第51卷 第1期　　气 象 学 报　　Vol.51,No.1
1993年2月　　ACTA METEOROLOGICA SINICA　　Feb.,1993

基于Bayes准则的时间序列判别预报模式*

丁裕国　江志红

（南京气象学院，210044）

根据Bayes准则下的多元线性判别分析和时间序列的线性自回归模式，本文提出一种时间序列的判别预报模式。该模式采用两种不同的变量筛选方案，对于气象时间序列的数量记录，由过去的记录判别未来记录的趋势(如正负距平、旱涝等)。在一定的自相关结构下，其判别效果较好。文献〔1—4〕曾论述用(0,1)两值时间序列建立$AR(p)$模式，但$AR(p)$模式有其局限性。将时间序列与多元判别分析结合，建立时间序列基础上的判别模式，用以往各时刻变量作为线性判别因子对未来各时刻的变量取值类型作出判别，既可保留时间序列线性模式的优点，又可利用多元逐步判别筛选因子的计算方法。从气象状况演变的物理机制来看，考虑前期状态演变比单纯考虑前期某一时刻的状态更有意义。

1　时序判别模式的建立

设有时间序列$\{x_t\}$，$1\leqslant t\leqslant N$，约定为零均值平稳序列。且若有

$$\left.\begin{array}{ll} x_{t_i}\in A & x_{t_i}>x_c=0 \\ x_{t_i}\in B & x_{t_i}\leqslant x_c=0 \end{array}\right\} i=1,2,\cdots,N$$

则可将x_t的值域分为两组(例如A和B类或G_1与G_2类)。对于G_1类，$x_{t_i}>0$，总可有资料向量

$$X_i=(x_{t_i}^{(1)}\quad x_{t_i-1}^{(1)}\quad\cdots x_{t_i-p}^{(1)}),i=1,\cdots,n_1 \tag{1}$$

对于G_2类，$xx_{t_i}\leqslant 0$，总可有资料向量

$$X_2=(x_{t_j}^{(2)}\quad x_{t_j-1}^{(2)}\quad\cdots x_{t_j-p}^{(2)}),j=1,\cdots,n_2 \tag{2}$$

上式中$x_{t_i-1}^{(1)}$代表属于G_1的样本资料，而相应的前p个时刻的取值分别记为$x_{t_i-1}^{(1)}$，$x_{t_i-2}^{(1)}$，$\cdots x_{t_i-p}^{(1)}$，同样，$x_{t_j-p}^{(2)}$，代表属于G_2的样本资料，而相应的前p个时刻的取值分别记为$x_{t_j-1}^{(2)}$，$x_{t_j-2}^{(2)}$，$\cdots x_{t_j-p}^{(2)}$，对于p的选取，将在后文中说明。一般说，p就是要建立的时序判别函数的阶数。由于抽样的随机性，向量式(1)和(2)中的样品个数n_1和n_2不一定相等。

根据式(1)和(2)，可得相应于G_1或G_2组的均值向量

$$\overline{x_1}=(\overline{x_t}^{(1)}\quad\overline{x}_{t-1}^{(1)}\quad\cdots\overline{x}_{t-p}^{(1)})' \tag{3}$$

$$\overline{x_2}=(\overline{x_t}^{(2)}\quad\overline{x}_{t-1}^{(2)}\quad\cdots x_{t-p}^{(2)})' \tag{4}$$

假定序列为正态平稳，且G_1和G_2类前期各时刻（设有p个时刻)变量的协方差阵相同，则可建立Bayes准则下的时间序列判别函数

* 1990年10月11日收到原稿，1991年3月14日收到修改稿。本文为气象基金资助课题的论文之一。

$$Y_K(x_t)=\ln P_K+c_0^{(K)}+c_1^{(K)}x_{t-1}+\cdots+c_p^{(K)}x_{t-p} \tag{5}$$

K=G_1或G_2(为叙述方便，设$G_1=1$，$G_2=2$)

其中
$$c_i^{(K)}=\sum_{j=1}^{p}S^{t-i,t-j}\,\bar{x}_{t-j}^{(K)} \qquad i=1,2,\cdots,p$$

$$c_0^{(K)}=-\frac{1}{2}\sum_{i=1}^{p}c_i^{(K)}\bar{x}_{t-i}^{(K)}$$

上式中$S^{t-i,t-j}$为后文定义的（组内）协方差阵的逆矩阵相应元素，上角标$t-i$，$t-j$表示元素的行列位置。

为了确定式(5)中变量及其阶数p，本文采用两种筛选变量的方法。一种是对前期所有可供选择的时刻$t-1$，$t-2$，…，$t-p$按某一准则逐个选入，其判别函数形式上类似于通常的$AR(p)$模式（即所谓"逐点法"）；另一种则是对前期所有可供选择的时刻$t-1$，$t-2$，…，$t-p$按逐步判别准则筛选，这样就有可能逐步"引进"某些时刻的变量而"剔除"不符合筛选准则的另一些时刻的变量，最终建立非等间隔的前期各时刻变量组成的线性判别函数。其判别函数在形式上类似于疏系数自回归模式（即所谓"选点法"）。

对两组变量分别有组内离差(组内协方差)和组间离差(组间协方差)

$$W_{t,t-\tau}=\sum_{i=1}^{n_1}(x_{it}^{(1)}-\bar{x}_t^{(1)})(x_{i,t-\tau}^{(1)}-\bar{x}_{t-\tau}^{(1)})$$

$$+\sum_{i=1}^{n_2}(x_{it}^{(2)}-\bar{x}_t^{(2)})(x_{i,t-\tau}^{(2)}-\bar{x}_{t-\tau}^{(2)}) \tag{6}$$

$$q_{t,t-\tau}=n_1(\bar{x}_t^{(1)}-\bar{x}_t)(\bar{x}_{t-\tau}^{(1)}-\bar{x}_{t-\tau})+n_2(\bar{x}_t^{(2)}-\bar{x}_t)(\bar{x}_{t-\tau}^{(2)}-\bar{x}_{t-\tau}) \tag{7}$$

其中$\tau=1,2,\cdots,p$；$\bar{x}_t^{(1)}$，$\bar{x}_t^{(2)}$，分别为时刻t的G_1类、G_2类的类均值；$\bar{x}_{t-\tau}^{(1)}$，$\bar{x}_{t-\tau}^{(2)}$分别为时刻$t-\tau$的相应于G_1，G_2类的类均值；而$\bar{x}_t$则为不分类的序列均值，$\bar{x}_{t-\tau}$亦同。显然，由于

$W_{t,t-\tau}=W_{t-\tau,t}$，$q_{t,t-\tau}=q_{t-\tau,t}$的缘故，组间协方差矩阵和组内协方差矩阵与总协方差矩阵应有关系式

$$\boldsymbol{T}=\boldsymbol{W}+\boldsymbol{Q} \tag{8}$$

上式中，$\boldsymbol{T}$为全序列的总协方差矩阵，$\boldsymbol{W}$和$\boldsymbol{Q}$分别为组内和组间协方差矩阵。类似于文献〔5〕，引入$Wilks$统计量$U=|W|/|T|$，对给定的前p个时刻，有近似公式

$$\chi^2(p)=-\left[(N-1)-\frac{1}{2}(p+2)\right]\ln U \tag{9}$$

这里N为x_t的序列长度(总样本容量)，p即为式(5)中规定的前期可供选择的p个时刻变量个数(判别函数的阶数)。根据χ^2检验，就可确定逐次引进判别因子的分类效果，以便确定是否继续引进判别因子。一旦确定了判别因子数p，就可建立式(5)的线性判别函数。为计算方便，本文仍类似于多元线性判别，假定式(5)右端第1项P_K为等概率情况。

判别函数式(5)的计算步骤可仿文献〔5〕，最终等价于求解方程

$$\boldsymbol{C}_K\boldsymbol{W}=\bar{\boldsymbol{X}}_t \tag{10}$$

中的判别函数系数向量

$$\boldsymbol{C}_K=(C_1^{(K)}\ C_2^{(K)},\ \cdots,\ C_p^{(K)})$$

在两分类判别时$K=1,2$，上述方程只要借助于多元回归的正规方程求解计算即可得到相应各阶判别系数$C_i^{(K)}$，$i=1,2,\cdots,p$。代入式(5)中得到相应的判别函数$Y_K(x_t)$，就可作出Bayes准则下的归类

决策。

采用"选点法"建立判别模式，原则上类似于"逐点法"。所不同的是，要将协方差阵(W_{ij})和总协方差阵(t_{ij})按逐步判别方案作"引进"或"剔除"的双重筛选，即利用一维 $Wilks$ 统计量作检验，并按逐步回归算法作消去变换，从而最终获得判别函数(参见文献〔5〕)。

2 计算实例与效果分析

取南方涛动指数(SOI)月际序列(1969—1978年)为例，采用"逐点法"建立判别模式($N=120$)。据表1中的资料，首先将前面 $n=108$ 的 SOI 序列划分为正、负距平两类样品，求得式(1)和(2)相应的资料阵，然后分别计算样本均值向量和序列协方差阵(组内与不分组两种)，最终应用 $Wilks$ 统计量及 x^2 检验，逐次建立各阶($p=1,2,\cdots,p$)时间序列判别模式并筛选出最佳模式。

图1为各阶判别模式拟合率(前108个记录)和预报准确率(后12个记录)随阶数的变化。除 $p=2$ 时，拟合率小于0.85以外，一般都大于0.90，当 $p>14$ 时，拟合率虽有下降但也在0.80左右；从最后12个记录(未建模记录)的判别预报效果来看，当 $p<12$ 时，准确率均在0.85以上，仅当 $p=12$，准确率降至0.75，以后又有回升，总的说，这样的预报准确率已相当好。不过，试验表明，并非所

表1 SOI月际记录序列(1969—1978年)

年	1	2	3	4	5	6	7	8	9	10	11	12
1969	−0.40	−0.40	−0.60	−0.45	−0.35	−0.50	−0.60	−0.70	−0.60	−0.55	−0.60	−0.60
1970	−0.30	−0.45	−0.47	−0.02	0.00	0.15	0.40	0.60	0.70	0.80	1.00	0.95
1971	1.00	1.20	1.25	1.15	0.85	0.75	0.70	0.85	0.90	0.85	0.75	0.50
1972	0.30	0.00	−0.20	−0.60	−0.80	−1.10	−1.40	−1.10	−1.00	−0.90	−0.85	−0.80
1973	−0.60	−0.60	−0.50	0.25	0.00	0.20	0.40	0.60	0.90	1.10	1.25	1.40
1974	1.75	1.35	1.30	1.00	0.90	0.75	0.78	0.65	0.60	0.50	0.25	0.00
1975	0.10	0.20	0.40	0.50	0.60	0.90	1.25	1.60	1.60	1.50	1.40	1.25
1976	1.20	1.00	0.90	0.70	0.50	0.00	−0.50	−0.70	−0.65	−0.60	−0.40	0.10
1977	0.00	−0.50	−0.55	−0.70	−1.20	−1.22	−1.20	−1.25	−1.20	−1.20	−1.10	−1.22
1978	−1.20	−0.90	−0.60	−0.50	0.10	0.35	0.20	0.00	−0.10	−0.20	−0.25	−0.22

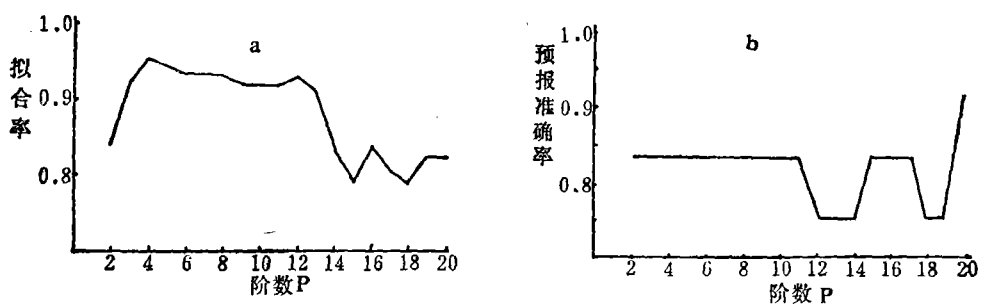

图1 判别模式的拟合率 (a)和预报准确率 (b)随阶数的变化

有时序判别模式都有如此好的效果，无论是"逐点法"还是"选点法"建模，其判别效果优劣主要取决于序列本身的自相关结构。作者以 SOI 序列与南京(1950—1951年)逐候降水量距平序列(简记 R·P 序列)为例做了对比试验。前者自相关较好，且自相关系数呈负指数衰减，而后者自相关较差，且呈高频起伏(图略)，其时序判别模式的判别效果前者高于后者。这是因为，影响判别效果的主要因素与组内离差阵行列式$|W|$及总离差阵行列式$|T|$之比值有关[6]。显然，前者表明组内不同时刻之间的自相

关性；后者表明序列本身不分组时的自相关性，由多元分析知，无论$|T|$或$|W|$(又称广义方差)，其值愈小，意味着相关性愈大，反之则相关性愈小[7]。因此，当广义总方差$|T|$一定时，即序列给定时，据式(9)，$|W|$愈小，必然判别效果愈好。换言之，判别效果取决于组内自相关结构。另一方面，从表2给出的对比数据还可发现另一个影响因素：分类条件平均差。由表2可见，时序判别模式的效果

表2　SOI与R·P序列各阶判别模式拟合率对比

P	统计量 U_e	分类平均差	拟合率	统计量 U_R	分类平均差	拟合率
2	0.3465	0.7483	0.89	0.9912	0.3358	0.54
4	0.3444	1.1579	0.90	0.9864	0.3400	0.57
6	0.3312	1.0415	0.89	0.9840	0.0184	0.56
8	0.3325	0.7556	0.90	0.9748	0.0848	0.57
⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	0.3096	0.2880	0.92	0.8935	0.1563	0.65
平均	0.3287	0.7557	0.908	0.9534	0.1574	0.589

优劣，主要取决于序列内部的自相关结构和分类条件平均差值。

此外，类似于文献〔1〕，我们还论证了时序判别在两分类时与(0,1)自回归的等价性。利用SOI序列资料分别建立(0,1)自回归和时序判别模式，结果表明，两者在建模拟合和预报方面效果虽不相上下，但在高阶情况下，采用时序判别模式要比(0,1)自回归模式计算更为方便。

3 小　结

(1) 时间序列判别分析可类似于多元线性判别分析方法，将前期逐个时刻或任意几个时刻作为判别因子选入模式，即"逐点法"或"选点法"。

(2) 时间序列判别预报效果主要受序列本身自相关结构的影响，当组内自相关性比较大时，判别效果较好；其次，分组后的条件平均值的差异大小，也是一个重要影响因素。

(3) 时序判别模式在两分类时与(0,1)序列自回归有一定的等价意义，但在计算方法上，特别是高阶情况下，时序模式比(0,1)自回归更方便。

参 考 文 献

[1] Kedem B. Binary time series. Marcel Dekker, New York and Basel, 1980, 45—63.

[2] 项静恬，杜金观，史久恩. 动态数据处理. 气象出版社，1986. 398—426.

[3] 丁裕国. (0,1)两值时间序列分析及其气象应用. 广西气象，1987，(5—6)，10—12.

[4] 么枕生. 气候统计学的研究展望. 气象科技，1984，(6)：1—8.

[5] 屠其璞，丁裕国等. 气象应用概率统计学. 气象出版社，1984.308—319.

[6] 中科院计算中心概率统计组. 概率统计计算. 科学出版社，1979. 213—215.

[7] 丁裕国，吴息. 经验正交函数展开气象场收敛性的研究. 热带气象，1988，4(4)，316—326.

TIME SERIES FORECAST MODEL BY DISCRIMINATORY ANALYSIS ACCORDING TO BAYES CRITERION

Ding Yuguo Jiang Zhihong
(*Nanjing Institute of Meteorology*, 210044)

Abstract

A discriminant prediction model of time series is presented by using of multiple linear discriminant and linear autoregression models of time series. There are two selection rules of the discriminative predictor, that is (1) the stepwise induction method by schedular order; (2) the choice method by schedular non-order, and both types of selection rules may be used in this model. Thus, this discriminant prediction model may be applied to the digital records of weather time series and this model is effective on the time series under certain autocorrelation structure.