应用概率统计 第七卷
第三期 1991年8月

Chinese Journal of Applied Probability
and Statistics Vol. 7. No. 3. Aug 1991

多母体判别中理论错判率的估计

孙 尚 拱
(北京医科大学)

摘 要

在m个正态等协方差母体条件下，本文把多重积分的理论错判率公式转化为求矩阵特征根及特征向量问题，讨论了公式适用的条件，进而找出小样本时估计理论错率的公式.

§1. 理 论 公 式

设有m个母体H_r, $r=1, 2, \cdots, m$, $X=(x_1, x_2, \cdots, x_p)'$, X在H_r中的分布为$N_p(M_r, \Sigma)$, 先验率为q_r. 令

$$u_{jk}(X)=\left[X-\frac{1}{2}(M_j+M_k)\right]'\Sigma^{-1}(M_j-M_k) \tag{1.1}$$

$$R_j=\left\{X;\ u_{jk}(x)\geqslant \ln\frac{q_k}{q_j},\ k=1, 2, \cdots, m,\ k\neq j\right\} \tag{1.2}$$

$$j=1, 2, \cdots, m,\ k=1, 2, \cdots, m,\ j\neq k.$$

对于指定的X向量，我们知道多母体Bayes判别准则为：

$$\text{当 } X\in R_j \text{ 时判 } X \text{ 属于 } H_j \tag{1.3}$$

记

$$\delta_{ji}=M_j-M_i,\ \Delta_{ji}^2=\delta_{ji}'\Sigma^{-1}\delta_{ji} \tag{1.4}$$

$$\Delta_{j,k,i}=\delta_{jk}'\Sigma^{-1}\delta_{ji} \tag{1.5}$$

$$M_{(j)}=\frac{1}{2}(\Delta_{j1}^2, \Delta_{j2}^2, \cdots, \Delta_{j,j-1}^2, \Delta_{j,j+1}^2, \cdots, \Delta_{jm}^2)' \tag{1.6}$$

$$U_{(j)}(X)=(u_{j1}(x), \cdots, u_{j,j-1}(x), u_{j,j+1}(x), \cdots, u_{jm}(x))'$$

$$V_{(j)}=\begin{pmatrix} \Delta_{j1}^2, & \Delta_{j,1,2}, & \cdots, & \Delta_{j,1,j-1}, & \Delta_{j,1,j+1}, & \cdots, & \Delta_{j,1,m} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \Delta_{j,j-1,1}, & \Delta_{j,j-1,2}, & \cdots, & \Delta_{j,j-1}^2, & \Delta_{j,j-1,j+1}, & \cdots, & \Delta_{j,j-1,m} \\ \Delta_{j,j+1,1}, & \Delta_{j,j+1,2}, & \cdots, & \Delta_{j,j+1,j-1}, & \Delta_{j,j+1}^2, & \cdots, & \Delta_{j,j+1,m} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \Delta_{j,m,1}, & \Delta_{j,m,2}, & \cdots, & \Delta_{j,m,j-1}, & \Delta_{j,m,j+1}, & \cdots, & \Delta_{j,m}^2 \end{pmatrix}$$

$$=(\delta_{j1}, \cdots, \delta_{j,j-1}, \delta_{j,j+1}, \cdots, \delta_{jm})'\Sigma^{-1}(\delta_{j1}, \cdots, \delta_{j,j-1}, \delta_{j,j+1}, \cdots, \delta_{j,m}) \tag{1.7}$$

显然$u_{(j)}(X)$是$m-1$维列向量，当参数已知时，Anderson (1958) 的§6.7已得出：

本文1988年3月22日收到，1989年11月17日收到修改稿.

$$U_{(j)}(X)\sim N_{m-1}(M_{(j)},\ V_{(j)}),\ j=1,\ 2,\ \cdots,\ m \tag{1.8}$$

记

$$U^0_{(j)}=\left(\ln\frac{q_1}{q_j},\ \cdots,\ \ln\frac{q_{j-1}}{q_j},\ \ln\frac{q_{j+1}}{q_j},\ \cdots,\ \ln\frac{q_m}{q_j}\right)' \tag{1.9}$$

当先验率全相同，则 $U^0_{(j)}=(0,\ 0,\ \cdots,\ 0)'$，由 [1] 的 § 6.7 还知，$H_j$ 中样品被正确地分类入 H_j 中去的概率 $P(j/j)$ 为：

$$P(j/j)=\int_{u^0_{(j)}}^{+\infty}N_{m-1}(M_{(j)},\ V_{(j)})du_{(j)} \tag{1.10}$$

上式是一个 $m-1$ 重无穷积分，积分下限为(1.9)向量，被积函数为

$$N_{m-1}(M_{(j)},\ V_{(j)})du_{(j)}=(2\pi)^{-\frac{m-1}{2}}|V_{(j)}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(u_{(j)}-M_{(j)})'V^{-1}_{(j)}(u_{(j)}-M_{(j)})\right]du_{(j)}$$

作积分变换，令

$$Z_{(j)}=V^{\frac{-1}{2}}_{(j)}(u_{(j)}-M_{(j)}), \tag{1.11}$$

则

$$P(j/j)=\int_{Z^0_{(j)}}^{+\infty}N_{m-1}(O,\ I)dZ_{(j)} \tag{1.12}$$

(1.12)中下限为

$$Z^0_{(j)}=V^{\frac{-1}{2}}_{(j)}(u^0_{(j)}-M_{(j)})\triangleq(v_1^{(j)},\ \cdots,\ v_{m-1}^{(j)})' \tag{1.13}$$

当 $Z^0_{(j)}$ 用分量表示后，显然有

$$P(j/j)=\prod_{i=1}^{m-1}[1-\Phi(v_i^{(j)})],\quad j=1,\ 2,\ \cdots,\ m \tag{1.14}$$

其中 $\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left(-\frac{y^2}{2}\right)dy$

下面求 $v_i^{(j)}$ 的表达式：

记 $V_{(j)}$ 的特征根为 $\lambda_1^{(j)},\ \cdots,\ \lambda_{m-1}^{(j)}$，对应的标准化特征向量为 $\alpha_1^{(j)},\ \cdots,\ \alpha_{m-1}^{(j)}$，分量表示为 $\alpha_i^{(j)}=(\alpha_{1i}^{(j)},\ \cdots,\ \alpha_{m-1,i}^{(j)})'$，令

$$\Lambda^{\frac{-1}{2}}_{(j)}=\mathrm{Diag}\left(\frac{1}{\sqrt{\lambda_1^{(j)}}},\ \cdots,\ \frac{1}{\sqrt{\lambda_{m-1}^{(j)}}}\right)$$

$$\alpha^{(j)}=(\alpha_1^{(j)},\ \cdots,\ \alpha_{m-1}^{(j)})$$

则显然有

$$V^{\frac{-1}{2}}_{(j)}=\alpha^{(j)}\Lambda^{\frac{-1}{2}}\alpha^{(j)\prime} \tag{1.15}$$

展开(1.15)，利用(1.6)及(1.9)，得

$$v_i^{(j)}=\sum_{l=1}^{m-1}\sum_{k=1}^{m-1}\alpha_{il}^{(j)}\alpha_{kl}^{(j)}C_k^{(j)}/\sqrt{\lambda_l^{(j)}} \tag{1.16}$$

$$j=1,\ 2,\ \cdots,\ m,\ i=1,\ 2,\ \cdots,\ m-1.$$

其中

$$C_k^{(j)}=\begin{cases}\ln\dfrac{q_k}{q_j}-\dfrac{1}{2}\Delta^2_{jk} & 1\leqslant k<j\\ \ln\dfrac{q_{k+1}}{q_j}-\dfrac{1}{2}\Delta^2_{j,k+1} & j\leqslant k\leqslant m-1\end{cases}$$

把(1.16)代入(1.14)，即得理论分类符合率的 m 个值.

当 $m=2$ 时，由于 $M_{(1)}=\Delta^2_{21}/2$，$V_{(1)}=\Delta^2_{21}$，所以

$$Z^0_{(1)}=\frac{1}{\Delta_{21}}\left(\ln\frac{q_2}{q_1}-\frac{\Delta^2_{21}}{2}\right)$$

当 $q_1=q_2$ 时，上式为

$$Z^0_{(1)}=-\Delta_{21}/2.$$

这时代入(1.14)得 $P(1/1)=1-\Phi(Z^0_{(1)})$，这是早就已知的公式.

§2. 不能应用错判率公式的情况

公式 (1.14) 是建立在 $X=(x_1,\cdots,x_p)'$ 在m个母体中是正态等协方差 Σ 且 Σ 逆阵存在的基础上，除了这些基本条件外，从(1.10)及(1.11)可见，还必须假定协方差阵 $V_{(j)}$ 是正定的，因此当 $V_{(j)}$ 具有零特征根时，公式(1.14)显然不成立. 而零特征根的出现必然是下面三种情况之一.

1. $p<m-1$，即变量数少于母体数减1时不能用 (1.14) 式. 因为这时 $\mathrm{rank}(V_{(j)})=\mathrm{rank}(\Sigma^{-1})=p$，所以 $V_{(j)}$ 中至少有 $m-1-p$ 个零特征根.

2. 当一个母体的均数(不妨设为第一个母体的均数 M_1)可由另外母体均数的加权线性表示时，即当存在常数 $\{C_i\}$ 使

$$M_1=\sum_{i=2}^{m}C_iM_i,\quad 而\quad \sum_{i=2}^{m}C_i=1 \tag{2.1}$$

则 $V_{(1)}$ 必有零特征根，反之也成立(设 $p\geqslant m-1$).

证　如果 $V_{(1)}$ 有零特征根，这时

$$\mathrm{rank}(V_{(1)})=\mathrm{rank}(\delta_{12},\delta_{13},\cdots,\delta_{1m})<m-1.$$

于是得 $m-1$ 个均数差 $\{\delta_{12},\cdots,\delta_{1m}\}$ 中至少有一个 (比如为 δ_{12}) 可由另外均数差线性表示，即存在常数 $\{l_i\}$，使下式成立：

$$\delta_{12}=\sum_{i=3}^{m}l_i\delta_{1i},\quad 即\quad M_1-M_2=\sum_{i=3}^{m}l_i(M_1-M_i)$$

令

$$C_i=\begin{cases}\dfrac{1}{1-\sum\limits_{j=3}^{m}l_j} & i=2\\ -l_i\Big/\left(1-\sum\limits_{j=3}^{m}l_j\right) & i\geqslant 3\end{cases}$$

则上式即为(2.1).

反之，如(2.1)成立，显然有

$$\sum_{i=2}^{m}C_i\delta_{1i}=0$$

于是 $\{\delta_{12},\delta_{13},\cdots,\delta_{1m}\}$ 线性相关，即 $V_{(1)}$ 为奇异.

3. 变量$(x_1,\cdots,x_p)$中如存在一个变量 (不妨设为 x_1)，它在每一个母体内的均数都可以由该母体内另外变量的均数线性表示，即存在常数 $\{C_i\}$ 使下式对一切母体成立：

$$M_{k1}=C_0+\sum_{j=2}^{p}C_jM_{kj},\quad k=1,2,\cdots,m \tag{2.2}$$

其中均数向量的分量表示式为 $M_k=(M_{k1},\cdots,M_{kp})'$，则 $V_{(j)}$(此处为 $V_{(1)}$) 必有零特征根，反之亦然.

证　设 $V_{(1)}$ 奇异，由于

$$\delta \triangleq (\delta_{12}, \cdots, \delta_{1m}) = \begin{pmatrix} M_{11}-M_{21}, & M_{11}-M_{31}, & \cdots, & M_{11}-M_{m1} \\ M_{12}-M_{22}, & M_{12}-M_{32}, & \cdots, & M_{12}-M_{m2} \\ \cdots\cdots & & & \\ M_{1p}-M_{2p}, & M_{1p}-M_{3p}, & \cdots, & M_{1p}-M_{mp} \end{pmatrix}$$

要使 $\mathrm{rank}(\delta)<m-1$（设 $p\geqslant m-1$），则 δ 中至少有一个行向量（不妨设为第一行）是另外行向量的线性组合，即存在常数 $C_2, \cdots, C_p$ 使下式成立：

$$M_{11}-M_{i1}=\sum_{j=2}^{p} C_j(M_{1j}-M_{ij}), \quad i=2, 3, \cdots, m$$

此即

$$M_{11}-\sum_{j=2}^{p} C_j M_{1j}=M_{i1}-\sum_{j=2}^{p} C_j M_{ij} \quad i\geqslant 2$$

令

$$C_0=M_{11}-\sum_{j=2}^{p} C_j M_{1j} \tag{2.3}$$

显然得

$$M_{i1}=C_0+\sum_{j=2}^{p} C_j M_{ij}, \quad i=1, 2, \cdots, m$$

此即(2.2)式.

反之，如果有一组常数$(C_0, C_2, \cdots, C_p)$使(2.2)成立. 取 $k=1$，则 C_0 如 (2.3) 所示. 由此很易得 $\mathrm{rank}(\delta)<m-1$.

§3. 理论分类符合率的样本估计

第一节中公式仅适用于大样本. 由大样本得出比较可信的参数估计再用公式. 而对于中小样本，下面给出参数的无偏估计.

记 H_r 母体中抽得的样本为 $\{X_1^{(r)}, \cdots, X_{n_r}^{(r)}\}$, $r=1, \cdots, m$. 令 $N=n_1+\cdots+n_m$, $X_i^{(r)}\sim N_p(M_r, \Sigma)$,

$$d_{ji}=\overline{X}^{(j)}-\overline{X}^{(i)}, \quad L=\sum_{r=1}^{m}\sum_{k=1}^{n_r}(X_k^{(r)}-\overline{X}^{(r)})(X_k^{(r)}-\overline{X}^{(r)})' \tag{3.1}$$

其中 $\overline{X}^{(r)}$ 为 H_r 中样本均数向量，令 $S=L/(N-m)$，显然 $\overline{X}^{(r)}$ 及 S 分别是 M_r 及 Σ 的无偏估计，令

$$D_{lf}^2=d_{lf}'s^{-1}d_{lf} \quad l, f=1, 2, \cdots, m, l\neq f \tag{3.2}$$

$$D_{j,i,k}=d_{ji}'s^{-1}d_{jk}, \quad i, j, k=1, 2, \cdots, m, i\neq j\neq k \tag{3.3}$$

我们知道

$$d_{lf}\sim N_p(\delta_{lf}, C_{lf}^{-1}\Sigma), \quad L\sim W_p(K, \Sigma)$$

其中 $C_{lf}^{-1}=(n_l+n_f)/n_l n_f$, $K=N-m$, W 为 Wishart 分布，且 d_{lf} 与 s^{-1} 独立，令

$$T_{lf}^2=C_{lf}K d_{lf}'L^{-1}d_{lf}=C_{lf}D_{lf}^2 \tag{3.4}$$

由[2]中的公式(86.2.13)，得

$$\frac{K-p+1}{p}\cdot\frac{T_{lf}^2}{K}=F(p, K-p+1, C_{lf}\Delta_{lf}^2) \tag{3.5}$$

此处 F 为非中心 F 分布，以 $K=N-m$ 代入得

$$D_{lf}^2=\frac{(N-m)p(n_l+n_f)}{(N-m-p+1)n_l n_f}F\left(p, N-m-p+1, \frac{n_l n_f}{n_l+n_f}\Delta_{lf}^2\right) \tag{3.6}$$

由于非中心分布 $F(p, q, \lambda)$ 的均值为 $q/(q-2)+q\lambda/p(q-2)$，所以

$$E(D_{ij}^2)=\frac{(N-m)p(n_i+n_j)}{(N-m-p-1)n_in_j}+\frac{N-m}{N-m-p-1}\Delta_{ij}^2 \tag{3.7}$$

于是得 Δ_{ij}^2 的无偏估计量(记为 $\hat{\Delta}_{ij}^2$)为

$$\hat{\Delta}_{ij}^2=\frac{N-m-p-1}{N-m}D_{ij}^2-\left(\frac{1}{n_i}+\frac{1}{n_j}\right)p \tag{3.8}$$

现求 $\Delta_{j,i,k}$ 的无偏估计($k\neq i\neq j$)，利用

$$E(s^{-1})=\frac{N-m}{N-m-p-1}\Sigma^{-1} \tag{3.9}$$

(见[5]中 p. 74 上的定理 2.4.6)，可证得：

$$E(d_{ji}d'_{jk})=\Sigma/n_j+\delta_{ji}\delta'_{jk},\quad i\neq k\neq j \tag{3.10}$$

由此得

$$\begin{aligned}E(D_{j,i,k})&=E(d'_{jk}s^{-1}d_{ji})=E\,\mathrm{tr}(d'_{jk}s^{-1}d_{ji})=E\,\mathrm{tr}(s^{-1}d_{ji}d'_{jk})\\&=\mathrm{tr}(E(S^{-1})\cdot E(d_{ji}d'_{jk}))=\mathrm{tr}\left\{\frac{N-m}{N-m-p-1}\Sigma^{-1}\cdot(\Sigma/n_j+\delta_{ji}\delta'_{jk})\right\}\\&=\mathrm{tr}\left\{\frac{N-m}{N-m-p-1}\Sigma^{-1}\delta_{ji}\delta'_{jk}\right\}+\mathrm{tr}\left\{\frac{N-m}{N-m-p-1}\cdot\frac{I}{n_j}\right\}\end{aligned}$$

即得

$$E(D_{j,i,k})=\frac{N-m}{N-m-p-1}\Delta_{j,i,k}+\frac{N-m}{N-m-p-1}\cdot\frac{p}{n_j} \tag{3.11}$$

于是得 $\Delta_{j,i,k}$ 的无偏估计(记为 $\hat{\Delta}_{j,i,k}$) 为：

$$\hat{\Delta}_{j,i,k}=\frac{N-m-p-1}{N-m}D_{j,i,k}-\frac{p}{n_j} \tag{3.12}$$

$$j\neq i\neq k,\ i=j=k=1, 2, \cdots, m.$$

由(3.8)、(3.12)即可得公式(1.7)中 $V_{(j)}$ 的无偏计量，记为 $\hat{V}_{(j)}$. 但这个 $\hat{V}_{(j)}$ 却不一定是正定，所以很自然的令

$$\hat{\Delta}_{ij}^2=\begin{cases}\dfrac{N-m-p-1}{N-m}D_{ij}^2-\left(\dfrac{1}{n_i}+\dfrac{1}{n_j}\right)p, & \text{当 } \hat{V}_{(j)} \text{ 正定时}\\[2ex]\dfrac{N-m-p-1}{N-m}D_{ij}^2, & \text{当 } \hat{V}_{(j)} \text{ 非正定时}\end{cases} \tag{3.13}$$

而对于 $\hat{V}_{(j)}$ 中非对角线元素还是用(3.12)为好，除非(3.13)取代后仍然使 $\hat{V}_{(j)}$ 非正定. 我们可以把(3.12)中的尾部去掉，这时 $\hat{V}_{(j)}$ 应该是非负定的.

用(3.8)及(3.12)可见，直接用 $D_{j,i,k}$ 及 D_{ij}^2 代替 $\Delta_{j,i,k}$ 及 Δ_{ij}^2 则估计出来的正确符合率总会高于理论符合率.

例：数值例(取自[6]中 p. 226).

表 1 数值例样本均数

	x_1	x_2	x_3	x_4
母 体 H_1	−14.4286	−17.3429	12.7143	31.1429
母 体 H_2	0.80	−17.4250	17.5000	0
母 体 H_3	−6.65	−17.3333	20.1667	−15

此例中 $n_1=7$, $n_2=4$, $n_3=6$, $N=17$, $p=4$, $q_1=7/17$, $q_2=4/17$, $q_3=6/17$. 四个变量的样本均数见表1. 采用 Bayes 逐步判别，规定临界值 $F_1=F_2=2.0$ 时，选中二个变量(x_2, x_4)，用它建立判别函数时其样本回代仅一例错判，使用刀切法则三个母体中各有一例被错判.

当估计理论符合率时，$\hat{V}_{(1)}$, $\hat{V}_{(2)}$, $\hat{V}_{(3)}$ 的特征根中都有一个负值，所以采用公式(3.13)修正. 计算得(1.16)值为:

$$\{v_i^{(1)}\}=\{43.6876,\ -30.7857\}$$
$$\{v_i^{(2)}\}=\{-25.1123,\ -47.1727\}$$
$$\{v_i^{(3)}\}=\{-18.5118,\ 50.1212\}$$

于是由(1.14)得理论符合率 $\{P(j/j)\}$ 的估计分别为

$$\{P(j/j)\}=\{0,\ 1,\ 0\}$$

此三值与拟合回代法(仅错判一例)与刀切法的符合率 $\{6/7,\ 3/4,\ 5/6\}$ 差别很大. 探讨理论符合率为何有这种极端现象，这可由表1中的均数结构作解释. 表1中 x_2 的均数在三母体中几乎都是相同的 $(-17.3,\ -17.4,\ -17.3)$，这符合公式 (2.2) 中情况，那里令 $C_0=-17.35$ $C_j=0(j\geqslant 2)$. 这说明本例中用(x_2, x_4)作判别时由它们计算理论符合率是不妥当的.

在上例中如改用 F 下界值 $F_1=F_2=3.0$，这时选中的变量为 x_3 及 x_4，而理论符合率的估计值仍为 $\{0,\ 1,\ 0\}$，究其原因，从表1的 (x_3, x_4) 可见，在三母体中同时近似地有

$$\bar{x}_3^{(i)}=17.5-0.16\bar{x}_4^{(i)},\quad i=1,\ 2,\ 3 \tag{3.14}$$

而特征根及特征向量是均数的连续函数[7]，因此特征根自然也近似于零. 即 (x_3, x_4) 时也符合公式(2.2)的条件.

总之，由第二节及上例可见，当理论符合率出现0或1时，它很可能是由于特征根为零而引起，其结果应被怀疑. 我们应仔细考察它是否属于第二节中三情况之一，如是则说明不可使用该公式.

参 考 文 献

[1] Anderson, T. W., *Introduction to Multivariate statistical Analysis*, John Wiley New York 1958.

[2] Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley, New York 1973.

[3] 方开泰、许建伦，*统计分布*，科学出版社，1987.

[4] Urbakh, V. Y., Linear discriminant analysis: Loss of discriminanting power when a variate is omitted, *Biometrics*, **27** (1971), 531—534.

[5] Minoru Siotani, Takesi Hayakawa, Yasunori Fujikoshi., *Modern Multivariate Statistical Analysis*, American Sciences Press, Inc., 1985.

[6] 中国科学院计算中心概率统计组编著，*概率统计计算*，科学出版社，1979.

[7] 孙尚拱，特征值及特征向量微商的解析表达式，*数学进展*，**17** (1988)，391—397.

AN ESTIMATOR OF THEORETICAL MISCLASSIFICATION RATE IN MANY POPULATIONS

SUN SHANGGONG
(Beijing Medical University)

Suppose there are m normal distribution Populations with the same covariance matrix. We turned the multiple integration formula of the theoretical misclassification rate into the problem of eigenvalues and eigenvectors of matrices. We further obtained conditions and their sample estimators.