第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)10—0240—03 文献标识码：A 中图分类号：TP391

基于AR模型的股指结构化特征研究

许长龙，蔡世民，周佩玲
(中国科学技术大学电子科学与技术系，合肥 230026)

摘 要：应用AR模型分析股价趋势变化的二元符号序列，得到该二元符号序列具有线性模型的特征，通过有限阶的条件概率分析该二元符号序列，得到该股指序列具有明显的结构化特征，从不同的角度揭示金融市场的不完全有效性。通过对比新兴和成熟股指的结构化特征，发现新兴的金融市场确实存在明显的非理性投资行为。

关键词：结构化特征；AR模型；条件概率；二元符号序列

Research on Structural Characteristics in Stock Index Based on AR Model

XU Chang-long, CAI Shi-min, ZHOU Pei-ling
(Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230026)

【Abstract】 The changing trends of binary sequences are analysed with AR model(the autoregressive model), to obtain the linear model characteristics. And through a limited order of conditional probability analysis of the binary sequences, the stock index has a remarkable structural feature. This paper is from another perspective to reveal that the financial markets are not completely effective. By comparing the structural characteristics of the stock index of the developing and developed markets, it finds that the non-rational investment behavior is in the immature emergent financial markets.

【Key words】 structural characteristics; AR model; conditional probabilities; binarized symbol series

1 概述

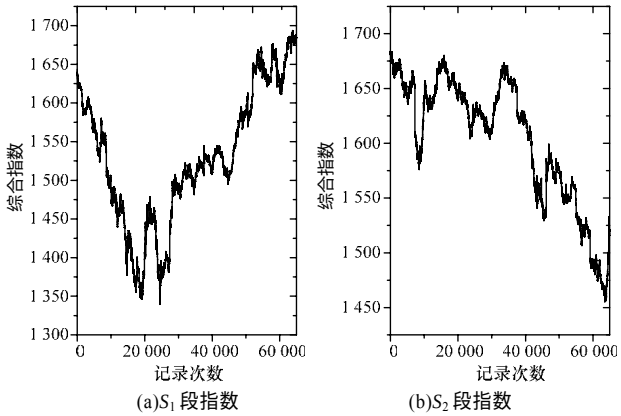
市场有效性假说(Efficient Market Hypothesis, EMH)成为现代经济学家的理论基石，它认为经济市场具有不可预测性，即过去的信息对预测未来经济的状况是无效的，经济市场中价格的变化属于随机游走行为[1]。但是近年来实证研究表明，金融市场中确实存在一些可以预测的现象[2]。这些现象的发现，说明真实的金融市场跟传统经济理论假说是相违背的，真实的金融市场中的每个参与者并不都是理性地、清楚地看到市场中的每一个利益点。张翼成教授通过对熵的分析提出边界有效市场(Marginally Efficient Market, MEM)的概念[3]。Sazuka 等人通过对美元兑换日元(USD/JPY)外汇率的及时交易数据(tick-by-tick data)进行条件概率的实证研究，得出外汇市场中存在结构化的特征，进一步从实证的角度验证了外汇市场并不完全有效[4]。股指结构化特征是指不同时期的股指序列在只考虑指数趋势变化的情况下，产生的二元符号序列进行多阶的条件概率分析，得出具有明显一致性的特征。本文从反映金融市场整体行为的综合指数为研究对象，分析得到股指序列结构化特征的普适性，另外对比新兴的中国金融市场和美国成熟的金融市场的结构化特征，反映出新兴的中国金融市场投资行为的特点。

2 基于AR模型的定阶

2.1 研究数据介绍

由于单一股票受到很多人为因素的影响，很难反映一个市场的总体行为，因此在本文中采用了反映金融市场整体特征的综合指数为研究对象，包括中国经济趋于缓和增长的2002年上证综合指数；1994年经济低迷至1997年东南亚经济危机之前经济泡沫的恒生指数；代表全球经济走势，成为世界经济晴雨表的纳斯达克(NASDAQ)综合指数、道琼斯(DJIA)综合指数、标准普尔 500(S&P500)综合指数，下面对各数据进行分别介绍：

上证综合指数总共有 N=324 676 个数据点，采样周期为10 s。为了分析对于不同时期指数的结构化特征，将该数据点均分为5段(S_1~S_5)。图1(a)和图1(b)分别表示 S_1 和 S_2 的综合指数走势图，可以看出它们的走势具有明显的不同，S_1 具有急跌抽低缓慢反弹，而 S_2 具有缓跌抽低急涨的走势。



(a)S_1 段指数 (b)S_2 段指数

图1 2002年上证综合指数

基金项目：国家自然科学基金资助项目(70571075)
作者简介：许长龙(1981－)，男，硕士研究生，主研方向：时间序列分析；蔡世民，博士研究生；周佩玲，教授
收稿日期：2008-11-10 **E-mail：**chlxu@mail.ustc.edu.cn

香港恒生指数一共采样的数据点 N=190 821，采样时间间隔为 1 min。将这段数据按年份区分为 4 段：1994 年数据点 N_1=53 699；1995 年数据点 N_2=56 105；1996 年数据点 N_3=56 830；1997 年上半年数据点 N_4=24 442。

纳斯达克综合指数、道琼斯综合指数、标准普尔 500 综合指数总共的采样 N=196 776 数据点，采样时间间隔 1 min。将这段数据按照年份分为 3 段：2005 年数据点 N_1=71 555；2006 年数据点 N_2=71 485；2007 年前 3 个季度数据点 N_3=53 736。

2.2 二元符号序列

假定 $\{X_t\}$ 是股指序列，定义股指收益 $\{Y_t\}$ ：

$$Y_t = X_{t+1} - X_t \tag{1}$$

按照如下的方法构造二元符号序列 $\{Y(t)\}$ ：

$$Y(t) = \begin{cases} + & (Y_t > 0) \\ - & (Y_t < 0) \end{cases} \tag{2}$$

为了完全表现股指波动的行为，在这里忽视了 Y_t=0(对所有的数据一共忽略了 8.12%的数据量)这种情况。通过研究 $\{Y(t)\}$ 中出现 + 或者 – 的条件概率 $P(+|x)$， $P(-|x)$，反映市场中的投资者对当前的选择是沿着市场的趋势进行投资还是背离市场的趋势进行投资。由于当前状态出现 + 和 – 的概率是互补的，因此在本文中只对当前状态为 + 的予以考虑。$P(+|x)$ 表示在先前 n 个状态后，下一个状态为 + 的概率。

2.3 模型定阶

$\{X_t\}$ 的自协方差函数(ACVF)为

$$\gamma(k) = Cov(X_{t+k}, X_t),\ k = 0, \pm1, \pm2, \cdots \tag{3}$$

$\{X_t\}$ 的自相关函数(Auto Correlation Function, ACF)为

$$\rho(k) = \gamma(k)/\gamma(0) = Corr(X_{t+k}, X_t),\ k = 0, \pm1, \pm2, \cdots \tag{4}$$

其中，$\rho(k)$ 度量了 X_t 和 X_{t-k} 间的相关，而不考虑它们与其中间变量 $X_{t-1}, X_{t-2}, \cdots, X_{t-k+1}$ 的关系。在拟合 AR(Auto Regression)模型中，定阶依赖于给定中间变量 X_{t-k} 与 X_t 的条件相关。只有 X_{t-k} 对 X_t 的拟合做出新的、不能被 $X_{t-1}, X_{t-2}, \cdots, X_{t-k+1}$ 所代替的作用时，才将其引入模型中。偏自相关系数(Partial Auto Correlation Function, PACF)被用来度量 X_t 和 X_{t-k} 与 $X_{t-1}, X_{t-2}, \cdots, X_{t-k+1}$ 之间的关系，其定义为

$$\pi(1) = Corr(X_1, X_2) = \rho(1) \tag{5}$$

$$\pi(k) = Corr(R_{1|2,\cdots,k}, R_{k+1|2,\cdots,k}),\ \text{对} k \geq 2 \tag{6}$$

其中，$R_{j|2,\,j|3,\cdots,k}$ 是由 X_j 关于 $X_2, X_3, \cdots, X_j$ 线性回归所得残差，即

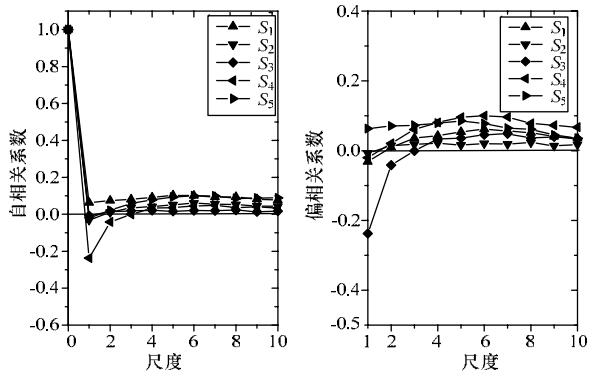
$$R_{j|2,j|3,\cdots,k} = X_j - (\alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k) \tag{7}$$

且

$$(\alpha_2, \alpha_3, \cdots, \alpha_k) = \arg\min_{\beta_2,\beta_3,\cdots,\beta_k} E\left\{X_j - (\beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)\right\}^2 \tag{8}$$

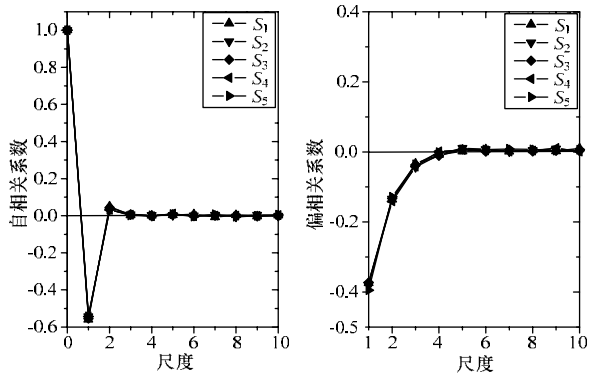
对于因果的 AR(p) 模型中，有 $\pi(k) = 0$，对一切 $k > p$，从而确定 p 为该 AR 模型的阶数[5]。

通过对比图 1 中 2002 年上证综合指数的 2 段数据，并不能从中发现一些相似性或者关联性的性质，同时图 2 显示的是股指收益的自相关和偏自相关系数，可发现股指收益不能通过简单的、同一的 AR 模型进行拟合。然而对上证 2002 年综合指数 5 段数据进行构造后得到二元符号序列，却得到自相关系数与偏自相关系数几乎重合，如图 3 所示。通过对股指收益二元符号序列的自相关系数，偏自相关系数跟 AR 模型关系的分析，得出该二元符号序列可以通过四阶 AR 模型很好地拟合[6]。



(a)自相关系数　　(b)偏自相关系数

图 2　上证 2002 年综合指数收益



(a)自相关系数　　(b)偏自相关系数

图 3　上证 2002 年综合指数二元符号序列

3　股指的结构化特征实证及分析

由于二元符号序列可以通过四阶 AR 模型很好地拟合，四阶以上的条件概率可以由四阶以内的条件概率推导得到，本文只对四阶以内的条件概率进行分析。表 1 表示上证指数中 2002 年 5 段的条件概率分布，容易发现对于不同的时期，相同的情况下得到非常相似的条件概率(在置信区间：*Error bar*/*S*=95%以上，它们都具有很好的相似形)，这种相似的特征不仅存在于上证 2002 年的综合指数，同时也存在本文研究的所有指数中，为了方便于观察，将表 1 的结果通过图 4 表示出来。图 5、图 6 分别表示香港恒生 1994 年~1997 年综合指数数据和纳斯达克综合指数、道琼斯综合指数、标准普尔 500 综合指数 2005 年~2007 年数据的条件概率分布图，称这种相似性为结构化特征。

表 1　上证 2002 年综合股指二元符号序列的条件概率对比

Divided data	*S*1	*S*2	*S*3	*S*4	*S*5	*S*	*Error bar*
Number of point	64 935	64 935	64 935	64 935	64 936	32 4676	
P(+)	0.48	0.48	0.49	0.48	0.48	0.48	0.00
P(+|+)	0.47	0.49	0.48	0.47	0.49	0.48	± 0.01
P(+|++)	0.51	0.53	0.53	0.46	0.53	0.51	± 0.02
P(+|-+)	0.53	0.54	0.54	0.52	0.53	0.53	0.00
P(+|+++)	0.56	0.59	0.59	0.50	0.59	0.57	± 0.02
P(+|-++)	0.58	0.61	0.60	0.55	0.59	0.59	± 0.01
P(+|+-+)	0.48	0.51	0.48	0.45	0.49	0.48	± 0.01
P(+|++-)	0.45	0.48	0.46	0.43	0.47	0.46	± 0.01
P(+|++++)	0.61	0.63	0.65	0.55	0.64	0.61	± 0.02
P(+|-+++)	0.64	0.68	0.66	0.58	0.65	0.64	± 0.02
P(+|+-++)	0.54	0.57	0.54	0.49	0.56	0.54	± 0.02
P(+|++-+)	0.50	0.53	0.51	0.45	0.52	0.50	± 0.02
P(+|+++-)	0.50	0.52	0.51	0.44	0.52	0.50	± 0.02
P(+|--++)	0.57	0.59	0.58	0.55	0.56	0.57	0.00
P(+|-+-+)	0.55	0.55	0.56	0.53	0.54	0.54	0.00
P(+|-++-)	0.54	0.54	0.55	0.53	0.54	0.54	0.00

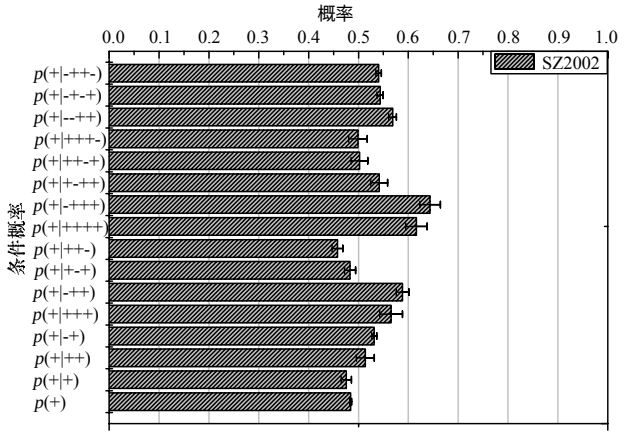


图 4　2002 年上证指数二元符号序列的条件概率分布

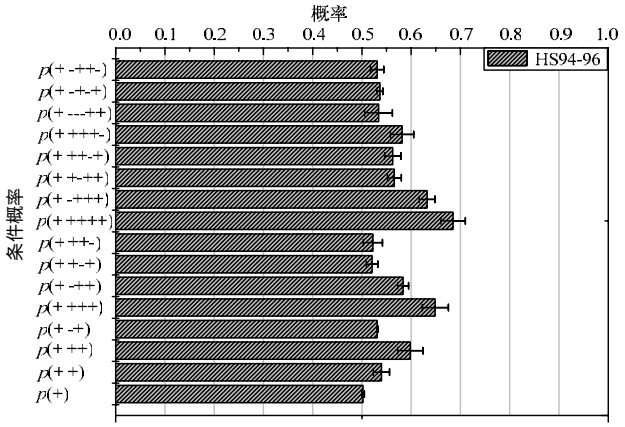


图 5　1994 年~1996 年恒生指数二元符号序列的条件概率分布

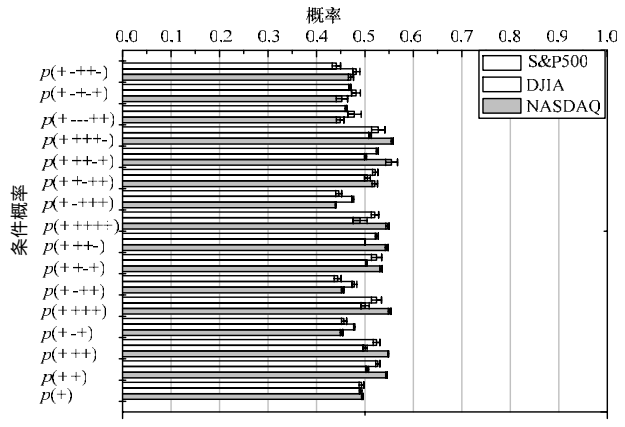


图 6　2005 年~2007 年 NASDAQ,DJIA,S&P500 二元符号序列的条件概率分布

对比 2002 年的上证指数、1994 年~1997 年的恒生指数与 2005 年~2007 年纳斯达克综合指数、道琼斯综合指数、标准普尔 500 综合指数，可以更多地发现市场中总体投资行为的取向，在国外成熟的证券市场中，总体的投资行为并不会因为前期的涨跌而出现明显的跟风现象，当前出现 + 的情况被前期投资行为的影响很小，而在新兴的上海证券市场、香港证券市场中，可以明显看到 $P(+|++++)$, $P(+|+++)$ 都远远高于 $P(+)$，同时 $P(+|++++) > P(+|+++)$，说明总体投资行为存在很大的跟风现象，另外 $P(+|x)$ 明显大于 0.5 的，说明投资者在股市中并不具有很强的风险意识，更多地是采取买的策略，这一点与中国证券市场是一个只能做多不能做空的市场相一致。通过跟成熟证券市场的对比，发现上证 2002 年、恒生 1994 年~1997 年存在很多非理性投资的行为。通过对比新兴和成熟的股指序列的结构化特征，进一步从实证的角度验证了金融市场并不完全有效。

4　结束语

本文基于 AR 模型通过对综合指数二元符号序列的研究，发现具有明显的结构化特征，进一步从实证的角度验证了金融市场是不完全有效的，同时综合比较国内外的证券市场，发现 2002 年上海证券市场和 1994～1997 年香港证券市场存在很多非理性投资的行为。在本模型中为了完全表现股指波动的行为，没有考虑指数不变的情况，但这种情况在证券市场中同样存在，在对标准普尔 500 中 345 只股票的三元符号序列一阶条件概率的研究中发现，同样具有明显的结构化特征。

参考文献

[1] Jefferies P, Hart M L, Hui P M, et al. From Market Games to Real-world Market[J]. Eur. phys. J. B, 2001, 20(1): 493-501.

[2] 王光强, 周佩玲. 神经网络算法在股指预测中的应用[J]. 计算机工程, 2006, 32(1): 211-212.

[3] Zhang Y C. Toward a Theory of Marginally Efficient Market[J]. Physica A, 1999, 269: 30-44.

[4] Sazuka N, Ohira T, Marumo K, et al. A Dynamical Structure of High Frequency Currency Exchange Market[J]. Physica A, 2003, 366: 1-2.

[5] 范剑青, 姚琦伟. 非线性时间序列-建模、预报及应用[M]. 北京: 高等教育出版社, 2005 .

[6] 田　铮. 时间序列的理论与方法[M]. 北京: 高等教育出版社, 2001.

编辑　索书志

(上接第 237 页)

参考文献

[1] 高金刚, 陈建春, 刘雄伟. 数控系统的软 PLC 系统开发[J]. 计算机测量与控制, 2004, 12(3): 254-256.

[2] 黄延延, 林　跃, 于海彬. 软 PLC 技术研究及实现[J]. 计算机工程, 2004, 30(1): 165-167.

[3] John K H. Tigelkamp M. IEC61 131-3: Progarmming Industrial Automation Systems[M]. [S. 1.]: Spring-Verlag, 2001.

[4] 肖世广, 李　彦, 吉　华. Linux 环境下基于 Qt 库的软件 PLC 环境下开发系统[J]. 计算机工程与设计, 2007, 28(7): 1663-1668.

[5] Mavati. Classic ladder[EB/OL]. (2008-06-02). http://mat.sourceforge.net.

编辑　索书志