二次型判别函数在中期晴雨判别中的应用

曲晓波　孙　欣　（沈阳中心气象台　110015）

1　引言

判别分析是气象上常用的统计方法之一。在样本服从正态分布，且类间样本协方差阵相等（$\Sigma_1 = \Sigma_2$）的假设下，基于 Bayes 准则可以得到线性判别函数，而在 $\Sigma_1 \neq \Sigma_2$ 时，同样基于 Bayes 准则的判别函数自然会出现二次型。由于二次型函数的出现使得计算量增加很大，因而二次型判别模型的产生困难很大，人们在以往的工作中较多地采用线性判别函数，这是不符合实际的。据文献〔1〕分析，若 Σ_1 与 Σ_2 有较大差异，二次型判别函数优于线性判别函数。施能〔2〕将类似的二次型判别函数用于长期天气的洪涝预报，取得一定成效。本文采用 BC^{-1} 统计量方法来决定是否接受 $\Sigma_1 = \Sigma_2$ 的假设，在拒绝该假设时，通过求得变换矩阵 A，对预报因子向量进行线性变换后，使得二次型判别函数大大简化，并在辽宁中期晴雨判别 MOS 方案中成功地使用该模型。结果表明：二次型判别函数相比，确有一定的优越性。

2　二次型判别函数的数学模型

对于检验假设 $\Sigma_1 = \Sigma_2$ 是否成立，采用 G·E·P·BOX 提出的方案，取 BC^{-1} 统计量，且

$$B = (n-2)\ln|\Sigma| - (n_1-1)\ln|\Sigma_1| - (n_2-1)\ln|\Sigma_2| \tag{1}$$

$$C^{-1} = 1 - \frac{2P^2+3P-1}{6(P+1)}\left[\frac{1}{n_1-1} + \frac{1}{n_2-1} - \frac{1}{n-2}\right] \tag{2}$$

其中 n、n_1 和 n_2 分别表示总的和两类样本数 Σ、Σ_1 和 Σ_2 分别表示总的和两类样本协方差矩阵，P 为预报因子维数。采用信度 α，当 $BC^{-1} > X^2_{\frac{P}{2}}(\alpha)$ 时，拒绝假设 $\Sigma_1 = \Sigma_2$，必须使用二次型判别函数；否则接受假设 $\Sigma_1 = \Sigma_2$，可以使用线性判别函数。

据 Bayes 准则的判别函数为

$$D(X) = \frac{1}{2}\left[(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2) - (X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|}\right] \tag{3}$$

式(3)中，μ_1、μ_2 为两类样本的平均向量。

从式(3)可以看到，在 $\Sigma_1 \neq \Sigma_2$ 时，二次型判别函数比线性判别函数加 P 个平方项和 C_P^2 个乘积项，除各项判别系数的计算过程更加繁杂外，这些系数的显著性检验需要更多的时间才能完成。存在上述问题的主要原因在于各个预报因子之间不相互独立。文献〔3〕中提出，将预报因子向量 X 经过线性变换成为 Y，使 Y 和各个分量之间相互独立，即需求出变换矩阵 A，使

$$A^T\Sigma_2 A = \Lambda \tag{4}$$

$$A^T\Sigma_1 A = I \tag{5}$$

成立。其中 Λ 为对角矩阵，I 为单位矩阵。

若由 Σ_1 的特征值构成的对角矩阵为 Λ_1，与其对应的特征向量构成的矩阵 T_1。如果这些 P 个特征值是不等的，则 T_1 是正交矩阵，即有 $T_1^{-1} = T_1^T$，则

$$\Sigma_1 = T_1\Lambda_1 T_1^{-1} \tag{6}$$

对矩阵 Σ_2 作如下变换得到 $\tilde{\Sigma}_2$，有

$$\tilde{\Sigma}_2 = (T_1\Lambda_1^{-\frac{1}{2}})^T\Sigma_2(T_1\Lambda_1^{-\frac{1}{2}}) \tag{7}$$

$\tilde{\Sigma}_2$ 的各个特征值 $\lambda_1, \lambda_2, \cdots \lambda_P$，构成对角矩阵 Λ，其对应特征向量构成的矩阵为 $\tilde{T}_2$。则

$$\tilde{\Sigma}_2 = \tilde{T}_2\Lambda\tilde{T}_2^{-1} \tag{8}$$

若设 Λ 元素互不相等，则 $\tilde{T}_2$ 是正交矩阵，即有 $\tilde{T}_2^{-1} = \tilde{T}_2^T$。得到转换矩阵

$$A = T_1\Lambda_1^{-\frac{1}{2}}\tilde{T}_2 \tag{9}$$

可以证明矩阵 A 满足式(4)～(5)。

进行线性变换

$$Y = (y_i) = A^T(X-\mu_1) \tag{10}$$

并设

$$M = (m_i) = A^T(\mu_2 - \mu_1) \tag{11}$$

式(3)中各项可变换为

$$(X - \mu_1)^T\Sigma_1^{-1}(X - \mu_1) = \sum_{i=1}^{P} y_i^2 \tag{12}$$

$$(X - \mu_2)^T\Sigma_2^{-1}(X - \mu_2) = \sum_{i=1}^{P} \frac{1}{\lambda_i}(y_i - m_i)^2 \tag{13}$$

$$-\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} = \frac{1}{2}\sum_{i=1}^{P}\ln\lambda_i \tag{14}$$

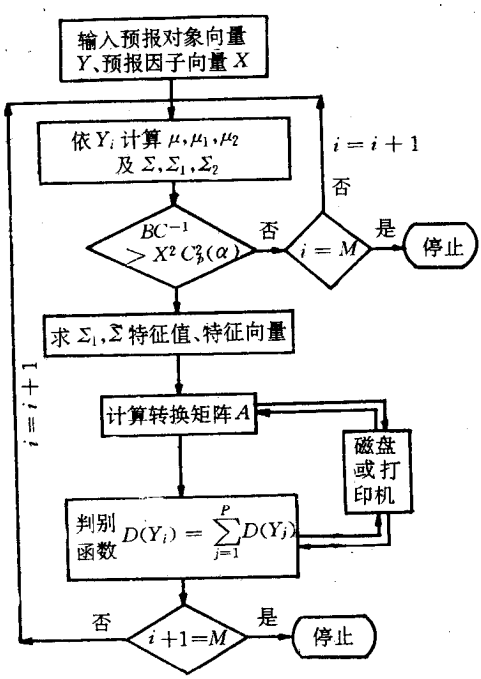
因此,式(3)可改写为

$$D(Y) = \frac{1}{2}\sum_{i=1}^{P}\left[\ln\lambda_i - y_i^2 + \frac{1}{\lambda_i}(y_i - m_i)^2\right] \tag{15}$$

对于上述两类判别问题,确定预报因子向量的重要性可用发散度

$$div(\text{Ⅰ},\text{Ⅱ}) = \sum_{i=1}^{P}(div)_i \tag{16}$$

其中,$(div)_i = \frac{1}{2}\left[m_i^2(1 + \frac{1}{\lambda_i}) + (\lambda_i + \frac{1}{\lambda_i} - 2)\right]$,$(div)_i$ 可表示 Y 中的第 i 个分量 y_i 对判别的重要性,若 y_i 依 $(div)_i$ 的大小进行排列,则可依据一定信度标准,取前面几个 y_i 建立二次型判别函数模型,从而实现对因子维数 P 的缩减。



附图　二次型判别函数计算流程

3　二次型判别函数计算流程

为使上述计算过程更加明确,本文给出二次型判别函数的计算流程图(附图),以更便于理解该模型和编制计算程序

4　辽宁中期晴雨二次型判别函数的应用

二次型判别函数为我们提供了一种较为简单可行的非线性判别分析方法,但如何选择预报因子使 $\Sigma_1 \neq \Sigma_2$ 的检验得以通过,是该方法能否应用的关键。我们采用统计量 F 初选因子,在众多预报因子中取前10个 F 值最大者,进入二次型判别函数计算。

$$F = \frac{\sum_{g=1}^{G} N_g(\bar{X}_{pg} - \bar{X}_p)^2/(G-1)}{\sum_{g=1}^{G}\sum_{k=1}^{Ng}(X_{pgk} - \bar{X}_{pg})^2/(L-G)} \tag{17}$$

式中,L 为样本数,N_g 为各类样本数,G 为类数。

我们将本模型用于辽宁省夏季逐站3～6天晴雨MOS判别预报中,建模资料为1990～1991年中7～9月的ECMWF资料,经过统计量 F 筛选后,所选取的预报因子均可建立二次型判别模型(一般 $BC^{-1} > 2.5X^2_{C_p^2}(0.05)$)。

由于当因子增加到一定程度时,增加因子并不能使判别效果提高,反而影响方程的稳定性。因此,我们选择当历史拟合率最高时的因子数为二次型判别函数的因子数,故方程的因子数为5～10个。

我们用相同的资料相应地建立了线性判别模型,附表给出辽宁省10个代表站即阜新、绥中、彰武、西丰、沈阳、营口、新宾、宽甸、丹东、大连的平均统计结果。

附表　两种判别模型对比结果

预报时效	第3天	第4天	第5天	第6天
因子数	8.8 (8.5)	9.1 (8.0)	9.4 (8.0)	8.2 (7.8)
准确率(%)	83.17 (80.12)	81.02 (80.00)	82.71 (80.57)	77.63 (79.44)

注:括号内数字为二次型判别函数结果。

从对比分析结果来看,在判别拟合效果相

差不大时，二次型函数判别模型所含因子数一般比线性判别函数所含因子数少近 1 个左右。另外，线性判别函数随着预报时效增加判别准确率下降，最大变差为 5.57%，而二次型判别函数则相互稳定，最大变差仅为 1.13%。

1992 年夏季，我们用二次型判别模型制做各地 3～4、5～6 天降水过程客观预报，经业务评定，其预报准确率分别达 63%和 60%。

5 结语与讨论

5.1 从上述计算可知，$\Sigma_1=\Sigma_2$ 的假设常常是不成立的。因此，通常的线性判别模型在实际中往往不能取得满意的结果，用二次型判别函数代替线性判别函数是必要的。

5.2 采取因子初选过程是必要的。一方面可以保障计算的稳定性，另一方面可以保证拒绝 $\Sigma_1=\Sigma_2$ 的假设，一般可采用 F 统计量。

5.3 虽然二次型判别函数计算比线性判别函数计算复杂，但由于计算条件的改善，这方面已不成问题。采用本方法在实际业务中是可行的。

5.4 在相同拟合前提下，二次型函数比线性函数所含因子数少，而且其判别效果比较稳定。

本文曾得到南京气象学院已故王得民教授生前的指导，在此深表敬意。

6 参考文献

1 吕纯濂等. Logistic 判别及其在气象上的应用. 南京气象学院学报，1982；1

2 施能等. 正交变换的二次判别方法及其在长江中下游旱涝预报中的应用. 南京气象学院学报，1983；1

3 中国科学院计算中心概率统计组. 概率统计计算. 北京：科学出版社. 1979：96