第 2 卷 第 2 期　　大 气 科 学　　Vol. 2, No. 2
1978 年 6 月　　SCIENTIA ATMOSPHERICA SINICA　　June, 1978

研究简报

最小方差准则的判别分析

曹 鸿 兴
（中央气象局气象科学研究所）

提　　要

从引入判别参数的概念出发，导出了在最小均方误差准则下的线性判别函数，证明了在二类问题中等价于 Fisher 准则的判别，且作了实例计算和讨论。

判别分析是解决在天气类别事先可划定的情况下确定预报对象属于何类的问题，一般可构造一个关于预报因子的函数，用历史资料定出这函数的参数，从而用它来判别未来天气的类别。构造这一函数的准则是可以不同的。目前气象中常用的是 Fisher 准则[1][6]，此外尚有 Bayes 准则、不确定性准则，Kullback 准则以及最小二乘准则[2][3]等，广义图象法[4]其实也是另一形式的判别。在天气预报中，趋势预报、定性预报、分类预报占着重要地位，为此，我们从另一角度来讨论判类问题，导出了在最小均方误差意义下的判别函数并作了相应的讨论。

一、原　　理

设由预报因子 $x_0, x_1, x_2, \cdots x_p$ 构成线性判别函数

$$F(\boldsymbol{X}) = \boldsymbol{U}^T\boldsymbol{X} = u_0x_0 + u_1x_1 + u_2x_2 + \cdots + u_px_p, \tag{1}$$

x_0 为恒等于 1 的单位因子，$\boldsymbol{X}$ 为 $p+1$ 维列向量，$\boldsymbol{U}$ 为待定系数列向量，T 表示转置。计有 N 个样本，预报因子的第 q 个样本的取值为

$$\boldsymbol{X}_q^T = (x_{0q}, x_{1q}, \cdots x_{pq}) \qquad (q = 1, 2, \cdots N)$$

相应，预报对象为 $y_q\ (q = 1, 2, \cdots N)$，根据预报要求依预报对象将总样本集划分成 G 类，这时样本资料相应被分成 G 组，以 $\boldsymbol{X}_{gn}$ $(g = 1, 2, \cdots G, n = 1, 2, \cdots N_g)$ 表示第 g 类的第 n 个样本的因子向量，N_g 为第 g 类的样本数。

构造泛函

$$J = \sum_{g=1}^{G}\sum_{n=1}^{N_g}[\boldsymbol{U}^T\boldsymbol{X}_{gn} - K_g]^2, \tag{2}$$

其中，K_g 为第 g 类的判别参数，在本文中视为给定常数。

我们的目的是使判别函数 F 具有这样的性能，对某类而言，它的每个样本的 F 值尽可能与该类的判别参数接近，也就是在最小方差意义下使 J 极小。由

$$\frac{\partial J}{\partial u_i} = 0 \qquad (i = 0, 1, 2, \cdots p)$$

1977 年 3 月 7 日收到修改稿。

求得正规方程

$$\boldsymbol{SU}=\boldsymbol{W} \tag{3}$$

$\boldsymbol{S}$ 为 $(p+1)\times(p+1)$ 矩阵，每一个元素

$$s_{ij}=s_{ji}=\sum_{g=1}^{G}\sum_{n=1}^{N_g}x_{ign}\,x_{jgn}\qquad(i,j=0,1,2,\cdots p)$$

x_{ign} 表示第 i 个因子在第 g 类中第 n 个样本的值，x_{jgn} 表示第 j 个因子在第 g 类中第 n 个样本的值。

$\boldsymbol{W}$ 为 $(p+1)\times 1$ 向量，每一个元素

$$w_i=\sum_{g=1}^{G}\sum_{n=1}^{N_g}K_g\,x_{ign}\qquad(i=0,1,\cdots p)$$

$\boldsymbol{U}$ 为 $(p+1)\times 1$ 待定系数向量。

求解(3)得判别函数，按下列规则进行判别，若

$$|F(\boldsymbol{X})-K_h|<|F(\boldsymbol{X})-K_g|\qquad(g=1,2,\cdots G,\ g\neq h) \tag{4}$$

则判定 $\boldsymbol{X}$ 样本属于第 h 类。

判别参数可经验给定或试算确定之。如取某类 y 值的平均值，即

$$K_g=\bar{y}_g,\quad \bar{y}_g=\frac{1}{N_g}\sum_{n=1}^{N_g}y_n;$$

或给以整数，如分四类，K_g 分别给定为 -3，-1，1，3。显然，前者是不等距的给法，后者是等距的。在多类判别中本方法之效果与给定判别参数的技巧有关。

如果将共有 N 个样本的总体分成 N 类，且令 $K_g=y_q$ $(g,q=1,2,\cdots N)$，则由(3)式求得的 $\boldsymbol{U}$ 就是回归系数，(1)式变为回归方程，因此原则上可以用逐步回归方法来进行逐步判别，以筛选因子。

用(3)求得的判别函数式只有一个，换句话说，得用一个判别函数来判定 G 类，当类别数增大时其效果显然会降低，因判别函数是一个线性连续函数，而判别参数是间断阶梯函数，用前者去逼近后者当 G 增大时越发困难。为此又建立了求 G 个判别函数的方法。

设对每一类都有一个线性判别函数

$$F_g=\boldsymbol{U}_g^{\mathrm{T}}\boldsymbol{X}=u_{g0}x_0+u_{g1}x_1+\cdots+u_{gp}x_p\qquad(g=1,2,\cdots G) \tag{5}$$

我们希望对 g 类而言，它的每个样本的 F 值与 g 类的判别参数 K_g 尽可能接近，而将其他类的样本资料代入(5)式后其 F 值与相应类别的判别参数 K_m $(m=1,2,\cdots G,\ m\neq g)$ 尽可能差得大，据此构造泛函

$$J_g=\frac{(G-1)\sum_{n=1}^{N_g}[\boldsymbol{U}_g^{\mathrm{T}}\boldsymbol{X}_{gn}-K_g]^2}{\sum_{\substack{m=1\\ m\neq g}}^{G}\sum_{n=1}^{N_m}[\boldsymbol{U}_g^{\mathrm{T}}\boldsymbol{X}_{mn}-K_m]^2} \tag{6}$$

上式如分子小，分母大则其值变小，为使 J_g 达极小，令

$$\frac{\partial J_g}{\partial u_{gi}}=0\qquad(g=1,2,\cdots G,\ i=0,1,2,\cdots p)$$

得正规方程

$$\boldsymbol{L}\boldsymbol{U}_g = \boldsymbol{B} \tag{7}$$

$\boldsymbol{L}$ 为 $(p+1)\times(p+1)$ 矩阵，其每一个元素为

$$l_{ij} = \sum_{n=1}^{N_g} x_{jgn}x_{jgn} - \mu\sum_{\substack{m=1\\ m\neq g}}^{G}\sum_{s=1}^{N_m} x_{imn}x_{jmn}$$

$\boldsymbol{B}$ 为 $(p+1)\times 1$ 向量，其每一个元素为

$$b_{ij} = K_g\sum_{n=1}^{N_g} x_{ign} - \mu\sum_{\substack{m=1\\ m\neq g}}^{G}\sum_{n=1}^{N_m} K_m x_{imn}$$

其中　　$\mu = \dfrac{J_g}{G-1}$。

由构造(6)的要求知，J_g 值界于 0 和 1 之间，所以，对实际资料来说，$0 < \mu < \dfrac{1}{G-1}$，当 $G = 3$ 时，$0 < \mu < 0.5$，故容易给定初值，譬如令 $\mu^{(0)} = \dfrac{1}{G}$，然后求解 (7) 式得 $\boldsymbol{U}_g^{(0)}$ 检验判别准确率，计算 $\mu^{(1)}$，再解(7)式得 $\boldsymbol{U}_g^{(1)}$，如此迭代数次视其判别准确率不能提高时停止。精确解则需用最优化方法求得，由于判别函数是个相对数，精确解一般来说并非必须的。

这样共求得 G 个判别函数，判别规则变为

$$|F_g(\boldsymbol{X}) - K_g| < |F_g(\boldsymbol{X}) - K_m| \qquad (g = 1,2,\cdots G; m = 1,2,\cdots G, m \neq g)$$

则判定 $\boldsymbol{X}$ 样本属于第 g 类。

二、二 类 判 别

在二类判别的情况下，(2)式变为

$$J = \sum_{n=1}^{N_A}\left[\sum_{i=0}^{p} u_i x_{in} - K_1\right]^2 + \sum_{n=1}^{N_B}\left[\sum_{i=0}^{p} u_i x_{in} - K_2\right]^2 \tag{8}$$

令 $K_1 \equiv y_1$，$K_2 \equiv y_2$，这就直接证明了最小方差准则的判别等价于预报量只取二个不同值的回归。

若取 $K_1 = K$，$K_2 = -K$，这时正规方程变为

$$\boldsymbol{S}\boldsymbol{U} = K\boldsymbol{D} \tag{9}$$

$\boldsymbol{S}$ 为 $(p+1)\times(p+1)$ 矩阵，它的每一个元素为

$$s_{ij} = s_{ji} = \sum_{n=1}^{N_A} x_{in}x_{jn} + \sum_{n=1}^{N_B} x_{in}x_{jn} \qquad i,j = 0,1,2,\cdots p \tag{9.1}$$

$\boldsymbol{D}$ 为 $(p+1)\times 1$ 向量，它的每一个元素为

$$d_i = \sum_{n=1}^{N_A} x_{in} - \sum_{n=1}^{N_B} x_{in} \qquad i = 0,1,2,\cdots p \tag{9.2}$$

由(9)式知，K 对解 $\boldsymbol{U}$ 不起实质性作用，故可令 $K_1=1$，$K_2=-1$，容易明白，判别规则(4)这时变为 $F>0$ 属A类，$F<0$ 属B类。

下面证明在二类问题中最小方差准则与 Fisher 准则判别的等价性。

Fisher 准则的泛函 I

$$I=\frac{[\bar{y}(A)-\bar{y}(B)]^2}{\sum_{n=1}^{N_A}[y_n(A)-\bar{y}(A)]^2+\sum_{n=1}^{N_B}[y_n(B)-\bar{y}(B)]^2}$$

预报判据为

$$y_C=\frac{N_A\bar{y}(A)+N_B\bar{y}(B)}{N_A+N_B} \tag{10}$$

式中 y 为判别函数，—表示对A或B类的平均。对(10)作线性变换

$$\bar{y}(A)=\bar{F}(A)+\left(1+\frac{N_B}{N_A}\right)y_C$$

$$\bar{y}(B)=\frac{N_A}{N_B}\bar{F}(B)$$

得

$$\bar{F}(A)=-\bar{F}(B) \tag{11}$$

用(11)按 Fisher 准则，I 可写为

$$I=\frac{[\bar{F}(A)-(-\bar{F}(A)]^2}{\sum_{n=1}^{N_A}[F_n(A)-\bar{F}(A)]^2+\sum_{n=1}^{N_B}[F_n(B)-(-\bar{F}(A))]^2}$$

$$=\frac{4K^2}{\sum_{n=1}^{N_A}[F_n(A)-K]^2+\sum_{n=1}^{N_B}[F_n(B)-(-K)]^2} \tag{12}$$

其中 $K\equiv\bar{F}(A)$，在(8)中令 $K_1=K$，$K_2=-K$；假定 F 为 $\boldsymbol{X}$ 的线性函数，则对(12)和对(8)求导数令其等于0，其结果是一样的。

如果把 $\boldsymbol{X}$ 资料0—1化，令事件出现 $K_1=1$，不出现 $K_2=0$，则由本法求得的判别式等价于事件概率回归。

三、实例和讨论

为了便于比较，用文[5]中的资料，作大同冬季24小时有(A类)无(B类)高云出现预报。

用原数据按(9.1)和(9.2)计算 $\boldsymbol{S}$ 阵和 $\boldsymbol{D}$ 向量，代入(9)式解之得判别函数

$$F=0.2249-0.1343x_1+0.05761x_2+0.4847x_3 \tag{13}$$

式中，x_1 为流场特征因子，x_2 为东胜500毫巴24小时变温，x_3 为酒泉与东胜500毫巴相对涡度之差。用(13)式算出全部样本的 F 值，以错判率最低确定 $F_C=0.17$，这样拟合率 $p_1=100/119=84\%$，试报准确率 $p_2=23/26=89\%$。

将原数据 0—1 化同理可求得判别函数

$$F = -0.3351 - 0.3749x_1 + 0.2463x_2 + 1.0718x_3$$

$F_C = -0.09$，$p_1 = 100/119 = 84\%$，$p_2 = 22/26 = 85\%$。而 Fisher 准则的判别[5]效果为 $p_1 = 83\%$，$p_2 = 85\%$。

由此可见，在实用上，为便于计算，完全可用 0—1 数据的判别分析，因其具有同等的预报效果。上述三者拟合和试报效果略有差异是因定 F_C 不同而引起的。单位因子 x_0 起坐标移动作用，也就是使 F_C 理论上等于 0，实际上位于 0 附近，手算时为减少方程维数可将 x_0 略去。

Fisher 准则、Bayes 准则等判别函数，一般仅依赖于资料的平均值、(协)方差等统计特性，就是说，就某种判别法而言，预报因子一旦选定，其判别效果也就确定了。在本文所叙述的方法中，判别函数还依赖于给定的判别参数，这样就提供了一种供选择的机会，即通过试算来挑选最优的判别参数使错判率最低。这是引进判别参数带来的好处，当然本文对如何选取判别参数未详加讨论。可以考虑选取随样本和类别改变的动参数 $K(\boldsymbol{X}_{gn}, g)$ 或某种函数形式的参数，从而提高判别效果。对二类判别来说，几种准则的线性判别函数都是等价的[2][3]，但最小方差准则的判别分析计算量要小一些。

判别函数可采取多种表达形式，如

$$F = \sum_{i=1}^{m} u_i p_i \quad 或 \quad F = \sum_{i=1}^{m} u_i H_i$$

式中，m 为因子数，p_i 为关于第 i 个因子的概率，$H_i = -p_i \ln p_i$ 为第 i 个因子的熵；自变量 x_i 也能用其他初等函数形式。

用本判别法作 1974 年新安江 7 月降水量的长期预报，分干旱、偏旱、正常、偏涝、雨涝五级预报，用冰岛低压、阿留申低压，北太平洋高压等大气活动中心一月海平面气压值、太阳黑子、地磁 c_i 指数、地球自转角速度、西风指数、上海一月气温、W型一月日数，上一年新安江汛期降水量等 17 个因子，结果把历史频率仅为 5/42 的雨涝报对了，是用逐步剔除法筛选因子的，其原理如下。

在第一次求得判别函数后，如第 R 个因子的系数 u_R 满足

$$u_R < \alpha \max_i u_i \qquad (i = 1, 2, \cdots m)$$

式中，α 为权因子，可取 0.05—0.25，则剔除该因子，将所有满足上式的因子剔除后重求判别系数，再根据上式要求进行新的剔除，如此重复到不再有满足上式的因子时停止，这一方法原理简单但计算量大。

参　考　资　料

[1] 王宗皓、李麦村等，天气预报中的概率统计方法，科学出版社，1974 年。

[2] 杨自强，判别分析与逐步判别，全国概率统计会议文件，1975 年。

[3] 中国科学院数学研究所，多元分析资料汇编(I)，1974 年。

[4] А. Н. Багров, В. Н. Ващик, А. И. Снитковский, К прогнозу опасных явлений погоды, *Мете. и Гидро.*, 1974, No. 3, pp. 35—43.

[5] 7310 部队司令部气象处，用分辨法作大同地区冬季高云预报，气象，1975 年，第 1 期。

[6] R. A. Fisher, The use of multiple mesurements in taxonomic problems, *Ann. Eugenics*, 1936, 7, pp. 179—188.