公共安全突发事件的探测分析
——利用方差多变点分析技术对SARS疫情的研究

廖远甦，刘　弘
（上海财经大学 经济学院 上海 200083）

摘　要：突发事件对公共安全领域构成了严重的威胁。虽然预测突发事件几乎是不可能的，但是及时捕捉突发事件的异常变化却是完全有可能的。因为突发事件发展态势的突变通常表现为某种特征变量的方差发生突变，所以，利用方差多变点分析技术可以敏锐地监控突发事件的变动趋势。本文在Inclan研究工作的基础上，改进了方差参数的先验分布和求解方法，利用变点解的一致性构造了一个快速算法，并以公共卫生突发事件SARS为例展示了新算法的高效率和准确性，为检测突发事件提供了一种实用有效的新技术。

关键词：方差变点；公共安全突发事件；突变；贝叶斯方法
中图分类号：F224.7　**文献标识码**：A　**文章编号**：1001-9952(2003)11-0076-05

一、前言

20世纪下半叶以来，世界范围内的公共安全突发事件时有发生，切尔诺贝利核电站事故、日本地铁沙林事件、美国9.11恐怖袭击事件、韩国大邱地铁纵火事件以及石家庄特大爆炸案、非典型性肺炎SARS等突发事件对全世界的社会公共安全领域都构成严重威胁。为了迅速有效地应对突发事件，迫切需要运用科学定量的技术手段监控突发事件，尤其是捕捉突发事件发展态势的突变。方差多变点分析技术就是一种探测突变的有效技术。

Inclan(1993)运用贝叶斯方法研究了股价收益率的多变点分析问题，但是她的方法存在两个缺陷。第一个缺陷是计算量大，当时间序列较长或者变点较多时，实用性较差；第二个缺陷是先验分布的超参数的确定不稳健，突出地表现在变点个数的先验分布上。Inclan和Tiao(1994)用累加平方和(cumulative sum of squares)的方法来研究方差多变点问题，并给出了IT检验，该方法计算量小，但探测多变点时必须将整个时间序列样本分割，难以保证得到的变点是在全局意义上具有显著性。

本文借鉴了上述两篇论文，从两个方面对Inclan(1993)的贝叶斯方法作了改进。一方面，避免用穷举法求极大似然值，利用启发式算法提高变点分析效率；另一方面，综合利用多种统计方法克服超参数选择的随意性和不稳健。我们利用新算法研究了公共卫生突发事件SARS在北京地区发展的突变情况。

二、突变的发生机制

系统科学对突变的发生机制进行了大量研究。不论是自然界还是人类社会，我们都可以

收稿日期：2003-08-18
作者简介：廖远甦(1975－)，男，江苏盐城人，上海财经大学经济学院博士生；
刘　弘(1965－)，男，上海人，上海财经大学经济学院副教授。

看到系统演化的两种基本形式，渐变和突变。突变是非线性系统的普遍行为，它有两种含义：一是相对于渐变的骤变，强调变化发生的瞬时性，指的是在可以忽略的时间内完成的变化；二是突变论讲的突变，指的是非常剧烈的变化。突变论告诉我们，不论构成系统的基本特性和引起结构或形态变化的"力"的性质如何，只要控制参量变化到分岔点上，就会出现一种定态向另一种定态的突变。对于公共安全突发事件的突变，上述两种含义兼而有之，公共安全突发事件本身就有骤然发生，难以防范的特点，也正由于措手不及，可能会在短时间迅速恶化，表现为发展态势的突变。

混沌理论也对突变提出了另一种见解。混沌是确定的非线性系统表现的随机行为，它的一个重要特征是系统对初始条件的敏感性，即所谓的"蝴蝶效应"，对此有一个形象的说法：一只蝴蝶在巴西扇动几下翅膀，就有可能在美国的德克萨斯州引起一场龙卷风。由于公共安全突发事件对人们心理的巨大冲击，人们对一些微小的变化非常敏感，整个社会可能表现出类似混沌的行为。例如，1974 年，日本在石油危机的大背景下，因为有家庭主妇在超级市场前排队购买手纸，从而引发了整个东京的手纸危机。

虽然系统科学对突变的发生机制提供了深刻的见解，但应用到公共安全领域这样的大系统中存在着一定的困难。突变论和混沌理论虽然在自然科学领域得到了广泛验证，但在各种因素相互作用的社会领域中，更多的是提供一种理解，因为社会领域中，系统的方程是未知的，再加上系统结构的时变性，至多只能利用统计方法检验某些特性是否存在，例如通过混沌吸引子的分形维和李亚普诺夫指数的计算说明混沌的存在性，而对这些技术的有效性还存在一些争议。方差多变点技术则不同，不论系统内部的机制如何，突变常常表现为系统的某些特征变量的异常波动，利用方差的突变来发现系统行为的突变，在技术上简单而实用。

三、模型的描述和求解

考察时间序列$\{a_t\}$，t=1,2,L,T，$a_t\sim N(0,\sigma_t^2)$，并且

$$\begin{aligned}\sigma_t^2 &= \tau_0^2 \qquad 1\leqslant t\leqslant k_1\\ &= \tau_1^2 \qquad k_1\leqslant t\leqslant k_2\\ &\cdots\cdots\\ &= \tau_{N_T}^2 k_{N_T}\leqslant t\leqslant T\end{aligned} \tag{1}$$

这里 $1\leqslant k_1<k_2<\cdots\cdots<k_{N_T}<T$ 表示方差变点的下标。方差为 σ_j^2，$j=0,L,N_T$ 的观测数是 $d_j=k_{j+1}-k_j$。约定 $k_0=0$，$k_{N_{T+1}}=T$，尽管实际上 $k_0\leqslant 0$，$k_{N_{T+1}}\geqslant T$。我们的问题是如何通过 $a=(a_1,a_2,L,a_T)'$ 估计 $k=(k_1,k_2,\cdots\cdots,k_{N_T})'$，即利用收益率样本估计变点的个数和位置。由于我们把问题限定在方差变点上，而方差是样本分布的密度参数，所以很自然地会把方差变点分析看成是密度估计问题，然后利用极大似然方法估计变点位置。

$a=(a_1,a_2,L,a_T)'$ 的联合分布可以写成：$p(a|\sigma,k,N_T)\propto\prod_{t=1}^{T}\sigma_t^{-1}exp\{-\frac{1}{2\sigma_t^2}a_t^2\}$

或者 $p(a|\tau,k,N_T)\propto\prod_{j=0}^{N_T}\tau_j^{-d_j}exp\{-\sum_{t=k_j+1}^{k_{j+1}}a_t^2\}$ (2)

这里 $\sigma=(\sigma_1,\sigma_2,\cdots\cdots,\sigma_T)'$，$\tau=(\tau_0,\tau_1,\cdots\cdots\tau_{N_T})'$，而且 $k=(k_1,k_2,\cdots\cdots,k_{N_T})'$。

我们关心的参数是 k，即变点位置，而 τ 显得相对不重要。如果直接应用极大似然方法，存在一些困难，因为 σ 和 τ 我们都不知道，贝叶斯方法处理这种问题不存在原则上的困难。运用贝叶斯公式得到参数的后验分布 $p(k,N_T,\tau|a)$ 后，k,N_T 的边缘分布密度可以方便地通过对后验分布密度对 τ 积分而得到。要运用贝叶斯方法，通常先确定 N_T、τ 的先验分布。Inclan 假定变点个数 N_T 服从二项分布，τ 服从逆伽玛分布，但出现了超参数不稳健的问题。超参数不稳健源于我们对问题缺乏好的先验信息，在没有好的先验信息的情况下，经验贝叶斯方法是一个较好的选

择。于是，我们对 τ 的先验分布做如下改进：在 N_T 和 k 给定的条件下，对于每一个 τ_j，显然有：

$$\sum_{t=k_j+1}^{k_{j+1}} a_t^2 \Big/ \tau_j^2 \sim x^2(d_j) \tag{3}$$

我们可以根据(3)式导出 τ_j 的分布，并把它作为 τ_j 的先验分布。这样我们就得到了 τ 的先验分布 $p(\tau|k,N_T)$。这个先验分布实际上用到了样本的信息，所以我们用的是经验贝叶斯方法。结合(2)式得：

$p(a,\tau|k,N_T)=p(a|\tau,k,N_T)p(\tau|k,N_T)$，对 τ 积分得到：

$$\begin{aligned} p(a|k,N_T) &= \int p(a,\tau|k,N_T)p(\tau|k,N_T)d\tau \\ &= E_{\tau|k,N_T}[p(a|\tau,k,N_T)] \end{aligned}$$

令 $x_j=\sum_{t=k_j+1}^{k_{j+1}} a_t^1 \Big/ \tau_j^2$，$x=(x_0,x_1,x_2\Lambda x_{N_T})$，那么

$$\begin{aligned} p(a|k,N_T) &= E_x[p(a|\tau,k,N_T)] \\ &= (2\pi)^{-T/2}\prod_{j=0}^{N_T}\Big(\sum_{t=k_j+1}^{k_{j+1}} a_t^2\Big)^{-d_j/2}\int_0^{+\infty} x_j^{d_j/2}e^{-x_j/2}\frac{1}{2^{d_j/2}\Gamma(d_j/2)}e^{-x_j/2}x_j^{d_j/2-1}dx_j \\ &= (4\pi)^{-T/2}\prod_{j=0}^{N_T}\frac{\Gamma(d_j)}{\Gamma(d_j/2)}\Big(\sum_{t=k_j+1}^{k_{j+1}} a_t^2\Big)^{-d_j/2} \end{aligned} \tag{4}$$

要利用贝叶斯方法，还需要确定 N_T 的先验分布，由于对 N_T 我们实在没有什么好的先验信息，而且 N_T 的先验分布对问题求解又十分敏感，所以我们根据密度函数 $p(a|k,N_T)$ 转而用极大似然方法求解 N_T 和 k。利用极大似然方法求解 N_T 和 k，可以先给定 N_T，然后通过最大化 $p(a|k,N_T)$ 确定 $k'|N_T$ 作为变点数为 N_T 时变点位置的估计。

在这里我们并没有使用纯粹的贝叶斯方法，而是将贝叶斯方法和极大似然方法结合起来，先利用经验贝叶斯方法消去多余参数，然后再利用极大似然方法寻找变点。这样做有三个好处：其一，利用了贝叶斯方法处理多余参数的方便；其二，极大似然方法回避了变点先验分布选择的困难；其三，极大似然方法只需要在解空间中找出似然密度最大的解，所以可以利用启发式方法求解而无需穷举，使多变点快速求解成为可能。

在研究变点的过程中，我们发现一个很重要的现象。假设 $k'|N_T$ 和 $k'|N_T+1$ 分别为变点数 N_T 和 N_T+1 时的最优解。$k'|N_T+1$ 中有 N_T 个分量和 $k'|N_T$ 中的分量很接近，也就是变点数为 N_T+1 时得到的 N_T 个变点与变点数 N_T 得到的变点具有一致性。随着变点数的变化，我们把变点估计的这种一致性称为变点解的一致性。所以我们可以利用变点解的一致性构造更快速的算法。下面是求极大似然值的新算法：

(1)初始化 $N_T=1$，最大化 $p(a|k,N_T=1)$ 得到 $k'|N_T=1$。

(2)把 $k'|N_T$ 中的 N_T 个分量作为要求解的 $k'|N_T+1$ 中 N_T 个分量的估计。最大化 $p(a|k,N_T)$ 得到另一个分量的估计。

(3)分别固定 $k'|N_T+1$ 中的 N_T 个分量，通过最大化 $p(a|k,N_T)$ 修正另一个分量，直到每个分量都不会改变为止，这样就得到了变点数为 N_T+1 时变点位置的最终估计 $k'|N_T+1,N_T=N_T+1$。

(4)如果满足中止条件则结束，否则转到“(2)”。

其中，步骤(2)利用了变点解的一致性；步骤(3)则利用了最优解中的各分量必然是互为条件最优这一性质。步骤(3)在修正每个分量的时候，实际上并不需要在整个时序上搜索，因为初始估计其实已经很接近最优解了，所以可以在每个分量的某个邻域内搜索。本文是在相邻分量确定的区域中搜索。对于分量 k'_n，搜索区域为 (k'_{n-1},k'_{n+1})。

新算法比 Inclan 的算法快捷，可以很方便地求解长时序的多变点问题。它可以求出不同显

著水平的多变点。那么算法的中止条件是什么呢？首先中止条件取决于我们研究方差变点的出发点。对于投资者来说,方差变点可以为控制风险提供参考,所以中、低等显著水平的变点也不能放过。对于一般应用来说,我们需要的只是全局上最显著的少数变点。所以,我们主要讨论选择最显著的少数变点的中止条件。一般来说,当我们设定变点数 N_T 后计算出的变点就是全局最显著的 N_T 个变点。更为精细的方法是用新加入变点后对数似然密度的增量作为新变点显著水平的度量。加入新变点一般会增加解释能力,表现为对数似然密度的增加,而增加的显著水平则可以通过对数似然密度的增量来反映,较高的对数似然密度的增量反映了新加变点的解释能力。借助于对数似然密度增量,我们无需对变点数 N_T 设定什么先验分布,只要看到对数似然密度的增量稳定在一个较小的水平上,就可以中止计算,把前面的变点作为全局最显著的少数变点。

四、北京地区"非典"疫情突变分析

公共卫生突发事件"非典"给全人类带来了灾难,我国是当时非典肆虐的重灾区,北京则是全国疫情最严重的地区。虽然目前北京的疫情已经得到很好的控制,但早些时候国内外对当地控制疫情的有效性还是有所疑虑。事实上,台湾和多伦多的疫情就出现了反复。疫情的反复通常伴随着新增病例的异常波动,因此完全可以用方差多变点技术监控疫情的突变,并及时采取应对措施。由于"非典"是一种新型传染病,所以我们对它认识不够,影响了疫情数据的完整采集。下面我们仅对早期的新增确证病例疫情数据(从 4 月 21 日到 5 月 25 日)作一分析(见图 1)。

图 1　新增确诊病例

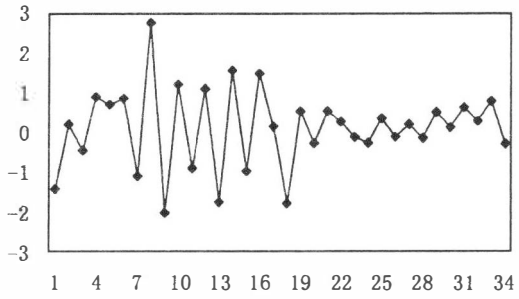

图 2　新增确诊病例差分

从新增病例图可以看出,新增病例是逐步下降的,并且存在明显的自相关现象。所以我们对新增病例作差分,并标准化得到新增病例的差分图。根据新增病例差分图,从直觉上可以看出,开始波动比较大,但后来波动比较小。表 1 是方差变点分析的计算结果。

表 1　北京当时"非典"新增病例的方差变点分析计算结果

变点数 N_T	变点位置 k	对数似然密度	对数似然密度增量
0		−48.085	
1	(18)	−36.701	11.384
2	(6,18)	−35.776	0.925
3	(1,6,18)	−35.312	0.464

由分析结果可见,变点数为 1 时的对数似然密度增量为 11.385,最为显著;后面两个变点的对数似然密度增量分别为 0.925 和 0.464,已经很小了,所以认为存在一个变点比较合适,该变点位置为 18,对应的日期是 5 月 9 日。该变点分成的两段数据样本方差分别为 1.827 和0.119,方差比为 15.353,说明 5 月 9 日以前新增病例变化的波动比较大,而 5 月 9 日以后的波动则小得多,缩小了近 15 倍。方差多变点分析说明疫情在 5 月 9 日发生了突变,新增病例变化的波动大幅减小,疫情发展态势迅速好转。疫情发生突变的主要原因是我国政府面对公共卫生突发事件"非典",积极应对,及时采取有效措施,并使疫情信息透明化,动员一切力量抗"非典",向全世界表明中

国是个负责任的大国。方差变点分析技术从定量角度说明了我国政府控制疫情的有效性。

五、结论

对于多变点问题，构造某种统计量来探测变点属于一种“粗”的方法，优点是计算量小，缺点是需要分割时序样本，因而找到的变点可能不是全局意义上的变点，而且无法灵活地寻找各种显著水平的变点。贝叶斯方法属于一种“细”的方法，优点是可以寻找各种显著水平的变点，缺点是一般需要组合数阶的运算量，而结合变点解的一致性的启发式算法则可以显著提高计算效率，做到既快又准地探测多变点。

本文将方差多变点分析技术运用于探测公共卫生突发事件 SARS 的突变，从定量角度说明我国政府控制疫情的有效性，该技术实际上可以推广到整个公共安全领域对突变的探测。在本例中，我们只对 SARS 的主要特征变量新增病例进行探测，根据公共安全事件的特点，也可以针对多个特征变量进行监控，互相印证，互为补充。方差多变点分析技术目前还属于事后探测技术，要使它能够实时地监控突发事件，可以考虑引入序贯分析方法，这将在我们以后的研究中逐步完善。

参考文献：

[1]Inclan C. Detection of multiple changes of variance using posterior odds[J]. Journal of Business & Economic Statistics, July 1993, 11, 289～300.

[2]Broemeling L. Econometrics and structural change[M]. Marcel Dekker, 1987, INC.

[3]Carla Inclan and George C. Tiao. Use of cumulative sum of squares for retrosepctive detection of changes of variances [J]. Journal of American Statistcal Association, September, 1994, Vol, 89, No, 427.

[4]项静恬，史久恩．非线性系统中数据处理的统计方法[M]. 北京：科学出版社，2000.

[5]吴喜之．现代贝叶斯统计学[M]. 北京：中国统计出版社，2000.

[6]许国志，顾基发，车宏安．系统科学[M]. 上海：上海科技教育出版社，2000.

[7]陈士华，陆君安．混沌动力学初步[M]. 武汉：武汉水利电力大学出版社，1998.

Detection of Analysis of Emergencies in Public Safety

——A Study on SASR by Using the Technology of Multiple Change of Variance Analysis

LIAO Yuan-shen, LIU Hong

(*School of Economics, Shanghai University of Finance and Economics, Shanghai* 200433, *China*)

Abstract: Emergencies seriously threaten the public safety field. It is almost impossible to forecast emergencies, but absolutely possible to detect emergencies in time. Because emergencies usually give rise to changes of variance of some characteristic variable, so multiple change of variance analysis can be used to acutely detect emergencies. Based on the Inclan research, the paper improves the prior distribution of variance parameter and the solution of maximum likelihood estimate for high efficiency in calculation. A fast algorithm is devised according to the continuousness of change point solutions. The example of SASR in public sanitary is used to show the high efficiency and accuracy of new algorithm. It's a new practical technology of detection of emergencies.

Key words: change point of variance; emergencies in public safety; sudden change; Bayesian statistics