判别分析与逐步判别

— 地质科学院地矿所四室方法组 —

地质科学院地矿所四室方法组
一九七六年十月

目　录

§1 引 言

众所周知，到目前为止，地质工作的基本方法之一是分类。分类包括两个方面的内容：其一是在我们进行地质研究的范畴内存在多少类型；其二是在类型的数目和内容已知的情况下，我们所研究的某个具体地质对象应该属于哪种类型。分类的第二个内容就是我们现在介绍的方法——判别分析——所要解决的问题。

在地质工作中，大量存在着对类型的判别问题，例如，确定一块岩石标本应该属于哪一类，判断一个侵入岩体是含矿岩体还是不含矿岩体，决定一个古生物的种属，等等。这样，从地质的观点来说，判别分析可以归结为图 1.1 的形式，即一个归属未定的样品，根据它的各种地质特征，究竟应该导入 A、

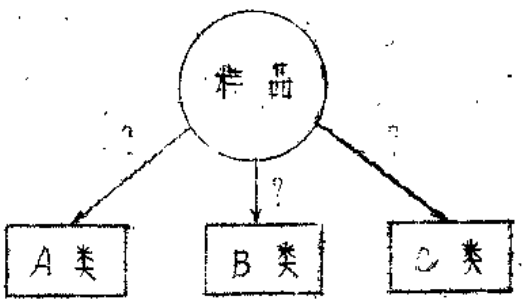

图 1.1

B、C 三类中的哪一类？当然，由于具体的地质问题不同，这个样品也可能只是在两类中作抉择，也可能是在更多的类型中作抉择，这里仅以在三类中作抉择为例而已。

判别是一个根据已知的类型特征对未知对象进行推论的过程。我们可以把判别问题理解为将某个样品的各种地质特征同

它可能归属的各种类型的地质特征进行对比，以决定应该将这个样品分入到哪个类型。但是，当涉及到的地质特征很多时，这种对比并不是一件容易的事，会出现许多错综复杂、相互矛盾的现象，例如，根据某一个特征，该样品应分入A类；而根据另一个特征，它又应该分入B类，这就往往使得地质人员左右为难。

减少样品的地质特征，当然会大大便利上述的对比和判别。当减少到只剩下一个地质特征时，这种对比就非常容易了。如如：根据斜长石中钙长石的分子含量，可以毫不费力地确定斜长石的类型（见图 1.2）

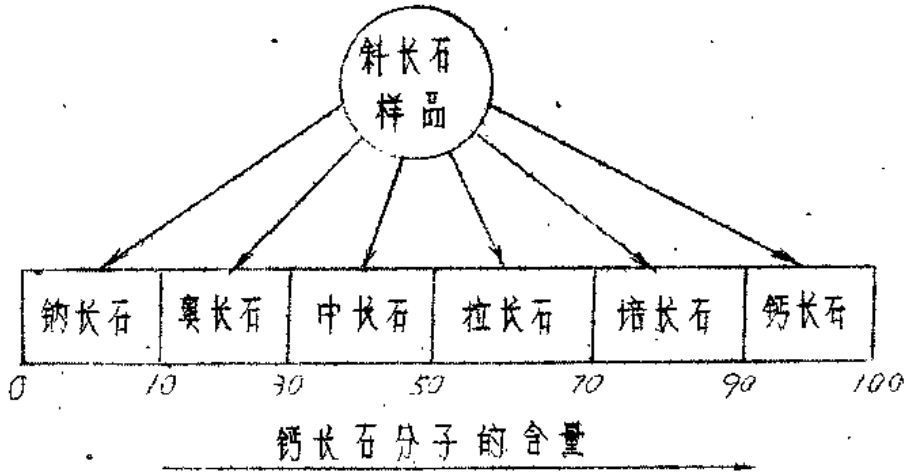

图 1.2　根据斜长石中钙长石的分子含量
划分斜长石的类型

但是，对于大多数地质问题来说，地质特征太少会给对比和判别带来片面性，增加地质特征当然意味着结论更全面可靠，但又遇到了前面提及的困难，于是，传统地质学方法在这里遇到了增加地质特征而带来的一系列矛盾。

统计学中的判别分析就是在这种背景下被引入地质学的。

我们在前面所提到的“判别”，实际上是传统地质学中用地质人员的头脑进行的判别；下面要介绍的“判别分析”，则是通过数学方法进行的判别。

对于地质人员来说，判别分析的最大意义是它把一个样品的许多地质特征综合成一个指标，根据这个综合指标来决定该样品究竟应分到哪一类就同前面所说的划分斜长石的类型一样容易了。

判别分析是多变量统计方法之一。它用于解决下述分类问题：假设所研究的全体对象按其某一属性总共分成G(≥ 2)组，现在要根据与这一属性有关的P个数量指标，即变量的取值x_1，x_2，…，x_p将某一未知组属的对象归为G组之一。用统计的术语，组即总体，对象即个体，(x_1，x_2，…，x_p)即多变量观测值，现在要根据这多变量观测值来决定相应的个体到底来自G个总体中的哪一个。

若$G=2$，则属于两组判别分析问题；若$G>2$，则属于多组判别分析问题。无论是两组或多组判别分析，根据假设条件的不同，其数学模型又分线性的和非线性两种。

本文首先将分别在费歇（Fisher）准则和巴叶斯（Bayes）准则底下建立两组线性判别分析的数学模型，并给出一个有趣的结果，即两组线性判别模型[5]线性回归方程的等价性。然后，在巴叶斯准则底下将两组线性判别分析的数学模型推广到多组的情况。

非线性的模型，本文将不做详细讨论。

与判别分析的实际应用有关的一个重要问题是所谓‘重要’变量的挑选，这就是本文最后介绍的逐步判别分析，其基本思想和具体做法都类似于逐步回归分析（见地质学中多变量统计方法之四：逐步回归分析）。

本文的附录是一个用DJS—6标法语言（二版）编写的依多组（因此也可做两组）逐步判别分析用的计算机程序。

§2. 费歇准则下两组线性判别模型的建立

这里所要讲的是两组线性判别分析。困难在于同时得依据 p 个变量 x_1，x_2，…，x_p。费歇的思想在于把 p 个变量进行线性组合，归结为一个量

$$y = c_1x_1 + c_2x_2 + \cdots + c_px_p \qquad (2.1)$$

其中诸 c_i 是待定的系数，它们满足"两组间'最大'分离"的准则，即费歇准则。在利用这个准则确定诸 c_i 之前，我们先介绍有关的几个术语。线性函数 (2.1) 是判别分析所依据的数学模型，称之为两组线性判别函数。y 称之为判别量。对某一特定的对象 $X_i=(x_{i1}, x_{i2}, \cdots, x_{ip})$ 相对应的 Y 的取值 y_i 称之为该对象的判别得分。某一特定的对象到底应该归为两组中的哪一组取决于它的判别得分的大小，是大于还是小于某个临界值 y_0，当判别得分恰好等于临界值 y_0 时，则可任意规定该特定对象的归属。

把 (2.1) 写成矩阵的形式，即就是

$$y = X'C, \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad y=(x_1, x_2, \cdots x_p)\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = (c_1x_1 + c_2x_2 + \cdots + c_px_p) \qquad (2.2)$$

此处

$$X = (x_1, x_2, \cdots, x_p)$$

$$C = (c_1, c_2, \cdots, c_p) = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}$$

下面我们就来确定(2.1)中诸系数 C_i，为此，假设第 g 组被取自本身的容量为 n_g 的子样

$$X_{\beta}^{(g)}=(x_{1\beta}^{(g)}, x_{2\beta}^{(g)}, \cdots, x_{p\beta}^{(g)}), \quad (\beta=1,2,\cdots,n_g;\ g=1,2)$$

所描述；另它们分别相对应的判别得分为 $y_{\beta}^{(g)}$ $(\beta=1,2,\cdots,n_g;\ g=1,2)$，那末 $C_1, C_2, \cdots, C_p$ 满足"两组间'最大'分离"的准则其数学含意就是应当使得它们的函数

$$I(C_1, C_2, \cdots, C_p)=\frac{(\bar{y}^{(1)}-\bar{y}^{(2)})^2}{a(y)} \qquad (2.3)$$

的值为最大。在这里 $\bar{y}^{(g)}$ 为第 g 组判别得分的样本平均值，它们由下述公式所确定：

$$\bar{y}^{(g)}=\frac{1}{n_g}\sum_{\beta=1}^{n_g} y_{\beta}^{(g)}, \quad (g=1,2)$$

$a(y)$ 为两组判别得分的组内平方和，它们由公式

$$a(y)=\sum_{g=1}^{2}\sum_{\beta=1}^{n_g}(y_{\beta}^{(g)}-\bar{y}^{(g)})^2$$

所确定，I 的分子反映组间的差异，分母反映组内的差异。自然，我们希望组间的差异尽可能地大，而组内的差异尽可能地小，这就是所谓"两组间'最大'分离"准则的由来，它首先由费歇(1936年)提出，所以就命名为费歇准则。

根据数学分析中的极值原理，I 为最大的必要条件是 I 对诸 C_i 的偏导数为零。为了方便起见，我们令，

$$\bar{x}_i^{(g)}=\frac{1}{n_g}\sum_{\beta=1}^{n_g} x_{i\beta}^{(g)}, \quad (i=1,2,\cdots,p;\ g=1,2) \qquad (2.4)$$

$$a_{ij}=\sum_{g=1}^{2}\sum_{k=1}^{n_g}(x_{ikg}^{(g)}-\bar{x}_i^{(g)})(x_{jkg}^{(g)}-\bar{x}_j^{(g)}),\quad (i,j=1,2,\cdots,p)$$

(2.5)

$$d_i=\bar{x}_i^{(1)}-\bar{x}_i^{(2)},\quad (i=1,2,\cdots,p)$$

$$E=(\bar{y}^{(1)}-\bar{y}^{(2)})^2=\Big[\sum_{j=1}^{p}c_j(\bar{x}_j^{(1)}-\bar{x}_j^{(2)})\Big]^2=\Big(\sum_{j=1}^{p}c_jd_j\Big)^2,$$

$$F=a(1)=\sum_{g=1}^{2}\sum_{k=1}^{n_g}\Big[\sum_{j=1}^{p}c_j(x_{jkg}^{(g)}-\bar{x}_j^{(g)})\Big]^2.$$

于是有

$$\frac{\partial I}{\partial c_i}=\frac{\partial}{\partial c_i}\Big(\frac{E}{F}\Big)=\frac{F\frac{\partial E}{\partial c_i}-E\frac{\partial F}{\partial c_i}}{F^2}=0,\quad (i=1,2,\cdots,p)$$

或者

$$\frac{\partial E}{\partial c_i}=\frac{1}{I}\frac{\partial E}{\partial c_i},\quad (i=1,2,\cdots,p)\qquad (2.6)$$

其中

$$\frac{\partial E}{\partial c_i}=2d_i\sum_{j=1}^{p}c_ja_{ij},\quad (i=1,2,\cdots,p)\qquad (2.7)$$

$$\frac{\partial F}{\partial c_i}=2\sum_{g=1}^{2}\sum_{k=1}^{n_g}\Big\{(x_{ikg}^{(g)}-\bar{x}_i^{(g)})\Big[\sum_{j=1}^{p}c_j(x_{jkg}^{(g)}-\bar{x}_j^{(g)})\Big]\Big\}$$

$$=2\sum_{j=1}^{p}c_j\Big[\sum_{g=1}^{2}\sum_{k=1}^{n_g}(x_{ikg}^{(g)}-\bar{x}_i^{(g)})(x_{jkg}^{(g)}-\bar{x}_j^{(g)})\Big]$$

$$=2\sum_{j=1}^{p}c_ja_{ij},\quad (i=1,2,\cdots,p)\qquad (2.8)$$

把 (2.7) 和 (2.5) 代入 (2.6)，可得

$$\sum_{j=1}^{p} a_{ij}c_j = b\alpha_i \quad (i=1,2,\cdots,p) \tag{2.9}$$

此处

$$b=\frac{1}{I}\left(\sum_{j=1}^{p} c_j\alpha_j\right).$$

(2.9) 是关于诸 c_j 的线性代数方程组，其中 b 与下标 i 无关，因此，对不同的 b，方程组 (2.9) 的解 $c_j(j=1,2,\cdots,p)$ 之间仅相差一个比例因子，由判别函数 (2.1) 可以看出，只要对临界值 y_0 作适当的平移，将不影响判别结果，也就是说，对不同的 b，方程组 (2.9) 的解 $c_j(j=1,2,\cdots p)$ 之间实质上是相等价的。这样一来，为简单起见，可令 $b=1$，这时 (2.9) 成为

$$\sum_{j=1}^{p} a_{ij}c_j = \alpha_i \quad (i=1,2,\cdots,p) \tag{2.10}$$

写成矩阵的形式，即就是

$$AC=D, \tag{2.11}$$

其中 C 同前，$A=(a_{ij})_{p\times p}$，$D'=(\alpha_1,\alpha_2,\cdots,\alpha_p)$。由 (2.11) 可解得

$$C=A^{-1}D,$$

其中 A^{-1} 是 A 的逆阵。

在判别分析中，可能要犯两类错误 第一类错误是尽管个体属于第一个总体却把它错误地分到第二个总体中去了；第二类错误则相反，把实际上属于第二个总体的个体错误地分到第一个总体中去了。

假定个体来自两个总体的先验概率相等，并且由两类错误所造成的损失相等，那末，临界值可由公式

$$y_C = \frac{\sigma_2 \bar{y}^{(1)} + \sigma_1 \bar{y}^{(2)}}{\sigma_1 + \sigma_2} \qquad (2.12)$$

确定，其中

$$\sigma_g = \sqrt{\frac{1}{n_g - 1} \sum_{k=1}^{n_g} (y_k^{(g)} - \bar{y}^{(g)})^2}, \quad (g = 1, 2)$$

是第 g 组判别得分的标准差。

公式（2.12）可解释如下：我们假定当 $X = (x_1, x_2, \cdots, x_p)$ 属于第 g 个总体时相应的判别量 $y^{(g)}$ 遵循正态分布（事实上，如果 X 对两个总体而言都遵循正态分布，那末与它们分别相对应的两个判别量 $y^{(g)}$ 必然遵循正态分布），那末，对于大的 n_1 和 n_2，它们的数学期望和方差可分别用 $\bar{y}^{(1)}$，$\bar{y}^{(2)}$ 和 σ_1^2，σ_2^2 估计。于是，上述两个正态分布便被完全确定。如图 2.1 所示，它们的概率密度函数曲线分别是曲线 I 和 II，则由公式（2.12）所给出的临界值 y_C 便是[illegible]点 Q 所对应的横坐标。这时犯两类错误的概率分别由 y_C 左右两侧的曲线 I 和 II 以下的阴影部分面积所表示。而这两部分的面积相等：

$$S_{面积(I)} = \int_{-\infty}^{y_C} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(y - \bar{y}^{(1)})^2}{2\sigma_1^2}} dy$$

$$= \int_{y_C}^{+\infty} \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y - \bar{y}^{(2)})^2}{2\sigma_2^2}} dy$$

$$= S_{面积(II)}$$

由图 2.1 可以看出，如果 $y^{(2)}$ 的方差 σ_2^2 越小，则临界值 y_c 就越接近于 $\bar{y}^{(2)}$，反之亦然。

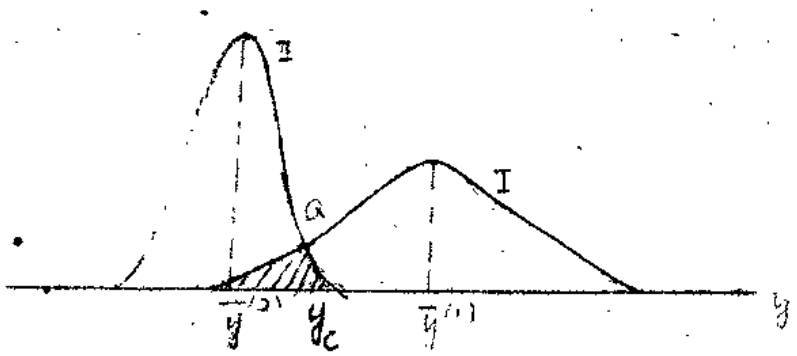

图 2.1

令

$$R_1^*: \quad y = C_1x_1 + C_2x_2 + \cdots + C_px_p \geq y_c, \qquad (2.13)$$

$$R_2^*: \quad y = C_1x_1 + C_2x_2 + \cdots + C_px_p < y_c, \qquad (2.14)$$

则不等式 (2.13) 和 (2.14) 把 p 维样本空间划分为两个互斥且完备的子空间 R_1^* 和 R_2^*。于是根据上面已经建立起来的判别准则，任一特定的个体 $X = (x_1, x_2, \cdots, x_p)$ 到底来自哪一个总体取决于该个体到底落在哪一个子空间中。以两个变量 x_1 和 x_2 的简单情况为例，如图 2.2 所示，直线

$$L: \quad C_1x_1 + C_2x_2 = y_c$$

把 $x_1 - x_2$ 平面划分为两个半平面 R_1^* 和 R_2^*。这两个半平面即是判别分析的依据。

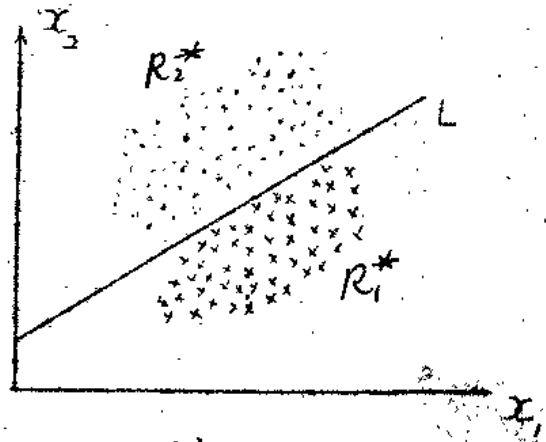

图 2.2

因此，两组判别分析实质上是如何把样本空间 R 划分为两个互斥且完备

前子空间 R_1^* 和 R_2^* 的问题。

§3. 巴叶斯准则下两组线性判别模型的建立

我们用 $R=(R_1, R_2)$ 表示对样本空间的任一分法，则在§2中指出的两类错误都在所难免。因此，我们只能希望由此造成的平均损失或者损失的数学期望尽可能地小，使得平均损失为最小的准则称为巴叶斯准则。

我们假定两个总体的分布密度函数分别为 $p_{iq}(x)=p_i(x_1, x_2, \cdots, x_p)$ $(i=1,2)$；个体来自两个总体的先验概率分别为 q_1 和 q_2；由两类错误造成的损失分别为 $C(2/1)$ 和 $C(1/2)$，则平均损失等于

$$C(2/1)P(2/1)q_1+C(1/2)P(1/2)q_2 \qquad (3.1)$$

其中

$$P(2/1)=\int_{R_2} p_1(x)\,dx=\int_{R_2}\cdots\int p_1(x_1,\cdots,x_p)\,dx_1\cdots dx_p,$$

$$P(1/2)=\int_{R_1} p_2(x)\,dx=\int_{R_1}\cdots\int p_2(x_1,\cdots,x_p)\,dx_1\cdots dx_p$$

分别是犯第一、二类错误的概率。

现令

$$R_1^*:\ [C(2/1)q_1]p_1(x)\geqslant[C(1/2)q_2]p_2(x)$$
$$R_2^*:\ [C(2/1)q_1]p_1(x)<[C(1/2)q_2]p_2(x) \qquad (3.2)$$

或者

$$R_1^*\quad \frac{p_1(x)}{p_2(x)}\geqslant\frac{C(1/2)q_2}{C(2/1)q_1}.$$

$$R_2^*: \quad \frac{p_1(X)}{p_2(X)} \geqslant \frac{C(1/2)q_2}{C(2/1)q_1} \tag{3.3}$$

则可以证明上述分法 $R=(R_1^*, R_2^*)$ 遵循巴叶斯准则，也即能使平均损失 (3.1) 为最小。现证明如下：

对任一分法 $R=(R_1, R_2)$，其相应的平均损失

$$[C(2/1)q_1]\int_{R_2} p_1(X)\,dX+[C(1/2)q_2]\int_{R_1} p_2(X)\,dX$$

$$=\int_{R_2}\{[C(2/1)q_1]p_1(X)-[C(1/2)q_2]p_2(X)\}\,dX$$

$$+\int_{R}[C(1/2)q_2]p_2(X)\,dX\ .$$

上式右端第二项与分法无关，等于常数。如果 R_2 包含所有 $[C(2/1)q_1]p_1(X)<[C(1/2)q_2]p_2(X)$ 的点，不包含任何 $[C(2/1)q_1]p_1(X)>[C(1/2)q_2]p_2(X)$ 的点，则上式右端第一项取到最小值，也即整个式子取到最小值。而分法 $R=(R_1^*, R_2^*)$ 正好满足上述要求，这就说明它遵循巴叶斯准则。

应当指出，当 $[C(2/1)q_1]p_1(X)=[C(1/2)q_2]p_2(X)$ 对个体 X 可以划归两个总体中的任何一个。在此我们规定把它划归与 R_i^* 相对应的那个总体。

另外，还应当指出，由 (3.3) 可知，分法 $R=(R_1^*, R_2^*)$ 仅与 $C(2/1)$ 和 $C(1/2)$ 的比值有关。应用判别分析的一个非常重要的特殊情况就是 $C(2/1)=C(1/2)$。通常在难以确切知道 $C(2/1)$ 和 $C(1/2)$ 或者两者之比的情况下，都作为这种特殊的情况来处理。这时我们不妨假设 $C(2/1)=C(1/2)=1$。于是 (3.1) 简化成为

$$P(2/1)q_1+P(1/2)q_2, \tag{3.4}$$

这也就是犯两类错误的平均概率。因此在这种情况下，巴叶斯

准则就是要使得把两类错误的平均概率最小。同时（3.2）和（3.3）简化成为

$$\begin{aligned} R_1^*: &\quad q_1 p_1(X) \geqslant q_2 p_2(X), \\ R_2^*: &\quad q_1 p_1(X) < q_2 p_2(X) \end{aligned} \tag{3.5}$$

和

$$\begin{aligned} R_1^*: &\quad \frac{p_1(X)}{p_2(X)} \geqslant \frac{q_2}{q_1} \\ R_2^*: &\quad \frac{p_1(X)}{p_2(X)} < \frac{q_2}{q_1} \end{aligned} \tag{3.6}$$

注意到在个体 X 给定的情况下，它来自两个总体的条件概率（或先验概率）分别为

$$\frac{q_1 p_1(X)}{q_1 p_1(X)+q_2 p_2(X)} \text{ 和 } \frac{q_2 p_2(X)}{q_1 p_1(X)+q_2 p_2(X)} \tag{3.7}$$

这里两个分式的分母相同，而分子分别是（3.5）中不等式的两边。所以，在这种情况下，实际上是以指定具有较高的条件概率的总体作为予法 $R=(R_1^*, R_2^*)$ 的依据，也即作为判别的准则。在 X 给定的情况下，它来自两个总体的条件概率（3.7）实际上就是它来自两个总体的可能性大小，我们以此作为判别的准则，认为给定的 X 来自可能性较大的总体，也是理所当然的事情。

当两个总体分别遵循 p 维正态分布 $N(\mu_g, \Sigma)$，$g=1,2$）时，在这里我们假定它们具有相同的协方差矩阵。于是就有

$$p_g(X)=\frac{1}{(2\pi)^{\frac{1}{2}p}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(X-\mu_g)'\Sigma^{-1}(X-\mu_g)\right], \quad (g=1,2)$$

这时，（3.3）中不等式的左边部分等于

$$\frac{\exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right]}{\exp\left[-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]}$$

对上式取以 e 为底的对数（自然对数），展开，并注意到 Σ^{-1} 的对称性，整行合并，可得

$$\begin{aligned}
&-\frac{1}{2}\left[(x-\mu_1)'\Sigma^{-1}(x-\mu_1)-(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right] \\
&=-\frac{1}{2}\left[x'\Sigma^{-1}x-x'\Sigma^{-1}\mu_1-\mu_1'\Sigma^{-1}x+\mu_1'\Sigma^{-1}\mu_1\right. \\
&\quad \left.-x'\Sigma^{-1}x+x'\Sigma^{-1}\mu_2+\mu_2'\Sigma^{-1}x-\mu_2'\Sigma^{-1}\mu_2\right] \\
&=x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2).
\end{aligned}$$

这样一来 (3.3) 就为

$$R_1^*:\ x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\geqslant t, \quad (3.8)$$

$$R_2^*:\ x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)<t;$$

相等时，其中 $t=\ln[C(1/2)q_2/C(2/1)q_1]$.

当参数 μ_1，μ_2 和 Σ 为未知时，对大的 n_1 和 n_2，它们的'最佳'估计量分别是

$$\begin{aligned}
\bar{x}^{(1)}&=(\bar{x}_1^{(1)},\ \bar{x}_2^{(1)},\cdots,\bar{x}_p^{(1)})', \\
\bar{x}^{(2)}&=(\bar{x}_1^{(2)},\ \bar{x}_2^{(2)},\cdots,\bar{x}_p^{(2)})'
\end{aligned} \quad (3.9)$$

和

$$S=(S_{ij}),\ (i,j=1,2,\cdots,p)$$

此地

$$S_{ij}=\frac{1}{n_1+n_2-2}a_{ij},\ (i,j=1,2,\cdots,p)$$

而 $\bar{x}_i^{(g)}$ $(i=1,2\cdots,p;\ g=1,2)$ 和 $a_{ij}(i,j=1,2\cdots,p)$ 由§2

中公式（2.4）和（2.5）确定。将这些参数的估计量代入（3.8）中不等式的左边，可得

$$X'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})-\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

上式第一项

$$X'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})=X'[(n_1+n_2-2)A^{-1}]D$$

$$=(n_1+n_2-2)X'C$$

即是在§2中所定义的两组线性判别函数。

§4. 判别函数与回归方程

本节将证明一个有趣的结果，即两组线性判别函数（2.1）与变量 $x_1, x_2, \cdots, x_p$ 关于只取两个不同值的伪变量的线性回归方程相等价，也即它们的系数，判别系数和回归系数成比例。为此，我们假定

$$x_{p+1}=\begin{cases} x^{(1)}_{p+1,k}=\dfrac{n_2}{n_1+n_2}, & \text{若 } X=X^{(1)}_k \ (k=1,2,\cdots,n_1) \\ x^{(2)}_{p+1,k}=\dfrac{-n_1}{n_1+n_2}, & \text{若 } X=X^{(2)}_k \ (k=1,2,\cdots,n_2) \end{cases}$$

（事实上，x_{p+1} 可以取任意两个不同的值）；伪变量 x_{p+1} 关于变量 $x_1, x_2, \cdots, x_p$ 的回归方程为

$$x_{p+1}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p,$$

则回归系数 $b_1, b_2, \cdots, b_p$ 是以下正规方程组的解（见地质学中多变量统计方法之四：逐步回归分析）

$$\sum_{j=1}^{p} t_{ij} b_j = t_{i,p+1}, \quad (i=1,2,\cdots,p) \tag{4.1}$$

其中

$$t_{ij} = \sum_{g=1}^{2} \sum_{k=1}^{n_g} (x_{ik}^{(g)} - \bar{x}_i)(x_{jk}^{(g)} - \bar{x}_j), \quad (i=1,2,\cdots,p;\ j=1,2,\cdots,(p+1))$$

又其中

$$\bar{x}_i = \frac{n_1 \bar{x}_i^{(1)} + n_2 \bar{x}_i^{(2)}}{n_1 + n_2}, \quad (i=1,2,\cdots,p) \tag{4.2}$$

$\bar{x}_i^{(g)}$ $(i=1,2,\cdots,p;\ g=1,2)$ 由§2中公式(2.4)确定。

现在的问题在于证明方程组(4.1)的解和方程组(2.9)的解的等价值，因为

$$\sum_{k=1}^{n_g} (x_{ik}^{(g)} - \bar{x}_i)(x_{jk}^{(g)} - \bar{x}_j)$$

$$= \sum_{k=1}^{n_g} [(x_{ik}^{(g)} - \bar{x}_i^{(g)}) + (\bar{x}_i^{(g)} - \bar{x}_i)][(x_{jk}^{(g)} - \bar{x}_j^{(g)}) + (\bar{x}_j^{(g)} - \bar{x}_j)]$$

$$= \sum_{k=1}^{n_g} (x_{ik}^{(g)} - \bar{x}_i^{(g)})(x_{jk}^{(g)} - \bar{x}_j^{(g)}) + \sum_{k=1}^{n_g} (x_{ik}^{(g)} - \bar{x}_i^{(g)})(\bar{x}_j^{(g)} - \bar{x}_j)$$

$$+ \sum_{k=1}^{n_g} (\bar{x}_i^{(g)} - \bar{x}_i)(x_{jk}^{(g)} - \bar{x}_j^{(g)}) + \sum_{k=1}^{n_g} (\bar{x}_i^{(g)} - \bar{x}_i)(\bar{x}_j^{(g)} - \bar{x}_j)$$

上式右边第二、三项等于零，所以

$$\sum_{k=1}^{n_g} (x_{ik}^{(g)} - \bar{x}_i)(x_{jk}^{(g)} - \bar{x}_j)$$