

面板数据下序贯模型变点的渐近检验法

陈卓恒^{1,2}, 胡亦钧²

(1. 华侨大学数学科学学院, 福建 泉州 362021)

(2. 武汉大学数学与统计学院, 湖北 武汉 430072)

摘要: 本文研究了面板数据模型下变点是否存在的检验问题. 对于面板序贯模型下的变点, 利用构造的检验统计量及其极限分布的方法, 获得了关于变点的一种渐近检验法. 随后进行 Monte-Carlo 数值模拟, 在模拟中对检验方法的经验检验水平、检验功效以及变点后的停时三个判断标准进行了考察. 结果显示无论在理论还是数值模拟上, 我们提出的渐近检验法均表现优良. 推广了现有文献的关于单序列变点的序贯检验方法.

关键词: 变点; 面板数据; 序贯检测; 渐近法

MR(2010) 主题分类号: 62E20; 62F05; 62L12

中图分类号: O212.3

文献标识码: A

文章编号: 0255-7797(2018)04-0731-12

1 引言

变点问题起源于工业质量控制的领域, 在质量控制中非常重要的一点是如何快速检测出生产线上不合格产品比例的增加. 在建模和数据挖掘中, 一个很常见的问题是, 如何根据现有的样本数据, 来判断我们感兴趣的一些量是否发生变化. 这种问题在统计中我们称其为变点问题. 关于变点问题的建模和分析始于 Page 1954 年发表在《Biometrika》上的文献 [1], 文献主要考虑了利用分位数构造的简单检测方法去检验生产线上观察到数据的稳定性. 在此之后, 关于变点问题的研究变得日益活跃, 各种研究文献也日益增多. 而且随着学科的发展和深入, 变点问题在经济、金融学、生物医学、气候学、导航系统、图像处理、信号探测、计算机等很多领域都有广泛的应用背景.

很久以来大多数的变点检测的文献都是基于这样的背景: 对于一组固定容量的历史数据集, 我们设计一些检验方法来判断这组数据内我们感兴趣的参数是否发生变化. 这种检验即为我们常说的“事后检验”. 在这种检验中, 样本是静态的, 这种数据我们称为是离线数据. 离线数据分析主要在譬如历史文本分析, 图像分析等领域有应用. 但是在更多的领域譬如质量控制, 医疗监测, 金融风险控制, 我们使用的都是在线数据, 也就是样本采取连续抽样的方式得到的. 采取连续抽样的方式来得到样本进行检验, 这种检验称为“序贯检验”. 序贯检验也是最近几年变点检测中的研究热点. 具体来说, 在序贯检验中, 一般假定静止期的长度为 m , 也就是在 m 个观察数据 $X_1, X_2, \dots, X_m$ 内不存在变点, 这 m 个观察数据也常称为历史数据. 在考虑渐近性质的时候一般令 $m \rightarrow \infty$. 在原假设成立即认为没有变点存在的情况下

*收稿日期: 2017-04-10

接收日期: 2017-07-10

基金项目: 国家自然科学基金青年项目基金资助 (11401227); 国家自然科学基金面上项目基金资助 (11371284); 华侨大学中青年教师科技创新计划资助 (ZQN-YX401); 华侨大学校级项目基金资助 (600005-Z17Y0078).

作者简介: 陈卓恒 (1980-), 女, 湖北孝感, 讲师, 主要研究方向: 变点检测.

$X_{m+1}, X_{m+2}, \dots$ 的参数是相同的; 若备则假设成立即认为变点存在的情况下, 则存在一个整数 $k^* \geq 0$, 使得 $X_{m+1}, X_{m+2}, \dots, X_{m+k^*}$ 的参数与历史数据相同, 而 $X_{m+k^*+1}, X_{m+k^*+2}, \dots$ 的参数不同, 这里 k^* 为未知变点. 接下来要做的就是如何构造合适的检验统计量并制定合理的停止规则, 从而判断变点是否存在以及在认为变点存在的情况下推断出变点的位置, 使得变点从产生到被检测出需要的时间尽量短, 并尽量消除误报, 即快速准确检测出变点.

文献 [2] 是较早提出并研究变点检测中序贯检验的代表性文献之一, 文中提出了两种检测方案, 一种利用波动差, 一种利用残差的累积和, 并拓宽了经典的不变原理得到所需要的收敛结果, 最后确定了临界值和停止规则, 找到最优停时, 并在模拟中得到了检验. 文献 [3] 进一步推广了文献 [2] 的结果, 同样对于线性回归模型, 文献 [3] 提出了两类检验统计量, 第一类统计量是建立在回归系数 β 取最小二乘估计时残差的加权累积和, 第二类是利用回归系数构造的迭代残差的部分和得到的统计量. 并制定了合适的检验规则, 使得在历史数据个数 $m \rightarrow \infty$ 时错误预报率在约定的水平内, 且检验功效趋于 1. 学者 Aue 在变点的序贯检验中也做了一系列工作, 譬如率先开展研究停止时刻 τ_n 的极限分布, 得到了文献 [4], 后续工作有文献 [5-7], 另外文献 [8] 得到了在对 ARMA 时间序列中结构变点进行序贯检验时, 停止时刻的极限分布. 在近几年中, 也有把自助抽样技术引入到序贯模型中的变点检测问题中, 可参考文献 [9, 10] 和 [11].

目前来说, 现有的文献主要研究单序列情形下的序贯模型, 那么, 如何在面板数据下研究序贯模型的变点, 这是一个崭新的问题. 在本文中, 我们主要对这个问题进行研究. 本文剩余部分的结构是这样的: 第 2 节将描述模型并得到相关的理论结果, 即检验统计量的大样本性质; 第 3 节基于前面的理论结果提出面板数据下关于变点的渐近检验法, 并进行了 Monte-Carlo 模拟计算; 第 4 节给出了前面引理和定理的证明过程; 第 5 节对前面的结论进行了总结.

2 模型及相关理论结果

考虑面板均值变点模型

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad 1 \leq i \leq N, 1 \leq j \leq m + T_m, \quad (2.1)$$

这里 μ_{ij} 和 ε_{ij} 分别表示均值和面板扰动项. $\{\mu_{ij}, j = 1, 2, \dots, m + T_m\}_i$ 和 $\{\varepsilon_{ij}, j = 1, 2, \dots, m + T_m\}_i$ 是相互独立的. T_m 与 m 有关而且假定 $T_m \leq \infty$, 序列 $\{\varepsilon_{ij}\}$ 关于 i 相互独立, $\varepsilon_{ij} \sim (0, \sigma^2)$ 且在概率空间 $(\Omega, \mathcal{F}, P)$ 上关于 j 是独立同分布的 (i.i.d.). 假定 $0 < \sigma^2 < \infty$ 而且对于某个 $v > 2$ 有 $E|\varepsilon_{ij}|^v < \infty$.

假定对于每个 $1 \leq i \leq N$, 已经观察到长度为 m 的历史数据, 它没有变点发生, 即对于 $1 \leq i \leq N$,

$$\mu_{ij} = \mu_{i0}, \quad 1 \leq j \leq m. \quad (2.2)$$

现在我们感兴趣的是接下来到来的数据里是否存在均值发生变化的共同变点, 即想检验原假设

$$\mathcal{H}_0 : \mu_{ij} = \mu_{i0}, \quad m + 1 \leq j \leq m + T_m \quad (2.3)$$

和备则假设

$$\begin{aligned} \mathcal{H}_1 : & \text{存在 } k_m^0 \geq 1 \text{ 使得 } \mu_{ij} = \mu_{i0}, \quad m + 1 \leq j \leq m + k_m^0, \\ & \text{以及 } \mu_{ij} = \mu_{i0} + \delta_{i0} \neq \mu_{i0}, \quad m + k_m^0 < j \leq m + T_m, \end{aligned} \quad (2.4)$$

这里 $\sigma^2, \mu_{i0}, \delta_{i0}, k_m^0$ 的值均为未知. 检测变点的方案是基于 CUSUM 型检测统计量 $\Gamma(m, k, N)$ 和一个边界值, 这里

$$\Gamma_i(m, k) = \frac{\sum_{m < j \leq m+k} [Y_{ij} - \frac{1}{m} \sum_{l=1}^m Y_{il}]}{g(m, k, N)}, \quad (2.5)$$

此处

$$g(m, k, N) = (Nm)^{\frac{1}{2}} \left(1 + \frac{k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}, \quad (2.6)$$

其中 $0 \leq \gamma < \frac{1}{2}$, 而且面板数据变点的检验统计量是

$$\Gamma(m, k, N) = \sum_{i=1}^N \Gamma_i(m, k). \quad (2.7)$$

此外, 基于历史数据集 $\{Y_{i1}, Y_{i2}, \dots, Y_{im}\}_{i=1,2,\dots,N}$ 使用以下的方差估计

$$\hat{\sigma}_{m,N}^2 = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{m-1} \sum_{j=1}^m \left(Y_{ij} - \frac{1}{m} \sum_{l=1}^m Y_{il} \right)^2 \right]. \quad (2.8)$$

关于检验法则, 在

$$\tau_N(m) = \begin{cases} \inf \{k \geq 1 : \frac{1}{\hat{\sigma}_{m,N}} |\Gamma(m, k, N)| \geq c\}, \\ \infty, & \text{对一切 } k = 1, 2, \dots, \frac{1}{\hat{\sigma}_{m,N}} |\Gamma(m, k, N)| \leq c \text{ 总成立} \end{cases} \quad (2.9)$$

处停止并拒绝 H_0 . $\tau_N(m)$ 也被称为停时, 而且这里的 c 应该被选择成使我们能够控制错误预报率. 即在原假设 H_0 下, 对于某个给定水平 $0 < \alpha < 1$,

$$\lim_{m \rightarrow \infty} P\{\tau_N(m) < \infty\} = \alpha, \quad (2.10)$$

在备则假设 H_1 下, 要求

$$\lim_{m \rightarrow \infty} P\{\tau_N(m) < \infty\} = 1. \quad (2.11)$$

引理 2.1 当 $m, N \rightarrow \infty$ 时, $\hat{\sigma}_{m,N}^2$ 是 σ^2 的无偏且相合的估计.

定理 2.1 对于面板数据模型 (2.1), 在原假设 (i.e. (2.3) 式) 的假定下, 当 $m, N \rightarrow \infty$ 时, 有

$$\lim_{N, m \rightarrow \infty} P \left\{ \frac{1}{\hat{\sigma}_{m,N}} \sup_{1 \leq k < \infty} |\Gamma(m, k, N)| \leq c \right\} = P \left\{ \sup_{0 \leq t \leq 1} \frac{|W(t)|}{t^\gamma} \leq c \right\}, \quad (2.12)$$

这里 $\{W(t), 0 \leq t < \infty\}$ 是一个 Wiener 过程, γ 见 (2.6) 式.

定理 2.2 对于面板数据模型 (2.1), 如果 $k_m^0 + m = O(m) < m + T_m$, 则无论 N 是固定的还是 $N \rightarrow \infty$, 在备则假设 (i.e. (2.4) 式) 下, 都有当 $m \rightarrow \infty$ 时,

$$\frac{1}{\hat{\sigma}_{m,N}} \sup_{1 \leq k < \infty} |\Gamma(m, k, N)| \xrightarrow{P} \infty. \quad (2.13)$$

3 渐近检验法及其模拟计算

根据定理 2.1 知, 在原假设成立即没有变点时, 若 $m, N \rightarrow \infty$ 时,

$$\frac{1}{\hat{\sigma}_{m,N}} \sup_{1 \leq k < \infty} |\Gamma(m, k, N)|$$

的极限分布为

$$\sup_{0 \leq t \leq 1} \frac{|W(t)|}{t^\gamma},$$

这里 $\{W(t), 0 \leq t < \infty\}$ 是一个 Wiener 过程. 因此可以通过此处已知的极限分布求出在给定的显著性水平下检验的临界值 c , 然后利用这里的临界值 c 以及检验法则来确定是拒绝还是接受原假设, 并在原假设被拒绝, 即认为变点存在的情形下估计出变点的位置. 这种方法我们称它为“渐近检验法”.

关于渐近检验法, 首先要计算检验的临界值. 由于临界值只与 γ 有关, 所以可记作 c_γ . 为了计算经验检验水平, 考虑一个不带变点的模型, 并且计算

$$\frac{1}{\hat{\sigma}_{m,N}} \sup_{1 \leq k \leq T_m} |\Gamma(m, k, N)| > c_\gamma$$

的次数, 这样就得到了这个次数和 Monte-Carlo 模拟次数之比. 进一步, 为了计算经验检验功效, 考虑带一个变点的模型, 并进行 M 次 Monte-Carlo 模拟. 在这 M 次模拟中, 计算出检验统计量

$$\frac{1}{\hat{\sigma}_{m,N}} \sup_{1 \leq k \leq T_m} |\Gamma(m, k, N)|$$

超过临界值 c_γ 的频率, 然后来计算经验检验功效. 同时, $\frac{1}{\hat{\sigma}_{m,N}} |\Gamma(m, k, N)|$ 超过临界值 c_γ 的最早出现的那个 k 被认为是变点位置的估计值.

一般来说, 作为一个好的检验方法, 经验检验水平应该小于或等于给定的水平, 而且经验检验功效应该足够大. 另外, 在变点后的停时应该足够的短才好. 因此首先按照定理 2.1 来计算临界值 c_γ .

第一步 模拟随机变量 $U_\gamma = \sup_{0 \leq t \leq 1} \frac{|W(t)|}{t^\gamma}$ M 次, 这里 $\{W(t), 0 \leq t < \infty\}$ 是一个 Wiener 过程.

第二步 在这 M 次模拟的基础上, 计算临界值 c_γ , 使得 $P[U_\gamma > c_\gamma] = \alpha$.

表 1: 由 50000 次 Monte-Carlo 模拟计算出的临界值 c_γ

γ	α				
	0.01	0.025	0.05	0.10	0.25
0	2.7060	2.4036	2.1896	1.9822	1.5091
0.15	2.7589	2.4568	2.2285	2.0060	1.6128
0.25	2.8103	2.4799	2.2794	2.0936	1.6899
0.35	2.9376	2.5966	2.3595	2.1959	1.8455
0.45	3.1943	2.8869	2.6515	2.4448	2.1103
0.49	3.3745	3.1770	2.8970	2.6791	2.3501

接下来, 在表 2-8 中, 利用表 1 中计算出的临界值 c_γ , 分别计算了在 γ , T_m , N , m 取不同值, 变点 k_m^0 在不同位置 (左端, 中间或右端) 时, 渐近方法的经验检验水平, 检验功效以及停止时刻的各数字特征. 具体来说:

表 2: 利用 c_γ 经 1000 次 Monte-Carlo 模拟计算出的经验检验水平, $N = 200$, $T_m = 500$

γ	$\alpha=0.025$			$\alpha=0.05$			$\alpha=0.10$		
	$m=30$	$m=100$	$m=300$	$m=30$	$m=100$	$m=300$	$m=30$	$m=100$	$m=300$
0	0.045	0.039	0.009	0.071	0.065	0.020	0.112	0.097	0.033
0.25	0.036	0.020	0.002	0.065	0.042	0.011	0.092	0.070	0.022
0.45	0.013	0.010	0.001	0.024	0.021	0.002	0.039	0.027	0.006
0.49	0.003	0.005	0	0.011	0.010	0.001	0.022	0.017	0.002

表 3: 利用 c_γ 经 1000 次 Monte-Carlo 模拟计算出的经验检验水平, $T_m = 500$, $m = 100$

γ	$\alpha=0.025$			$\alpha=0.05$			$\alpha=0.10$		
	$N=50$	$N=100$	$N=300$	$N=50$	$N=100$	$N=300$	$N=50$	$N=100$	$N=300$
0	0.033	0.023	0.033	0.055	0.049	0.064	0.095	0.100	0.103
0.25	0.024	0.026	0.029	0.043	0.039	0.047	0.069	0.061	0.071
0.45	0.010	0.012	0.011	0.017	0.018	0.012	0.028	0.027	0.027
0.49	0.001	0.003	0.001	0.003	0.004	0.004	0.005	0.012	0.009

表 4: 利用 c_γ 经 1000 次 Monte-Carlo 模拟计算出的检验功效, $N = 200$, $T_m = 500$, $k_m^0 = 25$

γ	$\alpha=0.025$			$\alpha=0.05$			$\alpha=0.10$		
	$m=30$	$m=100$	$m=300$	$m=30$	$m=100$	$m=300$	$m=30$	$m=100$	$m=300$
0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.45	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.49	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

表 5: 利用 c_γ 经 1000 次 Monte-Carlo 模拟计算出的检验功效, $N = 200$, $T_m = 500$, $m = 100$

γ	$\alpha=0.025$			$\alpha=0.05$		
	$k_m^0 = 200$	$k_m^0 = 300$	$k_m^0 = 390$	$k_m^0 = 200$	$k_m^0 = 300$	$k_m^0 = 390$
0	1.000	1.000	0.258	1.000	1.000	0.371
0.25	1.000	1.000	0.235	1.000	1.000	0.323
0.45	1.000	1.000	0.079	1.000	1.000	0.141

γ	$\alpha=0.10$		
	$k_m^0 = 200$	$k_m^0 = 300$	$k_m^0 = 390$
0	1.000	1.000	0.505
0.25	1.000	1.000	0.421
0.45	1.000	1.000	0.226

从表 2 可以看出, 当 m 固定时, 随着 γ 的增大, 经验检验水平越小. 当 γ 固定时, 随着 m 的增大, 经验检验水平越小. 从第一行和第二行的数据可以看到, 当 $\gamma = 0$ 和 0.25 时, 若历史数据集 m 较少, 得到的经验检验水平是会超过给定水平的. 从而可以考虑在 $[0, 1/2)$ 的范围

内选取较大的 γ , 来降低检验犯第一类错误的概率. 而 m 越大, 即没有变点的历史数据越多, 得到的经验检验水平越小, 这与我们的直观感觉也是相符合的.

从表 3 可以看出, γ 的大小对经验检验水平影响很大, 类似表 2 的表现. 当 N 固定时, 随着 γ 的增大, 经验检验水平是降低的. 在 $\gamma = 0$ 时, 经验检验水平甚至超过了给定的水平. 同时注意到, N 的大小对经验检验水平的影响不大, 不过比较而言, 中等大的 N 似乎能产生相对较低的经验检验水平.

通过模拟得到的列出的, 譬如表 4, 以及一些未列出的表格发现, 当变点发生在 m 个经验数据之后不久时, 则无论 m 和 N 值如何变化, 对于不同的 α 和 γ , 检验的功效都达到了 1.

从表 5 可以看出, 在变点 k_m^0 在 m 之后不久或者是更远一些, 检验的功效都是 1. 但是, 在变点靠近右端点时, 检验的功效会降低, 最高不超过 0.35. 在 k_m^0 固定时, 随着 γ 的增大, 功效是越来越低的. 因此, 基于检验水平和检验功效的考虑, 应该选取不大不小的 γ , 譬如取 $\gamma = 0.25$, 从而使两者能达到一种平衡.

表 6: $N = 200$, $T_m = 500$, $k_m^0 = 25$, 模拟次数 1000 次计算出的停止时刻的各数字特征

γ	$\tau_m \downarrow; \alpha \rightarrow$	$m = 30$			$m = 100$		
		0.025	0.05	0.10	0.025	0.05	0.10
0	Min	9.00	8.00	6.00	26.0	16.00	13.00
	Q1	28.00	27.00	27.00	28.0	28.00	27.00
	Median	28.00	28.00	28.00	28.0	28.00	28.00
	Mean	28.44	28.06	27.58	28.5	28.22	27.92
	Q3	29.00	29.00	29.00	29.0	29.00	29.00
	Max	32.00	32.00	31.00	32.0	31.00	31.00
0.25	Min	5.00	4.00	4.00	26.00	26.00	11.00
	Q1	28.00	28.00	27.00	28.00	28.00	27.00
	Median	28.50	28.00	28.00	29.00	28.00	28.00
	Mean	28.51	28.17	27.73	28.53	28.27	27.94
	Q3	29.00	29.00	29.00	29.00	29.00	29.00
	Max	33.00	32.00	31.00	32.00	31.00	31.00
0.45	Min	26.00	20.00	10.00	27.00	26.00	26.00
	Q1	28.00	28.00	28.00	28.00	28.00	28.00
	Median	29.00	29.00	29.00	29.00	29.00	29.00
	Mean	29.21	28.84	28.5	29.12	28.83	28.55
	Q3	30.00	30.00	29.00	30.00	29.00	29.00
	Max	33.00	33.00	32.00	32.00	32.00	31.00
0.49	Min	27.00	26.00	19.00	27.00	26.00	26.00
	Q1	29.00	28.00	28.00	29.00	28.00	28.00
	Median	30.00	29.00	29.00	29.00	29.00	29.00
	Mean	29.57	29.17	28.88	29.47	29.12	28.83
	Q3	30.00	30.00	30.00	30.00	30.00	29.00
	Max	34.00	33.00	33.00	32.00	32.00	32.00

从表 6 看出, 如果以停止时刻 τ_m 的中位数和均值作为判断标准的话, 在 γ 固定时, m 值的变化对停止时刻的影响不大, 也就是说, 对变点 k_m^0 的估计值影响不大. 但是, 在 m 值固定时, 随着 γ 的增大, 变点估计值与真实值的偏差是越来越大的.

从表 7 可以看出, 在 N 固定时, 随着 γ 的增大, 变点值的估计偏差也是越来越大的, 不过这些偏差的相差量不会很大, 不超过 1 或者 2. 若 γ 固定, 当 N 从 50 变到 300 时, 变点估计值的精确度明显提高了, 在真实变点 $k_m^0 = 25$ 时, 精确度提高了 4 ~ 5 个值. 从这里可以看出, 增大面板数据的横截面个数 N , 是可以明显提高变点估计的精确度的.

表 7: $T_m = 500$, $m = 100$, $k_m^0 = 25$, 模拟次数 1000 次计算出的停止时刻的各数字特征

γ	$\tau_m \downarrow; \alpha \rightarrow$	$N = 50$			$N = 300$		
		0.025	0.05	0.10	0.025	0.05	0.10
0	Min	23.00	8.00	8.00	26.00	12.00	10.00
	Q1	30.00	30.00	29.00	27.00	27.00	27.00
	Median	32.00	31.00	31.00	28.00	28.00	28.00
	Mean	31.79	31.13	30.56	27.94	27.66	27.44
	Q3	33.00	32.00	32.00	28.00	28.00	28.00
	Max	41.00	40.00	39.00	30.00	30.00	30.00
0.25	Min	26.00	23.00	18.00	26.00	26.00	12.00
	Q1	31.00	30.00	30.00	27.00	27.00	27.00
	Median	32.00	31.00	31.00	28.00	28.00	28.00
	Mean	32.08	31.52	30.98	27.99	27.78	27.58
	Q3	33.00	33.00	32.00	28.00	28.00	28.00
	Max	42.00	41.00	41.00	30.00	30.00	30.00
0.45	Min	27.00	27.00	26.00	26.00	26.00	26.00
	Q1	32.00	31.00	30.00	28.00	28.00	27.00
	Median	33.00	32.00	32.00	28.00	28.00	28.00
	Mean	33.29	32.59	32.01	28.44	28.18	27.98
	Q3	35.00	34.00	33.00	29.00	29.00	28.00
	Max	45.00	44.00	42.00	31.00	31.00	30.00
0.49	Min	27.00	26.00	26.00	27.00	27.00	26.00
	Q1	32.00	32.00	31.00	28.00	28.00	28.00
	Median	34.00	33.00	33.00	29.00	28.00	28.00
	Mean	34.14	33.30	32.65	28.75	28.45	28.21
	Q3	36.00	35.00	34.00	29.00	29.00	29.00
	Max	46.00	45.00	43.00	31.00	31.00	31.00

最后看下在 N , T_m , m 都固定, 而变点在不同位置时, 变点位置估计值的情况. 在表 8 中, 若 $k_m^0 = 25$, 这种情形在前面已经探讨过. 我们来看下 $k_m^0 = 390$, 即变点处于右端点的情形. 从表格中的数据可以发现, 对于这种变点最好以变点的中位数而不是均值作为变点的估计值, 因为中位数会离真实变点接近的多, 估计的偏差在 5 到 6 之间. 同时 γ 的大小与估计量的精确度的关系类似表 7 中的表现, 在 N 固定时, γ 越大, 估计的偏差也会相对大一些.

通过以上表格的具体分析,发现参数 γ 该如何选取是个很有趣的事情.从表2看出,较小或者中等大的 γ ($\gamma = 0, 0.25$) 且 m 取较小或中等大 ($m = 30, 100$) 时,数据表明经验检验水平与给定的水平接近.类似地,在表3中, $\gamma = 0, N = 100$ 同样可使得经验检验水平接近给定的真实值 α .我们再来看看参数 γ 与检验功效的关系.表4中的检验功效对应着变点 k_m^0 较早发生的情形,这时,无论 γ 在 $[0, \frac{1}{2})$ 区间取何值,对应的检验功效均为1.从表5看,在变点靠近右端点时, γ 越小(其它参数都固定)时,检验功效越大.同样,对表6-8的模拟结果进行分析可知,越小的 γ 可使变点 k_m^0 的估计值与真实值的偏离程度越小.根据以上这些分析,若从经验检验水平,检验功效以及变点估计的精确度三方面来综合考虑,我们倾向于选取 $[0, \frac{1}{4})$ 区间中较小的 γ ,从而达到最佳的检验和估计效果.

表 8: $N = 200, T_m = 500, m = 100$, 模拟次数 1000 次计算出的停止时刻的各数字特征

γ	$\tau_m \downarrow; \alpha \rightarrow$	$k_m^0 = 25$			$k_m^0 = 390$		
		0.025	0.05	0.10	0.025	0.05	0.10
0	Min	26.0	16.00	13.00	18.0	16.0	10.0
	Q1	28.0	28.00	27.00	391.8	328.0	183.8
	Median	28.0	28.00	28.00	397.0	396.0	395.0
	Mean	28.5	28.22	27.92	341.8	324.1	307.4
	Q3	29.00	29.00	29.00	398.0	398.0	398.0
	Max	32.0	31.00	31.00	399.0	399.0	399.0
0.25	Min	26.00	26.00	11.00	26.0	25	13.0
	Q1	28.00	28.00	27.00	393.0	281	190.5
	Median	29.00	28.00	28.00	396.5	396	395.0
	Mean	28.53	28.27	27.94	345.5	328	313.4
	Q3	29.00	29.00	29.00	398.0	398	398.0
	Max	32.00	31.00	31.00	399.0	399	399.0
0.45	Min	27.00	26.00	26.00	35.0	31.0	29.0
	Q1	28.00	28.00	28.00	395.0	394.0	394.0
	Median	29.00	29.00	29.00	397.0	397.0	397.0
	Mean	29.12	28.83	28.55	363.4	353.4	358.4
	Q3	30.00	29.00	29.00	398.0	399.0	398.0
	Max	32.00	32.00	31.00	399.0	399.0	399.0
0.49	Min	27.00	26.00	26.00	343.0	70.0	40.0
	Q1	29.00	28.00	28.00	394.2	395.0	394.2
	Median	29.00	29.00	29.00	397.0	398.0	397.0
	Mean	29.47	29.12	28.83	393.3	374.9	360.6
	Q3	30.00	30.00	29.00	399.0	398.0	399.0
	Max	32.00	32.00	32.00	399.0	399.0	399.0

4 定理的证明

在这一节中,将给出所有引理和定理的证明.

引理 2.1 的证明 注意到

$$\hat{\sigma}_{m,N}^2 = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{m-1} \sum_{j=1}^m \left(Y_{ij} - \frac{1}{m} \sum_{l=1}^m Y_{il} \right)^2 \right], \quad (4.1)$$

那么

$$\begin{aligned} E(\hat{\sigma}_{m,N}^2) &= \frac{1}{N(m-1)} \sum_{i=1}^N \sum_{j=1}^m \left\{ E(\mu_{ij} - \mu_{i0})^2 + E\left(\varepsilon_{ij} - \frac{1}{m} \sum_{l=1}^m \varepsilon_{il}\right)^2 \right. \\ &\quad \left. + 2E\left[(\mu_{ij} - \mu_{i0})\left(\varepsilon_{ij} - \frac{1}{m} \sum_{l=1}^m \varepsilon_{il}\right)\right] \right\} \\ &= \frac{1}{N(m-1)} \sum_{i=1}^N \sum_{j=1}^m \left[E\left(\varepsilon_{ij} - \frac{1}{m} \sum_{l=1}^m \varepsilon_{il}\right)^2 \right] = \frac{1}{N(m-1)} N(m-1)\sigma^2 = \sigma^2. \end{aligned}$$

而且

$$D(\sigma_{m,N}^2) = E(\sigma_{m,N}^4) - \sigma^4. \quad (4.2)$$

经过计算可得

$$\begin{aligned} \sigma_{m,N}^4 &= \frac{1}{N^2(m-1)^2} \left[\sum_{i=1}^N \sum_{j=1}^m \varepsilon_{ij}^2 - \frac{1}{m} \sum_{i=1}^N \left(\sum_{l=1}^m \varepsilon_{il} \right)^2 \right]^2 \\ &= \frac{1}{N^2(m-1)^2} \left\{ \left(\sum_{i=1}^N \sum_{j=1}^m \varepsilon_{ij}^2 \right)^2 + \frac{1}{m^2} \left[\sum_{i=1}^N \left(\sum_{j=1}^m \varepsilon_{ij} \right)^2 \right]^2 \right. \\ &\quad \left. - \frac{2}{m} \left(\sum_{i=1}^N \sum_{j=1}^m \varepsilon_{ij}^2 \right) \left[\sum_{i=1}^N \left(\sum_{l=1}^m \varepsilon_{il} \right)^2 \right] \right\}, \end{aligned} \quad (4.3)$$

于是有

$$\begin{aligned} E\left[\sum_{i=1}^N \sum_{j=1}^m \varepsilon_{ij}^2 \right]^2 &= E\left[\sum_{i=1}^N \left(\sum_{j=1}^m \varepsilon_{ij}^4 + 2 \sum_{1 \leq s < t \leq m} \varepsilon_{is}^2 \varepsilon_{it}^2 \right) + 2 \sum_{1 \leq k < l \leq N} \left(\sum_{j=1}^m \varepsilon_{kj}^2 \right) \left(\sum_{j=1}^m \varepsilon_{lj}^2 \right) \right] \\ &= N[3m + m(m-1) + (N-1)m^2]\sigma^4 \end{aligned} \quad (4.4)$$

和

$$\begin{aligned} E\left[\sum_{i=1}^N \left(\sum_{j=1}^m \varepsilon_{ij} \right)^2 \right]^2 &= E\left[\sum_{i=1}^N \left(\sum_{j=1}^m \varepsilon_{ij} \right)^4 + 2 \sum_{1 \leq i < j \leq N} \left(\sum_{l=1}^m \varepsilon_{il} \right)^2 \left(\sum_{l=1}^m \varepsilon_{jl} \right)^2 \right] \\ &= N[3m + m(m-1) + (N-1)m^2]\sigma^4, \end{aligned} \quad (4.5)$$

以及

$$E\left\{ \left(\sum_{i=1}^N \sum_{j=1}^m \varepsilon_{ij}^2 \right) \left[\sum_{i=1}^N \left(\sum_{l=1}^m \varepsilon_{il} \right)^2 \right] \right\} = [Nm(m+2) + N(N-1)m^2]\sigma^4. \quad (4.6)$$

由 (4.2)–(4.6) 式以及 Chebyshev 不等式, 得到

$$\begin{aligned} P(|\sigma_{m,N}^2 - \sigma^2| \geq \varepsilon) &\leq \frac{D(\sigma_{m,N}^2)}{\varepsilon^2} \\ &\leq \frac{\frac{1}{N^2(m-1)^2} \left[\left(1 + \frac{1}{m^2}\right) N(Nm^2 + 2m) - \frac{2}{m} (2Nm + N^2m^2) \right] - 1}{\varepsilon^2} \cdot \sigma^4 \\ &= \frac{\left(\frac{m}{m-1}\right)^2 - \frac{2m}{(m-1)^2} - 1 + o\left(\frac{1}{m}\right)}{\varepsilon^2} \cdot \sigma^4, \end{aligned} \quad (4.7)$$

这里 $N, m \rightarrow \infty$. 并且令 $\varepsilon = \frac{\sigma^2}{\sqrt{m}}$, 则当 $m \rightarrow \infty$ 时,

$$P(|\sigma_{m,N}^2 - \sigma^2| \geq \frac{\sigma^2}{\sqrt{m}}) \leq \frac{\frac{-1}{(m-1)^2} + o\left(\frac{1}{m}\right)}{\frac{1}{m}} \rightarrow 0. \quad (4.8)$$

$\sigma_{m,N}^2$ 的相合性得到了证明.

定理 2.1 的证明 借用文献 [3] 中的思想来证明定理 2.1. 对于面板数据模型 (2.1), 在原假设 (i.e. (2.3) 式) 下, 注意到

$$\Gamma(m, k, N) = \frac{\left(\sum_{i=1}^N \sum_{j=m+1}^{m+k} \varepsilon_{ij} - \frac{k}{m} \sum_{i=1}^N \sum_{l=1}^m \varepsilon_{il} \right)}{g(m, k, N)}, \quad (4.9)$$

这里 $g(m, k, N) = (Nm)^{\frac{1}{2}} \left(1 + \frac{k}{m}\right) \left(\frac{k}{m+k}\right)^\gamma$.

对每个 m , 利用 $\left\{ \sum_{i=1}^N \sum_{j=m+1}^{m+k} \varepsilon_{ij}, 1 \leq k < \infty \right\}$ 和 $\left\{ \sum_{i=1}^N \sum_{l=1}^m \varepsilon_{il} \right\}$ 的独立性, 以及 K-M-T 估计 (见文献 [12, 13]), 可得到

$$\sup_{1 \leq k < \infty} \frac{\left| \sum_{i=1}^N \sum_{j=m+1}^{m+k} \varepsilon_{ij} - \sqrt{N} \sigma W_{1,m}(k) \right|}{k^{\frac{1}{v}}} = \sqrt{N} \cdot O_p(1) \quad (4.10)$$

和

$$\sum_{i=1}^N \sum_{l=1}^m \varepsilon_{il} - \sigma W_{2,m}(Nm) = o_p\left((Nm)^{\frac{1}{v}}\right), \quad (4.11)$$

此处 $\{W_{1,m}(t)\}$ 和 $\{W_{2,m}(t)\}$ 是相互独立的 Wiener 过程, 且 $v > 2$. 从而推得

$$\sup_{1 \leq k < \infty} \frac{\left| \sum_{i=1}^N \left(\sum_{j=m+1}^{m+k} \varepsilon_{ij} - \frac{k}{m} \sum_{l=1}^m \varepsilon_{il} \right) - \left(\sqrt{N} \sigma \cdot W_{1,m}(k) - \frac{k}{m} \sigma \cdot W_{2,m}(Nm) \right) \right|}{g(m, k, N)} = o_p(1). \quad (4.12)$$

接下来, 利用

$$\begin{aligned} &\sup_{1 \leq k < \infty} \frac{|\sqrt{N} \sigma \cdot W_{1,m}(k) - \frac{k}{m} \sigma \cdot W_{2,m}(Nm)|}{g(m, k, N)} \stackrel{\mathcal{D}}{=} \sup_{1 \leq k < \infty} \frac{|\sqrt{N} \sigma \cdot W_1(k) - \frac{k}{m} \sigma \cdot W_2(Nm)|}{g(m, k, N)} \\ &\stackrel{\mathcal{D}}{=} \sup_{1 \leq k < \infty} \frac{|\sqrt{N} \sigma \cdot W_1(k) - \frac{k}{m} \sigma \cdot \sqrt{N} W_2(m)|}{g(m, k, N)} = \sup_{1 \leq k < \infty} \frac{|\sigma (W_1(k) - \frac{k}{m} W_2(m))|}{g_1(m, k)}, \end{aligned} \quad (4.13)$$

这里 $g_1(m, k) = m^{\frac{1}{2}} \cdot (1 + \frac{k}{m}) (\frac{k}{m+k})^\gamma$.

由文献 [3] 中的定理 2.1 的证明, 有

$$\sup_{1 \leq k < \infty} \frac{|\sigma(W_1(k) - \frac{k}{m}W_2(m))|}{g_1(m, k)} \xrightarrow{\mathcal{D}} \sup_{0 \leq t < \infty} \frac{|W_1(t) - tW_2(1)|}{(1+t)(\frac{t}{1+t})^\gamma} \quad (4.14)$$

以及

$$\sup_{0 \leq t < \infty} \frac{|W_1(t) - tW_2(1)|}{(1+t)(\frac{t}{1+t})^\gamma} \stackrel{\mathcal{D}}{=} \sup_{0 \leq t \leq 1} \frac{|W(t)|}{t^\gamma}. \quad (4.15)$$

而且如前面所讨论的, 当 $m, N \rightarrow \infty$ 时, 有 $\hat{\sigma}_{m,N}^2 - \sigma^2 = o_P(1)$.

于是定理 2.1 得到了证明.

定理 2.2 的证明 令 $k = \tilde{k} = k_m^0 + m$, 则

$$\Gamma_i(m, \tilde{k}) = \sum_{m < j \leq m+\tilde{k}} \frac{[Y_{ij} - \frac{1}{m} \sum_{l=1}^m Y_{il}]}{g(m, \tilde{k}, N)} = \frac{\left(\sum_{j=m+1}^{m+\tilde{k}} \varepsilon_{ij} - \frac{\tilde{k}}{m} \sum_{l=1}^m \varepsilon_{il} + \delta_{i0} (\tilde{k} - k_m^0) \right)}{g(m, \tilde{k}, N)}. \quad (4.16)$$

由定理 2.1 以及 Chebyshev 不等式, 得到

$$\begin{aligned} \Gamma(m, \tilde{k}, N) &= \frac{\left[\sum_{i=1}^N \sum_{j=m+1}^{k_m^0+2m} \varepsilon_{ij} - \frac{k_m^0+m}{m} \sum_{i=1}^N \sum_{l=1}^m \varepsilon_{il} \right]}{g(m, k_m^0+m, N)} + \frac{m \cdot \sum_{i=1}^N \delta_{i0}}{g(m, k_m^0+m, N)} \\ &= O_p(1) + \frac{m \cdot \sum_{i=1}^N \delta_{i0}}{(Nm)^{\frac{1}{2}} \left(1 + \frac{k_m^0+m}{m} \right) \left(\frac{k_m^0+m}{k_m^0+2m} \right)^\gamma}. \end{aligned} \quad (4.17)$$

由于 $(m)^{\frac{1}{2}} \cdot \left(1 + \frac{k_m^0+m}{m} \right) \left(\frac{k_m^0+m}{k_m^0+2m} \right)^\gamma = O(m^{\frac{1}{2}})$, 因此无论 N 是固定的还是 $N \rightarrow \infty$, 都有 $\Gamma(m, \tilde{k}, N) \rightarrow \infty$, 从而完成了定理 2.2 的证明.

5 结论

本文针对面板数据的序贯模型下的可能变点, 提出了 CUSUM 型检验统计量并制定了检验规则. 随后得到了检验统计量的大样本性质, 并基于相关理论结果构造了一种检验方法: 渐近检验法. 接下来, 对渐近检验法进行了 Monte-Carlo 数值模拟, 在模拟中对检验方法的经验检验水平, 检验功效进行了考察, 并在变点存在的情况下估计出变点的位置. 模拟显示渐近检验法是一种优良的检验估计方法. 最后给出了理论结果的证明.

参 考 文 献

- [1] Page E S. Continuous inspection schemes [J]. Biometrika, 1954, 41(1/2): 100–114.
- [2] Chu C S J, Stinchcombe M, White H. Monitoring structural change [J]. Econometrica, 1996, 64(5): 1045–1065.

- [3] Horváth L, Hušková M, Kokoszka P, et al.. Monitoring changes in linear models [J]. J. Stat. Plan. Infer., 2004, 126(1): 225–251.
- [4] Aue A, Horváth L. Delay time in sequential detection of change [J]. Stat. Prob. Lett., 2004, 67(3): 221–231.
- [5] Aue A, Horváth L, Hušková M, et al.. Testing for changes in polynomial regression [J]. Bernoulli, 2008, 14(3): 637–660.
- [6] Aue A, Horváth L, Kokoszka P, et al.. Monitoring shifts in mean: asymptotic normality of stopping times [J]. Test, 2008, 17(3): 515–530.
- [7] Aue A, Horváth L, Reimherr M. Delay times of sequential procedures for multiple time series regression models [J]. J. Econ., 2009, 149(2): 174–190.
- [8] Aue A, Dienes C, Fremdt S, et al.. Reaction times of monitoring schemes for ARMA time series [J]. Bernoulli, 2015, 21(2): 1238–1259.
- [9] Ciuperca G. Two tests for sequential detection of a change-point in a nonlinear model [J]. J. Stat. Plan. Infer., 2012, 143(10): 1719–1743.
- [10] Hušková M, Kirch C. Bootstrapping sequential change-point tests for linear regression [J]. Metrika., 2012, 75(5): 673–708.
- [11] Kirch C. Bootstrapping sequential change-point tests [J]. Sequential Anal., 2008, 27(3): 330–349.
- [12] Komlós J, Major P, Tusnády G. An approximation of partial sums of independent R.V.'s and the sample D.F. I [J]. Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 1975, 32(1): 111–131.
- [13] Komlós J, Major P, Tusnády G. An approximation of partial sums of independent R.V.'s and the sample D.F. II [J]. Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 1976, 35(3): 213–220.

THE ASYMPTOTIC METHOD OF SEQUENTIAL CHANGE-POINT IN PANEL DATA

CHEN Zhuo-heng^{1,2}, HU Yi-jun²

(1.School of Mathematical Science, Huaqiao University, Quanzhou 362021, China)

(2.School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China)

Abstract: In this paper, we study the statistical inference of change-point for panel data. For the change-point in a sequential panel data model, using the test statistic and the limiting distribution, we obtain an asymptotic method of change-point. Then we make a Monte-Carlo simulation, in the simulation, we investigate the effectiveness of our test method from three aspects: empirical test size, test power and average run length. Both the theory results and the Monte-Carlo simulation show that the asymptotic method we obtained is very excellent, in addition, the existing sequential method about the change-point in univariate sequence is generalized.

Keywords: change-point; panel data; sequential detection; asymptotic method

2010 MR Subject Classification: 62E20; 62F05; 62L12