第44卷 第3期 气 象 学 报 Vol. 44, No. 3
1986 年 8 月 ACTA METEOROLOGICA SINICA Aug., 1986

二次 Logistic 判别分析及其气象应用*

吕纯濂 陈舜华

(南京气象学院)

提 要

在线性 Logistic 判别中，各母体的相关结构是相同的，但对理论研究和实际应用都很重要的是各母体的相关结构不相同的情况，这样就引出了二次 Logistic 判别的问题。若变量维数 p 不太小，如 $p>4$，则在完全 Logistic 判别中将出现太多的要计算的参数。利用拟牛顿迭代法解出由最大似然估计所得到的关于 Logistic 参数的超越方程组。并作出华北地区夏季旱涝预报的二次 Logistic 判别分析，以及北京 7 月上旬内最高气温大于 35°C 的日数不少于 2 d 与少于 2 d 的二次 Logistic 判别分析的实例。最后，还提出了在某些特殊情况下的近似方法。

一、引 言

判别分析方法，如 Fisher, Bayes 及逐步判别等，已在实标应用中广泛使用。严格地说，这些方法仅当变量为正态分布时才可应用。Logistic 判别在我国还较少应用[1,2]，它关于变量的基本假设条件较宽，尤其 Logistic 判别不仅可假设变量来自具有等协方差阵的多元正态分布，且变量可为相互独立或二值 0 — 1 变量，或一些是连续变量而另一些是多值离散变量，因而，对判别分析使用 Logistic 模型优点之一为其具稳健性(robust)，即对一些不同分布的变量都可得到相同的 Logistic 公式，故在实用中，对未经正态检验的变量应用本方法是可行的[1,2]。

线性 Logistic 判别是在把各母体中因子之间的相关结构认为是相同的假定下得到的[1,2]。但考虑把各因子之间的相关结构在各母体中认为是不同的，有时在理论研究或实际应用中都是很重要的，这就引出了二次 Logistic 判别函数；但它有过多的需要估计的参数，在一般情况下，当维数 p(因子数)大于 4 时，在计算上会产生一定的困难，故必须给出近似方法，使它比全部二次判别函数要估计的参数要少且又现实可行。

本方法最初由 D.R.Cox 提出设想[3]，后经 N.E.Day 和 D.F.Kerridge 发展(仅考虑两个母体)[4]，J.A.Anderson对此方法作出了重大的贡献[5-9](主要用于生物医学上)；本文在 J.A.Anderson 工作[8]的基础上使数学模型更为具体，以便能编出程序进行具体计算，并在我国首次将该方法应用于气象上，比线性 Logistic 判别及多级 Bayes 判别的效果都好。

* 本文于 1984 年 3 月 3 日收到，1984 年 12 月 10 日收到修改稿。

二、数 学 模 式

线性 Logistic 判别中，在母体 $H_i(i=1,\cdots,m)$ 上观测到 $X=(x_1,\cdots,x_p)'$ 的概率为

$$P(X|H_i)=a_i\exp\{-\frac{1}{2}(X-M_i)'A^{-1}(X-M_i)\}\phi(X),$$

其中 M_i 为对应于母体 H_i 的均值向量，为保证其概率性质，对任意函数 $\phi(X)$ 仅假定可积性与非负性足矣，a_i 为归一化常数，表各因子相关结构的 $p\times p$ 正定对称阵 A 与 i 无关，即对每个母体皆同；现假定 A 与 i 有关，即不同母体的 $A_i(i=1,\cdots,m)$ 不同，故有

$$P(X|H_i)=a_i\exp\{-\frac{1}{2}(X-M_i)'A_i^{-1}(X-M_i)\}\phi(X)。\tag{1}$$

由 Bayes 准则，已知先验概率 $P(H_i)\triangleq q_i(i=1,\cdots,m)$ 时，观测到 X 后，视后验概率

$$P(H_s|X)\triangleq P_{sx}=P(X|H_s)P(H_s)/\sum_{l=1}^{m}P(X|H_l)P(H_l),(s=1,\cdots,m)\tag{2}$$

中谁大，则判 X 属该相应之母体 H_s。

将(1)代入(2)，利用 $M'_lA_l^{-1}X=X'A_l^{-1}M_l$，令

$$\tilde{C}_l=\ln(a_lq_l)-\frac{1}{2}M'_lA^{-1}{}_lM_l(l=1,\cdots,m)$$

$$\alpha_s=A_s^{-1}M_s-Am_m^{-1}M_m,\ (\text{此时 }\alpha_m=0)$$

$$\alpha_{s0}=\tilde{C}_s-\tilde{C}_m=\ln\left(\frac{a_sq_s}{a_mq_m}\right)-\frac{1}{2}(M_s'A_s^{-1}M_s-M'_mA_m^{-1}M_m)\text{此时 }\alpha_{m0}=0)$$

$$\Omega_s=-\frac{1}{2}(A_s^{-1}-A_m^{-1}),\ (\text{此时 }\Omega_m=0),\ (s=1,\cdots,m-1),$$

则有后验概率为

$$P_{sx}=\frac{\exp\{X'\Omega_sX+X'\alpha_s+\alpha_{s0}\}}{1+\sum_{l=1}^{m-1}\exp\{X'\Omega_lX+X'\alpha_l+\alpha_{l0}\}},\ (s=1,\cdots,m)\tag{3}$$

对所有 $s=1,\cdots,m$ 上式分母皆同，故 P_{sx} 之大小只需看分子且只需看 e 的指数部分的大小，故得二次判别函数

$$\begin{cases}U_s(X)=X'\Omega_sX+X'\alpha_s+\alpha_{s0}=\sum_{i,j=1}^{p}\omega_{ij}^{(s)}x_ix_j+\sum_{j=1}^{p}\alpha_{sj}x_j+\alpha_{s0},(s=1,\cdots,m-1)\\ U_m(X)\equiv 0。\end{cases}\tag{4}$$

这里比 $(m-1)(p+1)$ 个线性 Logistic 判别系数 $\{\alpha_{sj}:s=1,\cdots,m-1;j=0,1,\cdots,p\}$ 多了 $(m-1)p(p+1)/2$ 个二次项系数 $\{\omega_{ij}^{(s)}:\ s=1,\cdots,m-1;i,\ j=1,\cdots,p\}(\omega_{ij}^{(s)}=\omega_{ji}^{(s)}$；当 p 较大时，在迭代运算中会产生麻烦，故用谱分解法以减少 Ω_s 中的二次参数的个数：

$$\Omega_s=\sum_{j=1}^{p}\lambda_{js}l_{js}l'_{js}, \tag{5}$$

其中 λ_{js}为 Ω_s 的特征值(按其模的大小排序，$|\lambda_{1s}|$最大)，l_{js}为相应的特 征向量，这样就可只取前几项近似表达 Ω_s，以减少二次参数的个数；最简 单的是当$|\lambda_{2s}/\lambda_{1s}|$很小时，只取一项，即

$$\Omega_s\approx\lambda_s l_s l_s',(\text{省略下标 1}), \tag{6}$$

此时二次参数仅$(m-1)p$ 个，其中 $l_s=(l_{s1},\ \cdots,l_{sp})'$，因$\sum_{j=1}^{p}l^2_{sj}=1,(s=1,\cdots,m-1)$,不便于计算，作变换，令

$$\mu_s=\operatorname{sgn}(\lambda_s)=\begin{cases}+1\ \text{当}\ \lambda_s>0,\\ -1\ \text{当}\ \lambda_s<0,\end{cases}\quad(s=1,\cdots,m-1) \tag{7}$$

$$d_{sj}=l_{sj}\sqrt{|\lambda_s|},\ (s=1,\cdots,\ m-1;\ j=1,\cdots,p),$$

则有

$$X'\Omega_s X\approx\lambda_s X'l_s l'_s X=\mu_s(d_s'x)^2=\mu_s\Big(\sum_{v=1}^{p}d_{sv}x_v\Big)^2,$$

因而判别函数为(取 $x_0\equiv1$)

$$\begin{cases}U_s(X)=\mu_s(d'_sX)^2+\alpha'_sX+\alpha_{s0}=\mu_s\Big(\sum_{v=1}^{p}d_{sv}x_v\Big)^2+\sum_{j=0}^{p}\alpha_{sj}x_j,(s=1,\cdots,m-1)\\ U_m(X)\equiv0;\end{cases} \tag{4}'$$

而后验概率为

$$P_{sx}=\exp\{U_s(X)\}/[1+\sum_{l=1}^{m-1}\exp\{U_l(X)\}],(s=1,\cdots,m), \tag{2}'$$

故需估计的参数由原来(4)中的$(m-1)(p+1)+(m-1)p(p+1)/2$ 个减少为$(m-1)\times(2p+1)$个。

从混合母体中随机独立地抽取$n=\sum_{s=1}^{m}n_s$个样本(n_s为从母体 H_s中抽得的样本数)，n 个样本总体似然函数为

$$\ln L=\ln\prod_{s=1}^{m}\prod_{x\in H_s}P(XH_s)=\text{常数}+\sum_{s=1}^{m}\sum_{x\in Hs}\ln P_{sx} \tag{8}$$

它通过 P_{sx}由(2)′与未知参数$\{\alpha_{sj}\}$及$\{d_{sj}\}$建立联系，用最大似然估计时，需先由(2)′对未知参数求偏导：

$$\begin{cases}\dfrac{\partial P_{sx}}{\partial\alpha_{sj}}=P_{sx}(1-P_{sx})x_j,\\ \dfrac{\partial P_{sx}}{\partial\alpha_{tj}}=-P_{tx}P_{sx}x_j,\ s\neq t,\end{cases}\quad\begin{pmatrix}s,t=1,\cdots,m-1;\\ j=0,\cdots,p;\end{pmatrix} \tag{9}$$

$$\begin{cases}\dfrac{\partial P_{sx}}{\partial d_{s,}}=2\mu_s(d'_sX)P_{sx}(1-P_{sx})x_{,}, \\ \dfrac{\partial P_{sx}}{\partial d_{t,}}=-2u_t(d'_tx)P_{tx}P_{sx}x_{,}, \end{cases} \quad \begin{pmatrix} s,t=1,\cdots,m-1; \\ j=1,\cdots,p; \end{pmatrix} \tag{10}$$

故由(8)、(9)、(10)而得$(m-1)(2p+1)$元超越方程组

$$\begin{cases}\dfrac{\partial \ln L}{\partial \alpha_{s,}}=\sum\limits_{x\in H_S}x_{,}-\sum\limits_{t=1}^{m}\sum\limits_{x\in H_t}P_{sx}x_{,}=0,(s=1,\cdots,m-1;\ j=0,\cdots,P) \\ \dfrac{\partial \ln L}{\partial d_{s,}}=2\mu_s\Big[\sum\limits_{x\in H_s}(d_s'X)x_{,}-\sum\limits_{t=1}^{m}\sum\limits_{x\in H_t}P_{sx}(d'X)x_{,}\Big]=0,(s=1,\cdots,m-1;j=1 \\ \qquad\qquad \cdots,\ p) \end{cases} \tag{11}$$

其中P_{sx}由(2)′给出，当观测到n个样本$X=(x_1,\cdots,x_p)'$后，解方程组(11)得$(m-1)(2p+1)$个未知参数$\{\alpha_{s,}\}$和$\{d_{s,}\}$，从而得二次判别函数(4)′，对某一欲判样本X代入(4)′，视m个$U_s(X)$值谁大，就判其属相应的母体H.——Logistic判别准则。

$\mu_s=\mathrm{sgn}(\lambda_s)=\pm1$的选取考虑如下：$\mu_s=\pm1,(s=1,\cdots,m-1)$，共有$2^{m-1}$组取法，其中仅一组正确，即使$\ln L$最大的那一组，这可在计算机上实现，$m$太大时计算上会有一定困难，但一般实际问题中母体数$m$不会太大，故是现实可行的。

用quasi-Newton迭代法求解$(m-1)(2p+1)$元超越方程组(11)的详细过程可见[10]。

三、气 象 应 用

为使更多微机用户使用本方法，我们在TRS-80微机上用Basic语言编制了程序，结合实标问题进行了计算和分析，现举二例如下：

1. 北京7月上旬内最高气温大于等于35℃的日数不少于2天与少于2天的二次Logistic判别分析

取1951—1970年的资料，$n=20$；因子选为

x_1：5月下旬到6月上旬内最高气温≥35℃的日数；

x_2：6月下旬内最高气温≥35℃的日数。

H_1：表示北京7月上旬最高气温≥35℃的日数少于2天的母体；$n_1=11$；

H_2：表示北京7月上旬最高气温≥35℃的日数不少于2天的母体；$n_2=9$。

当$\mu=+1$时，(8)式的对数似然函数$\ln L=-121.949$；当$\mu=-1$时，$\ln L=-122.028<-121.949$；(而线性Logistic判别的$\ln L=-121.97<121.949$)；故取使$\ln L$最大的$\mu=+1$以及对应的二次Logistic判别系数，得到$m-1=1$个二次Logistic判别函数：

$U(X)=6.3314-3.6822x_1-1.5368x_2+(0.0111x_1+0.0090x_2)^2$；判别矩阵为

计算分类 \ 原分类	1	2	小计
1	10	2	12
2	1	7	8
小计	11	9	20

拟合率为 17/20＝85%；原始数据、分类情况、判别函数值及后验概率列于表 1 。由表可见，判对的个例后验概率都比较大(如 1965，1968 年高达 1)，判错的个例后验概率都不太大(如 1963，1964 年都只有 0.54628)，这有助于我们对所判对象所作的判断的把握性程度的估计。

表 1

No.	X(1)	X(2)	U(1)	U(2)	D. C	C. C	P. P
I = 1	0.00	5.00	−1.3501E+00	0	1	2	0.79415
I = 2	0.00	0.00	6.3303E+00	0	1	1	0.99822
I = 3	0.00	0.00	6.3303E+00	0	1	1	0.99822
I = 4	1.00	1.00	1.1127E+00	0	1	1	0.75263
I = 5	0.00	0.00	6.3303E+00	0	1	1	0.99822
I = 6	0.00	1.00	4.7939E+00	0	1	1	0.99179
I = 7	1.00	0.00	2.6489E+00	0	1	1	0.93394
I = 8	0.00	1.00	4.7939E+00	0	1	1	0.99179
I = 9	1.00	1.00	1.1127E+00	0	1	1	0.75263
I =10	0.00	1.00	4.7939E+00	0	1	1	0.99179
I =11	0.00	0.00	6.3303E+00	0	1	1	0.99822
I =12	2.00	7.00	−1.1781E+01	0	2	2	0.99999
I =13	3.00	3.00	−9.3201E+00	0	2	2	0.99991
I =14	1.00	4.00	−3.4950E+00	0	2	2	0.97054
I =15	3.00	1.00	−6.2490E+00	0	2	2	0.99807
I =16	2.00	0.00	−1.0322E+00	0	2	2	0.73735
I =17	0.00	4.00	1.8563E−01	0	2	1	0.54628
I =18	0.00	4.00	1.8563E−01	0	2	1	0.54628
I =19	6.00	7.00	−2.6498E+01	0	2	2	1.00000
I =20	3.00	5.00	−1.2391E+01	0	2	2	1.00000

2．华北地区夏季旱涝的二次 Logistic 判别分析

华北地区夏季旱涝趋势的预报是关系到该地区农业生产、人民生活的一个重大课题。我们选择了太原、济南、天津三地 6—8 月总降水量为预报对象，分为旱、平、涝 三个 母体 ($m=3$)：

H_1(旱)：$R_{6-8}<999$(mm)； $n_1=8$；

H_2(平)：$999\leqslant R_{6-8}<1401$(mm)； $n_2=7$；

H_3(涝)：$R_{6-8}\geqslant 1401$(mm)； $n_3=7$。

用 1955—1976 年共 22 年的资料($n=22$)，选用两个因子($p=2$)：

x_1：前一年 7 月亚洲地区 500 hPa 月平均环流指数 I_2；

x_2：前一年 11 月 500 hPa 西太平洋月平均副高面积指数。

$\mu_1=+1$， $\mu_2=+1$： $\ln L=-233.397$；

$\mu_1=+1$， $\mu_2=-1$： $\ln L=-233.283$(最大)；

$\mu_1=-1$， $\mu_2=+1$： $\ln L=-233.625$；

$\mu_1=-1$， $\mu_2=-1$： $\ln L=-233.553$；

线性 Logistic 判别： $\ln L=-233.434$；

故选使 lnL 最大的 $\mu_1=+1$， $\mu_2=-1$ 及对应的二次 Logistic判别系数，而得 $m-1=2$个二次 Logistic 判别函数：

$$U_1(X)=46.4031-51.2641\,x_1-1.0318\,x_2+(-0.0458\,x_1+0.0228\,x_2)^2;$$

$$U_2(X)=39.1854-41.2654\,x_1-0.8742\,x_2-(0.0218\,x_1+0.0077\,x_2)^2。$$

判别矩阵为

计算分类 \ 原分类	1	2	3	小计
1	7	1	0	8
2	0	6	1	7
3	1	0	6	7
小计	8	7	7	22

拟合率为 19/22≅86.4%；而用 5 个因子的线性 Logistic判别对同样的问题用同样的资料得拟合率 82%[1]，可见用二次 Logistic 判别不仅减少了因子个数，而且还提高了拟合率，而 Bayes 多级逐步判别的拟合率仅为 77%[1]，这更说明了我们的方法的优越性；原始数据、分类情况、判别函数值及后验列于表 2；表中后验概率是指计算分类的后验概率，它是分属三个母体的后验概率中最大的一个，另二个后验概率未列出。

四、讨 论

近似式(6)$\Omega_s\approx\lambda_{1s}l_{1s}l'_{1s}$并非适用所有情况，若需要，可从 Ω_s的谱分解(5)式中取更多的项，如取二项

$$\Omega_s\approx\lambda_{1s}l_{1s}l_{1s}'+\lambda_{2s}l'_{2s}l_{2s};$$

有时可先把 Ω_s分解出一个对角阵 D_s，再进行谱分解，即

$$\Omega_s\approx D_s+\lambda_s l_s l'_s;$$

也可以使

$$\Omega_s\approx\lambda_s[l_s l'_s-\mathrm{diag}(l_s l_s')];$$

甚至直接按(3)式 $\Omega_s=-\frac{1}{2}(A_s^{-1}-A_m^{-1})$，将 $A_s(s=1,\cdots,m)$作为第 s 个母体 H_s中各

表 2

No.	X(1)	X(2)	U(1)	U(2)	U(3)	D. C	C. C	P. P
I = 1	0.57	6.00	1.0999 E +01	1.0414 E +01	0	1	1	0.64210
I = 2	0.54	10.00	8.4347 E +00	8.1519 E +00	0	1	1	0.57017
I = 3	0.35	14.00	1.4096 E +01	1.2491 E +01	0	1	1	0.83275
I = 4	0.48	0.00	2.1793 E +01	1.9375 E +01	0	1	1	0.91824
I = 5	0.77	10.00	−3.3622 E +00	−1.3395 E +00	0	1	3	0.77122
I = 6	0.40	3.00	2.2801 E +01	2.0053 E +01	0	1	1	0.93979
I = 7	0.47	12.00	9.9801 E +00	9.2898 E +00	0	1	1	0.66603
I = 8	0.55	9.00	8.9461 E +00	8.6145 E +00	0	1	1	0.58209
I = 9	0.70	10.00	2.2808 E −01	1.5492 E +00	0	2	2	0.67602
I =10	0.56	15.00	2.3012 E +00	2.9484 E +00	0	2	2	0.63456
I =11	0.55	12.00	5.8761 E +00	5.9884 E +00	0	2	2	0.52735
I =12	0.47	9.00	1.3049 E +01	1.1916 E +01	0	2	1	0.75646
I =13	0.66	9.00	3.3044 E +00	4.0752 E +00	0	2	2	0.67585
I =14	0.66	11.00	1.2560 E +00	2.3245 E +00	0	2	2	0.69379
I =15	0.78	0.00	6.4173 E +00	6.9957 E +00	0	2	2	0.64034
I =16	0.79	7.00	−1.3098 E +00	4.6083 E −01	0	3	2	0.55525
I =17	0.72	17.00	−7.9463 E +00	−5.4067 E +00	0	3	3	0.99518
I =18	0.70	14.00	−3.8620 E +00	−1.9532 E +00	0	3	3	0.85996
I =19	0.68	17.00	−5.8936 E +00	−3.7560 E +00	0	3	3	0.97453
I =20	0.65	16.00	−3.3359 E +00	−1.6418 E +00	0	3	3	0.81352
I =21	0.88	10.00	−9.0039 E +00	−5.8788 E +00	0	3	3	0.99709
I =22	0.52	25.00	−5.7915 E +00	−4.1661 E +00	0	3	3	0.98177

因子间的样本相关阵而得 Ω_s，当然这仅在正态分布或近似正态分布的情况下才合理；也可把前述方法混合起来考虑，如把(4)式变为

$$U_s(X)=\lambda_s X'\Omega_s X+\alpha'_s X+\alpha_{s0},\quad (s=1,\cdots,m-1)$$

其中 α_{s0}，α'_s及 $\lambda_s(s=1,\cdots,m-1)$用正文中的迭代最大似然函数的方法估计，而 $\Omega_s(s=1,\cdots,m-1)$由本节中任一方法得到。

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，1，112—123，1982。

[2] 吕纯濂等，Logistic 判别及其应用(I)、(II)，数学的实践与认识，1983年 第 2，3 期，58—64，59—66。

[3] Cox, D.R., Some procedures associated with the logistic qualitative response curve. In Research Papers on Statistcs: Festschrift for J. Neyman, Ed. F. N. David, 55—71. Chichester: Wiley, 1966。

[4] Day, N.E., and D.F. Kerridge, A general maximum likelihood discrimination, *Biometrics*, **23**, 313—323, 1967.

[5] Anderson, A. A., Seperate sample logistic discrimination, *Biometrika*, **59**, 19—35, 1972.

[6] Anderson, J.A., Logistic discrimination with medical applications. In Discriminant Analysis and

Applications. T. Cacoullos (ed.), 1—15. New York: Academic Press. 1973.
[7] Anderson, J.A., Diagnoisis by logistic discriminant function: Further practical problems and results, *Appl. Statist*, **23**, 397—404, 1974.
[8] Anderson, J.A., Quadratic logistic discrimination, *Biometrika*, **62**, 149—154, 1975.
[9] Albert, A., & J.A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, **71**, 1—10, 1984.
[10] 吕纯濂等，用拟牛顿 (quasi-Newton) 迭代法解二次 Logistic 判别系数的超越方程组，数值计算与计算机应用，(未发表)。

QUADRATIC LOGISTIC DISCRIMINANT ANALYSIS AND ITS APPLICATIONS IN METEOROLOGY

Lu Chunlian Chen Shunhua

(Nanjing Institute of Meteorology)

Abstract

In the linear logistic discrimination, the dependence structures of various populations are the same. But it is important for theoretical researchs and practical applications that the dependence structures are different in various populations, thus the quadratic discrimination is produced in the paper. The full quadratic logistic discriminant approach has too many parameters to be estimated if the dimensionality, p, is not small, say $p>4$. An approximation is suggested here which gives a quadratic term in the discriminant function but with a greatly reduced number of parameters. A quasi-Newton iterative computation is used to solve the equations from maximum likelihood estimation. Examples of the prediction of the drought or flood in North China, and in which the number of days when the most high temperature 35°C in the 1st decade of July in Beijing is less than 2 or not, are provided. Other approximations are suggested in some special cases.