气候时间序列变点的推断

杨喜寿 杨洪昌

(山东大学管理科学系，济南 250100) (山东省气象台，济南 250031)

提 要 本文研究时间序列的变点问题，所给出的统计方法可用来推断一个时间序列是否存在变点，存在几个变点以及变点在什么位置。这种统计推断方法可用于气候阶段的划分。

关键词 变点 气候时间序列 似然比检验统计量

1 引言

设$\{X_1, X_2, \cdots, X_N\}$是独立正态随机变量序列，若存在$r(0<r<n)$，使得

$$X_i \sim N(\theta_1, \sigma^2), \quad i=1,2,\cdots,r$$
$$X_j \sim N(\theta_2, \sigma^2), \quad j=r+1,\cdots,n$$

其中$\theta_1 \neq \theta_2, 1 \leqslant r < n$，则称$r$是序列$\{X_1, \cdots, X_n\}$的一个变点。变点问题已被广泛应用于许多不同的领域，本文着重讨论气候时间序列变点的估计和推断问题，其内容编排如下：在第二节，讨论似然比检验统计量的原分布并且给出检验方法。在第三节，给出了关于几个变点的极大似然估计，建立了相应的递推算法和检验程序。

许多气象专家指出，气候可以发生明显的变化，表现出一定的阶段性[1~4]。研究气候的变化规律，其中的重要内容是划分气候变化的阶段。目前尚无科学的方法找出气候阶段的起点。变点统计分析方法可用来解决这一问题。本文在第四节应用变点统计分析技术研究了济南市自1919到1988年平均气温的变化规律。

2 变点的统计检验

设X_1，X_2，…，X_n是独立的正态随机变量，考虑原假设

$$H_0: \ X_i \sim N(\theta_1, \sigma^2), \quad i=1,2,\cdots,n$$

备选假设为

$$H_1: \ X_i \sim N(\theta_1, \sigma^2), \quad i=1,2,\cdots,r$$
$$X_j \sim N(\theta_2, \sigma^2), \quad j=r+1,r+2,\cdots,n$$

其中参数θ_1，θ_2，r是未知的$(\theta_1 \neq \theta_2, \ 1 \leqslant r < n)$。

当H_1成立时，参数r是一个变点，表明序列在r时刻之后其均值发生了变化。当

1993-02-11 收到，1993-11-21 收到修改稿

H_0成立时，表示序列无变点。

当σ已知时，不失一般性可令其为1。令

$$T_r=\sqrt{\frac{n}{r(n-r)}}\sum_{i=1}^{r}(X_i-\overline{X}),\quad \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1}$$

$$U=\max_{1\le r<n}|T_r|, \tag{2}$$

可以证明U是似然比检验统计量[5]。

当σ未知时，令

$$W=\max_{1\le r<n}[\sqrt{n-2}\,|T_r|/S_r], \tag{3}$$

式中

$$S_r=[\sum_{i=1}^{r}(X_i-\overline{X}_r)^2+\sum_{i=r+1}^{n}(X_i-\overline{X}'_r)^2]^{1/2},$$

$$\overline{X}_r=\frac{1}{r}\sum_{i=1}^{r}X_i,\quad \overline{X}'_r=\frac{1}{n-r}\sum_{i=r+1}^{n}X_i,$$

则W是似然比检验统计量。

我们已推导出U的原分布函数为[6]

$$F(x)=\left\{\prod_{r=2}^{n-1}[1-4T(x,x_r)-4T(x,\frac{1}{x_r})]\right\}\Big/[2\Phi(x)-1]^{n-3}, \tag{4}$$

式中

$$x_r=\sqrt{\frac{r(n-r+1)}{n}}-\sqrt{\frac{(r-1)(n-r)}{n}},$$

$$T(x,y)=\frac{1}{2\pi}\int_0^y\frac{1}{1+t^2}\exp\{-x^2(1+t^2)/2\}dt,$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left\{-\frac{t^2}{2}\right\}dt.$$

利用(4)式可以计算分布函数$F(x)$的数值。表1列出了U的某些重要临界值。

当σ未知时，可用W作为检验变点的统计量。但难于求出W的原分布。对于给

表1 $P\{U>x\}=\alpha$

n	α			n	α			n	α		
	0.10	0.05	0.01		0.10	0.05	0.01		0.10	0.05	0.01
4	2.06	2.34	2.90	13	2.48	2.75	3.26	30	2.70	2.95	3.45
5	2.15	2.43	2.98	14	2.51	2.77	3.28	35	2.75	2.99	3.48
6	2.23	2.51	3.05	15	2.53	2.78	3.30	40	2.79	3.00	3.50
7	2.30	2.56	3.09	16	2.55	2.80	3.31	45	2.80	3.04	3.52
8	2.34	2.60	3.13	17	2.58	2.81	3.32	50	2.82	3.06	2.54
9	2.36	2.63	3.17	18	2.59	2.84	3.34	55	2.85	3.09	3.56
10	2.40	2.67	3.19	19	2.60	2.85	3.35	60	2.88	3.10	3.58
11	2.44	2.70	3.22	20	2.60	2.86	3.36	70	2.90	3.13	3.60
12	2.46	2.72	3.24	25	2.66	2.91	3.41				

定的 r，我们注意到统计量$\sqrt{n-2}\,T_r / S_r$是自由度为$(n-2)$的 t-变量。所以可以应用 Bonferroni 不等式来计算 W 近似的 100(1−α)%分位点值，即有

$$P\{W>x\}=1-P\{W\leqslant x\}=1-P\left\{\max_{1\leqslant r<n}\frac{\sqrt{n-2}\,|T_r|}{S_r}\leqslant x\right\}$$

$$\approx 1-\sum_{r=1}^{n-1}P\left\{\frac{\sqrt{n-2}\,|T_r|}{S_r}\leqslant x\right\}.$$

Worsley[7]指出，当 n 和 α 较小时，Bonferroni 近似是相当精确的；当 n 很大时，Hawkins[5]证明当 $n\to\infty$时$W\to U$，其中$U=\max_{1\leqslant r<n}|T_r|/\sigma$。所以，对于大的 n，我们可以利用 U 的分位点作为 W 分位点的近似。

关于统计量 W 的分位点，Worsley[7]的计算结果如表 2 所示。

表 2　$P\{W>x\}=\alpha$

精确分位点

α	n							
	3	4	5	6	7	8	9	10
0.10	12.71	5.34	4.18	3.73	3.48	3.32	3.21	3.14
	(12.71)	(5.34)	(4.18)	(3.75)	(3.53)	(3.41)	(3.34)	(3.28)
0.05	25.45	7.65	5.39	4.60	4.20	3.95	3.78	3.66
	(25.45)	(7.65)	(5.39)	(4.60)	(4.22)	(4.00)	(3.86)	(3.76)
0.01	127.32	17.28	9.46	7.17	6.14	5.56	5.19	4.93
	(127.32)	(17.28)	(9.46)	(7.17)	(6.14)	(5.56)	(5.20)	(4.96)

近似分位点(Monte Carlo)

α	n							
	15	20	25	30	40	50		
0.10	2.97	2.90	2.89	2.86	2.88	2.87		
	(3.19)	(3.17)	(3.18)	(3.20)	(3.23)	(3.26)		
0.05	3.36	3.28	3.23	3.19	3.17	3.16		
	(3.55)	(3.49)	(3.47)	(3.47)	(3.48)	(3.50)		
0.01	4.32	4.13	3.94	3.86	3.77	3.79		
	(4.40)	(4.21)	(4.21)	(4.07)	(3.48)	(4.02)		

* 括号内为 Bonferroni 近似。

当使用 W 作为检验统计量时，比较表 1 和表 2，我们可发现：

(1) 如果检验水平 $\alpha=0.10$，那么当样本容量 $n<15$ 时可使用 Bonferroni 近似，当 $n\geqslant 15$ 时可使用 U 的分位点。

(2) 如果检验水平 $\alpha=0.05$，那么当样本容量 $n<30$ 时可使用 Bonferroni 近似，当 $n\geqslant 30$ 时可使用 U 的分位点。

(3) 如果检验水平 $\alpha=0.01$，那么当样本容量 $n<50$ 时可使用 Bonferroni 近似，当 $n\geqslant 50$ 时可使用 U 的分位点。

若检验水平 α 给定，则不难设计对于一个变点的检验程序。

例 用模拟的方法生成一个独立正态随机样本如下:

3.44, 3.91, 2.98, 7.26, 5.98, 6.19, 8.66

其样本1到3是取自母体 $N(3, 1)$; 样本4到7是取自母体 $N(7, 1)$。假定 θ_1, θ_2 和 σ^2 皆未知, 试作变点检验。

令

$$W_r = \frac{\sqrt{n-2}|T_r|}{S_r}, \quad r = 1, 2, \cdots, 6$$

计算结果如表3所示。

表3 模拟计算结果

r	1	2	3	4	5	6
$\|T_r\|$	2.2127	3.0347	4.6862	3.3333	3.2403	3.4255
S_r	4.6920	4.2073	2.2249	3.9749	4.0511	3.8957
W_r	1.0545	1.6129	4.7097	1.8751	1.7885	1.9662

由表3可知, $W=4.7097$ 对应的序列下标为 $r=3$, 再由表2可知, 95%显著水平的临界值为4.20, 于是下标3应当被认为是变点。

3 多个变点的统计推断

假定 $\{X_1, X_2, \cdots, X_N\}$ 具有 k 个变点, 记为 $l_1, l_2, \cdots, l_k$, 且假定

$$0 = l_0 < l_1 < \cdots < l_k < l_{k+1} = n,$$

令

$$G(l_1, l_2, \cdots, l_k) = \sum_{i=1}^{k+1} \frac{(X_j + X_{j+1} + \cdots + X_{l_i})^2}{n_i}$$

式中

$$n_i = l_i - l_{i-1}, \quad j = l_{i-1} + 1,$$

可以证明, 求 $(l_1, \cdots, l_k)$ 的极大似然估计等价于 $G(l_1, \cdots, l_k)$ 达到极大。

如果我们用 $(\hat{l}_1, \cdots, \hat{l}_k)$ 表示极大似然估计, 则有

$$G(\hat{l}_1, \cdots, \hat{l}_k) = \max_{0<l_1<\cdots<l_k<n} G(l_1, \cdots, l_k) \tag{5}$$

下面我们建立一个解(5)式的递推算法。

令

$$\begin{cases} S(i) = X_1 + \cdots + X_i, & i = 1, 2, \cdots, n \\ R_0(t, i) = \dfrac{S^2(t)}{t} + \dfrac{[S(i) - S(t)]^2}{i - t}, & 2 \leqslant i \leqslant n, \quad 1 \leqslant t \leqslant i-1 \end{cases} \tag{6}$$

第1步: 求 $t^*(i, 1)$, 使得 $R_0(t, i)$ 在 $t = t^*(i, 1)$ 处达到极大, 其极大值记为

$S^*(i,1)$，即

$$S^*(i,1)=R_0(t^*(i,1),i)=\max_{1\leqslant t\leqslant i-1}R_0(t,i),\qquad i=2,\cdots,n \tag{7}$$

假定已进行到第$(j-1)$步$(2\leqslant j\leqslant k)$，并且已求得$t^*(i,j-1)$、$S^*(i,j-1)$ $(j\leqslant i<n)$，令

$$R_1(t,i)=S^*(t,j-1)+\frac{[S(i)-S(t)]^2}{i-t},$$

那么，第j步：求$t^*(i,j)$，使得$R_1(t,i)$在$t=t^*(i,j)$处达到极大，其极大值记为$S^*(i,j)$，即

$$S^*(i,j)=R_1(t^*(i,j),i)=\max_{j\leqslant t\leqslant i-1}R_1(t,i),\quad j+1\leqslant i<n \tag{8}$$

这种递推算法只需进行k步。最后一步，即第k步，我们只要应用(8)式求出$S^*(n,k)$和$t^*(n,k)$，不必计算$t^*(k+1,k),S^*(k+1,k),\cdots,t^*(n-1,k),S^*(n-1,k)$。最后，所求得的极大似然估计为

$$\hat{l}_k=t^*(n,k),\quad \hat{l}_{k-1}=t^*(\hat{l}_k,k-1),\cdots,\quad \hat{l}_1=t^*(\hat{l}_2,1), \tag{9}$$

上述递推算法的证明并不困难，只是因为使用的字符过于复杂，使得书写冗长。为简单起见，我们仅对$k=2$的情况加以证明。至于$k>2$的情况，其证明方法是类似的。

$k=2$是指存在两个变点，序列被划分为三段。假定依照我们所设计的算法，得到两个变点的估计为

$$q=t^*(n,2),\quad p=t^*(q,1)$$

根据第1步的算法，对于任给的指标证$i\in\{2,3,\cdots,n\}$，有

$$G(t^*(i,1),i)\geqslant G(t,i),\quad t=1,2,\cdots,i-1 \tag{10}$$

又根据第2步的算法，有

$$G(t^*(q,1),q)\geqslant G(t^*(i,1),i),\quad i=2,3,\cdots,n \tag{11}$$

由(10)、(11)式可知

$$G(p,q)\geqslant G(t,i),\quad i=2,3,\cdots,n;t=1,2,\cdots,i-1$$

下面我们讨论变点的检验问题。目前关于变点问题的多数研究工作是假定变点个数已知。这里，我们假定变点个数是未知的。即，使变点个数是未知的，但它的最大可能值M可以根据实际问题来确定。一般说来，M应当远小于n。为了推导变点个数并且得到其估计值，我们利用第二节的检验方法设计了一种逐步检验程序。为简单起见，我们仅讨论$M=2$的情况。对于$M>2$，方法是类似的。

第1步：对$k=2$求序列$\{X_1,\cdots,X_n\}$变点的极大似然估计，记为$(\hat{l}_1,\hat{l}_2)$。对$\{X_1,\cdots,X_{\hat{l}_2}\}$检验$\hat{l}_1$是否可信为变点；对$\{X_{\hat{l}_1+1},\cdots,X_n\}$检验$\hat{l}_2$是否可信为变点。

如果$\hat{l}_1$和$\hat{l}_1$皆可信为变点，则认为$\{X_1, \cdots, X_n\}$具有两个变点，其估计为$(\hat{l}_1, \hat{l}_2)$；否则转下一步。

第2步：对$k=1$求序列$\{X_1, \cdots, X_n\}$的变点，记为$\hat{l}$。检验$\hat{l}$是否可信为变点。如果$\hat{l}$可信为变点，那么认为$\{X_1, \cdots, X_n\}$具有1个变点，其估计为$\hat{l}$；否则，认为$\{X_1, \cdots, X_n\}$无变点。

4 实例

济南市自1919至1988年的平均气温如表4和图1所示。因为1937、1938和1948年记录不全，我们只能省略这三年的数据。这并不严重地影响变点的推断。

表4 济南市年平均气温

年	0	1	2	3	4	5	6	7	8	9
	温度(℃)									
1910										15.2
1920	15.1	13.8	15.0	14.4	14.9	14.6	14.9	15.1	14.9	14.8
1930	14.5	14.2	15.1	14.1	14.2	15.1	14.0			15.3
1940	15.7	16.1	15.7	15.4	15.0	15.2	15.6	13.8		14.5
1950	13.9	14.0	14.0	14.5	13.5	14.6	13.2	13.5	14.3	14.6
1960	14.5	15.1	14.1	14.0	13.3	14.8	14.8	14.4	14.9	13.3
1970	13.8	14.0	14.1	14.5	14.1	14.7	13.8	14.9	14.9	14.6
1980	14.1	14.7	15.1	15.1	13.9	13.8	14.6	14.9	14.8	

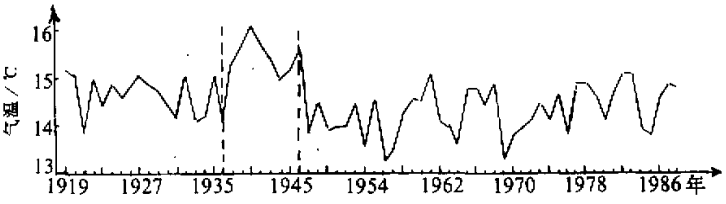

图1 济南市1919～1988年平均气温

经过初步分析，我们确定该序列的变点个数不超过4。

第1步：假定序列具有4个变点。应用递推算法得到变点的极大似然估计为

(1936, 1946, 1955, 1957)

首先考察序列$\{x_{1947}, \cdots, x_{1957}\}$；应用(3)式，我们得到检验统计量$W=2.53$，这里$n=11$。由表2可知，在90%的显著水平上，不能认为1955是变点。于是转向下一步。

第2步：假定序列具有3个变点。应用递推算法，我们得极大似然估计为

(1936, 1946, 1976)

首先考察$\{x_{1947}, \cdots, x_{1988}\}$。应用(3)式，我们得到检验统计量$W=2.70$，这里

$n=42$。由表 1 可知，在 90%的显著水平上，不能认为 1976 是变点(注：当 $n=42$ 时，U 的 90%分位点是 2.80，所以 W 的近似分位点是 2.80)。于是，再转向下一步。

第 3 步：假设序列有两个变点。应用递推算法我们得极大似然估计为

(1936，1946)

首先考察$\{x_{1919}, \cdots, x_{1946}\}$。应用(3)式得$W=4.71$，这里 $n=26$。由表 1 可知，在 99%的显著水平上可信 1936 为变点。

我们再考察$\{x_{1939}, \cdots, x_{1988}\}$。应用(3)式我们得 $W=6.15$，这里 $n=50$。由表 1 可知，在 99%的显著水平上可信 1946 为变点。

最后，得到的结论是：$\{x_{1919}, \cdots, x_{1988}\}$具有两个变点，它们是(1936，1946)。

从 1919 到 1988 年，济南市年平均气温可分为三个阶段。第 1 阶段是从 1919 到 1936，年平均气温的均值为 14.7℃；第 2 阶段是从 1939 年到 1946，均值为 15.5℃；第 3 阶段是从 1947 到 1988，均值为 14.3℃。这个结论和其他关于本世纪中国气温变化的研究结果是一致的[3] (注：对 1937 和 1938，我们无法确定它们属于哪个阶段)。

参 考 文 献

1 丁士晟，1983，吉林省气候变化对水资源的影响，气象学报，41，No.4，426–432。

2 冯九华，1987，简单气候模式的随机分析解，气象学报，45，No.3，297–303。

3 张先恭、李小泉，1982，本世纪我国气温变化的某些特征，气象学报，40，No.2，198–208。

4 王绍武、赵宗慈，1984，北半球对流层下部温度变化的研究，气象学报，42，No.2，238–245。

5 Hawkins, D.W., 1977, Testing a sequece of obervations for a shift in location, *J. Am. Statist. Assoc.*, **72**, 180–186.

6 杨喜寿，1994，正态总体位置参数移动的似然比统计量的分布，数理统计与应用概率，9，No.2，59–66。

7 Worsley, K.J., 1979, On the likelihood ratio test for a shift in location of normal populations, *J. Am. Statist. Assoc.*, **74**, 365–367.

Inference on the Change Points in a Climate Time Series

Yang Xishou

(Management Science Department of Shandong University, Jinan 250100)

Yang Hongchang

(Meteorological Observatory of Shandong Province, Jinan 250031)

Abstract　In this paper, the change point problems of time series are studied and a statistical inference method is proposed which can be used to infer whether there exist change points, how many change points there are, and where their positions are. This method can be applied to segmentation of the phases of climate variation.

Key words　change point　climate time series　likelihood ratio test statistics