鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2020,36(2):110—114

基于经验欧氏似然比的均值双变点检测

马岱君,李智航,张军舰

(广西师范大学 数学与统计学院,广西 桂林 541004)

摘要: 本文讨论了均值模型中双变点的检测问题,提出了一种利用经验欧氏似然比检测变点的方法,找到了该似然比的极限分布. 模拟结果表明本文提出的方法具有较好的检验功效,最后又通过飞机到达时间的实例进行了验证.

关键词: 均值模型; 双变点; 经验欧氏似然比

中图分类号: O212.7 **文献标志码:** A **文章编号:** 1673-8020(2020)02-0110-05

变点问题一直是统计学家们比较关注的一个研究方向. 自从 Page[1] 提出单变点问题后,变点问题得到了越来越多的关注. 比较常见的变点问题是均值变点问题,有大量学者讨论了更为简单的均值单变点模型的检验问题以及变点估计[2—5]. 关于参数方法研究变点问题的综述性文章有文献[6—8]等. 近年来,越来越多统计学者利用非参数方法研究变点问题,尤其是经验似然方法,如文献[9]提出用经验似然比对变点存在与否做假设检验; 文献[10]用截断经验似然比研究原点矩中的单变点问题,文献[11]推广至带有线性趋势变化的均值变点问题; 文献[12]则研究广义线性回归单变点模型. 经验似然方法是 Owen 在文献[13]中提出,并在文献[14]中给出系统的论述,经验似然相较于传统的参数方法具有一些较好的性质,如域保持性,置信域由样本决定等等; 不过经验似然计算比较复杂,其统计量一般无显式解. Owen[13—14] 提出经验欧氏似然,罗旭[15]则证明了经验欧氏似然具有与经验似然相类似的性质,但在一些场合下,该方法导出的统计量具有显示解. 基于以上研究现状,本文将利用经验欧氏似然方法对均值模型中的双变点进行检测,并通过数值模拟和实例分析说明该方法的有效性. 第 1 节介绍本文考虑的模型及检验方法,第 2 节是主要理论结果,第 3 节为数值模拟结果,第 4 节实例分析,第 5 节是结论.

1 模型与方法

设 $\{X_i\}_{i=1}^n$ 是 d 维随机变量序列,考虑如下均值双变点模型:

$$EX_i = \begin{cases} E_F X_i, i \in [1,p) \cup (q,n], \\ E_G X_i, i \in [p,q], \end{cases}$$

式中: $F, G = \lambda F + (1-\lambda) G^*$, G^* 均为一般的连续分布函数,$\lambda \in (0,1)$ 为常数. 当没有变点时,显然有 $F \equiv G^*$ 成立; 当变点存在时,且 $E_F X \neq E_{G^*} X$,$\lambda \in (0,1)$,即 G 为混合分布,λ 为混合比例,且均值发生改变. 可以建立如下假设检验问题:

$H_0: E_F X_i = E_G X_i = \mu_1, i = 1,2,\cdots,n;$

$H_1: \exists p,q \in \mathbf{Z}^+, 1 < p < q < n$,使得 $i \in [1,p) \cup (q,n]$ 时有 $E_F X_i = \mu_1$; 当 $i \in [p,q]$ 时,有

$$E_G X_i = \lambda\mu_1 + (1-\lambda)\mu_3 = \mu_2; \quad (1)$$

其中: μ_1, μ_2, μ_3 都是未知 d 维向量,$\lambda \in (0,1)$ 为未知常数. 对于检验问题(1)和固定的每个 $p,q \in \mathbf{Z}^+$,

收稿日期: 2020-01-02; **修回日期:** 2020-02-29

基金项目: 国家自然科学基金(11861017)

第一作者简介: 马岱君(1994—),女,山东临朐人,硕士研究生,研究方向为数理统计. E-mail: 957534920@qq.com

通信作者简介: 张军舰(1973—),男,河南内乡人,教授,硕士研究生导师,博士,研究方向为数理统计. E-mail: jjzhang@gxnu.edu.cn

$1 < p < q < n$, 可以构造如下两样本经验欧氏似然比检验函数:

$$l(p,q) = \sup\left\{-\frac{1}{2}\left(\sum_{i\in A}(n(1-\theta_{pq})p_i - 1)^2 + \sum_{i\in B}(n\theta_{pq}p_i - 1)^2\right) \mid \sum_{i\in A}p_i x_i = \mu_0, \sum_{i\in B}p_i x_i = \mu_0\right\}$$

其中: p_i 为对应的概率质量, $p_i \geqslant 0$, $\sum_{i\in A}p_i = 1$, $\sum_{i\in B}p_i = 1$; $A = \{k \mid k \in [1,p) \cup (q,n] \cup \mathbf{N}^+\}$; $B = \{k \mid k \in [p,q] \cup \mathbf{N}^+\}$; μ_0 为 d 维未知向量; $n\theta_{pq} = q - p$. 由拉格朗日乘子法容易求得(具体过程可参考文献[16]):

$$-2l(p,q) = n\theta_{pq}(1-\theta_{pq})(\bar{X}_{1pq} - \bar{X}_{2pq})^{\mathrm{T}}(\theta_{pq}S_{1pq} + (1-\theta_{pq})S_{2pq})^{-1}(\bar{X}_{1pq} - \bar{X}_{2pq}), \tag{2}$$

其中:

$$\bar{X}_{1pq} = \frac{1}{n-q+p}\sum_{i\in A}X_i, \bar{X}_{2pq} = \frac{1}{q-p}\sum_{i\in B}X_i,$$

$$S_{1pq} = \frac{1}{n-q+p}\sum_{i\in A}(X_i - \bar{X}_{1pq})(X_i - \bar{X}_{1pq})^{\mathrm{T}}, S_{2pq} = \frac{1}{q-p}\sum_{i\in A}(X_i - \bar{X}_{2pq})(X_i - \bar{X}_{2pq})^{\mathrm{T}}.$$

从式(2)来看,对整个序列 $\{X_i\}_{i=1}^n$ 而言,若不存在变点, $\bar{X}_{1pq}, \bar{X}_{2pq}$ 是总体均值的相合估计,故有较大的概率使得 $\bar{X}_{1pq} - \bar{X}_{2pq}$ 较小; 相反,若存在变点,有较大的概率使得 $\bar{X}_{1k} - \bar{X}_{2k}$ 较大. 故选取如下的检验函数来对假设检验问题(1)做检验:

$$M_n = \max_{1\leqslant p\leqslant q\leqslant n}\{-2l(p,q)\}.$$

但是当 $n\theta_{pq}$ 或 $n(1-\theta_{pq})$ 较小时,例如 $n\theta_{pq} < d$(d 为 X_i 的维数)时,样本协方差矩阵 S_{1pq} 或 S_{2pq} 的逆矩阵不存在,故 $n\theta_{pq}$ 至少应满足 $\min\{n\theta_{pq}, n(1-\theta_{pq})\} \geqslant d$,需要做截断处理. 文献[5]指出选取 $n\theta_{pq}$ 是比较任意的,所以调整 M_n 为如下形式:

$$M_n = \max_{p,q\in D}\{-2l_E(k)\}, D = \{p,q \mid 1 \leqslant p < q \leqslant n, q - p > \sqrt{n}\}. \tag{3}$$

使用 M_n 来对问题(1)做检验,需要明确该统计量的一些相关性质,下一节给出两个定理说明该检验函数的极限性质.

2　主要理论结果

定理 1　假设对 $\forall i, \mathrm{Var}X_i = \Sigma, \Sigma$ 正定, $\exists \delta \in (0,2) > 0$, s. t. $E_F \| X_i \|^{2+\delta} < \infty$,其中 $\| \cdot \|$ 为一般的二阶范数. 则在原假设成立时,有

$$\lim_{n\to\infty} P\{A(\log(u(n)))\sqrt{M_n}\} \leqslant x + D_d(\log(u(n))) = \exp\{-e^{-x}\},$$

其中:

$$A(x) = \sqrt{2\log x}, D_d(x) = 2\log x + \left(\frac{d}{2}\right)\log\log x - \log\left(\frac{d}{2}\right), u(n) = \frac{n^2 + (\sqrt{n})^2 - n\sqrt{n}}{(\sqrt{n})^2}.$$

证明　由 Marcinkiewicz - Zygmund 强大数律,可得

$$S_{1pq} = \Sigma + o_p((n(1-\theta_{pq}))^{-\frac{\delta}{2+\delta}}) = \Sigma + o_p(K^{-\frac{\delta}{2+\delta}}),$$

$$S_{2pq} = \Sigma + o_p((n\theta_{pq})^{-\frac{\delta}{2+\delta}}) = \Sigma + o_p(K^{-\frac{\delta}{2+\delta}}).$$

其中: $K = \min(n(1-\theta_{pq}), n\theta_{pq})$. 由重对数律,有

$$-2l_E(\theta_{pq}) = n\theta_{pq}(1-\theta_{pq})(\bar{X}_{1pq} - \bar{X}_{2pq})^{\mathrm{T}}(\Sigma + o_p(K^{-\frac{\delta}{2+\delta}}))^{-1}(\bar{X}_{1pq} - \bar{X}_{2pq}) =$$

$$n\theta_{pq}(1-\theta_{pq})((\bar{X}_{1pq} - \bar{X}_{2pq})^{\mathrm{T}}\Sigma^{-1}(\bar{X}_{1pq} - \bar{X}_{2pq}) + (\bar{X}_{1pq} - \bar{X}_{2pq})^{\mathrm{T}}(\bar{X}_{1pq} - \bar{X}_{2pq})o_p(K^{-\frac{\delta}{2+\delta}})) =$$

$$n\theta_{pq}(1-\theta_{pq})((\bar{X}_{1pq} - \bar{X}_2)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_{1pq} - \bar{X}_{2pq})) +$$

$$n\theta_{pq}(1-\theta_{pq})\left(O_p\left(\sqrt{\frac{\log\log((1-\theta_{pq})n)}{(1-\theta_{pq})n}}\right) - O_p\left(\sqrt{\frac{\log\log\theta_{pq}n}{\theta_{pq}n}}\right)\right)^2 o_p(K^{-\frac{\delta}{2+\delta}}) =$$

$$n\theta_{pq}(1-\theta_{pq})\left((\bar{X}_{1pq}-\bar{X}_2)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_{1pq}-\bar{X}_{2pq})+\left(O_p\left(\sqrt{\frac{\log\log K}{K}}\right)\right)^2 o_p(K^{-\frac{\delta}{2+\delta}})\right)=$$

$$n\theta_{pq}(1-\theta_{pq})(\bar{X}_{1pq}-\bar{X}_2)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_{1pq}-\bar{X}_{2pq})+o_p(K^{-\frac{\delta}{2+\delta}}).$$

注意到 $K>\sqrt{n}$,于是有

$$M_n=\max_{1\leqslant p+\sqrt{n}<q\leqslant n}\left\{n\theta_{pq}(1-\theta_{pq})(\bar{X}_1-\bar{X}_2)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_1-\bar{X}_2)\right\}+o_p(n^{-\frac{\delta}{4+4\delta}}),$$

然后由文献[6]推论A.3.1、类似文献[6]定理1.3.1的证明可得定理1,证毕.

定理1给出了在原假设成立的情况下,M_n 的极限分布为极值分布,利用此分布可以得到假设问题(1)的渐近判断方法,记 $\theta_{p_0q_0}=\frac{p_0-q_0}{n}$.

定理2 当 H_1 成立时,若 $\lim\limits_{n\to\infty}\theta_{p_0q_0}=\theta,\theta\in(0,1)$,$E_F(XX^{\mathrm{T}})-E_FX(E_FX^{\mathrm{T}})=\Sigma_1$,$E_G(XX^{\mathrm{T}})-E_GX(E_GX)^{\mathrm{T}}=\Sigma_2$,其中 Σ_1,Σ_2 正定有限,则有 $\frac{M_n}{n^{\omega}}\xrightarrow{p}+\infty$,其中 ω 为小于1的任意常数.

证明 当 $p=p_0,q=q_0$ 时,由 Marcinkiewicz-Zygmund 强大数律可得

$$\bar{X}_{1p_0q_0}=\mu_1+o(1),\bar{X}_{2p_0q_0}=\mu_2+o(1),$$

$S_{1p_0q_0}=\Sigma_1+o_p((n(1-\theta_{p_0q_0}))^{-\frac{\delta}{2+\delta}})=\Sigma_1+o_p(K^{-\frac{\delta}{2+\delta}})$,$S_{2p_0q_0}=\Sigma_2+o_p((n\theta_{p_0q_0})^{-\frac{\delta}{2+\delta}})=\Sigma_2+o_p(K^{-\frac{\delta}{2+\delta}})$.

于是有

$$-2l(p_0,q_0)=$$

$$n\theta_{p_0q_0}(1-\theta_{p_0q_0})(\mu_1-\mu_2+o_p(1))^{\mathrm{T}}(\theta_{p_0q_0}\Sigma_1+(1-\theta_{p_0q_0})\Sigma_2+o_p(1))^{-1}(\mu_1-\mu_2+o_p(1))=nO_p(1).$$

证毕.

由定理2可知:若变点存在,$\log\log M_n$ 趋于无穷,这与没有变点的情况有很大区别,因此用 M_n 来对问题(1)做假设检验是合适的.

3 数值模拟

下面通过模拟来说明第2节中的主要理论结果. 分别考虑2个模型:

模型1:$X_i\sim N(0,1),i\in A;X_i\sim\lambda N(0,1)+(1-\lambda)N(\mu,1),i\in B$;

模型2:$X_i\sim L(0,\sqrt{0.5}),i\in A;X_i\sim\lambda L(0,\sqrt{0.5})+(1-\lambda)L(\mu,\sqrt{0.5}),i\in B$.

其中 $A=\{1,\cdots,p-1,q+1,\cdots,n\}$,$B=\{p,\cdots,q\}$,$L(\cdot,\cdot)$ 为一般的拉普拉斯分布,相对于正态分布具有尖峰厚尾的特性. 给定显著性水平 $\alpha=0.05$,对于不同的样本量 n,分别在零假设下重复模拟一万次得到其经验水平,以此作为临界值;然后在不同的参数 n,p,q,λ,μ 下分别重复模拟1000次. 为获得该检验函数的真实检验效果,分别利用相应的经验水平来计算其检验功效,结果如表1~2.

表1为模型1的功效表,表内数据则是不同参数取值下的检验功效;表2则为模型2的功效表,其余类似. 如表1所示,总体来看,随着样本量增大,无论参数 p,q,λ,μ 如何取值,功效都在增大,到样本量为600时,检验功效都达到0.900以上,大部分都是1.000. 对于固定的 p,检验功效都分别随着参数 q,λ,μ 的增大而增大;对于相同的 $n,q-p$,在本文所考虑的 q,p 取值下,其检验功效都比较接近,由此说明,该检验对于 q,p 的位置是不敏感的,但是当 $q-p$ 越大,检验功效则越大. 观察 λ,μ 的取值,容易发现 λ,μ 越大,功效越大;当 λ 小 μ 大时,检验功效也比 λ 大 μ 小时的要高,由此看出,对于两变点之间的分布而言,即使混合比例较小,只要该混合分布的均值与两变点外的分布均值有较大差异,功效也不会受到太大影响. 从表2中可以看出与表1类似的结论,只是模型2下的功效表现要比模型1好. 综上可见,本文提出的方法具有一定的检验效果,且本文方法对数据服从厚尾分布相较于服从薄尾分布有较好的检验效果.

表 1　模型 1 功效表

Tab. 1 Power under model 1

n	(μ,λ)	$p/n=0.2$			$p/n=0.4$			$p/n=0.5$		
		$q/n=0.4$	$q/n=0.6$	$q/n=0.8$	$q/n=0.6$	$q/n=0.8$	$q/n=1.0$	$q/n=0.7$	$q/n=0.9$	$q/n=1.0$
200	(1,0.6)	0.364	0.574	0.627	0.346	0.596	0.619	0.381	0.602	0.614
	(1,0.8)	0.705	0.902	0.903	0.694	0.907	0.929	0.725	0.916	0.93
	(2,0.6)	0.885	0.994	0.997	0.878	0.997	0.997	0.894	0.996	0.998
	(2,0.8)	0.997	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
400	(1,0.6)	0.775	0.955	0.963	0.768	0.952	0.956	0.73	0.954	0.966
	(1,0.8)	0.981	1.000	1.000	0.979	1.000	1.000	0.983	0.998	1.000
	(2,0.6)	0.996	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
	(2,0.8)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
600	(1,0.6)	0.933	0.997	0.998	0.937	0.998	0.999	0.934	1.000	1.000
	(1,0.8)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(2,0.6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(2,0.8)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

表 2　模型 2 功效表

Tab. 2 Power under model 2

n	(μ,λ)	$p/n=0.2$			$p/n=0.4$			$p/n=0.5$		
		$q/n=0.4$	$q/n=0.6$	$q/n=0.8$	$q/n=0.6$	$q/n=0.8$	$q/n=1.0$	$q/n=0.7$	$q/n=0.9$	$q/n=1.0$
200	(1,0.6)	0.505	0.728	0.751	0.487	0.728	0.732	0.48	0.734	0.76
	(1,0.8)	0.838	0.961	0.971	0.832	0.963	0.972	0.843	0.963	0.968
	(2,0.6)	0.935	0.998	1.000	0.94	0.999	1.000	0.942	0.998	0.999
	(2,0.8)	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
400	(1,0.6)	0.886	0.985	0.988	0.874	0.987	0.981	0.884	0.983	0.986
	(1,0.8)	0.997	1.000	1.000	0.994	1.000	0.999	0.996	1.000	1.000
	(2,0.6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(2,0.8)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
600	(1,0.6)	0.957	0.998	1.000	0.964	0.997	0.999	0.967	0.998	0.999
	(1,0.8)	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(2,0.6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(2,0.8)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

4　实例分析

通过一个关于飞机到达时间的实例[11,16]对模型进行验证. 该实例一共有 213 个数据(数据来源于文献[16]),是一组于 1968 年 4 月 30 日中午至晚上 8 时从低空过渡控制区收集的飞机到达时间. 图 1 是该数据的一阶差分时序图,即到达时间间隔时序图. 图中虚线即为 M_n 取得最大值时的点 $p=115$, $q=131$,此时 M_n 为 18.51,由定理 1 中的极限分布可得 p 值为 0.0079,远比 0.05 小,这与文献[1,16]的研究结果一致.

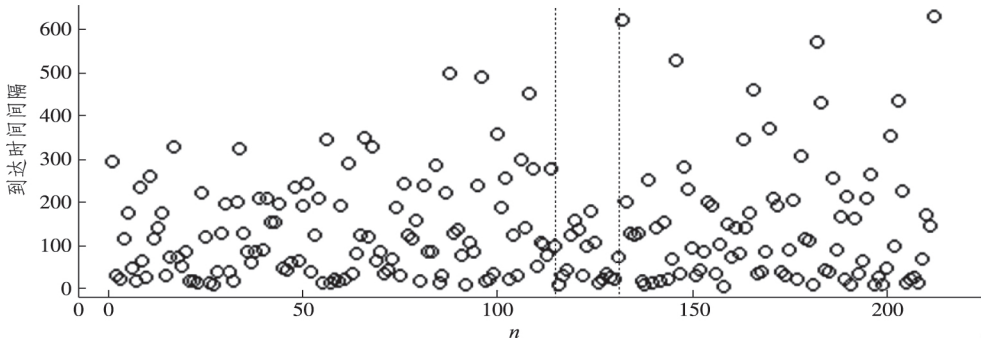

图 1　到达时间间隔时序图

Fig. 1 Sequence diagram for the inter – arrival time

5 结论

本文采用经验欧氏似然方法对双变点均值模型做假设检验,且两变点之间的分布为混合分布;根据该模型特点构建了经验欧氏似然比检验函数,找到其在零假设下的极限分布,并给出变点是否存在的判断方法;同时证明了该检验函数在有无变点时发散的阶数相差较大,从而随着样本量增大,检验效果越好;模拟结果及实例分析则进一步说明了该检验的优良性. 此外,本方法计算较为简单,具有较好的实用性.

参考文献:

[1] PAGE E S. Continuous Inspection Schemes [J]. Biometrika,1954,41(1/2):100-115.

[2] SEN A K,SIRVASTAVA M S. On tests for detecting changes in mean [J]. The Annals of Statistics,1975,3(1):98-108.

[3] CHERNOFF H,ZACKS S. Estimating the current mean of a normal which is subjected to changes in time [J]. Annals of Mathematical Statistics,1964,35(3):999-1018.

[4] HAWKINS D M. Testing a sequence of observations for a shift in location [J]. Journal of the American Statistical Association,1977,72(357):180-186.

[5] KIM H J,SIEGMUND D. The likelihood ratio test for a change-point in simple linear regression [J]. Biometrika,1989,76(3):409-423.

[6] CSÖRGÖ M,HORVÁTH L. Limit theorems in change-point analysis [M]. New York: John Wiley,1997.

[7] 陈希孺. 变点统计分析简介 [J]. 数理统计与管理,1991,10(1):55-58.

[8] CHEN J,GUPTA A K. Parametric statistical change point analysis [M]. Boston: Birkhäuser,2012.

[9] EINMAHL J H J,MCKEAGUE I W. Empirical likelihood based hypothesis testing [J]. Bernoulli,2003,9(2):267-290.

[10] ZOU C L,LIU Y K,PENG Q,et al. Empirical likelihood ratio test for the change-point problem [J]. Statistics and Probability Letters,2007,77(4):374-382.

[11] NING W. Empirical likelihood ratio test for a mean change point model with a linear trend followed by an abrupt change [J]. Journal of Applied Statistics,2011,39(5):947-961.

[12] 李云霞,刘伟棠. 基于经验似然的 Logistic 回归模型的变点检验 [J]. 高校应用数学学报,2015,30(3):367-378.

[13] OWEN A B. Empirical likelihood ratio confidence regions [J]. The Annals of Statistics,1990,18(1):90-120.

[14] OWEN A B. Empirical likelihood [M]. London: Chapman and Hall,2001.

[15] 罗旭. 半参数模型的经验欧氏似然估计的大样本性质 [J]. 应用概率统计,1994,10(4):344-352.

[16] HSU D A. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis [J]. Journal of the American Statistical Association,1979,74(365):31-40.

Mean Model with Two Change-points Detect via Empirical Euclidean Likelihood Ratio

MA Daijun,LI Zhihang,ZHANG Junjian

(College of Mathematics and Statistics,Guangxi Normal University,Guilin 541004,China)

Abstract: The empirical Euclidean likelihood ratio is proposed to detect two change-points in mean models. The asymptotic distribution of empirical Euclidean likelihoodratio is given. The simulation results indicate that our method has good test power. At last,the method is applied to the real example of aircraft arrival time.

Keywords: mean model; two change-points; empirical Euclidean likelihood ratio

(责任编辑 李秀芳)