

# 金融时序中异常数据挖掘算法设计及实证分析

杨 虎, 李 强

(重庆大学数理学院, 重庆 400044)

**摘 要:** 金融市场中的数据由于其内在联系, 通常表现为相互关联的时间序列。本文主要讨论如何将金融市场中时间序列模型简化为相应的线性模型, 继而用传统的线性模型方法去检验异常值的存在, 并且判断该异常值是加性异常值还是创新异常值。创新异常值的挖掘对于金融风险的研究不仅具有理论上的意义, 而且具有很强的现实意义。最后进行了算法的实证分析, 结果表明本文的两种方法在金融市场的研究中是可行的并且行之有效。

**关键词:** 金融时间序列; 创新异常; 信息准则

中图分类号: F830

文献标识码: A

## 1 引言

创新异常值 (innovation outliers, 简记为  $IO$ )<sup>[1-3]</sup> 在通常的意义下只是一个孤立的数据, 但在时间序列中, 由于内在的相关结构, 它的存在会波及到后面的数据, 从而使这些数据也表现出一定的异常, 容易出现成串的异常数据, 并从本质上改变未来的数据趋势。正是创新异常的存在会改变一段时间乃至今后的整体趋势, 故而成为金融统计分析 with 金融数据挖掘需要着重研究的问题<sup>[3-6]</sup>。本文就如何用传统的线性模型方法<sup>[7-8]</sup> 对时间序列中的创新异常进行行之有效的挖掘和识别进行了探讨。在金融统计分析中, 这种异常值往往携带重要的投资信息, 如何快速、有效的从这些时间序列中挖掘出这些重要的信息, 是实际中无法回避的问题。

## 2 时间序列中的异常值及其分类

异常值有多种定义方式, 在统计诊断中讨论较多的是残差相对非常突出的数据<sup>[9]</sup>, 这样的数据在统计推断中会引起大的失误, 从而影响到基于此的模型结构和预测效果。在理论和实际应用当中, 都存在如何探测异常值以及对检出的异常值如何处理的问题, 虽然异常值的概念很容易让人明白, 但要给它下一个精确的定义相当困难, 目前国际上有两种较为流行的看法: 一是把异常值看成与数据主体明

显不协调的单个或集团数据, 异常值可以解释为所假定分布中的极端值, 即落在分布的单侧或双侧  $\alpha$  分位点以外的数据; 二是把异常值视为杂质点, 它与数据的主体不是来源于同一分布<sup>[2]</sup>。

时间序列中的异常值的表现是多种多样的, 但是, 通过数据变换, 我们可以将其归纳为以下三类: 加性异常值  $AO$ , 创新异常值  $IO_1$ , 创新异常值  $IO_2$ 。(见图 1)。

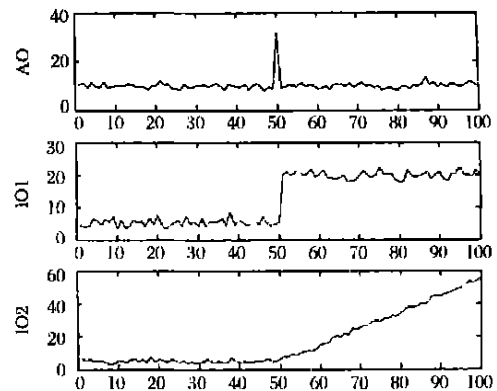


图 1 异常点分类举例

## 3 基于线性模型下异常值的检测与类型判别

### 3.1 时间序列的线性平滑和模型选择

因为我们要采用线性模型下的数据诊断的方法来检出和判别异常值, 所以我们首先应对原始数据作变换, 以弱化原时间序列的相关性并使之满足经典线性回归的各项假设条件<sup>[8]</sup>。记股票的收盘价为序列:  $P_1, P_2, \dots, P_t, P_{t+1}, \dots$  我们对数据作变换:

$$y_t = -R_t = \frac{P_{t+1} - P_t}{P_t}, t = 1, 2, \dots$$

收稿日期: 2003-05-27; 修订日期: 2004-03-31

作者简介: 杨虎 (1963-), 男 (汉族), 四川人, 重庆大学数理学院院长, 教授, 博士, 研究方向: 线性模型、金融统计、统计诊断与数据挖掘。

则  $y_t$  相互独立<sup>[10]</sup>, 并且当没有异常值时, 假定  $y_t$  具有相同的分布  $N(0, \sigma^2)$  是可以接受的(由于样本容量较大, 且对同一个品种连续的一段周期。当然在实际中可以视情况进行独立正态性检验),  $t = 1, 2, \dots$ , 其中  $\sigma$  未知。这样, 我们对序列  $y_1, y_2, \dots$  可以建立如下模型:

$$Y = \beta_0 + \beta_1 t + e \quad e \sim N(0, \sigma^2) \quad (3.1)$$

### 3.2 异常值的检测

在回归诊断的框架下, 可以从两个方面进行探讨: (1) 残差分析(Residual Analysis) 这主要是从模型假设的合理性方面进行研究, 考虑的统计量多为残差, 其中包括: 普通残差, 预测残差, 学生化残差, 递归残差和不相关残差等, 这样做是因为从残差中我们可以看出拟合的效果, 而异常数据就是那些拟合效果较差的点; (2) 影响分析(Influence analysis) 这主要是探索对统计推断(如估计或预测)有较大影响的试验数据, 我们期望每组数据对统计推断都有一定的影响, 但这种影响又不要过大, 如果某组数据的影响过大, 那么包含这组数据的经验回归方程与不包含这组数据的经验回归方程差异很大, 于是经验回归方程关于数据就不具有“稳定”性, 从而进行预测的基础就不复存在了<sup>[11]</sup>。

基于上面的回归诊断思想, 可以建立如下的检验和判别异常数据的方法:

#### 方法一: 从影响分析入手

传统的度量试验数据点对统计推断影响大小的量, 如 Cook 距离、AP 统计量<sup>[2]</sup>, 是以全部的数据点做出系数  $\beta$  的最小二乘估计  $\hat{\beta}$ , 并以此为基准, 考虑每次删除一个或多个数据点后系数的改变度量  $\frac{(\hat{\beta}(I) - \hat{\beta})' M(\hat{\beta}(I) - \hat{\beta})}{c}$ , 并以此作为判别异常数据的标准。但我们这里不能以此标准来探测异常数据, 因为股票市场的数据海量, 如果以全体数据点的回归系数为基准构造度量标准, 那么个别点的异常将被累积误差所“淹没”; 此外, 证券市场的异常点挖掘有极强的适时性要求, 通常需要在异常点出现后较短的时间内作出加性异常还是创新异常的判断, 因而原有的方法难以尽快对异常的属性作出判断。

传统的残差分析是为了检验试验数据是否满足我们对模型进行统计推断所进行的各项假设, 若不满足, 则对数据进行变换, 使之近似满足这些假设。这里则不然, 首先需要对模型(3.1)作统计推断, 如果统计量发生异常, 我们就认为该点为异常数据, 并进而对其跟踪以便尽快对类型加以判断。具体对模

型(3.1)而言, 在无异常值的回归模型中我们通常假设  $e_n = (y_n - \hat{y}_n) \sim N(0, \sigma^2)$  从而  $\hat{y}_n \sim N(X\hat{\beta}, \sigma^2 X L^{-1} X^T)$  对一维情形, 有  $X L^{-1} X^T = \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}}$  其中  $L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ 。

又因为  $y_n$  与  $y_{n-1}, \dots, y_1$  相互独立, 因此  $y_n - \hat{y}_n \sim N(0, \sigma^2(1 + X L^{-1} X^T))$ 。令  $Q_e = \sum_{i=1}^{n-1} (y_i - \hat{y}_i)^2$ , 由线性回归基本性质可知  $y_n - \hat{y}_n$  与  $Q_e$  独立, 且

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n - k - 1) \quad (3.2)$$

因此

$$\frac{\frac{y - \hat{y}}{\sigma \sqrt{1 + X L^{-1} X^T}}}{\sqrt{\frac{Q_e}{\sigma^2(n - k - 1)}}} \sim t(n - k - 1) \quad (3.3)$$

故  $y$  的  $1 - \alpha$  置信区间为  $(\hat{y}_1, \hat{y}_2)$ , 其中

$$\begin{cases} \hat{y}_1 = \hat{y} - \hat{\sigma}_{e|1-\alpha/2(n-k-1)} \sqrt{1 + X L^{-1} X^T} \\ \hat{y}_2 = \hat{y} + \hat{\sigma}_{e|1-\alpha/2(n-k-1)} \sqrt{1 + X L^{-1} X^T} \end{cases} \quad (3.4)$$

我们每次取  $n$  个数据点  $y_i, y_{i+1}, \dots, y_{i+n-1}$  进行回归, 其中  $i = 1, 2, \dots$ , 并由此预测出下一个点  $\hat{y}_{i+n}$  的  $1 - \alpha$  置信区间, 如果该点的真实值  $y_{i+n}$  落在该预测区间内, 则我们认为该点对下一次的回归直线不会起改变趋势的作用, 故而不是异常数据; 如果该点的真实值  $y_{i+n}$  落在该预测区间以上(对落在预测区间以下的情况可作类似的讨论), 那么我们有理由相信该点会对下一次的回归直线产生较大的影响, 故可认为该点为异常数据, 为了进一步判别其类型, 取点  $y_{i+1}, y_{i+2}, \dots, y_{i+n}$  做回归, 并由此预测出下一个点  $\hat{y}_{i+n+1}$  的  $1 - \alpha$  置信区间, 如果该点的真实值  $y_{i+n+1}$  落在该预测区间的上方, 则可判定点  $y_{i+n}$  为创新异常值, 如果该点的真实值  $y_{i+n+1}$  落在该预测区间以内或该预测区间的下方, 则点  $y_{i+n}$  的类型不能判定, 需进一步判断, 判断的准则为: 若下一点  $\hat{y}_{i+n+2}$  的真实值  $y_{i+n+2}$  落在该点的  $1 - \alpha$  置信区间以上, 则强烈支持点  $y_{i+n}$  为创新异常值, 否则  $y_{i+n}$  为加性异常值。

#### 方法二: 基于传统的残差分析方法

Akaike 于 1973 年提出的“信息统计量”在检测异常值方面取得了很好的进展<sup>[12]</sup>。在此基础之上, 可用改进的 SIC(Schwarz(or Bayesian) Information Criterion)方法<sup>[13]</sup>来检测和识别异常数据点。具体

如下:

对模型(3.1), 我们每次取  $n$  个数据点  $y_i, y_{i+1}, \dots, y_{i+n-1}, i = 1, 2, \dots$ , 进行回归, 其中  $e_i$  独立同服从正态分布  $N(0, \sigma^2)$ ,  $\sigma^2$  未知, 则  $y_i$  服从正态分布  $N(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, 2, \dots$ , 若取出的点当中存在异常数据, 则回归系数在该异常数据点以后会发生显著的变化。这就是说我们可以通过检验原假设:

$$H_0: \mu_{y_i} = \beta_0 + \beta_1 x_i, i = 1, 2, \dots, n$$

和对立假设:

$$H_1: \mu_{y_i} = \beta_0^1 + \beta_1^1 x_i, i = 1, \dots, k,$$

$$\mu_{y_i} = \beta_0^* + \beta_1^* x_i, i = k+1, \dots, n$$

其中  $k$  是异常数据点的位置, 在  $H_0$  似然函数为:

$$L_0(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp[-\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 / 2\sigma^2]$$

$\beta_0, \beta_1, \sigma^2$  的极大似然估计分别为:

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_x}, b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}, \hat{\sigma}^2 =$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2,$$

$$\text{其中: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, S_x = \sum_{i=1}^n (x_i - \bar{x})^2, S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

$$\text{此时, } SIC(n) = -2\log L_0(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) + 3\log n = n\log(2\pi) + n\log(\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2) + n + 3\log n - n\log n$$

在  $H_1$  之下,

$$L_1(\beta_0^1, \beta_1^1, \beta_0^*, \beta_1^*, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}$$

$$e^{-\sum_{i=1}^k (y_i - \beta_0^1 - \beta_1^1 x_i) / 2\sigma^2} e^{-\sum_{i=k+1}^n (y_i - \beta_0^* - \beta_1^* x_i) / 2\sigma^2}$$

$$\text{其中, } b_1^1 = \hat{\beta}_1^1 = S_{xy}^{(k)} / S_x^{(k)}, b_0^1 = \hat{\beta}_0^1 = \bar{y}_k - b_1^1 \bar{x}_k, b_1^* = \hat{\beta}_1^* = S_{xy}^{(n-k)} / S_x^{(n-k)}, b_0^* = \hat{\beta}_0^* = \bar{y}_{n-k} = b_1^* \bar{x}_{n-k},$$

$$\hat{\sigma}^2 = \frac{1}{n} [\sum_{i=1}^k (y_i - b_0^1 - b_1^1 x_i)^2 + \sum_{i=k+1}^n (y_i - b_0^* - b_1^* x_i)^2],$$

$$\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i, \bar{y}_k = \frac{1}{k} \sum_{i=1}^k y_i, \bar{x}_{n-k} =$$

$$\frac{1}{n-k} \sum_{i=k+1}^n x_i, \bar{y}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n y_i,$$

$$S_x^{(k)} = \sum_{i=1}^k (x_i - \bar{x}_k)^2, S_{xy}^{(k)} = \sum_{i=1}^k (x_i - \bar{x}_k)(y_i - \bar{y}_k),$$

$$S_x^{(n-k)} = \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2, S_{xy}^{(n-k)} = \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})(y_i - \bar{y}_{n-k}),$$

$$\text{此时, } SIC(k) = -2\log L_1(\hat{\beta}_0^1, \hat{\beta}_1^1, \hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\sigma}^2) + 5\log n = n\log 2\pi + n[\sum_{i=1}^k (y_i - b_0^1 - b_1^1 x_i)^2 + \sum_{i=k+1}^n (y_i - b_0^* - b_1^* x_i)^2] + 6\log n - n\log n$$

我们判别异常数据点的方法为: 如果对任意的  $k$  都有  $SIC(n) < SIC(k)$ , 则认为当前所选出的  $n$  个数据点当中没有异常数据点, 下一步取  $y_{i+1}, \dots, y_{i+n}, i = 1, 2, \dots$ , 进行判别, 准则同上。如果存在  $\hat{k}$  使得  $SIC(\hat{k}) = \min_{1 \leq k \leq n-1} SIC(k) < SIC(n)$ 。则认为第  $\hat{k}$  点为异常数据点。对以上的算法, 我们采用 Matlab 软件进行模拟, 检测和判别出的加性异常值(AO)和创新异常值(IO)的不同性态如下:

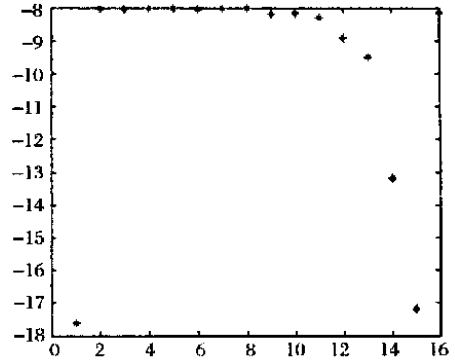


图2 加性异常值(AO)

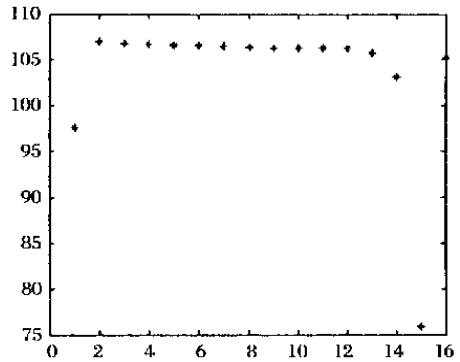


图3 创新异常值(IO)

图2和图3中的第一个点均为  $SIC(n)$  的值, 后面的各点为相应的  $SIC(k)$ , 从图中我们可以清楚的看到, 对加性异常值(AO), 此算法的选点结果认为不存在异常数据, 而对创新异常值(IO), 我们的算法清楚的描绘出了其存在的位置和异常属性。

注:以上两种方法当中回归步长  $n$  的选择可以根据具体的情况而定,比如考虑长线和短线就需要选择不同的  $n_0$ 。

#### 4 实证分析

以上我们所做的工作是从理论上探讨如何将时间序列中的创新异常值用线性模型理论的方法将其检测和识别出来。这样做的目的和意义在于能够避开时间序列研究的复杂性和算法上的设计问题,用简单而清楚的方法将结果表示出来,取得了较好的效果。下面,我们将结合证券市场的实例对我们所做出的算法进行演示。

中国证券市场是非常年轻的新兴市场,并且由于初设时遗留下来的全流通问题使得中国市场完全不同于周边证券市场<sup>[14]</sup>。尽管这些宏观面上的问题加上特有的小型市场对政策面的过敏感问题还会相对长期的存在,但从技术分析出发,很多问题是类似的,不同的仅仅是异常更加频繁一些,且会增加一些政策性的分类。

证券交易的目的在于以最小的风险取得最大的利润。我们所做的工作是通过分析证券品种的历史数据,从中寻找出符合一定条件的时间点,在这些点之后的预设天数内,要求最大收盘价能够达到目标利润。我们的选点算法选出了一些时间点,这些点正是我们在前面定义的创新异常值,若这些点满足前述的条件,则认为成功地预测了该点,否则,认为该点选取失败,然后我们还计算了该选点算法的成

功率。

下面我们以短线分析为例(这里选取的步长为 10,实际中可以通过另外的优选算法确定  $n$ )。测试数据截至 2003 年 3 月,选择了沪深证券市场中的 1159 支股票,因为在证券市场中,股票收盘价的异常定义会随时间长短选择的不同而发生改变,所以不能用统一的方法做时间序列中创新异常值的检测,故而采取对已有的可以判断类型的数据序列进行检验的方法来检验本文方法的效果,并以此说明确实可用线性模型方法代替时间序列里的异常值检测方法,其中有 1051 支股票发出了 15236 个指示,其中成功预测的点数为 12579 个,平均成功率为 82.82%,成功率达到 50% 的股票有 81.7%。以某只股票为例,我们的选点算法选出的点如下。其中图 4 的中部标记黑色小旗的点即为选出的符合预设条件的点,从图中我们可以看到,在前期该股一直在低位徘徊,在中部,连续三支阳线站在短期均线上方,有效的突破了前期的徘徊趋势,并且五日均线穿二十日均线,形成了金叉,预示着股价会有显著的上扬;而本文介绍的选点算法所选出的点恰好位于该处,从而及时而有效地选出了介入点,这表明:我们所做的线性模型下的创新异常值的判断方法能够有效的排除偶然脉冲点(AO)而把创新异常点(IO)显著的描述出来。我们通过上面所述的传统的线性模型方法实现了对时间序列中创新异常值的检测和识别。

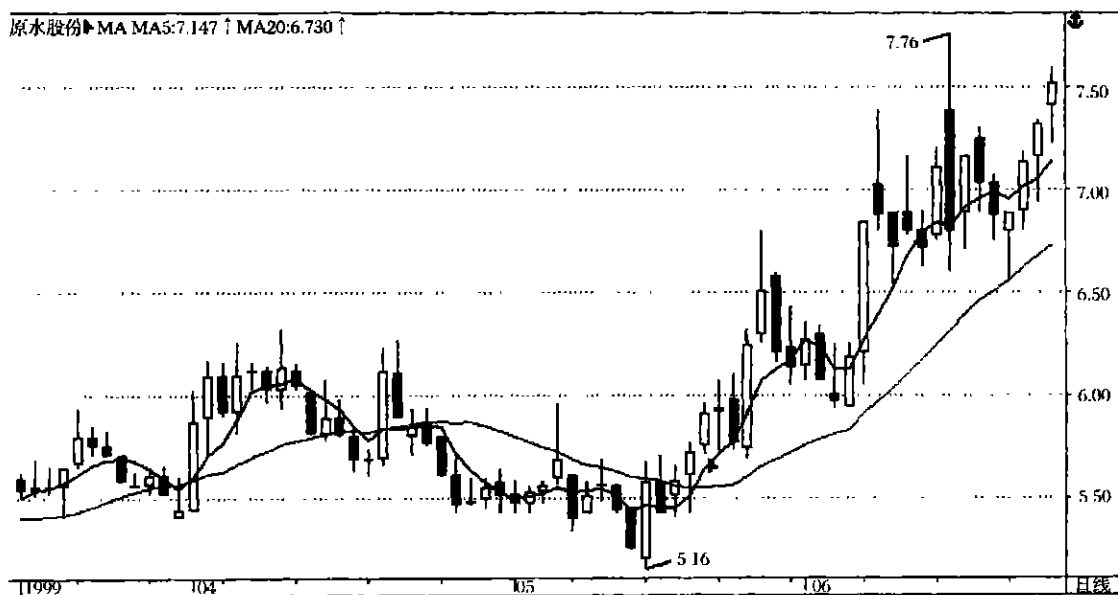


图 4 实际异常挖掘示意图

## 5 结论

金融时序中异常的表现是多种多样的, 这给实际的处理和分析带来很大的不便, 针对这些问题, 时间序列分析中发展了很多实用的探测异常的方法, 尤其是创新异常的刻画在金融时序里有着良好的应用预期, 但这样的方法往往比较复杂而不易为广大投资者所了解, 适时性要求更是难以满足, 面对海量和瞬息万变的金融市场, 必须寻求简单的替代工具。本文比较成功的进行了这方面的尝试, 采用传统的线性模型方法和统计诊断技术设计了同样可以对时间序列中创新异常值进行检测和识别的两种实用算法。最后结合沪深证券市场中的真实数据, 对算法进行了实证分析, 通过与原时间序列中创新异常值的比较, 表明了该方法基本实现了上述目的。

用简单而明确的线性模型方法对时间序列中创新异常值的有效检测和识别, 可以让普通投资者避开时间序列研究的复杂性和实时挖掘算法的设计问题, 可以大大地简化原有的金融时间序列研究方法, 尤其适合进行异常投资机会的盘中挖掘, 这对于金融市场和金融时间序列的研究与实践均具有指导意义。

### 参考文献:

- [1] Chen, C., Liu, L. M., Joint Estimation of Model Parameters and Outlier Effects in Time Series[J]. JASA, 1993, 88: 284– 297.
- [2] 韦博成, 等. 统计诊断引论[M]. 东南大学出版社, 1991.

- [3] Liu, L. M., Bhattacharyya, S., Sclove, S. L., Chen, R., Lattyak, W. J., Data mining on time series: an illustration using fast-food restaurant franchise data[J]. Computational Statistics & Data Analysis 2001, 37: 455– 476.
- [4] John, G. H., Miller, P., Building long/short portfolios using rule induction[M]. In: Computational Intelligence for Financial Engineering, Piscataway NJ: IEEE Press, 1996.
- [5] John, G. H., Stock selection using rule Induction[J]. IEEE Expert, 1996, 52– 58.
- [6] 张尧庭. 金融市场的统计分析[M]. 广西师范大学出版社, 1998.
- [7] Rao, C. R., Linear Models– Least Squares and Alternatives [M]. Springer– Verlag, 1995.
- [8] 王松桂. 线性模型的理论及其应用[M]. 安徽教育出版社, 1987.
- [9] Beckman, R. J., Cook, R. D., Outlier ...s[J]. Technometrics, 1983, 25: 119– 149.
- [10] Hsu, D. A., Tests for Variance Shifts at an Unknown Time Point[J]. Applied Statistics, 1977, 26: 179– 184.
- [11] 王松桂. 回归诊断发展综述[J]. 应用概率统计, 1988, 4 (3): 310– 319.
- [12] Akaike, H., Information Theory and an Extension of the Maximum Likelihood Principle[J]. In Proceedings of the 2nd International Symposium on Information, edited by B. N. Petrov and F. Czaki. Budapest: Akademiai Kiado, 1973: 267– 281.
- [13] Schwarz, G., Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6: 471– 464.
- [14] 金建栋, 许树信, 肖灼基. 中国证券市场[M]. 中国金融出版社, 1993.

## Linear Mining Algorithms Design for Outliers in Financial Time Series and its Authentic Proofs

YANG Hu, LI Qiang

( College of Science, Chongqing University, Chongqing 400044, China)

**Abstract:** Owing to the internal relations, the data in financial market usually manifest as the interrelated time series. This paper mainly discusses how to simplify time series models in financial market into relevant linear models and how to examine the existence of outliers and differentiate innovation outliers from additive outliers with traditional linear models. The mining of innovation outliers has not only the theoretical significance but also a great practical significance in the research on financial risk. Besides, the two algorithms proposed in this paper are analyzed with authentic proofs; in this way, the two methods in the study of financial market are proved feasible and effective.

**Key words:** financial time series; innovation outliers; informational criterion