac{c_j}{2} (\theta^T X_{ij})^2 \right) \right]^2 w_{ij} \right). \quad (2.8)$$

令 $G(\cdot) = (g(\cdot), g'(\cdot), g''(\cdot))^T$. 为了估计 θ_0 和 $G(\cdot)$, 给出如下的计算步骤:

第 0 步 计算 θ_0 的初值 $\tilde{\theta}_{(1)}$. 设置 $t = 1$.

第 1 步 令 $\theta = \tilde{\theta}_{(t)}$, 最小化 (2.8) 计算 $(a_j, b_j, c_j; j = 1, \dots, n)$, 得

$$\begin{pmatrix} \tilde{a}_j^{(t)} \\ \tilde{b}_j^{(t)} \\ \tilde{c}_j^{(t)} \end{pmatrix} = \begin{pmatrix} S_{n,0}(v_{X_j}) & S_{n,1}(v_{X_j}) & S_{n,2}(v_{X_j}) \\ S_{n,1}(v_{X_j}) & S_{n,2}(v_{X_j}) & S_{n,3}(v_{X_j}) \\ S_{n,2}(v_{X_j}) & S_{n,3}(v_{X_j}) & S_{n,4}(v_{X_j}) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(v_{X_j}) \\ T_{n,1}(v_{X_j}) \\ T_{n,2}(v_{X_j}) \end{pmatrix} =: S_n^{-1}(v_{X_j}) T_n(v_{X_j}), \quad (2.9)$$

其中,

$$\begin{aligned} S_{n,l}(v_{X_j}) &= \frac{1}{n} \sum_{i=1}^n K_{h_n}(\tilde{\theta}_{(t)}^T X_{ij}) \left(\frac{\tilde{\theta}_{(t)}^T X_{ij}}{h_n} \right)^l, \\ T_{n,l}(v_{X_j}) &= \frac{1}{n} \sum_{i=1}^n K_{h_n}(\tilde{\theta}_{(t)}^T X_{ij}) \left(\frac{\tilde{\theta}_{(t)}^T X_{ij}}{h_n} \right)^l Y_i, \\ v_{X_j} &= \tilde{\theta}_{(t)}^T X_j, \quad l = 0, 1, 2, 3, 4. \end{aligned}$$

定义 $\tilde{G}_{(t)}(\cdot) = (\tilde{g}_{(t)}(\cdot), \tilde{g}'_{(t)}(\cdot), \tilde{g}''_{(t)}(\cdot))^T = (\tilde{a}^{(t)}(\cdot), \tilde{b}^{(t)}(\cdot), \tilde{c}^{(t)}(\cdot))^T$.

第 2 步 计算

$$\tilde{\theta}_{(t+1)} = \left\{ \sum_{j=1}^n \sum_{i=1}^n \tilde{w}_{ij}^{(t)} \tilde{\rho}_j^{(t)} (\tilde{b}_j^{(t)})^2 X_{ij} X_{ij}^T \right\}^{-1} \times \sum_{j=1}^n \sum_{i=1}^n \tilde{w}_{ij}^{(t)} \tilde{\rho}_j^{(t)} \tilde{b}_j^{(t)} X_{ij} (Y_i - \tilde{a}_j^{(t)}),$$

其中, $\tilde{w}_{ij}^{(t)} = K_{h_n}(\tilde{\theta}_{(t)}^T X_{ij}) / \sum_{l=1}^n K_{h_n}(\tilde{\theta}_{(t)}^T X_{lj})$, $\tilde{\rho}_j^{(t)} = \rho_n(n^{-1} \sum_{j=1}^n K_{h_n}(\tilde{\theta}_{(t)}^T X_{ij}))$, $\rho_n(\cdot)$ 是调整函数.

第 3 步 计算 $\tilde{\theta}_{(t+1)} := \text{sgn}(\tilde{\theta}_{(t+1),1}) \tilde{\theta}_{(t+1)} / \|\tilde{\theta}_{(t+1)}\|_2$, 其中 $\tilde{\theta}_{(t+1),1}$ 表示 $\tilde{\theta}_{(t+1)}$ 的第一个分量. 设置 $t := t + 1$, 并返回至第 1 步.

第 4 步 重复第 1 步至第 3 步, 直到收敛. 称 θ 的最终估计量为精细最小平均条件方差估计量, 记作 $\tilde{\theta}_n$.

第 5 步 把 $\tilde{\theta}_n$ 代入第 1 步, 得到 $g(\cdot)$ 、 $g'(\cdot)$ 和 $g''(\cdot)$ 的最终估计量, 记作 $\tilde{G}_n(\cdot) = (\tilde{g}_n(\cdot), \tilde{g}'_n(\cdot), \tilde{g}''_n(\cdot))^T$.

注 1 第 0 步中 $\tilde{\theta}_{(1)}$ 的计算可参见文献 [1, 7, 18] 的方法. 一个合适的调整函数 $\rho_n(\cdot)$ 能够降低边界点处的估计偏差, 本文根据文献 [10] 的建议, 选择如下的调整函数:

$$\rho_n(v) = \begin{cases} 1, & v \geq 2c_0 n^{-\epsilon}, \\ \frac{\exp\{(2c_0 n^{-\epsilon} - v)^{-1}\}}{\exp\{(2c_0 n^{-\epsilon} - v)^{-1}\} + \exp\{(v - c_0 n^{-\epsilon})^{-1}\}}, & 2c_0 n^{-\epsilon} > v > c_0 n^{-\epsilon}, \\ 0, & v \leq c_0 n^{-\epsilon}, \end{cases}$$

其中, 调整参数 $0 < \epsilon < \frac{1}{20}$, $0 < c_0 < \frac{1}{20}$.

2.2 跳点检测

注意到上述精细最小平均条件方差估计方法中非参数函数 $g(\cdot)$ 的估计采用局部二次技术, 它在实施过程中会光滑掉 $g(\cdot)$ 中的跳点. 因此, 第 1 步中 $g(\cdot)$ 的估计存在不可忽略的偏差, 进而影响参数 θ 的估计, 致使 θ 的估计不准确. 为了提高 $g(\cdot)$ 和 θ 的估计精度, 需要估计 $g(\cdot)$ 中跳点.

由 (2.7) 知, 在给定点 v_0 的邻域内, 我们利用二次函数拟合 $g(\cdot)$. 直观来说, 若 v_0 是 $g(\cdot)$ 的跳点, 则 v_0 两侧的观测值显著不同, 由 (2.7) 拟合的二次曲线是陡峭的. 因此, $|\tilde{g}'_n(v_0)|$ 的值比较大. 进一步, $\tilde{g}''_n(v_0)$ 接近于 0, 且在 v_0 两侧, $\tilde{g}''_n(\cdot)$ 异号. 这被称作是带跳函数在跳点附近二阶导数的零穿越性质 (参见文献 [19]). 结合 $\tilde{g}'_n(\cdot)$ 和 $\tilde{g}''_n(\cdot)$ 在跳点处的信息, 我们提出下面的跳点检测方法:

第 1 步 对任一点 $\tilde{v}_i = \tilde{\theta}_n^T X_i$, 若满足 (i) $|\tilde{g}'_n(\tilde{v}_i)| \geq \phi_n$; (ii) $|\tilde{g}''_n(\tilde{v}_i)| \leq \psi_n$; (iii) 存在 $(\tilde{v}_i - h_n, \tilde{v}_i + h_n)$ 内两点 $\tilde{v}_{i_1}$ 和 $\tilde{v}_{i_2}$, 使得 $|\tilde{g}''_n(\tilde{v}_{i_1})| > \psi_n$, $|\tilde{g}''_n(\tilde{v}_{i_2})| > \psi_n$ 和 $\tilde{g}''_n(\tilde{v}_{i_1})\tilde{g}''_n(\tilde{v}_{i_2}) < 0$, 其中, ϕ_n 和 ψ_n 是非负阈值, 则 X_i 被标记为跳点.

第 2 步 假设 $\{\tilde{v}_i^*, i = 1, \dots, m\}$ 是第 1 步检测的跳点. 若存在整数 $1 \leq l_1 < l_2 \leq m$, 使得 $\tilde{v}_q^* - \tilde{v}_{q-1}^* \leq h_n$ ($q = l_1 + 1, \dots, l_2$), $\tilde{v}_{l_1}^* - \tilde{v}_{l_1-1}^* > h_n$ 且 $\tilde{v}_{l_2+1}^* - \tilde{v}_{l_2}^* > h_n$, 则称 $(\tilde{v}_{l_1}^* + \tilde{v}_{l_2}^*)/2$ 是一个估计的跳点, 代替跳点 $\{\tilde{v}_{l_1}^*, \dots, \tilde{v}_{l_2}^*\}$.

利用上述程序可检测出 $g(\cdot)$ 中跳点的位置和个数, 记作 $\{\hat{s}_1, \dots, \hat{s}_q\}$ 和 $\hat{q}$.

注 2 第 1 步中条件 (ii) 和 (iii) 是 $g''(\cdot)$ 的零穿越性质. 二阶导数的估计值随机, 阈值用来控制 $g(\cdot)$ 的连续点两侧异号的问题. 本文仅研究边界邻域内无跳点和相邻跳点的距离至少为 h_n 的情形. 第 2 步控制相邻跳点的距离至少为 h_n .

2.3 半参数跳点检测估计

基于第 2.2 小节检测到的跳点位置的估计量 $\{\hat{s}_1, \dots, \hat{s}_q\}$, 则跳点幅度 $\{d_k, k = 1, \dots, q\}$ 可以被估计为

$$\hat{d}_k = \check{g}_{n+}(\hat{s}_k) - \check{g}_{n-}(\hat{s}_k), \quad (2.10)$$

其中, $\check{g}_{n+}(\cdot)$ 和 $\check{g}_{n-}(\cdot)$ 分别表示 $g(\cdot)$ 在右邻域和左邻域内得到的局部二次估计量 (估计方法参见文献 [20]). 这两个估计量用到的带宽 b_n 与跳检测方法中的带宽 h_n 不同.

将估计的跳点位置、个数和幅度代入模型 (1.1), 得到伪响应变量 $\{Y_i^*, i = 1, \dots, n\}$,

$$Y_i^* = Y_i - \sum_{k=1}^{\hat{q}} \hat{d}_k I(\tilde{v}_i > \hat{s}_k), \quad i = 1, \dots, n.$$

此时, 模型 (1.1) 变为

$$Y_i^* = f(\theta^T X_i) + \varepsilon_i, \quad i = 1, \dots, n. \quad (2.11)$$

然后, 利用文献 [10] 的估计方法 (用带宽 l_n 代替 h_n), 得到模型 (2.11) 中 θ 和 $g(\cdot)$ 的精细最小平均条件方差估计量, 分别记作 $\hat{\theta}_n$ 和 $\hat{f}_n(\cdot)$. 因此, 基于估计量 $\hat{\theta}_n$ 、 $\hat{f}_n(\cdot)$ 、 $\hat{s}_k$ 、 $\hat{q}$ 和 $\hat{d}_k$, 令 $\hat{v} = \hat{\theta}_n^T X$, 则可定义 $g(\cdot)$ 的新估计量

$$\hat{g}_n(\hat{v}) = \hat{f}_n(\hat{v}) + \sum_{k=1}^{\hat{q}} \hat{d}_k I(\hat{v} > \hat{s}_k). \quad (2.12)$$

称 $\hat{\theta}_n$ 和 $\hat{g}_n(\cdot)$ 为半参数跳点检测估计量 (semiparametric jump-detection-based estimators, SJDE).

注 3 (2.12) 中半参数跳点检测估计量 $\hat{g}_n(\cdot)$ 由连续连接函数 $f(\cdot)$ 的精细最小平均条件方差估计量 $\hat{f}_n(\cdot)$ 和跳点估计量 (位置 $\hat{s}_k$ 、个数 $\hat{q}$ 和幅度 $\hat{d}_k$) 构成. 由第 3 节的定理 3-6 和文献 [10, 定理 2] 的讨论知, 半参数跳点检测估计量 $\hat{g}_n(\cdot)$ 在连续区间和跳点领域都是相合的.

2.4 参数选择

带宽在平衡偏差和方差方面扮演着重要角色. 跳点检测和参数估计中, 有两组参数 (h_n, ϕ_n, ψ_n) 和 (b_n, l_n) 需要选择. 根据上面的表述可知, 需要先选第一组参数, 再选第二组参数.

首先讨论跳点检测中的参数 h_n 、 ϕ_n 和 ψ_n 的选择. 参数 ϕ_n 是 $|\tilde{g}_n(\cdot)|$ 的阈值. 当选用的核函数是对称核且满足条件 $1/(nh_n) = o(1)$ 时, 由定理 2 知, 在 $g(\cdot)$ 的连续点 $v_x = \theta_n^T x$ 处, $\tilde{g}'_n(v_x)$ 的渐近均值为 $g'(v_x)$, 渐近方差为

$$\sigma_{\tilde{g}'_n}^2(v_x) = \frac{\nu_2 \sigma^2(v_x)}{nh_n \mu_2^2 p_\theta(v_x)}, \quad (2.13)$$

其中, $\mu_2 = \int u^2 K(u) du$, $\nu_2 = \int u^2 K^2(u) du$. 因此,

$$\lim_{n \rightarrow \infty} \frac{\{\tilde{g}'_n(v_x) - g'(v_x)\}^2}{\hat{\sigma}_{\tilde{g}'_n}^2(v_x)} \stackrel{\mathcal{D}}{=} \chi_1^2,$$

其中, “ $\stackrel{\mathcal{D}}{=}$ ” 表示等价分布, $\hat{\sigma}_{\tilde{g}'_n}^2(v_x)$ 是 $\sigma_{\tilde{g}'_n}^2(v_x)$ 的相合估计量. 因此, ϕ_n 的一个自然选择为

$$\phi_n = \hat{\sigma}_{\tilde{g}'_n}(v_x) \sqrt{\chi_{1,1-\alpha_n}^2(\hat{\eta}_n)}, \quad (2.14)$$

其中, $\chi_{1,1-\alpha_n}^2(\hat{\eta}_n)$ 表示 χ_1^2 分布的 $1 - \alpha_n$ 分位数, 非中心化参数 $\hat{\eta}_n = \{\tilde{g}'_n(v_x)\}^2 / \hat{\sigma}_{\tilde{g}'_n}^2(v_x)$, $\tilde{g}'_n(v_x)$ 是 $g'(v_x)$ 的相合估计量. 本文中, 用最小平均条件方差估计和文献 [20] 介绍的单侧函数估计思想计算 $\hat{\sigma}^2(\cdot)$ 和 $\tilde{g}'_n(\cdot)$. 其他的计算方法也可推广使用, 参见文献 [21, 22]. 另外, 为了利用 (2.14) 选择 ϕ_n , 需要选择 α_n . 数值模拟发现, 当 $\alpha_n = 0.05$ 时, 仿真结果比较合理. 因此, 本文建议使用 $\alpha_n = 0.05$.

参数 h_n 和 ψ_n 的选择需要合适的跳点检测准则. 令 $J = \{s_1, \dots, s_q\}$ 和 $\hat{J}_n$ 分别是真实的和估计的跳点集合, 那么, Hausdorff 距离是常用的评价准则,

$$d_H(\hat{J}_n, J; h_n, \psi_n) = \max \left\{ \max_{s \in \hat{J}_n} \min_{s' \in J} |s - s'|, \max_{s \in J} \min_{s' \in \hat{J}_n} |s - s'| \right\}. \quad (2.15)$$

由于 J 未知, $d_H(\hat{J}_n, J; h_n, \psi_n)$ 不能直接计算, 故建议使用 bootstrap 方法. 假设存在 B 个 bootstrap 样本, 根据第 k 个样本检测到的跳点为 $\hat{J}_n^{(k)}$, 则 $d_H(\hat{J}_n, J; h_n, \psi_n)$ 的估计为

$$\hat{d}_H(\hat{J}_n, J; h_n, \psi_n) = \frac{1}{B} \sum_{i=1}^B d_H(\hat{J}_n^{(k)}, \hat{J}_n; h_n, \psi_n), \quad (2.16)$$

h_n 和 ψ_n 的最优值可通过求解

$$\min_{h_n > 0} \min_{\psi_n > 0} \hat{d}_H(\hat{J}_n, J; h_n, \psi_n) \quad (2.17)$$

逼近.

为了产生 bootstrap 样本, 本文使用下面的数据生成方法:

第 1 步 计算函数 $g(\cdot)$ 的估计量 $\tilde{g}(\cdot)$, 定义残差 $\hat{\varepsilon}_i = Y_i - \tilde{g}(\hat{\theta}_n^T X_i)$.

第 2 步 从 $\{\hat{\varepsilon}_i, i = 1, \dots, n\}$ 中抽取样本 $\{\varepsilon_i^*, i = 1, \dots, n\}$, 再定义 bootstrap 样本 $\{(X_i, Y_i^*), i = 1, \dots, n\}$, 其中, $Y_i^* = \tilde{g}(\hat{\theta}_n^T X_i) + \varepsilon_i^*$.

在检测出跳点后, 根据 (2.12) 计算非参数函数估计量 $\hat{g}(\cdot)$. 在 (2.10) 和 (2.11) 中的使用带宽的 b_n 和 l_n 可通过最小化 “去一” 交叉验证 (cross-validation, CV) 量

$$CV(b_n, l_n) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i^* - \sum_{j=1}^p \hat{g}^{-i}(\hat{\theta}_n^T X_i) \right\}^2$$

获得, 其中, $\hat{g}^{-i}(\cdot)$ 表示 $f(\cdot)$ 的估计量, 它的计算用到除第 i 次观测外的所有重抽样样本.

注 4 数值模拟发现, 上述获取 bootstrap 样本的机制比从原始数据 $\{(X_i, Y_i), i = 1, \dots, n\}$ 直接抽样有更好的跳点检测表现.

3 渐近性质

本节给出新方法的渐近性质. 为了证明定理, 需要假设如下正则条件:

(C1) 对所有 $\theta \in \{\theta: \|\theta - \theta_0\| < \delta\}$, $\theta^T X$ 的密度函数 $p_\theta(v)$ 及其直到三阶导数在 $\mathbb{R}$ 上有界, 其中, $\delta > 0$, $E\|X\|^6 < \infty$, $E\|Y\|^3 < \infty$.

(C2) 连接函数 $g(\cdot)$ 有界, 在 s_k ($k = 1, \dots, q$) 处有有限个跳点, 且在跳点处右连续. 对所有 $\theta \in \{\theta: \|\theta - \theta_0\| < \delta\}$ ($\delta > 0$), 条件均值 $g_\theta(v) = E(Y | \theta^T X = v)$ 及其直到三阶导数在连续区间内有界; 在跳点邻域内有左、右三阶导数.

(C3) $K(\cdot)$ 是对称密度函数, 有有界二阶导数和紧支撑.

(C4) 带宽 $h_n \propto n^{-1/7}$, $b_n \propto n^{-1/7}$, $l_n \propto n^{-1/5}$, 调整参数 $\epsilon < 1/20$.

注 5 条件 (C1) 说明, 对于 θ_0 邻域中 θ , 如果 $\theta^T X$ 连续, 则 X 可有离散的分量 (参见文献 [2]). 关于 X 矩条件的要求不算强. 若将证明复杂化并使用更小的带宽, 条件 (C2) 中连接函数光滑性的要求可放宽至二阶导数存在. 条件 (C3) 包含 Gauss 核函数和二次核函数. 条件 (C4) 包含了估计连接函数和指标参数的最优带宽 (参见文献 [8]).

我们感兴趣的是估计量的渐近有效性. 首先将支撑 $\mathbb{U}$ 划为两部分: (i) 连续区域 D_1 ; (ii) 跳点邻域 D_2 , 即

$$D_1 = [a + h_n, b - h_n] \setminus D_2, \quad D_2 = \bigcup_{k=1}^q (s_k - h_n, s_k + h_n).$$

令 $v_x = \theta^T x$, $H = \text{diag}\{1, h_n, h_n^2\}$,

$$S_n(v_x) = p_\theta(v_x) \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}, \quad \Sigma_n(v_x) = \sigma^2(v_x) p_\theta(v_x) \begin{pmatrix} \nu_0 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_3 & \nu_4 \end{pmatrix},$$

其中, $\mu_l = \int u^l K(u) du$, $\nu_l = \int u^l K^2(u) du$, $l = 0, 1, 2, 3, 4$. 记 $\mu_\theta(x) = E(X | \theta^T X = \theta^T x)$, $\nu_\theta(x) = \mu_\theta(x) - x$, $w_\theta(x) = E(X X^T | \theta^T X = \theta^T x)$, $W_0(x) = \nu_{\theta_0}(x) \nu_{\theta_0}^T(x)$, $W(x) = w_{\theta_0}(x) - \mu_{\theta_0}(x) \mu_{\theta_0}^T(x)$.

定理 1 假设正则条件 (C1)–(C4) 成立, 对 $\forall v_x \in D_1$, 当 $n \rightarrow \infty$ 时,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

其中,

$$\Sigma = [E(g'(\theta_0^T x)^2 W(x))]^+ E\{g'(\theta_0^T x)^2 W_0(x) \sigma^2(\theta_0^T x)\} [E(g'(\theta_0^T x)^2 W(x))]^+,$$

A^+ 表示矩阵 A 的 Moore-Penrose 逆.

定理 2 假设正则条件 (C1)–(C4) 成立, θ 是已知向量 θ_0 或被估计至阶 $O_p(n^{-1/2})$, 对 $\forall v_x \in D_1$, 我们有

$$\sqrt{nh_n} \left[H^{-1}(\tilde{G}_n(v_x) - G(v_x)) - \frac{h_n^2}{6} \begin{pmatrix} 0 \\ g'''(v_x) \frac{\mu_4}{\mu_2} \\ 0 \end{pmatrix} \right] \xrightarrow{\mathcal{D}} N(0_{3 \times 1}, S_n^{-1}(v_x) \Sigma_n(v_x) S_n^{-1}(v_x)),$$

其中, $0_{3 \times 1}$ 表示元素全为 0、长度为 3 的列向量, $\xrightarrow{\mathcal{D}}$ 表示依分布收敛. 特别地,

$$\sqrt{nh_n} \{\tilde{g}_n(v_x) - g(v_x)\} \xrightarrow{\mathcal{D}} N\left(0, \frac{\sigma^2(v_x)}{p_\theta(v_x)} V\right),$$

其中, $V = (\mu_4^2\nu_0 - 2\mu_2\mu_4\nu_2 + \mu_2^2\nu_4)/(\mu_4 - \nu_2^2)^2$.

定理 3 假设 s 是 $g(\cdot)$ 的跳点, v_x 是 $(s - h_n, s + h_n)$ 内一点, 与 s 的 Euclid 距离为 τh_n ($0 \leq \tau < 1$). 定义 $F_n^{(1)}(v_x)$ 和 $F_n^{(2)}(v_x)$ 是由 s 所划分 $[v_x - h_n, v_x + h_n]$ 的左、右两个区间, $F^{(m)} = \{u : u \in [-1, 1], v_x + uh_n \in F_n^{(m)}(v_x)\}$ ($m = 1, 2$) 是 $K(\cdot)$ 对应的支撑. 记 $g(\cdot)$ 在 s 处跳点的幅度为 d , 若定理 2 的条件成立且 $\log(1/h_n)/(nh_n^3) = o(1)$, 则有

(i)

$$\tilde{g}'(v_x) = \frac{d}{h_n\mu_2} \int_{F^{(2)}} uK(u) du + o_p\left(\frac{1}{h_n}\right),$$

(ii)

$$\tilde{g}''(v_x) = \begin{cases} O_p\left(\frac{1}{h_n}\right), & \tau = 0, \\ \frac{C}{h_n^2} + O_p\left(\frac{1}{h_n}\right), & 0 < \tau < 1, \quad v_x \in F_n^{(1)}(v_x), \\ -\frac{C}{h_n^2} + O_p\left(\frac{1}{h_n}\right), & 0 < \tau < 1, \quad v_x \in F_n^{(2)}(v_x), \end{cases}$$

其中,

$$C = \frac{2d}{\mu_4 - \mu_2^2} \left[\mu_2 \int_0^\tau K(u) du - \int_0^\tau u^2 K(u) du \right].$$

定理 4 假设定理 1-3 的条件成立, 若 $\alpha_n = o(1)$, $U_{1-\alpha_n/2}/\sqrt{n h_n} = o(1)$, $1/(\psi_n h_n) = o(1)$, $\psi_n h_n^2 = o(1)$, 则有

$$d_H(\hat{J}_n, J; h_n, \psi_n) = O_p(h_n), \quad \text{a.s.},$$

其中, $U_{1-\alpha_n/2}$ 表示标准正态分布的 $1 - \alpha_n/2$ 分位数.

定理 5 假设定理 1-4 的条件成立, 若满足条件 $h_n/b_n = o(1)$, 则有

$$|\hat{d}_k - d_k| = O_p\left(\frac{h_n}{b_n}\right) + O_p\left\{\left(\frac{\log(1/h_n)}{nb_n}\right)^{1/2}\right\} + o_p(b_n^2), \quad \text{a.s.}$$

定理 6 假设正则条件 (C1)-(C4) 成立, 当 $n \rightarrow \infty$ 时,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^*),$$

其中,

$$\Sigma^* = [E\{f'(\theta_0^T x)^2 W(x)\}]^+ E[f'(\theta_0^T x)^2 W_0(x) \sigma^2(\theta_0^T x)] [E\{f'(\theta_0^T x)^2 W(x)\}]^+.$$

定理 7 假设定理 1-6 的条件成立, 若 θ 是已知向量 θ_0 或被估计至阶 $O_p(n^{-1/2})$, 对 $\forall v_x \in D_1 \cup D_2$, 有

$$\begin{aligned} \int \{\hat{g}_n(v_x) - g(v_x)\}^2 dv_x &= O_p(h_n + l_n) + O_p\left(\frac{\log(1/b_n)}{nb_n} + \frac{\log(1/l_n)}{nl_n}\right) \\ &\quad + O_p\left(\frac{1}{n^2 l_n^2}\right) + O_p\left(\frac{h_n^2}{b_n^2}\right) + o_p(b_n^4), \quad \text{a.s.} \end{aligned}$$

4 数值模拟

本节通过数值例子评价提出的跳点检测方法和半参数跳点检测估计量的有限样本性质. 由于跳点个数 J 未知, 不能直接用介绍的方法选择程序参数. 这里先用 bootstrap 方法估计跳点个数, 再用提出的方法估计跳点 (记作 JDP_{BST} (jump detection procedure 简称为 JDP), bootstrap 样本 $B = 200$). 为方便比较, 将真实参数下提出方法的表现作为参考 (记作 JDP_{TRU}).

考虑下面的带跳回归函数:

$$g(v) = \begin{cases} -3v + 2, & v \in [0, 0.3), \\ -3v + 3 - \sin\left(\frac{\pi(v - 0.3)}{0.2}\right), & v \in [0.3, 0.7), \\ \frac{v}{2} + 0.75, & v \in [0.7, 1]. \end{cases}$$

函数 $g(v)$ 有两个跳点, 在 0.3 和 0.7 处, 幅度分别是 1.0 和 0.2. 令 $\theta_0 = (4/9, -4/9, 7/9)^T$, $X = (X_1, X_2, X_3)^T$, 其中, X_1 、 X_2 和 X_3 分别是来自 $[0, 3/4]$ 、 $[0, 3/4]$ 和 $[0, 3/7]$ 的均匀分布. 随机误差 ε 服从均值为 0、方差为 σ^2 的正态分布. 选取样本量 $n = 100$ 或 200, 标准差 $\sigma = 0.25$ 或 0.50. 仿真模拟重复 200 次, 使用 Epanechnikov 核函数, $\epsilon = 0.04$, $c_0 = 0.01$, $\alpha_n = 0.05$.

首先, 评价 bootstrap 方法的表现. 计算 JDP_{TRU} 和 JDP_{BST} 两种方法检测的跳点个数, 以及计算 $\hat{J}$ 和 J 之间的平均 Hausdorff 距离和程序参数的平均值. 仿真结果见表 1. 由表 1 知, 使用 JDP_{TRU} 和 JDP_{BST} 方法检测的跳点个数在大多数情形下等于真实的跳点个数, JDP_{BST} 的表现比 JDP_{TRU} 稍差, 但差异不明显. 随着样本量 n 的增大, σ 或者 ρ_x 的减小, 仿真结果更精确. 这说明提出的 bootstrap 参数选择程序是有效的, 也验证了定理 4 的理论结果. 另外, $\hat{J}$ 和 J 之间 Hausdorff 距离的估计结果亦是合理的.

表 1 跳点个数频数表、平均 Hausdorff 距离和程序参数平均值

n	σ	估计方法	频数表	$d_H(h_n, \psi_n)$
100	0.25	JDP_{TRU}	2 3	0.026
			189 11	(0.14, 16.04)
		JDP_{BST}	2 3	0.028
	0.50	JDP_{TRU}	185 15	(0.14, 24.67)
			1 2 3	0.057
		JDP_{BST}	9 163 28	(0.14, 30.59)
200	0.25	JDP_{TRU}	1 2 3 4	0.060
			17 158 22 3	(0.14, 41.28)
		JDP_{BST}	2 3	0.020
	0.50	JDP_{TRU}	195 5	(0.12, 15.81)
			2 3	0.024
		JDP_{BST}	193 7	(0.12, 17.36)
	0.25	JDP_{TRU}	1 2 3	0.043
			14 181 5	(0.12, 26.70)
		JDP_{BST}	1 2 3	0.051
	0.50	JDP_{TRU}	17 177 6	(0.11, 34.89)
		JDP_{BST}		

其次, 研究提出的跳点检测方法检测跳点的能力. 表 2 列出了不同情形下在真实跳点 0.02 范围内检测出跳点的频数. 由表 2 知, 样本量 n 、噪声水平 σ 和跳的幅度 d_k 都会影响 JDP_{BST} 的表现. 样本量 n 增大, σ 减小, 或跳的幅度增大, JDP_{BST} 的表现更好 (0.3 处跳的幅度大于 0.7 处跳的幅度). 这些表现在第 3 节大样本性质中没有很好地体现出来.

最后, 研究半参数跳点检测估计量 $\hat{\theta}_n$ 和 $\hat{g}_n(\cdot)$ 的有限样本表现. 在获得跳点个数、位置和幅度的估计之后, 使用提出的跳点检测方法和精细最小平均条件方差方法估计参数向量和连接函数. 对于参数 θ , 表 3 给出了其在 rMAVE 和 JDP_{BST} 两种方法下的模拟结果. 由表 3 知, JDP_{BST} 方法有更小的平均绝对偏差 (average absolute deviation, AAD). 对于连接函数 $g(\cdot)$, 由 bootstrap 方法计算的估计量记作 $\hat{g}_n^{\text{BST}}(\cdot)$. 为了说明其有效性, 考虑其他两种估计方法. 一是基于局部分段线性估计和单侧核函数提出的保跳 (jump-preserving, JP) 估计方法 (参见文献 [20]). 二是自适应保跳回归 (adaptive jump-preserving, AJP) 方法 [23]. 我们修改这两种方法使其适应本文模型, 两个曲线估计量分别记作 $\hat{g}_n^{\text{JP}}(\cdot)$ 和 $\hat{g}_n^{\text{AJP}}(\cdot)$. 对于这三个估计量, 重复 200 次模拟计算平均积分平方误差 (mean integral squared error, MISE) 和局部 MISE (真实跳点 0.05 范围内的 MISE), 即

$$\text{MISE} = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^{n-1} \frac{\text{SE}_j(u_i) + \text{SE}_j(u_{i+1})}{2n} \right),$$

$$\text{MISE}_{s_k} = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^{n_0-1} \frac{\text{SE}_j(u_i) + \text{SE}_j(u_{i+1})}{2n_0} I_{[s_k - 0.05 < u_i < u_{i+1} < s_k + 0.05]} \right),$$

其中, $\text{SE}_j(u_i) = (g(u_i) - \hat{g}_n^c(u_i))^2$, c 取 JP、AJP 或 BST, $u_i = \hat{\theta}_n^T X_i$, N 是重复模拟次数, n_0 是跳点 0.05 范围内设计点个数. 表 3 给出了三个非参数估计量的仿真结果. 由表 4 知, $\hat{g}_n^{\text{BST}}(\cdot)$ 方法的表现一致优于 $\hat{g}_n^{\text{JP}}(\cdot)$ 和 $\hat{g}_n^{\text{AJP}}(\cdot)$. 以上两个模拟说明了提出的估计方法可以有效地估计参数向量和连接函数.

表 2 200 次重复模拟下真实跳点 0.02 范围内检测出跳点的频数

n	σ	0.3 $\pm$ 0.02	0.7 $\pm$ 0.02
100	0.25	189	143
	0.50	153	98
200	0.25	200	166
	0.50	179	127

表 3 参数向量的平均绝对偏差 $p^{-1} \sum_{j=1}^p |\hat{\theta}_j - \theta_j|$ 的随机模拟结果

n	σ	估计方法	AAD
100	0.25	rMAVE	0.0135
		JDP_{BST}	0.0123
	0.50	rMAVE	0.0267
		JDP_{BST}	0.0249
200	0.25	rMAVE	0.0090
		JDP_{BST}	0.0074
	0.50	rMAVE	0.0146
		JDP_{BST}	0.0134

表 4 连接函数的 MISE、局部 MISE 和标准差 (括号内) 的随机模拟结果

n	σ	估计量	MISE	局部 MISE	
				0.3 ± 0.05	0.7 ± 0.05
100	0.25	$\hat{g}_n^{\text{BST}}$	0.0108 (0.0008)	0.0121 (0.0009)	0.0127 (0.0010)
		$\hat{g}_n^{\text{JP}}$	0.0403 (0.0019)	0.0934 (0.0032)	0.0824 (0.0028)
		$\hat{g}_n^{\text{AJP}}$	0.0322 (0.0018)	0.0829 (0.0030)	0.0746 (0.0032)
	0.50	$\hat{g}_n^{\text{BST}}$	0.0254 (0.0017)	0.0212 (0.0019)	0.0238 (0.0021)
		$\hat{g}_n^{\text{JP}}$	0.0819 (0.0027)	0.1867 (0.0076)	0.2019 (0.0087)
		$\hat{g}_n^{\text{AJP}}$	0.0726 (0.0025)	0.2029 (0.0076)	0.2286 (0.0095)
	0.25	$\hat{g}_n^{\text{BST}}$	0.0052 (0.0004)	0.0061 (0.0005)	0.0068 (0.0004)
		$\hat{g}_n^{\text{JP}}$	0.0258 (0.0011)	0.0782 (0.0031)	0.0834 (0.0033)
		$\hat{g}_n^{\text{AJP}}$	0.0224 (0.0011)	0.0826 (0.0030)	0.0892 (0.0036)
200	0.50	$\hat{g}_n^{\text{BST}}$	0.0157 (0.0010)	0.0116 (0.0015)	0.0125 (0.0017)
		$\hat{g}_n^{\text{JP}}$	0.0631 (0.0027)	0.1178 (0.0059)	0.1259 (0.0068)
		$\hat{g}_n^{\text{AJP}}$	0.0586 (0.0025)	0.1280 (0.0062)	0.1353 (0.0071)

5 实证分析

本节利用提出的方法来处理南京师范大学的教师年薪数据. 该数据由该校人力资源办公室提供, 记录了 2016 年 11 月份 1459 名教师的收入情况 (删除异常值, 如院士、杰青、长江学者的数据及缺失值). 我们将年薪作为响应变量 Y , 工作年限、职称级别、最后职称等级评定时间至今的年限作为响应变量, 分别记作 X_1 、 X_2 、 X_3 . 由于高校教师收入与职称密切相关, 这里考虑讲师、副教授和教授三个级别, 每个级别分为 3 个等级, 因此, 职称由一个有序变量来描述, 取值从 1 到 9, 表示从讲师的最低级别到教授的最高级别. 我们的兴趣是研究三个协变量对教师年薪的影响. 图 1 给出 (X_{ij}, Y_i) ($j = 1, 2, 3$) 的散点 - 核光滑曲线图. 由核光滑曲线知, 三个协变量与响应变量均有较强的线性关系, 因此利用如下模型:

$$Y_i = g(\theta_1 X_{i1} + \theta_2 X_{i2} + \theta_3 X_{i3}) + \varepsilon_i, \quad i = 1, \dots, 1459 \quad (5.1)$$

对数据建模, 其中, X_j ($j = 1, 2, 3$) 是标准化的解释变量. 为方便起见, 使用相同符号标记解释变量.

我们使用精细最小平均条件方差估计和本文提出的半参数跳点检测估计两种方法拟合模型 (5.1). 表 5 列出两种方法下指标系数的估计结果, 括号内的标准差根据 100 次 bootstrap 抽样计算获得. 由表 5 可知, 指标系数均为正值, X_2 的系数远大于其他的系数, SJDE 方法的标准差更小. 这说明工作年限、职称级别、最后职称等级评定时间至今的年限对教师年薪有正向作用, 并且职称级别对年薪有很大影响, 这也与现实情况相符; 标准差小说明 SJDE 方法对参数的估计更加稳健. 图 2 给出了两种方法下连接函数的估计曲线. 由图 2 可知, SJDE 方法更好地保持连接函数中 6 个可能的跳点 (箭头标识), 这说明不同职称级别的教师年薪差异明显, 与现实情况一致; 而 rMAVE 方法将这些跳点光滑掉了, 不能很好地反应实际情况. 以上分析说明本文提出的模型能合理地解释该年薪数据, 提出的 SJDE 方法更加稳健.

图 1 工作年限、职称级别、最后职称等级评定时间至今的年限对年薪的散点 - 核光滑曲线图

表 5 指标系数的估计结果

估计方法	θ_1	θ_2	θ_3
rMAVE	0.1059 (0.0072)	0.9934 (0.0329)	0.0443 (0.0146)
SJDE	0.0982 (0.0056)	0.9945 (0.0295)	0.0364 (0.0121)

图 2 三个指标变量组合后的散点图以及拟合后的回归图: (a) rMAVE 方法; (b) SJDE 方法

6 定理的证明

为了证明定理的结果, 先给出一些符号. 令 $v_x = \theta^T x$,

$$N(\theta^T X_i) = g(\theta_0^T X_i) - g(\theta^T X_i),$$

$$M(\theta^T X_i) = g(\theta^T X_i) - g(v_x) - g'(v_x)\theta^T X_i - \frac{1}{2}g''(v_x)(\theta^T X_i)^2 - \frac{1}{6}g'''(v_x)(\theta^T X_i)^3 - N(\theta^T X_i),$$

$$T_{n,l}^*(v_x) = \frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ij}}{h_n} \right)^l \varepsilon_i,$$

$$P_{n,l}(v_x) = \frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ix}}{h_n} \right)^l N(\theta^T X_i),$$

$$R_{n,l}(v_x) = \frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ix}}{h_n} \right)^l M(\theta^T X_i),$$

$$B_{n,l}(v_x) = \frac{1}{6} S_{n,l+3}(v_x) h_n^3 g'''(v_x),$$

则有

$$\begin{aligned} T_{n,l}(v_x) &= S_{n,l}(v_x)g(v_x) + S_{n,l+1}(v_x)h_n g'(v_x) + \frac{1}{2}S_{n,l+2}(v_x)h_n^2 g''(v_x) + B_{n,l}(v_x) \\ &\quad + P_{n,l}(v_x) + R_{n,l}(v_x) + T_{n,l}^*(v_x). \end{aligned}$$

结合第 2.1 小节算法第 1 步, 有

$$\hat{G}(v_x) - G(v_x) - S_n^{-1}(v_x)[B_n(v_x) + P_n(v_x) + R_n(v_x)] = S_n^{-1}(v_x)T_n^*(v_x), \quad (6.1)$$

其中,

$$\begin{aligned} B_n(v_x) &= \begin{pmatrix} B_{n,0}(v_x) \\ B_{n,1}(v_x) \\ B_{n,2}(v_x) \end{pmatrix}, \quad P_n(v_x) = \begin{pmatrix} P_{n,0}(v_x) \\ P_{n,1}(v_x) \\ P_{n,2}(v_x) \end{pmatrix}, \\ R_n(v_x) &= \begin{pmatrix} R_{n,0}(v_x) \\ R_{n,1}(v_x) \\ R_{n,2}(v_x) \end{pmatrix}, \quad T_n^*(v_x) = \begin{pmatrix} T_{n,0}^*(v_x) \\ T_{n,1}^*(v_x) \\ T_{n,2}^*(v_x) \end{pmatrix}. \end{aligned}$$

引理 1 假设对任意 $\theta \in \Theta$, $E(Z \mid \theta^T X = \theta^T x) = m(\theta^T x)$ 和其直至三阶导数有界, 并且对某一个 $r > 3$, $E\|Z\|_2^r < \infty$. 令 (X_i, Z_i) ($i = 1, \dots, n$) 是来自总体 (X, Z) 的样本, $X_{ix} = X_i - x$, 如果条件 (C1)、(C3) 和 (C4) 成立, 则

$$\frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ix}}{h_n} \right)^l Z_i = \mu_l p_\theta(\theta^T x) m(\theta^T x) + \mu_{l+1} h_n \{p_\theta(\theta^T x) m(\theta^T x)\}' + O(\tau_n),$$

其中, $\mu_l = \int K(u) u^l du$, $l = 0, 1, \dots$, $\tau_n = h_n^2 + (nh_n / \log n)^{-1/2}$.

引理 1 的证明参见文献 [10].

令 $S_{n,l}(v_x) = \frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) (\theta^T X_{ix} / h_n)^l$, 由引理 1 可推出

$$S_{n,l}(v_x) = \mu_l p_\theta(v_x) + o_p(1), \quad l = 0, 1, 2, 3, 4. \quad (6.2)$$

引理 2 假设条件 (C1)–(C4) 成立, 则有

$$nh_n \text{var}\{T_n^*(v_x)\} = p_\theta(v_x) \sigma^2(v_x) \begin{pmatrix} \nu_0 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_3 & \nu_4 \end{pmatrix} + o_p(1).$$

证明 由于

$$\begin{aligned} \text{var}\{T_{n,0}^*(v_x)\} &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \varepsilon_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n K_{h_n}^2(\theta^T X_{ix}) \sigma^2(\theta^T X_i) \\ &= \frac{1}{n} \mathbb{E}\{K_{h_n}^2(\theta^T(X-x)) \sigma^2(\theta^T X)\} + o(1) \\ &= \frac{1}{nh_n^2} \int K^2\left(\frac{\theta^T(X-x)}{h_n}\right) \sigma^2(\theta^T X) p_\theta(\theta^T X) d(\theta^T X) + o(1) \\ &= \frac{1}{nh_n^2} \int K^2\left(\frac{v-v_x}{h_n}\right) \sigma^2(v) p_\theta(v) dv + o(1) \\ &= \frac{1}{nh_n} \int K^2(w) \sigma^2(v_x + wh_n) p_\theta(v_x + wh_n) dw + o(1) \\ &= \frac{1}{nh_n} \sigma^2(v_x) p_\theta(v_x) \nu_0 + o\left(\frac{1}{nh_n}\right), \end{aligned}$$

其中, $v = \theta^T X$, $w = (v - v_x)/h_n$, 故有

$$nh_n \text{var}\{T_{n,0}^*(v_x)\} = p_\theta(v_x) \sigma^2(v_x) \nu_0 + o(1).$$

类似地, 可以证明

$$\begin{aligned} nh_n \text{var}\{T_{n,1}^*(v_x)\} &= p_\theta(v_x) \sigma^2(v_x) \nu_2 + o(1), \\ nh_n \text{var}\{T_{n,2}^*(v_x)\} &= p_\theta(v_x) \sigma^2(v_x) \nu_4 + o(1), \\ nh_n \text{cov}\{T_{n,0}^*(v_x), T_{n,1}^*(v_x)\} &= p_\theta(v_x) \sigma^2(v_x) \nu_1 + o(1), \\ nh_n \text{cov}\{T_{n,0}^*(v_x), T_{n,2}^*(v_x)\} &= p_\theta(v_x) \sigma^2(v_x) \nu_3 + o(1), \\ nh_n \text{cov}\{T_{n,1}^*(v_x), T_{n,2}^*(v_x)\} &= p_\theta(v_x) \sigma^2(v_x) \nu_4 + o(1). \end{aligned}$$

即可证明引理 2. □

定理 1 的证明 此定理的证明与文献 [10, 定理 2] 的证明类似, 不再赘述. □

定理 2 的证明 根据 (6.2) 和定理 1 的结果, 经过简单计算, 有

$$B_n(v_x) = \frac{h_n^3}{6} \begin{pmatrix} 0 \\ \mu_4 p_\theta(v_x) g'''(v_x) \\ 0 \end{pmatrix} + o_p(h_n^3), \quad (6.3)$$

$$P_n(v_x) = 1_{3 \times 1} O_p(n^{-1/2}), \quad (6.4)$$

$$R_n(v_x) = 1_{3 \times 1} o_p(n^{-1/2}), \quad (6.5)$$

其中, $1_{3 \times 1}$ 表示元素全为 1 长度为 3 的列向量. 因此, 由 (6.1) 和 (6.3)–(6.5) 可知,

$$H^{-1}\{\tilde{G}(v_x) - G(v_x)\} - \frac{h_n^2}{6} \begin{pmatrix} 0 \\ \mu_2^{-1} \mu_4 p_\theta(v_x) g'''(v_x) \\ 0 \end{pmatrix} \approx H^{-1} S_n^{-1}(v_x) T_n^*(v_x). \quad (6.6)$$

要建立 $\tilde{G}(v_x)$ 在的渐近正态性, 需要证明

$$\sqrt{nh_n} T_n^*(v_x) \xrightarrow{\mathcal{D}} N(0, \Sigma_n(v_x)). \quad (6.7)$$

显然,

$$E(T_n^*(v_x)) = 0_{3 \times 1}.$$

结合引理 2 的结果以及 Slutsky 定理和中心极限定理即证 (6.7).

最后, 根据 (6.1)、(6.3)–(6.7) 和引理 2 的结果以及 Slutsky 定理和中心极限定理, 即可证得定理 2. $\square$

定理 3 的证明 根据引理 1 和 (6.1), 对任意的 v_x , 有

$$S_n^{-1}(v_x) \approx p_\theta^{-1}(v_x) \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}^{-1} = p_\theta^{-1}(v_x) \begin{pmatrix} \frac{\mu_4}{\mu_4 - \mu_2^2} & 0 & \frac{-\mu_2}{\mu_4 - \mu_2^2} \\ 0 & \frac{1}{\mu_2} & 0 \\ \frac{-\mu_2}{\mu_4 - \mu_2^2} & 0 & \frac{1}{\mu_4 - \mu_2^2} \end{pmatrix}. \quad (6.8)$$

对 $l = 0, 1, 2$, 利用积分的 Riemann 和逼近方法, 有

$$\begin{aligned} E(T_{n,l}(v_x)) &= E\left\{\frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ix}}{h_n}\right)^l Y_i\right\} \\ &= E\left\{\frac{1}{n} \sum_{i=1}^n K_{h_n}(\theta^T X_{ix}) \left(\frac{\theta^T X_{ix}}{h_n}\right)^l \varepsilon_i\right\} \\ &= E\left\{K_{h_n}(\theta^T(X-x)) \left(\frac{\theta^T(X-x)}{h_n}\right)^l g(\theta^T X)\right\} + o(1) \\ &= \frac{1}{h_n} \int K(\theta^T(X-x)) \left(\frac{\theta^T(X-x)}{h_n}\right)^l g(\theta^T X) p_\theta(\theta^T X) d(\theta^T X) + o(1) \\ &= \int K(w) w^l g(v_x + wh_n) p_\theta(v_x + wh_n) dw + o(1) \\ &= p_\theta(v_x) \int K(w) w^l g(v_x + wh_n) dw + o(1), \end{aligned} \quad (6.9)$$

其中, $w = \theta^T(X-x)/h_n$. 因此, 根据 (6.6)、(6.8) 和 (6.9) 可得

$$E \begin{pmatrix} \tilde{g}_n(v_x) \\ \tilde{g}'_n(v_x) \\ \tilde{g}''_n(v_x) \end{pmatrix} = \begin{pmatrix} \int \frac{\mu_4 - \mu_2 w^2}{\mu_4 - \mu_2^2} K(w) g(v_x + wh_n) dw + O_p\left(\frac{1}{nh_n}\right) \\ \frac{1}{h_n \mu_2} \int w K(w) g(v_x + wh_n) dw + O_p\left(\frac{1}{nh_n^2}\right) \\ \frac{2}{h_n^2 (\mu_4 - \mu_2^2)} \int (w^2 - \mu_2) K(w) g(v_x + wh_n) dw + O_p\left(\frac{1}{nh_n^3}\right) \end{pmatrix}. \quad (6.10)$$

为证明结论 (i), 根据 (6.10) 的第二项和条件 $\int wK(w)dw = 0$, 有 $\int_{F^{(1)}} uK(u)du = -\int_{F^{(2)}} uK(u)du$. 进而,

$$\begin{aligned} E(\tilde{g}'_n(v_x)) &= \frac{1}{h_n\mu_2} \left\{ \int_{F^{(1)}} + \int_{F^{(2)}} \right\} g(v_x + wh_n)wK(w)dw + O_p\left(\frac{1}{nh_n^2}\right) \\ &= \frac{1}{h_n\mu_2} \left\{ g_-(v_x) \int_{F^{(1)}} wK(w)dw + g_+(v_x) \int_{F^{(2)}} wK(w)dw \right\} + o_p\left(\frac{1}{h_n}\right) \\ &= \frac{d}{h_n\mu_2} \int_{F^{(2)}} wK(w)dw + o_p\left(\frac{1}{h_n}\right). \end{aligned}$$

再根据条件 $1/(nh_1^4) = o(1)$, 有

$$\tilde{g}'_n(v_x) - E(\tilde{g}'_n(v_x)) = O_p\left\{\left(\frac{\log(1/h_n)}{nh_n^3}\right)^{1/2}\right\}. \quad (6.11)$$

因此, 当满足条件 $\log(1/h_n)/(nh_n^3) = o(1)$ 时,

$$\tilde{g}'_n(v_x) = \frac{d}{h_n\mu_2} \int_{F^{(2)}} wK(w)dw + o_p\left(\frac{1}{h_n}\right). \quad (6.12)$$

下面证明结论 (ii). 由于

$$\int_{F^{(1)}} K(w)dw = 1 - \int_{F^{(2)}} K(w)dw, \quad \int_{F^{(1)}} w^2K(w)dw = \mu_2 - \int_{F^{(2)}} w^2K(w)dw,$$

则根据 (6.10) 的第二项, 有

$$\begin{aligned} E(\tilde{g}''_n(v_x)) &= \frac{2}{h_n^2(\mu_4 - \mu_2^2)} \left\{ \int_{F^{(1)}} + \int_{F^{(2)}} \right\} (-\mu_2 + w^2)g(v_x + wh_n)K(u)dw + O_p\left(\frac{1}{nh_n^3}\right) \\ &= \frac{2d}{h_n^2(\mu_4 - \mu_2^2)} \left[\int_{F^{(2)}} w^2K(w)dw - \mu_2 \int_{F^{(2)}} K(w)dw \right] + O_p\left(\frac{1}{h_n}\right). \end{aligned}$$

根据 Hölder 不等式, 有 $\mu_4 - \mu_2^2 = \int w^4K(w)dw - (\int w^2K(w)dw)^2 > 0$. 令 $\xi(\tau) = \frac{\int_0^\tau w^2K(w)dw}{2 \int_0^\tau K(w)dw}$, 则 ξ 是 $(0,1)$ 上关于 τ 的严格增加函数. 因此, 对任意的 $\tau \in (0,1)$, 有 $\xi(\tau) < \xi(1)$, 这说明

$$\int_0^\tau w^2K(w)dw < 2 \int_0^\tau K(w)dw \int_0^1 w^2K(w)dw = \mu_2 \int_0^\tau K(w)dw.$$

接下来, 对任意的 $\tau \in (0,1)$, 有

$$C = \frac{2d}{(\mu_4 - \mu_2^2)} \left[\mu_2 \int_0^\tau K(w)dw - \int_0^\tau w^2K(w)dw \right] > 0. \quad (6.13)$$

根据 (6.10) 第三项和条件 $1/(nh_n^4) = o(1)$, 可得

$$\tilde{g}''_n(v_x) - E(\tilde{g}''_n(v_x)) = O_p\left\{\left(\frac{\log(1/h_n)}{nh_n^5}\right)^{1/2}\right\}. \quad (6.14)$$

因此, 根据 (6.9)、(6.13) 和 (6.14) 即证定理 3(ii) 的结论. 证毕. $\square$

定理 4 的证明 根据定理 3 的证明可知,

$$\tilde{g}'_n(v_x) - E(\tilde{g}'_n(v_x)) = O_p \left\{ \left(\frac{\log(1/h_n)}{nh_n^3} \right)^{1/2} \right\}.$$

由 (2.14) 知,

$$\hat{\sigma}_{\tilde{g}'_n}^2(v_x) = \frac{\nu_2 \hat{\sigma}^2(v_x)}{nh_n^3 \mu_2^2 \hat{p}_\theta(v_x)}, \quad (6.15)$$

其中, $\hat{\sigma}^2(\cdot)$ 和 $\hat{p}_\theta(\cdot)$ 分别是 $\sigma^2(\cdot)$ 和 $p_\theta(\cdot)$ 的相合估计量. 进而, 有

$$\psi_n = \hat{\sigma}_{\tilde{g}'_n}(v_x) \sqrt{\chi_{1,1-\alpha_n}^2} = \hat{\sigma}_{\tilde{g}'_n}(v_x) U_{1-\alpha_n/2}. \quad (6.16)$$

在条件 $U_{1-\alpha_n/2}/\sqrt{nh_n} = o(1)$ 下, 有 $g'(v_x) < \psi_n$, a.s. 因此, 当 n 足够大时,

$$\hat{J}_n \subseteq U_n(J), \quad \text{a.s.}, \quad (6.17)$$

其中, $U_n(J) = \bigcup_{k=1}^q [s_k - h_n, s_k + h_n]$ 表示 $g(\cdot)$ 所有跳点的邻域集合.

若 s_k 是跳点, 根据定理 3(i), 有 $|\tilde{g}'_n(s_k)| = \frac{d_k}{h_n \mu_2} \int_{F^{(2)}} u K(u) du + o(1/h_n)$. 因此, 当 n 足够大时, 根据 (6.15)、(6.16) 和条件 $U_{1-\alpha_n/2}/\sqrt{nh_n} = o(1)$, 有 $|\tilde{g}'_n(s_k)| > \phi_n = o(1/h_n)$, a.s. 再由定理 3(ii)、条件 $1/(\psi_n h_n) = o(1)$ 和 $\psi_n^2 = o(1)$, 可得 $|\tilde{g}''_n(s_k)| \leq \psi_n$, $|\tilde{g}''_n(v_{l_1})| > \psi_n$, $|\tilde{g}''_n(v_{l_2})| > \psi_n$, $|\tilde{g}''_n(v_{l_1})\tilde{g}''_n(v_{l_2})| < \psi_n$, 其中, v_{l_1} 和 v_{l_2} 是 s_k 邻域内两点, 满足 v_{l_1} 在 s_k 的左侧, v_{l_2} 在 s_k 的右侧, 并且 v_{l_1} 和 v_{l_2} 距离 s_k 为 κh_n ($\kappa > 0$). 因此, 当 n 足够大时, s_k 会被检测为跳点, 即

$$J \subseteq \hat{J}_n, \quad \text{a.s.} \quad (6.18)$$

结合 (6.17) 和 (6.18) 可得定理 4 的结果. $\square$

定理 5 的证明 由定理 4 可知, 当 n 足够大时, $\hat{s}_k \in [s_k - h_n, s_k + h_n]$, a.s. 不失一般性, 我们仅假设 $k = 1$ 的情形, 并且假设 $\hat{s}_1 = s_1 - \delta h_n$ 和 $\delta \in [0, 1]$.

因为 $\hat{d}_1 = \check{g}_{n+}(\hat{s}_1) - \check{g}_{n-}(\hat{s}_1)$, $d_1 = g_+(s_1) - g_-(s_1)$, 所以,

$$\begin{aligned} |\hat{d}_1 - d_1| &= |[\check{g}_{n+}(\hat{s}_1) - \check{g}_{n-}(\hat{s}_1)] - [g_+(s_1) - g_-(s_1)]| \\ &\leq |\check{g}_{n+}(\hat{s}_1) - g_+(s_1)| + |\check{g}_{n-}(\hat{s}_1) - g_-(s_1)|. \end{aligned} \quad (6.19)$$

在此情形下, 有

$$\begin{aligned} \check{g}_{n+}(\hat{s}_1) &= (1 \ 0 \ 0) \begin{pmatrix} S_{n,0}(\hat{s}_1) & S_{n,1}(\hat{s}_1) & S_{n,2}(\hat{s}_1) \\ S_{n,1}(\hat{s}_1) & S_{n,2}(\hat{s}_1) & S_{n,3}(\hat{s}_1) \\ S_{n,2}(\hat{s}_1) & S_{n,3}(\hat{s}_1) & S_{n,4}(\hat{s}_1) \end{pmatrix} \begin{pmatrix} T_{n,0}(\hat{s}_1) \\ T_{n,1}(\hat{s}_1) \\ T_{n,2}(\hat{s}_1) \end{pmatrix} \\ &= g_+(s_1) + O_p \left(\frac{h_n}{b_n} \right) + O_p \left\{ \left(\frac{\log(1/h_n)}{nh_n} \right)^{1/2} \right\} + o_p(b_n^2). \end{aligned} \quad (6.20)$$

另一方面, 易得

$$\begin{aligned} |\check{g}_{n-}(\hat{s}_1) - g_-(s_1)| &\leq |\check{g}_{n-}(\hat{s}_1) - g_-(\hat{s}_1)| + |g_-(\hat{s}_1) - g_-(s_1)| \\ &= O_p(h_n) + O_p \left\{ \left(\frac{\log(1/h_n)}{nh_n} \right)^{1/2} \right\} + o_p(b_n^2). \end{aligned} \quad (6.21)$$

因此, 根据 (6.19)–(6.21) 即得证明定理 5 的结论. 证毕. $\square$

定理 6 的证明 此定理的证明与文献 [10, 定理 2] 的证明类似, 不再赘述. $\square$

定理 7 的证明 由定理 4 和 5 知, 当 n 足够大时, 有 $\hat{q} = q$, a.s., 则

$$\begin{aligned}
 \hat{f}_n(v_x) &= \hat{g}_n(v_x) - \sum_{k=1}^q \hat{d}_k I(v_x > \hat{s}_k) \\
 &= (1 \ 0) \begin{pmatrix} S_{n,0}(v_x) & S_{n,1}(v_x) \\ S_{n,1}(v_x) & S_{n,2}(v_x) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(v_x) \\ T_{n,1}(v_x) \end{pmatrix} - \sum_{k=1}^q \hat{d}_k I(v_x > \hat{s}_k) \\
 &= \sum_{i=1}^n \frac{K_{l_n}(\theta^T(X_i - x))f(v_x)}{\sum_{l=1}^n K_{l_n}(\theta^T(X_l - x))} + \sum_{i=1}^n \frac{K_{l_n}(\theta^T(X_i - x))\varepsilon_i}{\sum_{l=1}^n K_{l_n}(\theta^T(X_l - x))} \\
 &\quad + \sum_{i=1}^n \frac{K_{l_n}(\theta^T(X_i - x))\{\sum_{k=1}^q d_k I(v_x > s_k) - \sum_{k=1}^q \hat{d}_k I(v_x > \hat{s}_k)\}}{\sum_{l=1}^n K_{l_n}(\theta^T(X_l - x))} \\
 &=: A_1(v_x) + A_2(v_x) + A_3(v_x).
 \end{aligned} \tag{6.22}$$

对于 $A_1(v_x)$, 经过简单的计算, 有 $|A_1(v_x) - f(v_x)| = O_p(l_n^2) + O_p(\frac{1}{nl_n})$. 从而,

$$\int \{A_1(v_x) - f(v_x)\}^2 dv_x = O_p(l_n^4) + O_p\left(\frac{1}{n^2 l_n^2}\right). \tag{6.23}$$

类似地, 可得

$$A_2(v_x) = O_p\left\{\left(\frac{\log(1/l_n)}{nl_n}\right)^{1/2}\right\},$$

进而,

$$\int \{A_2(v_x)\}^2 dv_x = O_p\left\{\frac{\log(1/l_n)}{nl_n}\right\}. \tag{6.24}$$

对于 $A_3(v_x)$, 由定理 5 知, 对 $\forall v_x \in D_2$ 和 $\pi = \max\{h_n, l_n\}$, 有

$$\begin{aligned}
 |A_3(v_x)| &\leq \left| \sum_{i=1}^n \frac{K_{l_n}(\theta^T(X_i - x))}{\sum_{l=1}^n K_{l_n}(\theta^T(X_l - x))} \right| \times \left| \sum_{k=1}^q [d_k I(v_x > s_k) - \hat{d}_k I(v_x > \hat{s}_k)] \right| \\
 &\leq \sum_{k=1}^q |d_k - \hat{d}_k| \\
 &= O_p\left(\frac{h_n}{b_n}\right) + O_p\left\{\left(\frac{\log(1/l_n)}{nl_n}\right)^{1/2}\right\} + o_p(b_n^2).
 \end{aligned} \tag{6.25}$$

另一方面, 当 $v_x \in \bigcup_{k=1}^q [s_k - \pi, s_k + \pi]$ 时, 有

$$A_3(v_x) \leq \sum_{k=1}^q (|d_k| + |\hat{d}_k| + |d_k - \hat{d}_k|).$$

从而,

$$A_3(v_x) = O_p(1), \quad \text{a.s.,} \quad v_x \in \bigcup_{k=1}^q [s_k - \pi, s_k + \pi]. \tag{6.26}$$

由 (6.24) 和 (6.25) 可得

$$\int \{A_3(v_x)\}^2 dv_x = \left[O_p\left(\frac{h_n}{b_n}\right) + O_p\left\{\left(\frac{\log(1/l_n)}{nl_n}\right)^{1/2}\right\} + o_p(b_n^2) \right]^2 + O_p(l_n). \quad (6.27)$$

因此, 根据 (6.23)、(6.24)、(6.27) 和不等式 $(a+b)^2 \leq 2(a^2+b^2)$, 即可证明定理 7. $\square$

7 结束语

本文基于精细最小平均条件方差估计和函数二阶导数的零穿越性质提出了一种新的半参数三步方法估计带跳单指标模型, 给出了参数向量和连接函数的半参数跳点检测估计量, 讨论了程序参数的实际选择, 并建立了各步估计量的渐近性质. 随机模拟的结果显示, 本文提出的跳点检测方法和半参数跳点检测估计量在一系列设计下均有着良好的有限样本表现. 最后通过一个实际例子验证了所提模型和估计方法的有效性和可行性. 本文只考虑了单指标模型中有跳点的情形, 类似的方法可推广至部分线性单指标模型和变指标模型. 结合参数向量可存在变点, 本文的方法也可以用于分析带跳变点单指标模型. 这将成为我们后续研究的问题.

致谢 感谢审稿人对本文提出的宝贵意见.

参考文献

- 1 Powell J L, Stock J H, Stoker T M. Semiparametric estimation of index coefficients. *Econometrica*, 1989, 57: 1403–1430
- 2 Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J Econometrics*, 1993, 58: 71–120
- 3 Cook R D, Li B. Dimension reduction for conditional mean in regression. *Ann Statist*, 2002, 30: 455–474
- 4 Ma Y Y, Zhu L P. A review on dimension reduction. *Int Statist Rev*, 2013, 81: 134–150
- 5 Sheng W H, Yin X R. Direction estimation in single-index models via distance covariance. *J Multivariate Anal*, 2013, 122: 148–161
- 6 Zhu L P, Wang T, Zhu L X, et al. Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 2010, 97: 295–304
- 7 Härdle W, Stoker T M. Investigating smooth multiple regression by method of average derivatives. *J Amer Statist Assoc*, 1989, 84: 986–995
- 8 Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Ann Statist*, 1993, 21: 157–178
- 9 Xia Y C, Tong H, Li W K, et al. An adaptive estimation of dimension reduction space. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 363–410
- 10 Xia Y C. Asymptotic distributions for two estimators of the single-index model. *Econom Theory*, 2006, 22: 1112–1137
- 11 Cui X, Härdle W K, Zhu L X. The EFM approach for single-index models. *Ann Statist*, 2011, 39: 1658–1688
- 12 Wang L, Yang L. Spline estimation of single-index models. *Statist Sinica*, 2009, 19: 765–783
- 13 Peng H, Huang T. Penalized least squares for single index models. *J Statist Plann Inference*, 2011, 141: 1362–1379
- 14 Lai P, Li G, Lian H. Semiparametric estimation of fixed effects panel data single-index model. *Statist Probab Lett*, 2013, 83: 1595–1602
- 15 Zhu L P, You J H, Xu Q F. Statistical inference for single-index panel data models. *Scand J Statist*, 2014, 41: 830–843
- 16 Cao G Q, Wang Z F, Wu Y H, et al. Inference of change-point in single index models. *Sci China Ser A*, 2008, 51: 1855–1870
- 17 Jin B S, Dong C L, Tan C C, et al. Estimator of a change point in single index models. *Sci China Math*, 2014, 57: 1701–1712
- 18 Horowitz J L, Härdle W. Direct semiparametric estimation of single-index models with discrete covariates. *J Amer Statist Assoc*, 1996, 91: 1632–1640
- 19 Joo J H, Qiu P. Jump detection in a regression curve and its derivation. *Technometrics*, 2009, 51: 289–305
- 20 Qiu P. A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *J Nonparametr Stat*, 2003, 15: 437–453

- 21 Chu C K, Glad I K, Godtliebsen F, et al. Edge-preserving smoothers for image processing. *J Amer Statist Assoc*, 1998, 93: 526–541
- 22 Gijbels I, Lambert A, Qiu P. Jump-preserving regression and smoothing using local linear fitting: A compromise. *Ann Inst Statist Math*, 2007, 59: 235–272
- 23 Zhao Y Y, Lin J G, Huang X F, et al. Adaptive jump-preserving estimates in varying-coefficient models. *J Multivariate Anal*, 2016, 149: 65–80

Semiparametric jump-detection-based estimation for single-index models with jumps

Yanyong Zhao, Jinguan Lin & Xiuli Du

Abstract Single-index models are common dimension reduction models in statistics. In practical applications, the link function may have singularities, including jumps at some unknown positions and structural changes of the related process. Detection of such singularities is important for estimating the coefficients and understanding the structural changes. This article proposes a jump detection method based on the refined minimum average conditional variance estimation and zero-crossing properties of second-order derivative of a function. Using the detected jumps, we give semiparametric jump-detection-based estimators for the parameter vector and the link function. Then the selection of procedure parameters is discussed. Under some mild conditions, we establish large sample properties of the jump detection procedure and the proposed estimators. Numerical studies and a real data analysis are conducted to assess the finite sample property of the proposed methods.

Keywords single-index models with jumps, zero-crossing properties, semiparametric jump-detection-based estimators, salary data

MSC(2010) 62G08, 62H12

doi: 10.1360/N012018-00019