文章编号:1002—1566(2004)05—0068—10

用时间序列方法预测股票价格初探

谢衷洁,王 弛

(北京大学数学科学学院 北京 100871)

摘 要:本文提出了描述股价变动的一种模型,讨论了模型建立、模型预测的一整套方法及改进算法,比较了算法与改进算法间的优劣,并通过实证分析说明整套理论有一定的可行性。大部分在本文中讨论的算法对于可用核模型描述的其他时间序列的预测问题也同样适用。

关键词: 股价;滤波;非参数估计;向量自回归;成交量

中图分类号:O212 **文献标识码**:A

A Fundamental Step for Stock Price Forecasting using Time Series Methods

XIE Zhong - jie, WANG Chi

(School of Mathematical Sciences Peking University, Beijing 100871 , China)

Abstract: A model to describe the stock price movement is discussed in this paper, which involves model construction and model forecasting. An empirical research shows that our methods can be used in practical applications. Most methods introduced in this paper are also useful for other time series forecasting problems.

Key words: stock price; filtering; non - parametric estimate; vector autoregression; volume

第一部分:引 言

预测股票市场走势一直是人们关心的问题。Franses 和 Van(1996),Choo 等(1999)讨论了用 GARCH 模型预测股市的方法。Frances 和 David(1998),Kaboudan(1999)比较了多种模型的预测效果。随着我国股票市场的日益规范,股票价格的变化受人为因素的影响也逐渐减小,能否对我国股市做预测也越来越为人们所关注。技术分析是目前证券业中常用的预测方法之一。它从开盘价、收盘价、最高价、最低价、交易量等五个基本数据出发,通过观察图形或者计算指标的方法对股价的走势进行预测。虽然技术分析缺乏数学理论根据,但它是人们长期实践经验的总结,其基本思想值得借鉴。本文希望从技术分析的基本原理中获得启发,选用更有说服力的数学方法实现对股价的预测。为达到这一目的,首先对技术分析理论做一下简单回顾。人们在长期实践的基础上提出了股票市场的三个基本假设(参见杨健和蔡红宇(1999)),作为技术分析的理论前提:1. 市场行为涵盖一切;2. 价格朝趋势移动;3. 历史往往重演。假设 1 指出影响股价的政治、经济、企业状况、股民心理等多方面因素都会在股票价格中反映出来,因此在预测股价时只需从股价数据本身出发,而不必考虑各种因素的影响。假设 2 指出了趋势的存在性。价格是随着趋势运行的方向变动的。假设 3 指出了股价变化存在波

收稿日期:2002 年 11 月 20 日

基金项目:国家自然科学基金(10171005)资助项目。

动起伏和周期性。这是因为投资者的投资心理具有一定的不变性,投资手段也有不变性。技术分析在此基础上提出的道氏理论、波浪理论等,都是用来刻画股价变动的趋势性和周期性。从上述理论中得到启示,可以改用时间序列的方法从股价曲线中提取出趋势和周期成分,实现对股票价格的预测。根据上述分析,本文在第二部分提出了一个时间序列模型用来描述股价变动,并指出运用该模型预测股价的一般方法。第三部分讨论了方法的可行性、存在的问题及改进方案。作为算法的实例,第四部分以2000年7月3日到2001年5月31共218天的沪市A股指数(1A0002)数据为基础,尝试对2001年6月的前5个交易日的A股指数进行预测。第五部分简要列出了对4支股票的收盘价预测的结果,表明算法有一定的普适性。第六部分对算法的优缺点进行了评价。

第二部分:模型的建立

一、股价变动模型

从第一部分的分析中可以得出,股票中含有趋势成分和周期成分。用时间序列方法处理这类问题,实际中通常使用 $X-11$ 方法。该方法使用事先给定的滤波器从数据中滤出周期项和趋势项,剩余部分则认为是随机误差。

从 $X-11$ 方法得到启示,可以建立如下模型:

$$x_t = tt_t + \sum_{j=1}^{q} st(j)_t + \xi_t + \varepsilon_t \tag{1}$$

其中 x_t 表示 t 时刻的股票价格;tt_t 表示股价中包含的趋势成分;$st(j)_t$ 表示股价中包含的周期成分,因为股价变化的复杂性,可能包含不止一个周期,所以用 $j=1,\cdots,q$ 表示可能有 q 个周期;ξ_t 表示具有相关结构的可用 $AR(p)$ 模型拟合的部分,添加这一项也是因为股价数据的复杂性;ε_t 表示误差项。

现在问题就转化为如何选取适当的方法从数据中提取出趋势、周期成分。对于股票数据而言,使用 $X-11$ 方法是不恰当的。$X-11$ 方法要求数据具有单一、固定的周期,而股价的周期显然不一定是单一的,也不容易预先知道。下面就来介绍一些方法以解决这个问题。

二、极大极小准则下的最优双边滤波器

滤波器是人们提取特定频段信号时常用的工具。这里如果我们把趋势成分看成是包含在 x_t 中的变化很慢的低频信号,就可以利用低通滤波器从数据中过滤出趋势成分了。

谢衷洁(1993)给出了在极大极小准则下的最优双边滤波器:

$$H_0(\lambda) = \begin{cases} \delta Cos(MCos^{-1}\Psi(\lambda)) & \alpha \leqslant \lambda \leqslant \pi \\ \dfrac{\delta}{2}\{[\Psi(\lambda)+\sqrt{\Psi^2(\lambda)-1}]^M + [\Psi(\lambda)-\sqrt{\Psi^2(\lambda)-1}]^M\} & 0 \leqslant \lambda \leqslant \alpha \end{cases} \tag{2}$$

其中 $\Psi(\lambda)=\dfrac{2Cos\lambda - Cos\alpha + 1}{Cos\alpha + 1}$,$\alpha$ 为截止频率,M 为滤波项数长度,$\delta>0$ 为给定误差。

实际计算中应将 $H_0(\lambda)$ 转化为时域的滤波系数 $\{h_k^{(0)}\}$,转换公式为:

$$h_k^{(0)} = \frac{1}{2M+1}\sum_{j=-M}^{M} H_0(j\times\lambda_0)Cos(j\times k\times\lambda_0), k = 0, \pm 1, \cdots, \pm M \tag{3}$$

其中 $\lambda_0=\dfrac{2\pi}{2M+1}$。

可以认为 x_t 过(3)滤波后的输出就是趋势成分,有

$$\hat{tt} = \sum_{k=-M}^{M} h_k^{(0)} x_{t-k} \tag{4}$$

此方法在实际应用中得到了成功的应用,见谢衷洁(1995)。

三、潜周期分析

这一小节我们来讨论如何从减掉趋势成分后的数据中获得周期成分。设数据有 q 个周期,相应频率值分别为 $\omega_1,\cdots,\omega_q$,在一定的模型假设下,何书元(1984)给出如下结果:

令

$$J_N(\lambda) = N^{-\frac{31}{16}} \left| \sum_{t=1}^{N} x_t e^{it\lambda} \right|^2, \quad -\pi \le \lambda \le \pi \tag{5}$$

则当 N 充分大时,概率1的存在不倚赖于 N 的正常数 K_1,K_2,使得

$$\inf_{|\lambda-\omega_j| \le \frac{\pi}{N}} (J_N(\lambda)) \ge K_1 N^{1/16}, j = 1,2,\cdots,q$$

$$\sup_{\lambda \in \Lambda} (J_N(\lambda)) \le K_2 N^{-1/16}$$

其中 $\Lambda = \bigcup_{j=1}^{q} \{\lambda: N^{-15/16} \le |\lambda - \omega_j| \le 2\pi - N^{-15/16}\}$

这说明对于任意的 $\gamma>0$,当 N 充分大时 ω_j 邻域中的每一点 λ 都有 $J_N(\lambda)>\gamma$;反之,任意的 $\delta>0$,当 N 充分大时不在 ω_j 邻域中的每点 λ 都有 $J_N(\lambda)<\delta$。选取适当的 γ,计算 $J_N(\lambda)$,$\lambda \in (-\pi,\pi)$,画出图形,则图形中 $>\gamma$ 的区间的个数即为 x_t 所含的潜周期个数 q 的估计值 $\hat{q}$,每个区间的最大值所对应的 λ 即为 ω_j 的估计值,记为 $\hat{\omega}_j$。进而也可算出复振幅 A_j 的估计值

$$\hat{A}_j = \frac{1}{N} \sum_{t=1}^{N} x_t e^{-it\hat{\omega}_j}, j = 1.,2,\cdots,\hat{q} \tag{6}$$

何书元(1984)还证明了这些估计量都具有强相合性。利用这种方法,我们可以很容易的通过观察图形得出频率(周期)的估计值。

综合而言,对股票数据建立模型(1)。设观测数据长度为 N,先用极大极小原则下的最优滤波器滤出趋势项,再用潜周期分析的方法找到潜周期。对减掉趋势、周期成分后的数据,用AR(p)模型拟合,并用AIC准则确定阶数。这样就可以得到 tt_t、$st(j)_t$、ξ_t 在 $t=1,\cdots,N$ 时刻的估计值,完成建模。就模型预测而言,可以对 tt_t 项用多项式拟合并估计得 $\hat{tt}_{N+1}$,对 $st(j)_t$ 项做周期拓展得估计值 $\hat{st}(j)_{N+1}$,对 ξ_t 项按拟合的AR(p)模型估计得 $\hat{\xi}_{N+1}$,进而就可以得出 x_{N+1} 的预测值 $\hat{x} = \hat{tt}_{N+1} + \sum_{j=1}^{q} \hat{st}(j)_{N+1} + \hat{\xi}_{N+1}$。

第三部分:算法的改进

用第二部分的方法预测股价,效果不够好。原因是方法本身有很多值得探讨的地方。首先,用双边滤波器分离趋势项时,需要对"过去"(1时刻以前)和"未来"(N时刻以后)的股价做人为的挺拓,虽然人们发明了多种挺拓方法(见谢衷洁(1995)),但还是无法避免与真实数据的差异,这就会产生一定的误差;其次,在对趋势项做预报时,第二部分采用了多项式拟合的方法,但这种方法的预测误差有时很大,极大的影响了取终的预测效果;另外,第二部分的方法针对性不强,没有用到股票市场的特殊信息。针对这些缺欠,本部分对第二部分的方法做了以下几方面的改进:

一、用单边滤波器代替双边滤波器

从理论上讲，双连滤波器 $H_0(\lambda)$ 是极大极小准则下最优的滤波器。但为求得 t 时刻的值，它需要 $t-M,\cdots t+M$ 时刻的数据，而当 $t=N$ 时，$t+1,\cdots t+M$ 时刻的数据是未知的，无论用什么方法近似，都与真实值有差异，造成滤出的趋势项在尾部与真实趋势偏离。

采用单边滤波器可以在一定程度上解决这个问题。单边滤波器为求得 t 时刻的值，只需要 $t-N,\cdots t$ 时刻的数据，而这些都是已知的，所以求出的趋势项是精确值。当然，单边滤波器也有缺点，就是其滤波效果肯定不如双边滤波器。但如果选取的单边滤波器的频率响应函数与 $H_0(\lambda)$ 足够接近，保证滤波效果与双边滤波的效果也比较接近，那么这样的单边滤波器就是解决问题的理想选择。当 M 较大时，可以近似用 $\sum_{k=0}^{M}|H_0(k\lambda_0)-H^*(k\lambda_0)|^2$ 定义 $H_0(\lambda)$ 与 $H^*(\lambda)$ 间的距离，所以只须找到适当的 $H^*(\lambda)$ 使得 $\sum_{k=0}^{M}|H_0(k\lambda_0)-H^*(k\lambda_0)|^2$ 取最小值即可，我们把这个标准称为**均方误差准则**。

定理：设 $H_0(\lambda)$ 为最优双边滤波器，则在所有满足 $\sum_{m=0}^{M}c_m=1$，c_m 为实数的单边滤波器 $H^*(\lambda)=\sum_{m=0}^{M}c_m e^{-im\lambda}$，$0\le\lambda\le\pi$ 中，当

$$c_m=\frac{1}{M+1}+\sum_{k=1}^{M}\left\{\frac{2\cos mk\lambda_0}{2M+1}-\frac{1}{(2M+1)(M+1)}\right\}H_0(k\lambda_0),m=0,\cdots,M \qquad (7)$$

其中 $\lambda_0=\frac{2\pi}{2M+1}$。

时，$\sum_{k=0}^{M}|H_0(k\lambda_0)-H^*(k\lambda_0)|^2$ 达到最小值，即 $H^*(\lambda)$ 与 $H_0(\lambda)$ 在近似均方意义下最接近。记此时的 $H^*(\lambda)$ 为 $H_0^*(\lambda)$，称为均方误差准则下最优的单边滤波器：

证明：见附录。

下面让我们从脉冲响应函数(IRF)和滤波效果两个方面来考察单边滤波器：

1. 脉冲响应函数图形比较

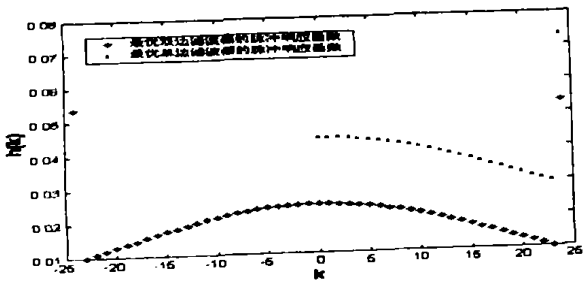

图1：最优单、双边滤波器脉冲响应函数比较

注：这里取 $M=24,\delta=0.1,\alpha=\frac{\pi}{25}$

从图1中可以看出，最优双、单边滤波器的时滤波系数图形形状差别不大，只是单边滤踊器的时滤波系数要大一些。

2. 滤波效果检验

在已知函数 $f(t)$ 上添加噪声 $\varepsilon(t)$ 得 $y(t)=f(t)+\varepsilon(t)$，给定 $y(t)$ 在 $t=1,\cdots,N$ 时刻的值，测试最优单边滤波器能否从所给数据中分离出原函数。用最优单边滤波器对函数 $y(t)=f(t)+\varepsilon(t)$ 滤波，将滤波的结果与原函数 $f(t)$ 比较(见图2)，可见滤波结果与原函数比较接近，且在尾部亦可正确反映原函数的趋势。

注：各图中噪声 $\varepsilon(t)\sim iidN(0,0.5)$ 分布，样本量为 $N=205$。最优单边滤波器的参数为 $M=24,\delta=0.1,\alpha=\frac{\pi}{25}$。

本文以下计算均采用最优单边滤波器。

二、用非参数回归方法对趋势项做预测

第二部分对趋势项做预测时用的是多项式拟合的方法，其效果常常不能令人满意。下面

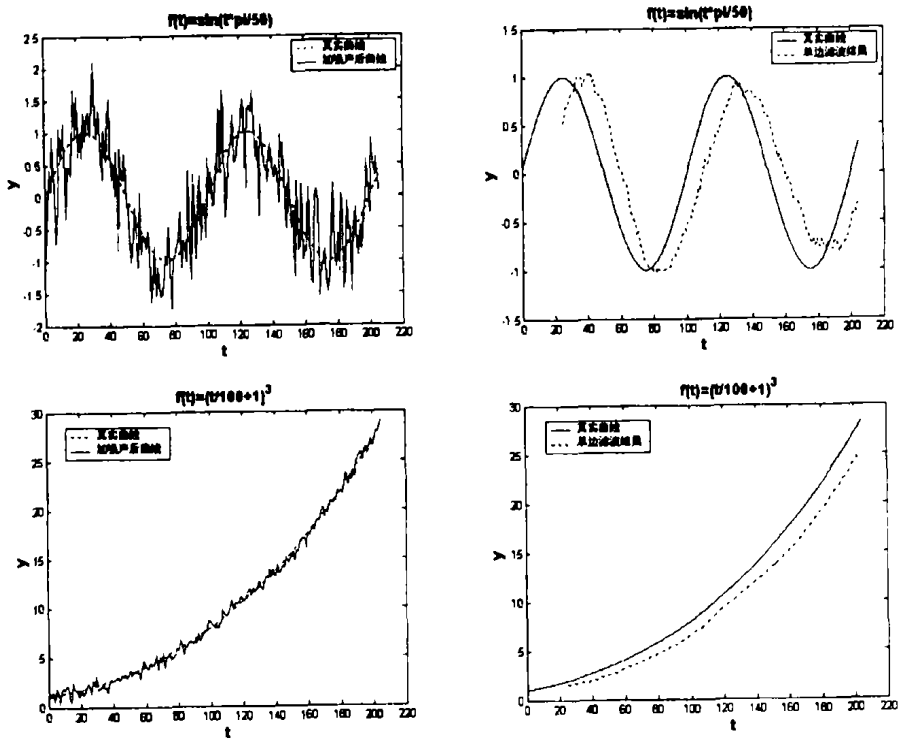

图2:最优单边滤波器的滤波效果

讨论如何用非参数回归的方法解决这个问题。设已知时间序列为$\{\xi_t\}$,$t=1,\cdots,N$,问题是如何预测ξ_{t+1}。J. Fan(1996)指出,可以设$\eta_t=\xi_{t+1}$,$t=1,\cdots,N-1$,则(ξ_t,η_t),$t=1,\cdots,N-1$为二维空间中的一系列点,将ξ_t视为横坐标,η_t为核点的取值,则问题转化为求ξ_N点取值,可用非参数方法求得。但。J. Fan的方法没有区分向上和向下的趋势,会造成错的预报。例如ξ_N处在一个向上的趋势中,我们有理由认为ξ_{N+1}应大于ξ_N;而如果N的时刻之间与ξ_N值相近的点都处在向下的趋势中,会导致预测值小于ξ_N。可以对J. Fan的方法做改进,考虑变化率。改令$\zeta_t=\frac{\xi_{t+1}-\xi_t}{\xi_t}$,$t=1,\cdots,N-1$,$\eta_t=\zeta_{t+1}$,$t=1,\cdots,N-2$,则$(\zeta_t,\eta_t)$,$t=1,\cdots,N-2$是二维空间中的一系列点,先求出$\zeta_N$的估计值$\hat{\zeta}_N$,再由它求$\hat{\xi}_{N+1}$,就可以解决这个问题。具体的非参数估计可采用如下做法:

1. 用k-近邻法,认为ζ_{N-1}点的取值只与和它最接近的k个点的取值有关,设这k个点为$\zeta_{t_1},\cdots,\zeta_{t_k}$。

2. 计算距离权重:$w1_{t_i}=\frac{1}{|\zeta_{t_i}-\zeta_{N-1}|}$,$i=1,\cdots,k$。

3. 计算时间权重:$w2_{t_i}=\frac{1}{\frac{N-1-t_i}{5}+1}$,$i=1,\cdots,k$。考虑时间权重是基于时间序列的特殊性。时间序列各项之间有相关性,时间距离越短的两项相关性越强,故此应引时时间权重反映两观测点在时间上的接近程度。

4. 计算ζ_{N-1}点取值的估计值$\hat{\eta}_{N+1}=\frac{\sum_{i=1}^{k}\eta_{t_i}\times w1_{t_i}\times w2_{t_i}}{\sum_{i=1}^{k}w1_{t_i}\times w2_{t_i}}$,即为$\zeta_N$的估计值$\hat{\zeta}_N$。

5. 由$\hat{\xi}_{N+1}=\hat{\zeta}_N\times\xi_N+\xi_N$可求出$\hat{\zeta}_{N+1}$。

对已知函数 $y=f(t)$，给定 $1,\cdots,N$ 时刻的值，分别用非参数方法与多项式方法预测 $N+1,\cdots,N+5$ 时刻的值，如图 3 所示。可见用非参数方法对初等函数做预测可以收到很好的效果。

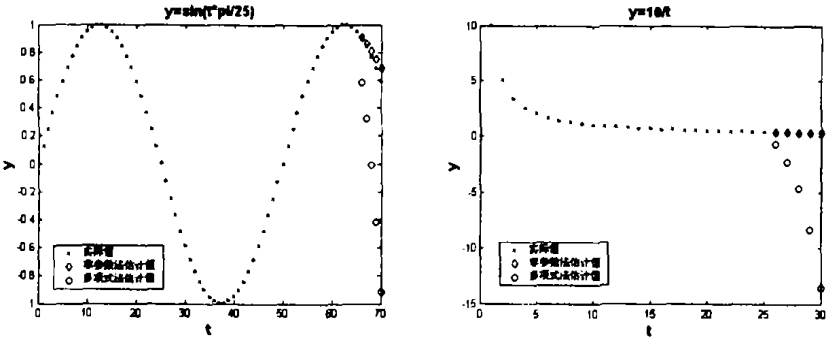

图 3：非参数方法与多项式法预测效果比较

注：图中采用的非参数方法未考虑时间权重。多项式法使用的是 5 阶多项式拟合。样本量分别为 $N=65$ 和 $N=25$。

三、向量自回归

前面讨论的所有方法都具有普遍性，如果分析数据不是股票价格而是其他数据，也可用前面的方法进行分析。但我们知道，普遍性很强的方法，其精确度往往不够高，所以如果能结合股市的特殊性，则有望得到更好的结果。在股市中，与股价联系最紧密的要数成交量了，吴冲锋和吴文锋(2001)讨论了结合成交量分析股价的一种方法。这里我们将股价和成交量看成一个二维向量，用向量自回归模型拟合，这样即可以体现出股价和成交量间的相互作用关系，处理起来又比较简单。向量自回归模型的一般形式为(以下用下划线表示向量)：

$$\underline{x}_t = \underline{\varphi}_1\underline{x}_{t-1} + \underline{\varphi}_2\underline{x}_{t-2} + \cdots + \underline{\varphi}_n\underline{x}_{t-n} + \underline{\varepsilon}_t \tag{8}$$

其中 $\underline{x}_t$ 为 q 维向量，$E\underline{x}_t=\underline{0}$，$\{\underline{\varepsilon}_t\}$ 为白噪声序列。特别地，当 $q=1$ 时，即为变通的 AR(p)模型。

关于向量自回归模型的参数估计，也有与 AR(p)类似的递维公式。同时也有相应的 AIC 准则用于确定阶数，有关结果见顾岚(1994)。

第四部分：算法实现与大盘预报

下面以沪市 A 股指数(1A0002)的预测为例说明整套算法的实现过程。我们选用 2000 年 7 月 3 日到 2001 年 5 月 31 日共 218 天的数据，尝试对 2001 年 6 月的前 5 个交易日的 A 股指数进行预测。数据来源：http://www.sohu.com/。

一、数据预处理

对股票价格预测来说，由于有涨、跌停板(即一日内涨、跌幅度不能超过 10%)制度，可以认为不会有异常点，所以可不做预处理。

二、滤出趋势项

根据单边滤波公式(7)滤出趋势项。滤波结果如图 4 所示。

注：这里取 $\delta=0.1$，$T=50$，$\alpha=\dfrac{2\times\pi}{T}$。

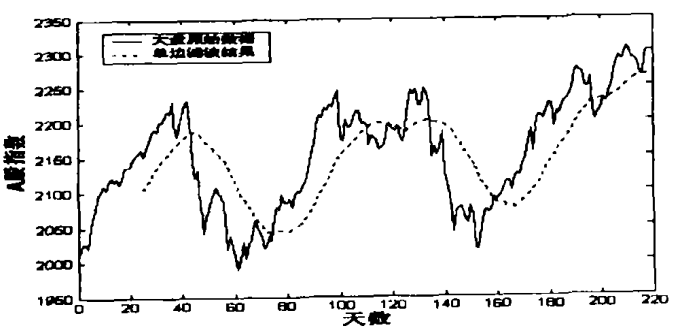

图 4：最优单边滤波器对大盘数据滤波效果

三、周期成分滤出

对于滤去趋势项后的数据,用何书元算法结合极大熵法找出潜周期。极大熵法的具体计算公式见顾岚(1994)。

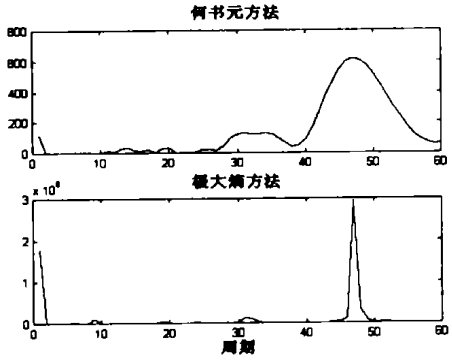

图5:潜周期的判定

综合两种方法,从图5中可以看出数据有两个明显的周期为47、31个交易日,相应的振幅和初相位估计分别为(41.6966, -1.4293),(18.3283, -2.4388)。

四、用AR(p)拟合

滤去趋势、周期成分之后的大盘数据基本上可以认为是平稳的,用AR(p)拟合并用AIC准则确定阶数,得 $p=5$,模型参数估计为0.9935, -0.1847, 0.2080, -0.1866, 0.1015。

五、预报

根据模型(1),对趋势项用非参数回归方法做预报,对周期项 $st(j)_t$ 可以直接做周期延拓,对 ξ_t 项可根据AR(p)直接计算,就可以得出 $\hat{x}_{N+1}$ 的预测值。

如果如下:

表1:沪市A股指数预报结果

天数	1	2	3	4	5
预测值	2304.6	2311.3	2311.9	2314.9	2316.7
实际值	2307.6	2318.3	2321.9	2328.3	2319.3
相对误差	0.0013	0.0030	0.0043	0.0058	0.0011
平均相对误差:0.0031					

注:其中相对误差的计算公式为 $\frac{|x_t-\hat{x}_t|}{x_t}$。

六、向量自回归

如第三部分(三)中分析的那样,上面的方法没有涉及股票自身的特殊性,为了弥补这一缺欠,可以用向量自回归模型代替AR(p)模型。令 $c_t=\log C_t$,其中 C_t 为交易量。计算 ξ_t 与 c_t 的相关性,得 $\rho(\xi_t,c_t)=0.4651$,可见引入交易量的信息是有价值的。

将 ξ_t 与 c_t 组成二维向量,建立向量自回归模型(8),估计模型阶数及参数,再重新做预测,可见预测结果略有改进。

表2:应用向量自回归后的A股指数预报结果

天数	1	2	3	4	5
预测值	2305.0	2311.1	2314.3	2317.3	2318.9
实际值	2307.6	2318.3	2321.9	2328.3	2319.3
相对误差	0.0011	0.0031	0.0033	0.0047	0.0002
平均相对误差:0.0025					

第五部分:个股预报结果

选取4支股票:浙江东方(600120)、成量股份(600673)、陆家嘴(600663)和天目药业(600671),先用2000年7月月3日到2001年5月31日共218天的收盘价数据,尝试对2001

年6月的前5个交易日的收盘价进行预测。结果见表3。

表3:个股预报结果

股票名称	预测结果						平均相对误差
浙江东方	天数	1	2	3	4	5	0.0019
	预测值	24.4145	24.6141	24.7962	24.9503	25.0697	
	实际值	24.3600	25.4000	25.1000	25.2900	25.0500	
	相对误差	0.0022	0.0309	0.0121	0.0134	0.0008	
成量股份	天数	1	2	3	4	5	0.0086
	预测值	15.1402	15.2302	15.3461	15.4595	15.3879	
	实际值	15.2900	15.3700	15.3800	15.7100	15.4800	
	相对误差	0.0098	0.0091	0.0022	0.0159	0.0059	
陆家嘴	天数	1	2	3	4	5	0.0191
	预测值	15.9065	15.8630	15.8358	15.8211	15.8177	
	实际值	16.0500	16.2000	16.2700	16.2100	16.0600	
	相对误差	0.0089	0.0208	0.00267	0.0240	0.00151	
天目药业	天数	1	2	3	4	5	0.0130
	预测值	14.8652	14.9142	14.9656	14.9905	14.9875	
	实际值	14.9100	15.2000	15.2100	15.2000	15.1900	
	相对误差	0.0030	0.0188	0.0161	0.0138	0.0133	

第六部分:模型评价及方法比较

从上面的分析可以看出,用模型(1)来描述股价的变动有一定的可行性。利用第二部分的一般算法及第三部分的改进算法,可以较好的建立模型并进行一定精度的预报。而且,本文的大部分方法都具有普遍性,一般可以用模型(1)描述的问题都可以用这些方法求解。当然,本文提出的算法只是一个雏形,以下几点值得做进一步的讨论:

1. 用非参数法作趋势预测,需要同时考虑时间权数和距离权数,一些参数的取值需要人为确定,缺乏严格的数学理论保证。

2. 本文试图考虑股市的特殊性,也通过向量自回归模型引入了交易量的信息,但预测效果改善不显著。如果建立其他模型或引入股市的其它特有信息,或许可以得到更好的结果。

3. 从图2中可以看出,单边滤波结果与原函数间似乎有一个相位差。如果用最小二乘法估计出相位差,再对尾部用非参数方法拓展,或许可以改善拟合效果。

[参考文献]

[1] 谢衷洁. 滤波及其应用[M]. 湖南:湖南教育出版社,1995.

[2] Xie,Zhongjie. Case Studies in Time Series Analysis[M]. World Scientific, Singapore, 1993.

[3] 谢衷洁. 时间序列分析[M]. 北京:北京大学出版社,1990.

[4] 顾岚. 时间序列分析在经济中的应用[M]. 北京:中国统计出版社,1994.

[5] 杨健,蔡红宇. 中国股市实证技术分析指南[M]. 北京:中国人民大学出版社,1999.

[6] 吴冲锋. 吴文峰. 基于成交量的股价序列分析[J]. 系统工程理论方法应用,2001,10(1):1-7.

[7] J. Fan and I. Gijbels. Local Polynomial Modeling and Its Applications[M]. Chapman&Hall,London,1996.

[8] M. A. kaboudan. A Measure of Time Series' Predictability Using Genetic Programming Applied to Stock

Returns[J]. Journal of Forecasting, 1999, 18:345 - 357.

[9] Choo Wei Chong, Muhammad Idrees Ahmad and Mat Yusoff Abdullah. Performance of GARCH Models in Forecasting Stock Market Volatility[J]. Journal of Forecasting, 1999,18:333 - 343.

[10] Franses, P. H. and Van Dijk, R. Forecasting Stock Market Volatility Using (non - linear) GARCH models [J]. Journal of Forecasting, 1996,15:229 - 235.

[11] Frances B. Shin and David H. Kil. Classification Cramer - Rao Bounds on Stock Price Prediction[J]. Journal of Forecasting, 1998,17:389 - 399.

附录:单边滤波公式推导

引理: $\sum_{k=1}^{M}\cos k\lambda=\begin{cases}M & \lambda=1\\ \dfrac{\sin\frac{2M+1}{2}-\sin\frac{\lambda}{2}}{2\sin\frac{\lambda}{2}} & 0<\lambda\le\pi\end{cases}$,

特别地,当 $\lambda=\frac{2m\pi}{2M+1}$, $m=1,\cdots,M$ 时, $\sum_{k=1}^{M}\cos k\lambda=-\frac{1}{2}$。

证明:当 $\lambda\neq0$ 时,

$$\sum_{k=1}^{M}\cos k\lambda=\frac{1}{2\sin\frac{\lambda}{2}}\left(2\sin\frac{\lambda}{2}\cos\lambda+\cdots+2\sin\frac{\lambda}{2}\cos M\lambda\right)$$

$$=\frac{1}{2\sin\frac{\lambda}{2}}\left\{\sin\frac{3\lambda}{2}-\sin\frac{\lambda}{2}+\sin\frac{5\lambda}{2}-\sin\frac{3\lambda}{2}+\cdots+\sin\frac{(2M+1)\lambda}{2}-\sin\frac{(2M-1)\lambda}{2}\right\}$$

$$=\frac{\sin\frac{2M+1}{2}\lambda-\sin\frac{\lambda}{2}}{2\sin\frac{\lambda}{2}}$$

当 $\lambda=0$ 时,显然 $\cos k\lambda=1$, $k=1,\cdots,M$, $\sum_{k=1}^{M}\cos k\lambda=M$。

定理证明:设 $y=\sum_{k=0}^{M}|H_0(k\lambda_0)-H^*(k\lambda_0)|^2$

由 $\sum_{l=0}^{M}c_l=1$ 得, $c_0=1-\sum_{l=1}^{M}c_l$,又 $H_0(0)=H^*(0)=1$,有

$$y=\sum_{k=1}^{M}(H_0(k\lambda_0)-H^*(k\lambda_0))(H_0(k\lambda_0)-\overline{H^*(k\lambda_0)})$$

$$=\sum_{k=1}^{M}\left[H_0(k\lambda_0)-\sum_{l=1}^{M}c_le^{-ilk\lambda_0}+1-\sum_{l=1}^{M}c_l\right]\left[H_0(k\lambda_0)-\left(\sum_{l=1}^{M}c_le^{-ilk\lambda_0}+1-\sum_{l=1}^{M}c_l\right)\right]$$

$$\because\frac{\partial^2y}{\partial c_m\partial c_n}=-\sum_{k=1}^{M}\left[2\cos nk\lambda_0-2\cos(n-m)k\lambda_0+2\cos mk\lambda_0-2\right]=\begin{cases}2M+1 & m\neq n\\ 4M+2 & m=n\end{cases}$$

易见海色矩阵正定。

$\therefore y$ 有最小值。整理方程组 $\frac{\partial y}{\partial c_m}=0$, $m=1,\cdots,M$ 得

$$\sum_{k=1}^{M}(2-2cosmk\lambda_0)H_0(k\lambda_0)$$

$$=\sum_{k=1}^{M}\left[2\sum_{l=1}^{M}c_l\cos lk\lambda_0-2\sum_{l=1}^{M}c_l\cos(l-m)k\lambda_0+2\cos mk\lambda_0\sum_{l=1}^{M}c_l-2\sum_{l=1}^{M}c_l+2-2\cos mk\lambda_0\right]$$

由引理,得

$$\sum_{k=1}^{M}(2-2\cos mk\lambda_0)H_0(k\lambda_0)=-(2M+1)\sum_{l=1}^{M}c_l-(2M+1)c_m+2M+1,m=1,\cdots,M \quad ①$$

对 $m=1,\cdots,M$ 各式相加得:

$$\sum_{l=1}^{M}c_l=\frac{M}{M+1}-\frac{1}{M+1}\sum_{k=1}^{M}H_0(k\lambda_0),m=1,\cdots,M \quad ②$$

②代入①得

$$c_m=\frac{1}{M+1}+\sum_{k=1}^{M}\left\{\frac{2\cos mk\lambda_0}{2M+1}-\frac{1}{(2M+1)(M+1)}\right\}H_0(k\lambda_0),m=1,\cdots,M$$

又由 $c_0=1-\sum_{l=1}^{M}c_l$ 解得 $c_0=\dfrac{1+\sum_{k=1}^{M}H_0(k\lambda_0)}{M+1}$。

$$\therefore c_m=\frac{1}{M+1}+\sum_{k=1}^{M}\left\{\frac{2\cos mk\lambda_0}{2M+1}-\frac{1}{(2M+1)(M+1)}\right\}H_0(k\lambda_0),m=0,\cdots,M \quad \#$$

(上接第51页)

四、讨　论

4.1　非参数判别分析的具体方法、步骤未见书中介绍或杂志报道。本文使用 *SAS* 软件,采用不等带宽核密度估计的非参数判别分析,*SAS* 只是简单地输出显示了此方法的平方距离函数、判别函数、后验概率的笼统公式。作者通过推导分析、数据验算等,弄清了这些公式的意义,明确了公式中各符号的含义,掌握了此方法的具体步骤。

4.2　对本文78例训练样本,仍按后退法逐步判别分析筛选出的同样13个主要判别指标,进行参数法判别分析,*SAS* 运行得交叉证实结果:33例脑出血病人有22例判为脑出血,11例错判为脑缺血,判别正确率为66.67%;45例脑缺血病人有33例判为脑缺血,12例错判为脑出血,判别正确率为73.33%;总判别正确率为70.51%,判别效果差。说明对定性指标变量,应采用非参数判别分析。

4.3　对本文78例训练样本,以头颅 *CT* 和核磁共振检查为依据,医生临床分类诊断的结果为:33例脑出血病人有20例诊断为脑出血,10例误诊为脑缺血,3例未作出诊断,诊断正确率为60.61%;45例脑缺血病人有35例诊断为脑缺血,8例误诊为脑出血,2例未作出诊断,诊断正确率为77.78%;总诊断正确率为70.52%,诊断效果差。

4.4　在本文42例应用病例中,有12例病人在判别分析后作过颅 *CT* 或核磁共振检查,发现判别分析12例中,仅有1例脑缺血错判为脑出血,实际应用判别效果良好。

[参考文献]

[1] 邓祖新. *SAS* 系统和数据分析[M]. 北京:电子工业出版社,2002. 216-232.

[2] 金丕焕. 医用 *SAS* 统计分析[M]. 上海:复旦大学出版社,2002. 144-148.

[3] 孙振球. 医学统计学[M]. 北京:人民卫生出版社. 2002. 297-298.