$m_0 + m_1 + m_2 =$ 总类目数 p　$(p = r_1 + r_2 + \cdots + r_m)$.

表 6.2 给出两个样品的取“值”情况. 显然 $m_1=3, m_0=7, m_2=4$. 项目数 $m=4$, 总类目数 $p=4+2+5+3=14$.

表 6.2　两样品的取值情况

项目 / 样品	项目 1				项目 2		项目 3					项目 4		
	1	2	3	4	1	2	1	2	3	4	5	1	2	3
$X_{(i)}$	1	0	0	0	0	1	1	0	0	0	1	0	1	0
$X_{(j)}$	0	1	0	0	0	1	0	0	1	0	1	0	1	0

(1) 两样品 $X_{(i)}$ 和 $X_{(j)}$ 间的距离定义为

$$d_{ij} = m_2/(m_1 + m_2).$$

即为不配对的类目数与在有反应的类目(包括 1—1 配对和不配对)数的比值. 比如表 6.2, $d_{ij}=4/7$.

当项目只能取可能类目中的一类, 即在不能兼取情况下, 两样品的距离定义为: $d_{ij}=m_2^*/m$, 其中 m_2^* 是不配对的项目(变量)个数; m 为项目总个数.

类似于欧氏距离, 还可以定义距离为

$$d_{ij}^2 = \sum_{k=1}^{m} \sum_{l=1}^{r_k} (\delta_i(k,l) - \delta_j(k,l))^2,$$

即不配对的总数.

(2) 样品 $X_{(i)}$ 和 $X_{(j)}$ 间的相似性度量由表 6.3 给出几种定义方

表 6.3　两样品间相似性的几种度量

	匹配系数	说　明
1	$(m_0+m_1)/p$	配对的总数在总类目数中之比
2	m_1/p	1—1 配对的总数在总类目数中之比
3	$m_1/(m_1+m_2)$	不考虑 0—0 配对的情况
4	$2(m_0+m_1)/(p+m_0+m_1)$	对 1—1 和 0—0 配对数双倍加权
5	$(m_0+m_1)/(p+m_2)$	对不配对数双倍加权
6	$2m_1/(2m_1+m_2)$	对 1—1 配对数双倍加权, 不考虑 0—0 配对
7	$m_1/(m_1+2m_2)$	不考虑 0—0 配对, 且对不配对数双倍加权
8	m_1/m_2	不考虑 0—0 配对, 1—1 配对数与不配对数之比

法，这种相似性度量也称为**匹配系数**.

三、变量间的相似系数和距离

聚类分析方法不仅用来对样品进行分类，有时还需要对变量进行分类. 在对变量进行分类时，通常采用相似系数来表示变量之间的亲疏程度.

设 C_{ij} 表示变量 X_i 和 X_j 间的相似系数，一般要求：

(1) $C_{ij}=\pm 1 \Longleftrightarrow X_i=aX_j$ ($a\neq 0$ 为常数)；

(2) $|C_{ij}|\leqslant 1$，对一切 i,j 成立；

(3) $C_{ij}=C_{ji}$，对一切 i,j 成立.

$|C_{ij}|$ 越接近 1，则表示 X_i 和 X_j 的关系越密切，C_{ij} 越接近 0，两者关系越疏远.

对于定量变量，我们通常采用的相似系数有 X_i 和 X_j 间的夹角余弦和相关系数.

1. 夹角余弦

变量 X_i 的 n 次观测值 $(x_{1i},x_{2i},\cdots,x_{ni})$ 可以看成 n 维空间的向量，则 X_i 和 X_j 的夹角 α_{ij} 的余弦 $\cos\alpha_{ij}$ 称为两向量的相似系数，记为 $C_{ij}(1)$，即

$$C_{ij}(1)=\cos\alpha_{ij}=\frac{\sum_{t=1}^{n}x_{ti}x_{tj}}{\sqrt{\sum_{t=1}^{n}x_{ti}^2}\sqrt{\sum_{t=1}^{n}x_{tj}^2}}\quad (i,j=1,\cdots,m).$$

当 X_i 和 X_j 平行时，其夹角 $\alpha_{ij}=0°$，$C_{ij}(1)=1$，说明这两个向量完全相似；当 X_i 和 X_j 正交时，其 $\alpha_{ij}=90°$，$C_{ij}(1)=0$，说明这两个向量不相关.

2. 相关系数

相关系数就是对数据作标准化处理后的夹角余弦. 变量 X_i 和 X_j 的相关系数常用 r_{ij} 表示，在这里我们记为 $C_{ij}(2)$，即

$$C_{ij}(2)=\frac{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}}\quad(i,j=1,\cdots,m).$$

当 $C_{ij}(2)=1$ 时表示两变量线性相关，一般情况，$|C_{ij}|\leqslant 1$.

3. 变量间的距离

(1) 利用相似系数来定义变量间的距离，即

$$d_{ij}=1-|C_{ij}|\quad 或\quad d_{ij}^2=1-C_{ij}^2\quad(i,j=1,2,\cdots,m).$$

(2) 利用样本协方差阵 S 来定义距离. 设 $S=(s_{ij})_{m\times m}>0$，变量 X_i 和 X_j 间的距离可定义为

$$d_{ij}=s_{ii}+s_{jj}-2s_{ij}\quad(i,j=1,2,\cdots,m).$$

(3) 把变量 X_i 的 n 次观测值看成为 n 维空间的点. 在 n 维空间中按小节二“样品间的距离和相似系数”中介绍的方法可类似定义 m 个变量间的各种距离.

4. 定性变量间的相似系数

当变量 X_i 是定性(属性)变量时，也可以定义多种相似系数. 设变量 X_i 的 p 种取值记为 $r_1,r_2,\cdots,r_p$(或称项目 X_i 有 p 个类目)；X_j 的 q 种取值记为 $t_1,t_2,\cdots,t_q$. n 个样品中两个定性变量的实际观测结果经整理后列成表 6.4，其中 n_{kl} 表示在 n 个样品中 X_i 取第 k 个值 r_k，且 X_j 取第 l 个值 t_l 的频数. 通常称表 6.4 为**列联表**.

表 6.4　列联表

变量 X_i \ 变量 X_j	t_1	t_2	$\cdots$	t_q	求行和
r_1	n_{11}	n_{12}	$\cdots$	n_{1q}	$n_{1+}=\sum_l n_{1l}$
r_2	n_{21}	n_{22}	$\cdots$	n_{2q}	$n_{2+}=\sum_l n_{2l}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r_p	n_{p1}	n_{p2}	$\cdots$	n_{pq}	$n_{p+}=\sum_l n_{pl}$
求列和	n_{+1}	n_{+2}	$\cdots$	n_{+q}	总和 $n_{++}=\sum_i\sum_j n_{ij}$

注：表中 $n_{+l}=\sum_k n_{kl}(l=1,2,\cdots,q)$，$n_{++}=n$ 为观测个数总和.

在利用列联表对两个属性变量的独立性检验中，经常要用到 χ^2 统计量：

$$\chi_{ij}^2 = n_{++}\left(\sum_{k=1}^{p}\sum_{l=1}^{q}\frac{n_{kl}^2}{n_{k+}n_{+l}} - 1\right),$$

建立在 χ^2 统计量基础上的相似系数有：

(1) 联列系数：

$$C_{ij}(3) = \sqrt{\chi_{ij}^2/(\chi_{ij}^2 + n)},$$

(2) 连关系数(有三种)：

$$C_{ij}(4) = \sqrt{\frac{\chi_{ij}^2}{n \cdot \max(p-1, q-1)}},$$

$$C_{ij}(5) = \sqrt{\frac{\chi_{ij}^2}{n \cdot \min(p-1, q-1)}},$$

$$C_{ij}(6) = \sqrt{\frac{\chi_{ij}^2}{n \cdot \sqrt{(p-1)(q-1)}}}.$$

如果 X_i 和 X_j 只取两个值(即 $p=q=2$，这两个值不妨分别记为 0 和 1)，则表 6.4 的列联表可简化为表 6.5.

表 6.5 二值变量的列联表

X_i \ X_j	1	0	求行和
1	a	b	$a+b$
0	c	d	$c+d$
求列和	$a+c$	$b+d$	总和 $n=a+b+c+d$

常用的相似系数有：

(3) 点相关系数：

$$C_{ij}(7) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad (i, j = 1, \cdots, m). \tag{6.2.2}$$

这是与定量变量的相关系数相对应的量.

(4) 四分相关系数：

$$C_{ij}(8)=\sin\left(90^\circ\frac{(a+d)-(b+c)}{a+b+c+d}\right).$$

(5)夹角余弦：

$$C_{ij}(9)=\left(\frac{a}{a+b}\frac{a}{a+c}\right)^{1/2}=\frac{a}{\sqrt{(a+b)(a+c)}}. \quad (6.2.3)$$

考虑到 $C_{ij}=C_{ji}$，改进的量是

$$C_{ij}(10)=\left(\frac{a}{a+b}\frac{a}{a+c}\frac{d}{d+b}\frac{d}{d+c}\right)^{1/2}=\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}.$$

§6.3　系统聚类法

系统聚类法是目前在实际应用中使用最多的一类方法，它是将类由多变到少的一种方法.

本节考虑 n 个样品的聚类问题，其观测数据列为表 6.1 的形式，n 个 m 元样品记为 $X_{(i)}$ $(i=1,2,\cdots,n)$.

一、系统聚类法的基本思想和基本步骤

设有 n 个样品，每个样品测得 m 项指标. 系统聚类方法的基本思想是：首先定义样品间的距离(或相似系数)和类与类之间的距离. 初始将 n 个样品看成 n 类(每一类包含一个样品)，这时类间的距离与样品间的距离是等价的；然后将距离最近的两类合并成为新类，并计算新类与其他类的类间距离(类间距离的几种定义将在下面介绍)，再按最小距离准则并类. 这样每次缩小一类，直到所有的样品都并成一类为止. 这个并类过程可以用谱系聚类图形象地表达出来.

由以上系统聚类法的基本思想，即可得出它的基本步骤如下：

(0) 数据变换：使用上节介绍的方法对数据进行变换. 数据变换目的是为了便于比较和计算，或改变数据的结构.

下面的步骤是选择度量样品间的距离定义(如欧氏距离)及度量类间的距离定义(如最短距离法，参见下面小节二“系统聚类分析的

方法”)：

(1) 计算 n 个样品两两间的距离，得样品间的距离矩阵 $D^{(0)}$.

(2) 初始(第一步：$i=1$)n 个样品各自构成一类，类的个数 $k=n$，第 i 类 $G_i=\{X_{(i)}\}$ $(i=1,\cdots,n)$. 此时类间的距离就是样品间的距离(即 $D^{(1)}=D^{(0)}$). 然后对样品 $X_{(i)}(i=2,\cdots,n)$执行并类过程的步骤(3)和(4).

(3) 对步骤(2)得到的距离矩阵 $D^{(i-1)}$，合并类间距离最小的两类为一新类. 此时类的总个数 k 减少 1 类，即

$$k=n-i+1.$$

(4) 计算新类与其他类的距离，得新的距离矩阵 $D^{(i)}$. 若合并后类的总个数 k 仍大于 1，重复步骤(3)和(4)；直到类的总个数为 1 时转到步骤(5).

(5) 画谱系聚类图；

(6) 决定分类的个数及各类的成员.

例 6.3.1 设有 5 个产品，分别对每个产品测得一项质量指标 X，其值如下：1，2，4.5，6，8. 试对这 5 个产品按质量指标进行分类.

解 设样品间的距离取为欧氏距离，类间的距离取为类间的最短距离，根据上面介绍的步骤，计算如下：

(1) 计算 5 个样品：$X_{(1)},X_{(2)},X_{(3)},X_{(4)},X_{(5)}$两两间的距离，得初始的类间距离矩阵 $D^{(1)}$(也就是样品间距离矩阵 $D^{(0)}$)：

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
$X_{(1)}$	0	$\boxed{1}$	3.5	5.0	7.0
$X_{(2)}$		0	2.5	4.0	6.0
$X_{(3)}$			0.0	1.5	3.5
$X_{(4)}$				0.0	2.0
$X_{(5)}$					0.0

(2) 初始 n 个样品各自构成一类，得 5 个类：$G_i=\{X_{(i)}\}$ $(i=1,\cdots,5)$，类的个数 $k=5$.

(3) 由 $D^{(1)}$可知，类间距离为 1 时最小，首先应合并 $X_{(1)}$和 $X_{(2)}$为一新类，记为 $CL4=\{X_{(1)},X_{(2)}\}$；此时类的总个数 k 减少 1，变为

$k=4$，故把此步骤得到的新类记为 $CL4$.

(4) 按最短距离法计算新类 $CL4$ 与其他类的距离，得新的距离矩阵 $D^{(2)}$：

	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$CL4$
$X_{(3)}$	0	1.5	3.5	2.5
$X_{(4)}$		0.0	2.0	4.0
$X_{(5)}$			0.0	6.0
$CL4$				0.0

因此时类的总个数 $k=4$ 大于 1，继续重复并类过程.

(5) 由 $D^{(2)}$ 可知，类间距离为 1.5 时最小，故合并 $X_{(3)}$ 和 $X_{(4)}$ 为一新类，记为 $CL3=\{X_{(3)},X_{(4)}\}$；此时类的总个数 k 又减少 1，变为 $k=3$，故把此步骤得到的新类记为 $CL3$.

(6) 按最短距离法计算新类 $CL3$ 与其他类的距离，得新的距离矩阵 $D^{(3)}$：

	$X_{(5)}$	$CL4$	$CL3$
$X_{(5)}$	0	6	2.0
$CL4$		0	2.5
$CL3$			0.0

因此时类的总个数 $k=3$ 大于 1，继续重复并类过程.

(7) 由 $D^{(3)}$ 可知，应合并 $X_{(5)}$ 和 $CL3$ 为一新类，记为 $CL2=\{X_{(5)},X_{(3)},X_{(4)}\}$；此时类的总个数 k 再减少 1，变为 $k=2$，故把此步骤得到的新类记为 $CL2$.

(8) 按最短距离法计算新类 $CL2$ 与其他类的距离，得新的距离矩阵 $D^{(4)}$：

	$CL4$	$CL2$
$CL4$	0	2.5
$CL2$		0.0

因此时类的总个数 $k=2$ 大于 1，继续重复并类过程.

(9) 由 $D^{(4)}$ 可知，最后应合并 $CL4$ 和 $CL2$ 为一新类，记为 $CL1$

$=\{X_{(1)},X_{(2)},X_{(5)},X_{(3)},X_{(4)}\}$；此时类的总个数 $k=1$，故把此步骤得到的新类记为 $CL1$.

(10) 此时所有样品全并成一类，得新的距离矩阵 $D^{(5)}$：

	$CL1$
$CL1$	0

并类过程至此结束.

(11) 画谱系聚类图(见图 6.1).

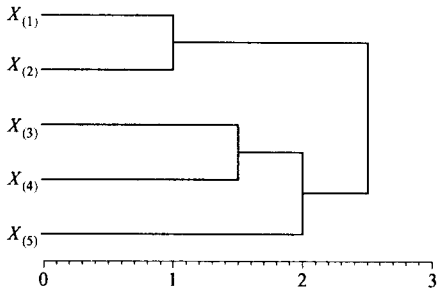


图 6.1 谱系聚类图

(12) 确定类的个数及各类的成员：

若分为两类，则 $G_1^{(2)}=\{X_{(1)},X_{(2)}\}$，$G_2^{(2)}=\{X_{(5)},X_{(3)},X_{(4)}\}$；

若分为三类，则 $G_1^{(3)}=\{X_{(1)},X_{(2)}\}$，$G_2^{(3)}=\{X_{(5)}\}$，$G_3^{(3)}=\{X_{(3)},X_{(4)}\}$；

若分为四类，则 $G_1^{(4)}=\{X_{(1)},X_{(2)}\}$，$G_2^{(4)}=\{X_{(5)}\}$，$G_3^{(4)}=\{X_{(3)}\}$，$G_4^{(4)}=\{X_{(4)}\}$；

若分为五类，则 $G_i^{(5)}=\{X_{(i)}\}$ $(i=1,2,3,4,5)$.

有了谱系聚类图，根据实际问题希望分为几类，都可以从谱系聚类图中得到分类结果. 到底分为几类最合适？这里并没有绝对正确的原则，一般可根据实际问题的不同，从谱系聚类图直观看出，或通过分界值(阈值)给出分类，也可以用一些近似的检验统计量来验证分类个数如何选取更合适(见 §6.4 中小节三“类个数的确定”).

以上并类过程可以使用 SAS/STAT 软件的 CLUSTER (系统聚类)过程对这组简单的数据进行分类.

二、系统聚类分析的方法

系统聚类法的聚类原则决定于样品间的距离(或相似系数)及类间距离的定义,类间距离的不同定义就产生了不同的系统聚类分析方法.本节介绍常用的几种系统聚类分析方法(各种方法名字后面括号中的英文是CLUSTER过程中相对应的名字).

以下用 d_{ij} 表示样品 $X_{(i)}$ 和 $X_{(j)}$ 之间的距离,当样品间的亲疏关系采用相似系数 C_{ij} 时,令 $d_{ij}=1-|C_{ij}|$ (或 $d_{ij}^2=1-C_{ij}^2$);用 D_{ij} 表示类 G_i 和 G_j 间的距离.

1. 最短距离法(SINgle linkage)

类与类之间的距离定义为两类中相距最近的样品之间的距离,即类 G_p 和 G_q 之间的距离 D_{pq} 定义为

$$D_{pq}=\min_{i\in G_p,j\in G_q} d_{ij}$$ (这里 $i\in G_p$ 表示 $X_{(i)}\in G_p$,以下同).

当某步骤类 G_p 和 G_q 合并为 G_r 后,按最短距离法计算新类 G_r 与其他类 G_k 的类间距离,其递推公式为

$$\begin{aligned} D_{rk}&=\min_{i\in G_r,j\in G_k} d_{ij} \quad (G_r=\{G_p,G_q\})\\ &=\min\{\min_{i\in G_p,j\in G_k} d_{ij},\min_{i\in G_q,j\in G_k} d_{ij}\}\\ &=\min\{D_{pk},D_{qk}\} \quad (k\neq p,q). \end{aligned}$$

在例6.3.1中的类间距离就是使用最短距离法定义的.

2. 最长距离法(COMplete method)

类与类之间的距离定义为两类中相距最远的样品间的距离,即类 G_p 和 G_q 之间的距离 D_{pq} 定义为

$$D_{pq}=\max_{i\in G_p,j\in G_q} d_{ij}.$$

当某步骤类 G_p 和 G_q 合并为 G_r 后,按最长距离法计算新类 G_r 与其他类 G_k 的类间距离,其递推公式为

$$D_{rk}=\max\{D_{pk},D_{qk}\} \quad (k\neq p,q).$$

最长距离法即为两类合并后的新类与其他类的距离是与原来两类的类间距离的最大者,它加大了合并后的类与其他类的距离,具有空间距离扩张性质(见§6.4中小节—"系统聚类法的简单性质").

3. 中间距离法(MEDian method)

如果类与类之间的距离既不采用两类之间的最近距离,也不采用最远的距离,而是采用介于这两者间的距离,这种方法称为**中间距离法**.

当某步骤类 G_p 和 G_q 合并为 G_r 后,按中间距离法计算新类 G_r 与其他类 G_k 的类间距离,其递推公式为

$$D_{rk}^2 = \frac{1}{2}(D_{pk}^2 + D_{qk}^2) + \beta D_{pq}^2 \quad (-1/4 \leqslant \beta \leqslant 0, k \neq p,q).$$

常取 $\beta=-1/4$,此时由初等几何知,D_{rk} 就是以 D_{qk}, D_{pk}, D_{pq} 为边的三角形中 D_{pq} 边上的中线.

4. 重心法(CENtroid method)

以上三种方法在定义类与类之间距离时,没有考虑每一类中所包含的样品个数.如果将两类间的距离定义为两类重心间的距离,这种聚类方法称为**重心法**.对样品分类时,每一类的重心就是属于该类样品的均值.

设某一步骤将 G_p 和 G_q 合并成 G_r 后,它们所包含的样品个数分别为 n_p, n_q 和 n_r $(n_r=n_p+n_q)$.各类的重心分别为 $\overline{X}^{(p)}, \overline{X}^{(q)}$ 和 $\overline{X}^{(r)}$.显然有:

$$\overline{X}^{(r)} = \frac{1}{n_r}(n_p\overline{X}^{(p)} + n_q\overline{X}^{(q)}). \tag{6.3.1}$$

设某一类 G_k $(k\neq p,q)$的重心为 $\overline{X}^{(k)}$,它与新类 G_r 的距离是

$$D_{rk} = d\{\overline{X}^{(r)}, \overline{X}^{(k)}\}.$$

如果样品间的距离定义为欧氏距离,把(6.3.1)式代入 D_{rk},则有

$$\begin{aligned} D_{rk}^2 &= (\overline{X}^{(k)} - \overline{X}^{(r)})'(\overline{X}^{(k)} - \overline{X}^{(r)}) \\ &= \left[\frac{n_p}{n_r}(\overline{X}^{(k)} - \overline{X}^{(p)}) + \frac{n_q}{n_r}(\overline{X}^{(k)} - \overline{X}^{(q)})\right]' \\ &\quad \cdot \left[\frac{n_p}{n_r}(\overline{X}^{(k)} - \overline{X}^{(p)}) + \frac{n_q}{n_r}(\overline{X}^{(k)} - \overline{X}^{(q)})\right] \\ &= \frac{n_p}{n_r}D_{pk}^2 + \frac{n_q}{n_r}D_{qk}^2 - \frac{n_p}{n_r}\frac{n_q}{n_r}D_{pq}^2 \quad (k \neq p,q). \end{aligned} \tag{6.3.2}$$

上式就是当样品间距离取为欧氏距离时,合并后新类与其他类距离

平方的递推公式. 如果样品间距离不是欧氏距离,根据不同情况可导出不同的递推公式.

5. 类平均法(AVErage linkage)

重心法虽有较好的代表性,但并未充分利用各个样品的信息,因而又有人提出用两类样品两两之间平方距离的平均作为类之间的距离,即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij}^2.$$

采用这种类间距离的聚类方法,称为**类平均法**.

当某步骤将类 G_p 和 G_q 合并为 G_r: $G_r = \{G_p, G_q\}$,且 $n_r = n_p + n_q$,则 G_r 与其他类 G_k 距离平方的递推公式为

$$D_{rk}^2 = \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2 \quad (k \neq p, q).$$

类平均法是一种使用比较广泛、聚类效果较好的方法.

6. 可变类平均法(FLExible-beta method)

类平均法的类间距离递推公式中,没有反映类 G_p 和 G_q 之间距离 D_{pq} 的影响,而**可变类平均法**是将合并后的新类 G_r 与其他类 G_k 的距离平方公式进一步推广为:

$$D_{rk}^2 = (1-\beta)\left[\frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2\right] + \beta D_{pq}^2 \quad (k \neq p, q),$$

其中 β 是可变参数,一般取 $\beta < 1$. 显然,可变类平均法是由类平均法和中间距离法适当推广得到的(当 $\beta = 0$ 时就是类平均法;当 $-1/3 \leqslant \beta \leqslant 0$ 且 $n_p = n_q$ 时就是中间距离法;当 $n_p = n_q$ 时就是下面将介绍的可变法).

可变类平均法的分类效果与 β 的选择关系极大,当 β 接近 1 时一般分类效果不好,在实用中 β 常取负值,如取 $\beta = -1/4$.

7. 可变法及 McQuitty 相似分析法(MCQ)

当某步骤将类 G_p 和 G_q 合并为 G_r 后,**可变法**把 G_r 与其他类 G_k 距离平方的递推公式定义为

$$D_{rk}^2 = \frac{(1-\beta)}{2}(D_{pk}^2 + D_{qk}^2) + \beta D_{pq}^2 \quad (k \neq p, q).$$

在 SAS/STAT 软件的 CLUSTER 过程中使用 $\beta = 0$ 的递推公式:

$$D_{rk}^2 = (D_{pk}^2 + D_{qk}^2)/2,$$

并把此方法称为 **McQuitty 相似分析法**.

8. 离差平方和法(WARD)

离差平方和法是 Ward (1936)提出的,也称为 **Ward 法**. 它基于方差分析思想,如果类分得正确,则同类样品之间的离差平方和应当较小,不同类样品之间的离差平方和应当较大.

假定已将 n 个样品分为 k 类,记为 $G_1, G_2, \cdots, G_k$, n_t 表示 G_t 类的样品个数, $\overline{X}^{(t)}$ 表示 G_t 的重心, $X_{(i)}^{(t)}$ 表示 G_t 中第 i 个样品($i=1, \cdots, n_t$),则 G_t 中样品的离差平方和为

$$W_t = \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \overline{X}^{(t)})' (X_{(i)}^{(t)} - \overline{X}^{(t)}),$$

其中 $X_{(i)}^{(t)}$, $\overline{X}^{(t)}$ 为 m 维向量, W_t 为一数值($t=1,2,\cdots,k$).

k 个类的总离差平方和为

$$W = \sum_{t=1}^{k} W_t = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \overline{X}^{(t)})' (X_{(i)}^{(t)} - \overline{X}^{(t)}).$$

当 k 固定时,要选择使 W 达到极小的分类.

Ward 法的基本思想是,先将 n 个样品各自成一类,此时 $W=0$;然后每次将其中某两类合并为一类,因每缩小一类离差平方和就要增加,每次选择使 W 增加最小的两类进行合并,直至所有样品合并为一类为止.

Ward 法把某两类合并后增加的离差平方和看成为类间的平方距离,即令

$$D_{pq}^2 = W_r - (W_p + W_q)$$

表示类 G_p 和 G_q 的平方距离,其中 $G_r = \{G_p, G_q\}$, W_r, W_p, W_q 分别为 G_r, G_p, G_q 类中样品的离差平方和. 利用 W_r 的定义,可得

$$\begin{aligned} W_r &= \sum_{t=1}^{n_r} (X_{(t)}^{(r)} - \overline{X}^{(r)})' (X_{(t)}^{(r)} - \overline{X}^{(r)}) \\ &= \sum_{i=1}^{n_p} (X_{(i)}^{(p)} - \overline{X}^{(r)})' (X_{(i)}^{(p)} - \overline{X}^{(r)}) \\ &\quad + \sum_{i=1}^{n_q} (X_{(i)}^{(q)} - \overline{X}^{(r)})' (X_{(i)}^{(q)} - \overline{X}^{(r)}), \end{aligned}$$

其中 $\overline{X}^{(r)}=\frac{1}{n_r}[n_p\overline{X}^{(p)}+n_q\overline{X}^{(q)}]$. 经整理可得

$$D_{pq}^2=\frac{n_pn_q}{n_r}(\overline{X}^{(p)}-\overline{X}^{(q)})'(\overline{X}^{(p)}-\overline{X}^{(q)}).$$

当样品间距离采用欧氏距离时，上式可表为

$$D_{pq}^2=\frac{n_pn_q}{n_r}d_{pq}^2,$$

其中 d_{pq}^2 表示 G_p，G_q 的重心 $\overline{X}^{(p)}$ 与 $\overline{X}^{(q)}$ 的平方距离：

$$d_{pq}^2=d^2(\overline{X}^{(p)},\overline{X}^{(q)}).$$

这表明此时 Ward 法定义的类间距离与重心法只相差一个常数倍.

当 G_p 和 G_q 合并为 G_r 后，G_r 与其他类 G_k 的距离有如下递推公式

$$D_{rk}^2=\frac{n_k+n_p}{n_r+n_k}D_{pk}^2+\frac{n_k+n_q}{n_r+n_k}D_{qk}^2-\frac{n_k}{n_r+n_k}D_{pq}^2. \quad (6.3.3)$$

在实际应用中，离差平方和法应用比较广泛，分类效果较好. 但它要求样品间距离必须采用欧氏距离.

除上述这些系统聚类法外，在 SAS/STAT 软件的 CLUSTER 过程中还给出 3 种系统聚类方法：最大似然谱系聚类法(EML)，密度估计法(DEN)和两阶段密度估计法(TWO).

三、系统聚类方法的统一

上面介绍的 8 种方法的聚类步骤完全一样，所不同的是类与类之间的距离用不同的方法定义，因而得到不同的递推公式，Lance 和 Williams 于 1967 年首先给出了统一公式，这样为编制统一的计算程序提供了很大的方便.

设 G_p 与 G_q 合并为 G_r：$G_r=\{G_p,G_q\}$，则新类 G_r 与其他类 $G_k(k\neq p,q)$ 的平方距离为

$$D_{rk}^2=\alpha_pD_{pk}^2+\alpha_qD_{qk}^2+\beta D_{pq}^2+\gamma|D_{pk}^2-D_{qk}^2|,$$

其中 α_p，α_q，β 和 γ 是参数，不同的系统聚类方法有不同的取值(见表 6.6).

表 6.6 系统聚类法的参数表

方 法	α_p	α_q	β	γ
最短距离法	1/2	1/2	0	−1/2
最长距离法	1/2	1/2	0	1/2
中间距离法	1/2	1/2	$-1/4\leqslant\beta\leqslant 0$	0
重心法*	n_p/n_r	n_q/n_r	$-\alpha_p\alpha_q$	0
类平均法	n_p/n_r	n_q/n_r	0	0
可变类平均法	$(1-\beta)n_p/n_r$	$(1-\beta)n_q/n_r$	$\beta<1$	0
可变法	$(1-\beta)/2$	$(1-\beta)/2$	$\beta<1$	0
离差平方和法*	$\frac{n_p+n_k}{n_r+n_k}$	$\frac{n_q+n_k}{n_r+n_k}$	$-\frac{n_k}{n_r+n_k}$	0

注：右上角有“*”的方法要求样品间的距离取欧氏距离.

§6.4 系统聚类法的性质及类的确定

对同一组数据用不同的聚类方法，一般会得到不尽相同的分类结果，我们应选择哪一个结果呢？应该把这组数据分为几类最合适？这是实际应用中很关心的问题. 本节将研究系统聚类法的一些简单性质及确定分类个数的准则.

一、系统聚类法的简单性质

系统聚类法具有下面两个简单性质：

(1) 单调性：设 D_k 表示系统聚类法中第 k 次并类时的距离. 如例 6.3.1，用最短距离时有：$D_1=1, D_2=1.5, D_3=2, D_4=2.5$，满足 $D_1\leqslant D_2\leqslant D_3\leqslant D_4$. 一个系统聚类法若能保证$\{D_k, k=1,2,\cdots,n-1\}$是单调上升的，则称它具有**单调性**. 并类距离具有单调性，这符合系统聚类法的基本思想. 可以证明，最短距离法、最长距离法、类平均法、可变类平均法，以及离差平方和法都具有单调性，只有重心法和中间距离法不具有单调性(见习题六的第 6-4 和 6-5 题).

(2) 空间的浓缩与扩张：以例 6.3.1 来说明该性质. 比较最短距

离法和最长距离法的并类过程，以及每一步骤相应的距离矩阵，可以看出，每一步骤都有：

$$d_{ij}(\text{短}) \leqslant d_{ij}(\text{长}) \quad (\text{对一切 } i,j).$$

这种性质称为**最长距离法比最短距离法扩张**；或称**最短距离法比最长距离法浓缩**.

对于系统聚类的各种方法，有如下结论：类平均法比最短距离法扩张，但比最长距离法浓缩；类平均法比重心法扩张，但比离差平方和法浓缩. 太浓缩的方法不够灵敏，太扩张的方法当样品容量大时容易失真. 类平均法比较适中，相对于其他方法不太浓缩也不太扩张，而且具有单调性. 因而是一种应用广泛、聚类效果较好的方法.

系统聚类各种方法的比较还可以从其他方面的性质来研究，比如系统结构性、最优化性等，各种方法的比较目前仍是值得研究的一个课题.

二、类的定义及特征

聚类分析的目的是对样品或变量进行分类，但至今对什么是类还没有给出定义. 在实际应用中，不同领域里类的含义是不尽相同的，要给出一个严格的统一定义是不容易的.

1. 类的几种定义

Rao 在 1977 年曾给出以下三种定义.

定义 6.4.1　设阈值 T 是给定的正数，若集合 G 中任两个元素的距离 d_{ij} 都满足：

$$d_{ij} \leqslant T \quad (i,j \in G),$$

则称 G 对于阈值 T 组成一个类.

定义 6.4.2　设阈值 T 是给定的正数，如果集合 G 中每个 $i \in G$ 都满足：

$$\frac{1}{n-1}\sum_{j \in G} d_{ij} \leqslant T,$$

其中 n 是集合 G 中元素的个数，则称 G 对于阈值 T 组成一个类.

定义 6.4.3　设 T 和 H $(H>T)$ 是两个给定的正数，如果集合

G 中两两元素距离的平均满足：

$$\frac{1}{n(n-1)}\sum_{i\in G}\sum_{j\in G}d_{ij}\leqslant T,\quad d_{ij}\leqslant H\quad (i,j\in G),$$

其中 n 是集合 G 中元素的个数，则称 G 对于阈值 T，H 组成一个类.

类似地还可以给出以下两种定义.

定义 6.4.4 设 T 是给定的正数，若对集合 G 中任一个 $i\in G$，一定存在 $j\in G$，使得这两个元素的距离 d_{ij} 满足：

$$d_{ij}\leqslant T\quad (i,j\in G),$$

则称 G 对于阈值 T 组成一个类.

定义 6.4.5 设阈值 T 是给定的正数，将集合 G 任意分为两类：G_1 和 G_2，这两类之间的距离 $D(G_1,G_2)$ 满足：

$$D(G_1,G_2)\leqslant T,$$

则称 G 对于阈值 T 组成一个类.

在系统聚类的方法中，我们介绍了类与类之间的 8 种距离及统一的递推公式. 类的定义 6.4.5 中可以用于 8 种类间距离的任一种；定义 6.4.1 可用于最长距离法；定义 6.4.4 可用于最短距离法；定义 6.4.2 可用于类平均法.

容易看出，前 4 种定义中，定义 6.4.1 要求是最高的，凡是符合它的类，一定也符合其后三种定义的类. 此外，凡是符合定义 6.4.2 的类，也一定是符合其后两种定义的类.

2. 类的特征

设类 G 包含的样品记为 $X_{(1)},X_{(2)},\cdots,X_{(n)}$，其中 $X_{(t)}$ $(t=1,2,\cdots,n)$ 为 m 元总体的样本. 可以从不同角度来刻画 G 的特征，常用的特征有以下三种：

(1) 均值(或称 G 的重心)：$\overline{X}_G=\dfrac{1}{n}\sum_{t=1}^{n}X_{(t)}$.

(2) 样本离差阵 A_G 及样本协方差阵 S_G 分别为：

$$A_G=\sum_{t=1}^{n}(X_{(t)}-\overline{X}_G)(X_{(t)}-\overline{X}_G)',\quad S_G=\frac{1}{n-1}A_G.$$

(3) 类的直径：用 D_G 表示类 G 的直径，常用的直径有

$$D_G=\sum_{t=1}^{n}(X_{(t)}-\overline{X}_G)'(X_{(t)}-\overline{X}_G)=\mathrm{tr}(A_G),$$

$$D_G = \max_{i,j \in G} d_{ij}.$$

三、类个数的确定

聚类分析中，类的个数如何确定的问题是一个十分困难的问题，人们至今仍未找到令人满意的方法；但这又是一个不可回避的问题.

迄今为止，我们只是从不同的角度直观地叙述了几种“类”的概念，并未给出严格的统一定义，要对各种不同形式的类给予统一的定义是比较困难的，“类”的概念是一个模糊的概念. 因此，在实际应用中，人们并不完全从类的定义来确定类. 下面介绍确定类个数的几种常见方法.

1. 由适当的阈值确定

选定某种聚类方法，按系统聚类的步骤并类后，得到一张谱系聚类图. 聚类图(或简称谱系图)只反映样品间(或变量间)的亲疏关系，它本身并没有给出分类，这就需要规定一个临界相似性尺度，用以分割谱系图而得到样品(或变量)的分类. 比如例 6.3.1，用最短距离法得谱系聚类图(见图 6.1)，给定临界值(阈值)$d=2$，根据定义 6.4.5，其含义为：当类间距离$\leqslant 2$时形成的各个类中所包含的样品间关系密切，应归属同一类. 这相当于在距离>2(比如 2.01)处切一刀，显见 5 个样品可分为两类：$X_{(1)}, X_{(2)}$为一类，$X_{(3)}, X_{(4)}, X_{(5)}$为一类.

2. 根据数据点的散布图直观地确定类的个数

如果考察的指标只有两个($m=2$)，则可通过数据点的散布图直观地确定类的个数. 如果有三个变量，可以绘制三维散布图并通过旋转三维坐标轴由数据点的分布来确定应分几个类(使用 SAS 软件). 当考察的指标在三个以上时，可以由这些指标综合出两个或三个综合变量(见第七章)后再绘制数据点在综合变量上的散布图，从而直观地确定分类个数.

3. 根据统计量确定分类个数

在 SAS/STAT 软件的 CLUSTER 过程中，提供以下一些统计量可以近似地检验分类个数如何选择更合适.

(1) R^2 统计量：假定已将 n 个样品分为 k 类，记为 $G_1, G_2, \cdots, G_k$，n_t 表示 G_t 类的样品个数($n_1+\cdots+n_k=n$)，$\overline{X}^{(t)}$表示 G_t 的重心. $X_{(i)}^{(t)}$表示 G_t 中第 i 个样品($i=1,\cdots,n_t$)，$\overline{X}$ 表示所有样品的重心，则 G_t 类中 n_t 个样品的离差平方和为

$$W_t = \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \overline{X}^{(t)})'(X_{(i)}^{(t)} - \overline{X}^{(t)}),$$

其中 $X_{(i)}^{(t)}, \overline{X}^{(t)}$和 $\overline{X}$ 均为 m 维向量，W_t 为一数值；所有样品的总离差平方和为

$$T = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \overline{X})'(X_{(i)}^{(t)} - \overline{X}),$$

T 又可以分解为

$$\begin{aligned} T &= \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \overline{X}^{(t)} + \overline{X}^{(t)} - \overline{X})'(\cdots) \\ &= \sum_{t=1}^{k} W_t + \sum_{t=1}^{k} n_t (\overline{X}^{(t)} - \overline{X})'(\overline{X}^{(t)} - \overline{X}) = P_k + B_k. \end{aligned}$$

令

$$R_k^2 = \frac{B_k}{T} = 1 - \frac{P_k}{T},$$

则 R_k^2 值越大，也就是 B_k/T 越大，表示 k 个类的类间偏差平方和的总和 B_k 在总离差平方和 T 中占的比例越大，这说明 k 个类越能够区分开. 因此 R_k^2 统计量可用于评价合并为 k 个类时的聚类效果. R_k^2 越大，聚类效果越好.

R_k^2 的值总是在 0 和 1 之间，当 n 个样品各自为不同的类时($T=B_n$)，$R_n^2=1$；当 n 个样品最后合并成同一类时($T=P_n$)，$R_1^2=0$，而且 R_k^2 的值总是随着分类个数 k 的减少而变小. 如果孤立地看每次合并后的 R_k^2 值，其意义是不大的. 如果希望通过分析 R_k^2 值来确定 n 个样品应分为几类最合适，则应该看 R_k^2 值的变化. 比如，在分为 4 个类之前的并类过程中 R_k^2 的值减少是逐渐的，改变不大；假定分为 4 类时的 $R_4^2=0.797$，而下一次合并后分为 3 类时 R_3^2 的值下降较多，比如 $R_3^2=0.402$，这时通过分析 R_k^2 统计量的变化可得出，分为 4 个类是较合适的.

(2) 半偏 R^2 统计量：

$$\text{半偏 } R_k^2 = B_{KL}^2/T = R_{k+1}^2 - R_k^2,$$

其中 $B_{KL}^2=W_M-(W_K+W_L)$，表示合并类 G_K 和 G_L 为新类 G_M 后类内离差平方和的增值，该统计量用于评价合并 G_K 和 G_L 的效果. 根据以上定义，半偏 R_k^2 的值是上一步骤 R_{k+1}^2 与该步骤 R_k^2 的差值，故查看 R_k^2 变化的大小可以得到半偏 R_k^2. 某步骤半偏 R_k^2 的值越大，说明上一次合并为 $k+1$ 个类后的效果好，该统计量用于评价一次合并的效果.

(3) 伪 F 统计量：

$$\text{伪 } F_k = \frac{(T-P_k)/(k-1)}{P_k/(n-k)} = \frac{B_k}{P_k}\frac{n-k}{k-1},$$

该统计量用于评价分为 k 个类的聚类效果，伪 F_k 值越大表示这 n 个样品可显著地分为 k 个类. 伪 F_k 统计量可以作为确定类个数的有用指标，但并不具有像 F 统计量的分布.

(4) 伪 t^2 统计量：

$$\text{伪 } t^2 = \frac{B_{KL}^2}{(W_K+W_L)/(n_K+n_L-2)},$$

该统计量用以评价此步骤合并类 G_K 和 G_L 的效果. 由伪 t^2 统计量的定义知，该值大，即表示 G_K 和 G_L 合并为 G_M 后类内离差平方和的增量 B_{KL}^2 相对于 G_K 和 G_L 两类的类内离差平方和大. 这表明上一次被合并的两个类是很分开的，也就是上一次聚类的效果是好的. 伪 t^2 统计量可以作为确定类个数的有用指标，但并不具有像随机变量 t^2 那样的分布.

4. 根据谱系图确定分类个数的准则

Bemirmen(1972)提出了应根据研究的目的来确定适当的分类方法，并提出了一些根据谱系图来分析的准则：

准则 A　各类重心之间的距离必须很大；

准则 B　确定的类中，各类所包含的元素都不要太多；

准则 C　类的个数必须符合实用目的；

准则 D　若采用几种不同的聚类方法处理，则在各自的聚类图中应发现相同的类.

应该指出，关于类个数如何确定问题，至今还没有一个合适的标

准，也就是说对任何观测数据都没有唯一正确的分类方法.

例 6.4.1 表 6.7 是我国 16 个地区农民在 1982 年支出情况的抽样调查数据的汇总资料，每个地区都调查了反映每人平均生活消费支出情况的六个指标. 试利用调查资料对 16 个地区进行分类.

解 对数据作标准化变换，样品间距离定义为欧氏距离，系统聚类的方法分别使用类平均法、中间距离法、可变类平均法和离差平方和法. 利用这几种方法所得到的并类过程及谱系聚类图都是相似的.

表 6.7 16 个地区农民生活水平的调查数据 （单位：元）

地区	食品 (X_1)	衣着 (X_2)	燃料 (X_3)	住房 (X_4)	生活用品及其他 (X_5)	文化生活服务支出 (X_6)
北京	190.33	43.77	9.73	60.54	49.01	9.04
天津	135.20	36.40	10.47	44.16	36.49	3.94
河北	95.21	22.83	9.30	22.44	22.81	2.80
山西	104.78	25.11	6.40	9.89	18.17	3.25
内蒙古	128.41	27.63	8.94	12.58	23.99	3.27
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上海	221.11	38.64	12.53	115.65	50.82	5.89
江苏	144.98	29.12	11.67	42.60	27.30	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5.00
安徽	153.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.26	16.96	19.52	21.75	6.73
江西	140.54	21.50	17.64	19.19	15.97	4.94
山东	115.84	30.26	12.20	33.61	33.77	3.85
河南	101.18	23.26	8.46	20.20	20.50	4.30

输出 6.4.1 给出使用类平均聚类法的并类过程；由输出 6.4.2 给出的谱系聚类图，易得出分为二类、三类、四类等的分类结果. 应该分为几类最合适？输出 6.4.1 中有如下几个统计量提供了有用的信息：

(1) R^2_{NCL} 统计量（列标题为 RSQ①）用于评价每次合并成 NCL 个类时的聚类效果. 现考察 R^2_{NCL} 的值随 NCL 的变化. 比如，在分为四个类之前（NCL>4）的并类过程中 R^2_{NCL} 的减少是逐渐的，改变不

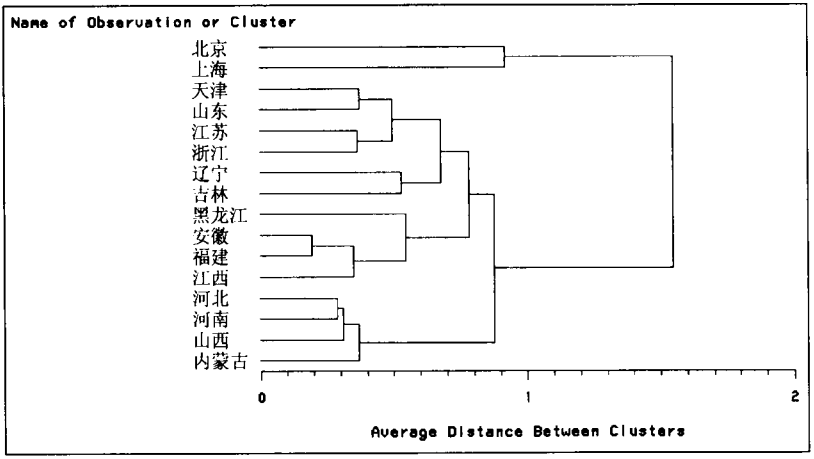
① 输出结果中每一列的列标题根据输出页的宽度不同，或为完整的英文标题，或为英文缩写，这里 RSQ 是英文缩写.

输出 6.4.1　类平均聚类法的并类过程

Average Linkage Cluster Analysis

NCL	-Clusters Joined-		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Norm RMS Dist	T i e
15	安徽	福建	2	0.00246	0.998	.	.	28.9	.	0.19216	
14	河北	河南	2	0.00548	0.992	.	.	19.2	.	0.28669	
13	CL14	山西	3	0.00670	0.985	.	.	16.8	1.2	0.30980	
12	CL15	江西	3	0.00991	0.975	.	.	14.4	4.0	0.34752	
11	江苏	浙江	2	0.00892	0.967	.	.	14.4	.	0.36575	
10	CL13	内蒙古	4	0.01055	0.956	.	.	14.5	1.7	0.36867	
9	天津	山东	2	0.00915	0.947	.	.	15.6	.	0.37044	
8	CL9	CL11	4	0.02369	0.923	.	.	13.7	2.6	0.49540	
7	辽宁	吉林	2	0.01849	0.905	.	.	14.2	.	0.52667	
6	黑龙江	CL12	4	0.02653	0.878	.	.	14.4	4.3	0.54427	
5	CL8	CL7	6	0.05563	0.822	.	.	12.7	3.7	0.67864	
4	CL5	CL6	10	0.12588	0.697	.	.	9.2	6.5	0.78178	
3	CL4	CL10	14	0.19460	0.502	0.674	-2.71	6.6	7.7	0.87404	
2	北京	上海	2	0.05614	0.446	0.507	-0.74	11.3	.	0.91764	
1	CL2	CL3	16	0.44587	0.000	0.000	0.00	.	11.3	1.54537	

输出 6.4.2　类平均聚类法的谱系聚类图



大;当分为四个类时的 $R_4^2=0.697$,而下一次合并后分为三个类时 R_3^2 下降较多($R_3^2=0.502$),由此通过对 R^2 统计量的变化分析可得出分为四个类是较合适的.

(2) 查看 R_{NCL}^2变化的大小可以由半偏 R_{NCL}^2(列标题为 SPRSQ)得到. 根据半偏 R_{NCL}^2的值是上一步骤 R_{NCL+1}^2与该步骤 R_{NCL}^2的差值,故某步骤的半偏 R_{NCL}^2值越大,说明上一步骤合并的效果越好. 此例半偏 R_{NCL}^2最大和次大分别为 NCL=1,3 和 4,说明根据半偏 R^2 准则

分为两个类、四个类或五个类是较合适的.

(3) 伪 F 统计量(列标题为 PSF)用于评价分为 NCL 个类的聚类效果. 伪 F_{NCL} 值越大表示这些观测样品可以显著地分为 NCL 个类. 此例中伪 F_{NCL} 最大和次大分别为 NCL=5 和 2(当 NCL<6),说明根据伪 F 准则分为五个类或两个类是较合适的.

(4) 伪 t^2 统计量(列标题为 PST2)用以评价此步骤合并类的效果. 由该统计量的定义知,伪 t^2 值大表明上一次合并的两个类是很分开的,也就是上一次聚类的效果是好的. 此例中伪 t^2 最大和次大分别为 NCL=1,3 和 4,说明根据伪 t^2 准则分为两个类、四个类或五个类是较合适的.

从以上分析可以看出:半偏 R^2 准则支持分为两类、四类和五类;伪 F 统计量支持分为两类或五类;伪 t^2 统计量支持分为两类、四类或五类. 综合分析,认为采用类平均法分类,将 16 个地区分为两类或五类较合适. 分为五类的结果为:$G_1^{(5)}=\{$北京$\}$,$G_2^{(5)}=\{$上海$\}$(各 1 个地区),$G_3^{(5)}=\{$天津,山东,江苏,浙江,辽宁,吉林$\}$(六个地区),$G_4^{(5)}=\{$黑龙江,安徽,福建,江西$\}$(四个地区),$G_5^{(5)}=\{$河北,河南,山西,内蒙古$\}$(四个地区). 若分为两类,则 $G_1^{(2)}=\{G_1^{(5)},G_2^{(5)}\}$,$G_2^{(2)}=\{G_3^{(5)},G_4^{(5)},G_5^{(5)}\}$.

用 Ward 法分类可以得出:R^2 准则支持分为两类、三类和四类;伪 F 统计量支持分为四类或五类;伪 t^2 统计量支持分为两类、三类或四类. 综合分析,认为采用 Ward 法分类,将 16 个地区分为两类或四类较合适. 分为四类的结果为:$G_1^{(4)}=\{$北京,上海$\}$(二个地区),$G_2^{(4)}=\{$天津,山东,江苏,浙江,辽宁,吉林$\}$(六个地区),$G_3^{(4)}=\{$黑龙江,安徽,福建,江西$\}$(四个地区),$G_4^{(4)}=\{$河北,河南,山西,内蒙古$\}$(四个地区). 若分为两类,则 $G_1^{(2)}=G_1^{(4)}$,$G_2^{(2)}=\{G_2^{(4)},G_3^{(4)},G_4^{(4)}\}$.

不同的聚类方法得到的结果或多或少都有些差别. 在实际应用中,应综合各种计算结果,提出合适的分类结果. 比如此例把 16 个地区的农民生活消费支出情况分为五类,并计算出各类地区农民生活平均消费水平(见表 6.8). 由表 6.8 可见第一类属低消费水平,第二类属中等消费水平,第三类消费水平较高,而北京市和上海市的农民

生活水平比较富裕，属高消费水平.

表 6.8　16 个地区的分类及平均消费水平

类　别	第一类	第二类	第三类	第四类	第五类
该类所包含的地区	河北，河南 山西，内蒙古	黑龙江，安徽 福建，江西	天津，山东，江苏 浙江，辽宁，吉林	北京	上海
食品平均消费	107.395	138.698	145.165	190.33	221.11
衣着平均消费	24.708	23.855	32.457	43.77	38.64
燃料平均消费	8.275	15.865	13.870	9.73	12.53
住房平均消费	16.278	19.000	34.431	60.54	115.65
生活用品以及其他平均消费	21.368	19.413	32.715	49.01	50.82
文化生活服务支出平均消费	3.405	6.025	4.537	9.04	5.89

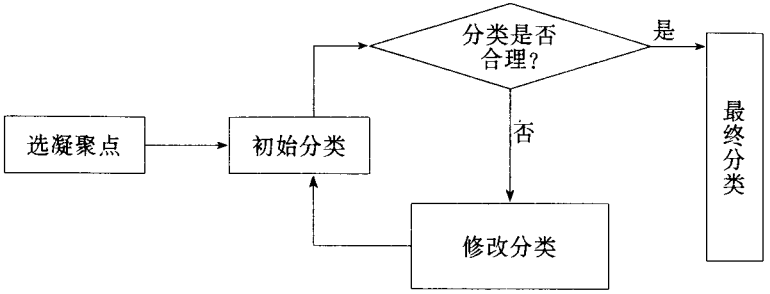
§6.5　动态聚类法

系统聚类法一次形成类以后就不能改变了，这就要求一次分类分得比较准确，对分类的方法就提出较高的要求，相应的计算量自然也比较大. 如 Q 型系统聚类法，聚类过程是在样品间的距离矩阵基础上进行的，当样本容量很大时，需要占据足够大的计算机内存空间；而且在并类过程中，需要将每类样品和其他类样品间的距离逐一加以比较，以决定应合并的类别，故需要较多的计算时间. 所以对于大样本问题，Q 型系统聚类法可能会因计算机内存或计算时间的限制而无法进行，这给应用带来一定的不便. 基于这种情况，产生了动态聚类，即**动态聚类法**.

动态聚类法又称为逐步聚类法，其基本思想是，开始先粗略地分一下类，然后按照某种最优的原则修改不合理的分类，直至类分得比较合理为止，这样就形成一个最终的分类结果. 该方法具有计算量较小，占用计算机内存空间较少，方法简单的优点，适用于大样本的 Q 型聚类分析.

为了粗略地分一下类(简称为**初始分类**)，有时需要先选一批“凝聚点”，然后让样品向最近的凝聚点聚集，这样由凝聚点聚集形成的

类，就得到初始分类. 动态聚类法的聚类过程可用下面框图描述.



一、选择凝聚点和确定初始分类

凝聚点就是一批有代表性的点，是欲形成类的中心的点. 凝聚点的选择直接决定初始分类，对分类结果也有很大的影响. 由于凝聚点的不同选择，其最终分类结果也将出现不同，故选择时要慎重. 通常选择凝聚点的方法有：

(1) 人为选择：当人们对所欲分类的问题有一定了解时，根据经验，预先确定分类个数和初始分类，并从每一类中选择一个有代表性的样品作为凝聚点.

(2) 将所有样品人为地分为 k 类，计算每一类的重心，并将这些重心作为凝聚点.

(3) 用密度法选择凝聚点：以某个正数 d 为半径，以每个样品为球心，落在这个球内的样品数(不包括作为球心的样品)就叫做这个样品的密度. 计算所有样品点的密度后，首先选择密度最大的样品作为第一凝聚点，并且人为地规定一个正数 D(一般 $D>d$，常取 $D=2d$). 然后选出次大密度的样品点，若它与第一个凝聚点的距离大于 D，则将其作为第二个凝聚点；否则舍去此点，再选密度次于它的样品. 这样，按密度大小依次考察，直至全部样品考察完毕为止. 此方法中，d 要给的合适，太大了使凝聚点个数太少，太小了使凝聚点个数太多.

(4) 人为地选择一正数 d. 首先以所有样品的均值作为第一凝

聚点，然后依次考察每个样品，若某样品与已选定的凝聚点的距离均大于 d，该样品作为新的凝聚点，否则考察下一个样品.

(5) 随机地选择. 如果对样品的性质毫无所知，可采用随机数表来选择，打算分几类就选几个凝聚点，或者就用前 k 个样品作为凝聚点(假设分 k 类). 这个方法一般不提倡使用.

确定初始分类常用的方法有：

(1) 人为分类，凭经验将样品进行初步分类.

(2) 选择凝聚点后，每个样品按与其距离最近的凝聚点归类.

(3) 选择一批凝聚点后，每个凝聚点自成一类，将样品依次归入其距离最近的凝聚点所在的类，并重新计算该类的重心，以代替原来的凝聚点，再考虑下一个样品的归类，直至所有样品都归类为止.

(4) 先将数据作标准化处理，仍用 x_{ij} 表示标准化后第 i 个样品的第 j 个指标. 令

$$x_{i\cdot} = \sum_{j=1}^{m} x_{ij}, \quad R = \max_{1\leqslant i\leqslant n} x_{i\cdot} - \min_{1\leqslant i\leqslant n} x_{i\cdot}.$$

如欲将全部样品分为 k 类，对每个样品 $X_{(i)}$ 计算：

$$\frac{(k-1)(x_{i\cdot} - \min x_{i\cdot})}{R} + 1.$$

假设与这个数接近的整数为 l，则将样品 $X_{(i)}$ 归入第 l 类($1\leqslant l\leqslant k$).

(5) 用某种聚类方法得到一个分类，将其作为初始分类. 当样本量大时，有时只用部分样品按某种聚类方法进行分类，如用每类重心作为凝聚点，再用(2)或(3)的方法对全部样品归类后得初始分类.

二、逐步聚类法

逐步聚类的不同方法主要是以修改分类的不同原则来区分的，常用方法有按批修改法，逐个修改法等. 本节重点介绍按批修改法.

1. 按批修改法

(1) 按批修改法的步骤：

步骤 1：选择一批凝聚点(个数人为指定)，并选定所用距离定义；

步骤 2：将所有样品按与其距离最近的凝聚点归类；

步骤 3：计算每一类的重心，将重心作为新的凝聚点，转到步骤 2. 如果某一步骤所有的新凝聚点与前一次的老凝聚点重合，则过程终止. 有时不绝对要求这个过程收敛，而是人为规定这个修改过程重复若干次就行了.

(2) 分类函数：假设 $X_{(i)}$ $(i=1,2,\cdots,n)$为 n 个样品点，初始分为 k 类：$G_1,G_2,\cdots,G_k$，它们的重心记为 $\overline{X}^{(1)},\overline{X}^{(2)},\cdots,\overline{X}^{(k)}$，每类的样品数记为 n_i $(i=1,2,\cdots,k)$. 用 t_i 表示样品 $X_{(i)}$所属类的标号，如 $t_1=2$ 表示第 1 个样品属第 2 类.

样品 $X_{(i)}$到类 G_j 的距离 D_{ij} 定义为

$$D_{ij}^2=(X_{(i)}-\overline{X}^{(j)})'(X_{(i)}-\overline{X}^{(j)}),$$

则分类函数定义为：

$$l(G_1,G_2,\cdots,G_k)=\sum_{i=1}^{n}D_{it_i}^2.$$

上式定义的分类函数，实质上就是系统聚类分析中的离差平方和 W，其中 $W=\sum_{i=1}^{k}W_i$，而 W_i 为 G_i 类样品的离差平方和.

按批修改法的修改原则就是使这个分类函数逐渐减小，直至不能再减小为止.

例 6.5.1 已知 5 个样品的观测值为：1,4,5,7,11. 试用按批修改的动态聚类法对 5 个样品进行分类.

解 (1) 选凝聚点：用密度法，取 $d=2,D=4$；采用欧氏距离. 各点的密度如下表所示：

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
密度	0	1	2	1	0

第一个凝聚点选 $X_{(3)}$，因 $X_{(2)},X_{(4)}$与 $X_{(3)}$的距离$<D$，第二个凝聚点选 $X_{(1)}$，第三个凝聚点选 $X_{(5)}$.

(2) 初始分类：按最小距离的原则将所有样品归类，结果是：

$$G_1^{(0)}=\{X_{(3)},X_{(2)},X_{(4)}\},\quad G_2^{(0)}=\{X_{(1)}\},\quad G_3^{(0)}=\{X_{(5)}\}.$$

(3) 修改分类：首先计算各类重心，它们分别是 $5\frac{1}{3}$,1,11. 再以

它们作为新凝聚点，按最小距离原则归类，结果是：

$$G_1^{(1)}=\{X_{(3)},X_{(2)},X_{(4)}\},\quad G_2^{(1)}=\{X_{(1)}\},\quad G_3^{(1)}=\{X_{(5)}\}.$$

这三类的重心仍然为 $5\frac{1}{3}$，1，11，因此过程终止. 最终分类就是 $G_i^{(1)}$ $(i=1,2,3)$.

附注 1　按批修改法的优点是计算量小，速度快，但其分类结果依赖于凝聚点的选择. 若选 $X_{(3)},X_{(5)}$ 作为凝聚点，则最终分类是：

$$G_1=\{X_{(1)},X_{(2)},X_{(3)},X_{(4)}\},\quad G_2=\{X_{(5)}\}.$$

若选 $X_{(1)},X_{(5)}$ 作为凝聚点，则最终分类是：

$$G_1=\{X_{(1)},X_{(2)},X_{(3)}\},\quad G_2=\{X_{(4)},X_{(5)}\}.$$

附注 2　在按批修改法中，有人将步骤 3 改为：计算每一类重心，取老凝聚点与重心联线的对称点作为新凝聚类，转到步骤 2. 如果某一步骤所有新凝聚点与前一次的老凝聚点重合，则过程终止. 这样做在某些场合会得到更好的分类结果.

2. 逐个修改法

按批修改法是当样品全部归类后才改变凝聚点. 另一种自然的想法是每对一个样品进行分类后，同时改变凝聚点，这就产生了逐个修改法. 逐个修改法在许多教科书上称做 K-均值法. 逐个修改的方法不止一种，以下介绍常见的一种，它的聚类步骤是：

步骤 1：规定样品间的距离，人为地定出三个数：K(分类数)，C(类间距离的最小值)和 R(类内距离的最大值)；取前 K 个样品点作为凝聚点.

步骤 2：计算这 K 个凝聚点两两之间的距离，如最小的距离$<C$，则将相应的两个凝聚点合并，用这两个点的重心作为新凝聚点，再重复步骤 2，直至所有凝聚点之间的距离均$\geqslant C$ 为止.

步骤 3：将剩下的 $n-K$ 个样品逐个归类，对每一个样品，计算该样品与所有凝聚点的距离，如最小距离$>R$，则该样品作为新凝聚点；如最小距离$\leqslant R$，则将该样品归入与它距离最近的凝聚点所在的类，随即重新计算这一类的重心，以重心作为新的凝聚点. 如凝聚点之间的距离都$\geqslant C$，则考虑下一个样品，否则用步骤 2 进行合并后再考虑下一个样品，直至所有样品都归了类.

步骤 4：将样品从头至尾再逐个按步骤 3 进行归类，不同之处是：某个样品归类后，如分类与原来一致，则重心不必计算；如分类与原来不同，则涉及到的两类重心要重新计算.

如果新的分类与上一次相同，则聚类过程结束，否则重复步骤 4.

例 6.5.2 对例 6.5.1 的 5 个样品用逐个修改法进行聚类.

解 (1) 用欧氏距离，取 $K=3, C=2, R=3$. 取 $X_{(1)}, X_{(2)}, X_{(3)}$ 作为凝聚点.

(2) 计算凝聚点之间的距离，$d(1,2)=3$，$d(1,3)=4$，$d(2,3)=1<C=2$，将 $X_{(2)}$ 和 $X_{(3)}$ 合并(记为新 2 类：2^*)，它们的重心是 $(4+5)/2=4.5$，它与 $X_{(1)}$ 的距离 $d(1,2^*)=3.5>2$. 凝聚点有两个：1 和 4.5；相应的两类记为 G_1 和 G_2.

(3) 考虑样品 $X_{(4)}$，它与两凝聚点的距离为：$d(G_1,4)=6$，$d(G_2,4)=2.5$，最小值 $2.5<3$，不能作为新凝聚点，归入 G_2；再考虑 $X_{(5)}$，因 $d(G_1,5)=10$，$d(G_2,5)=6.5$，最小值 $6.5>3$，故 $X_{(5)}$ 作为新凝聚点单独成一类.

至此我们得到三类：$G_1=\{X_{(1)}\}$，$G_2=\{X_{(2)}, X_{(3)}, X_{(4)}\}$，$G_3=\{X_{(5)}\}$.

(4) 将样品从头至尾按上面的(3)进行归类. 聚类结果同上. 故聚类过程结束，并得到最终分类为 G_1, G_2, G_3.

附注 1 逐个修改法的最终分类与样品的考虑顺序有关，一般按 $X_{(1)} \Rightarrow X_{(5)}$ 次序，如果按 $X_{(5)} \Rightarrow X_{(1)}$ 的次序考虑，K, C, R 不变，则分类结果为：

$$G_1=\{X_{(5)}\},\quad G_2=\{X_{(4)}\},\quad G_3=\{X_{(2)}, X_{(3)}\},\quad G_4=\{X_{(1)}\}.$$

由于分类结果与样品归类的次序有关，逐个修改法开始选凝聚点时最好选有代表性的点，而不是简单的取前 K 个样品，这样聚类结果会合理一些.

附注 2 逐个修改法的最终分类与三个参数有关，因此在计算过程中最好让这三个参数作适当变化，最后根据实际问题的要求取舍聚类结果.

例 6.5.3 试用 SAS/STAT 软件的 FASTCLUS（快速聚类）过程对 16 个地区农民生活水平的调查数据进行分类.

解　为使用动态聚类法对例 6.4.1 中表 6.7 给出的 16 个地区的农民在 1982 年支出情况的抽样调查资料进行分类，首先对数据进行标准化，然后使用 FASTCLUS（快速聚类）过程对标准化后的数据进行动态聚类.

输出 6.5.1　动态聚类的初始凝聚点（数据标准化）

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=5 Maxiter=1

Initial Seeds

Cluster	X1	X2	X3	X4	X5	X6
1	1.46034	2.17117	-0.77840	1.04836	1.88228	2.48234
2	2.38417	1.39269	-0.02619	3.12687	2.05005	0.55074
3	-1.39460	-1.00649	-0.89392	-0.38861	-0.54617	-1.34407
4	0.12021	0.51102	1.30688	-0.20569	0.96281	-0.93322
5	0.09740	-1.24473	1.16391	-0.49874	-0.64442	1.06583

输出 6.5.2　动态聚类的分类结果（数据标准化）

OBS	GROUP	CLUSTER	DISTANCE
1	北京	1	0.00000
2	上海	2	0.00000
3	河北	3	0.70731
4	山西	3	0.64693
5	内蒙古	3	0.84370
6	河南	3	0.64426
7	天津	4	1.30193
8	辽宁	4	1.19017
9	吉林	4	1.71508
10	浙江	4	1.10310
11	山东	4	1.16730
12	黑龙江	5	1.17739
13	江苏	5	1.45351
14	安徽	5	0.62637
15	福建	5	0.92034
16	江西	5	1.18491

由输出 6.5.1 中可以看出，5 个初始凝聚点是从标准化数据集中按指定规则选取的 5 个观测样品. 输出 6.5.2 给出了分类的结果，分类结果类似数据标准化后的系统聚类法.

§6.6　有序样品聚类法（最优分割法）

在以上几节讨论的问题中，样品是相互独立的，因而分类时彼此是平等的. 但在有些实际问题中，要求样品分类时不能打乱次序. 例如在油田勘探中，需要通过岩心了解地层的结构，故而要求对地层的不同结构进行分类. 这时岩心所在的位置（即样品的次序）在分类时

不能打乱. 如果用 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 表示 n 个有序的样品,则每一类必须是这样的形式: $\{X_{(i)},X_{(i+1)},\cdots,X_{(i+k)}\}$,其中 $1\leqslant i\leqslant n,k\geqslant 0$ 且 $i+k\leqslant n$,即同一类样品必须是互相邻接的. 这种分类问题称为**有序样品聚类法**,也称为**最优分割法**.

设有 n 个有序样品 $X_{(1)},X_{(2)},\cdots,X_{(n)}$,其中每个样品有 m 个指标,即 $X_{(i)}$ $(i=1,\cdots,n)$ 为 m 维向量. n 个样品分成 k 类的一切可能分法有 C_{n-1}^{k-1} 种,这比不相关样品的可能分法要少得多,故在 n 不大的情况下,有可能讨论所有可能的分类结果,并有可能在某种损失函数意义下,从中求得最优解. 本节介绍费希尔发展的一个算法,它可保证求得最优解,故此法称为**最优分割法**,又称为**费希尔算法**.

系统聚类法开始时将 n 个样品 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 分成 n 类,然后逐步并类,直到所有样品并成一类为止. 而最优分割法则相反,开始时将所有样品归为一类,然后分成两类、三类等等,直到分为 n 个类. 这是两种不同类型的聚类方法,但最优分割法定义分类的损失函数的思想类似于系统聚类方法中的 Ward 法,即要求分类后产生的离差平方和的增量最小.

一、最优分割法的聚类步骤

设有序样品依次为 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ ($X_{(i)}$ 为 m 维向量).

1. 定义类的直径

设某一类 G 包含的样品有 $\{X_{(i)},X_{(i+1)},\cdots,X_{(j)}\}$ $(j>i)$,记为 $G=\{i,i+1,\cdots,j\}$. 该类的均值向量 $\overline{X}_G$ 为

$$\overline{X}_G=\frac{1}{j-i+1}\sum_{t=i}^{j}X_{(t)}.$$

用 $D(i,j)$ 表示这一类的直径,常用的直径有:

$$D(i,j)=\sum_{t=i}^{j}(X_{(t)}-\overline{X}_G)'(X_{(t)}-\overline{X}_G). \qquad (6.6.1)$$

当 $m=1$ 时,也可以定义直径为

$$D(i,j)=\sum_{t=i}^{j}|X_{(t)}-\widetilde{X}_G|,$$

其中 $\tilde{X}_G$ 是这一类数据的中位数.

2. 定义分类的损失函数

用 $b(n,k)$ 表示将 n 个有序样品分为 k 类的某一种分法. 常记分法 $b(n,k)$ 为：

$$\begin{aligned} G_1 &= \{i_1, i_1+1, \cdots, i_2-1\}, \\ G_2 &= \{i_2, i_2+1, \cdots, i_3-1\}, \\ &\cdots\cdots\cdots\cdots \\ G_k &= \{i_k, i_k+1, \cdots, n\}, \end{aligned}$$

其中分点为 $1=i_1<i_2<i_3<\cdots<i_k<n=i_{k+1}-1$(即 $i_{k+1}=n+1$).

定义上述分类法的损失函数为

$$L[b(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1}-1). \tag{6.6.2}$$

当 n,k 固定时，$L[b(n,k)]$越小，即表示各类的离差平方和越小，分类是合理的，因此要寻找一种分法 $b(n,k)$，使分类损失函数 L 达最小. 记 $P(n,k)$ 是使(6.6.2)式达到极小的分类法.

3. $L[b(n,K)]$的递推公式

费希尔算法最核心的部分是利用以下两个递推公式：

$$\begin{cases} L[P(n,2)] = \min\limits_{2\leqslant j\leqslant n} \{D(1,j-1) + D(j,n)\}, \\ L[P(n,k)] = \min\limits_{k\leqslant j\leqslant n} \{L[P(j-1,k-1)] + D(j,n)\}. \end{cases} \tag{6.6.3}$$

以上两个公式由(6.6.2)式及 $P(n,k)$ 的定义即可证明. (6.6.3)式的第二式表明，若要寻找将 n 个样品分为 k 类的最优分割，应建立在将 $j-1$ 个样品分为 $k-1$ 类的最优分割基础上(这里 $j=2,3,\cdots,n$).

4. 最优解的求法

若分类数 k $(1<k<n)$ 已知，求分类法 $P(n,k)$，使它在损失函数意义下达最小，其求法如下：

首先找分点 j_k，使递推公式(6.6.3)达极小，即

$$L[P(n,k)] = L[P(j_k-1,k-1)] + D(j_k,n).$$

于是得第 k 类 $G_k=\{j_k, j_k+1, \cdots, n\}$. 然后找 j_{k-1}，使它满足

$$L[P(j_{k-1},k-1)]=L[P(j_{k-1}-1,k-2)]+D(j_{k-1},j_k-1),$$

得到第 $k-1$ 类 $G_{k-1}=\{j_{k-1},j_{k-1}+1,\cdots,j_k-1\}$. 类似的方法依次可得到所有类 $G_1,G_2,\cdots,G_k$,这就是我们欲求的最优解,即

$$P(n,k)=\{G_1,G_2,\cdots,G_k\}.$$

总之,为了求最优解,主要是计算

$\{D(i,j);1\leqslant i<j\leqslant n\}$ 和 $\{L[P(i,j)],1\leqslant i\leqslant n,i\leqslant j\leqslant n\}$.

下面通过一个例子来说明最优解的具体求法.

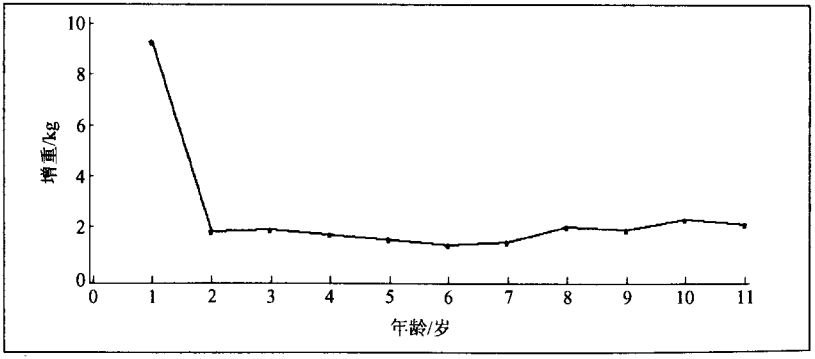
二、应用简例

例 6.6.1 为了了解儿童的生长发育规律,今统计了男孩从出生到 11 岁每年平均增长的重量如下表所示. 试问男孩发育可分为几个阶段?

年龄/岁	1	2	3	4	5	6	7	8	9	10	11
增加重量/kg	9.3	1.8	1.9	1.7	1.5	1.3	1.4	2.0	1.9	2.3	2.1

解 这是一个有序样品的聚类问题,我们用最优分割法来进行聚类. 输出 6.6.1 首先给出男孩每年体重的增加随年龄的变化规律图. 由此图可直观地看出,男孩发育确实可以分为几个阶段.

输出 6.6.1 男孩每年体重的增加随年龄的变化



有序样品聚类的具体步骤如下:

(1) 计算直径$\{D(i,j)\}$,结果见表 6.9. 因每个样品只有一个指

标(即 $m=1$),由(6.6.1)式的定义,故有

$$D(i,j)=\sum_{t=i}^{j}(X_{(t)}-\overline{X}_G)^2.$$

例如计算 $D(5,7)$,此时类 G 包含三个样品$\{X_{(5)},X_{(6)},X_{(7)}\}$,故有

$$\overline{X}_G=\frac{1}{3}(1.5+1.3+1.4)=1.4,$$

$$D(5,7)=(1.5-1.4)^2+(1.3-1.4)^2+(1.4-1.4)^2=0.02.$$

表 6.9　直径 $D(i,j)$

j \ i	1	2	3	4	5	6	7	8	9	10
2	28.125									
3	37.007	0.005								
4	42.208	0.020	0.020							
5	45.992	0.088	0.080	0.020						
6	49.128	0.232	0.200	0.080	0.020					
7	51.100	0.280	0.232	0.088	0.020	0.005				
8	51.529	0.417	0.393	0.308	0.290	0.287	0.180			
9	51.980	0.469	0.454	0.393	0.388	0.370	0.207	0.005		
10	52.029	0.802	0.800	0.774	0.773	0.708	0.420	0.087	0.080	
11	52.182	0.909	0.909	0.895	0.889	0.793	0.452	0.088	0.080	0.020

(2) 计算最小分类损失函数$\{L[P(l,k)],3\leqslant l\leqslant 11,2\leqslant k\leqslant 10\}$,即分别计算将 l 个样品分成两类、三类…时,最优分割的损失函数,所有结果列于表 6.10.

表 6.10　最小分类损失函数 $L[P(l,k)]$

l \ k	2	3	4	5	6	7	8	9	10
3	0.005(2)								
4	0.020(2)	0.005(4)							
5	0.088(2)	0.020(5)	0.005(5)						
6	0.232(2)	0.040(5)	0.020(6)	0.005(6)					
7	0.280(2)	0.040(5)	0.025(6)	0.010(6)	0.005(6)				
8	0.417(2)	0.280(8)	0.040(8)	0.025(8)	0.010(8)	0.005(8)			
9	0.469(2)	0.285(8)	0.045(8)	0.030(8)	0.015(8)	0.010(3)	0.005(8)		
10	0.802(2)	0.367(8)	0.127(8)	0.045(10)	0.030(10)	0.015(10)	0.010(10)	0.005(8)	
11	0.909(2)	0.368(8)	0.128(8)	0.065(10)	0.045(11)	0.030(11)	0.015(11)	0.010(11)	0.005(11)

首先,计算$\{L[P(l,2)],3\leqslant l\leqslant 11\}$(即表 6.10 的第一列),如 $l=3$ 时由递推公式(6.6.3)及表 6.9 可得:

$$L[P(3,2)]=\min_{2\leqslant j\leqslant 3}\{D(1,j-1)+D(j,3)\}$$
$$=\min\{D(1,1)+D(2,3),D(1,2)+D(3,3)\}$$
$$=\min\{0+0.005,28.125+0\}=0.005.$$

上式表示将三个样品分为两类,共有两种可能分法:{1},{2,3};或{1,2},{3},这两种分法的最小损失函数为 0.005(即前一种分法).因为这个最小值在 $j=2$ 时达到,记为 0.005(2). 类似的,当 $l=4$ 时有:

$$L[P(4,2)]=\min_{2\leqslant j\leqslant 4}\{D(1,j-1)+D(j,4)\}$$
$$=\min\{D(1,1)+D(2,4),D(1,2)+D(3,4),$$
$$D(1,3)+D(4,4)\}$$
$$=\min\{0.020,28.145,37.007\}=0.020.$$

最小值在 $j=2$ 时达到,记为 $L[P(4,2)]=0.020(2)$. 表 6.10 中 $k=2$ 的那一列,括弧中的数字都是 2,表示对一切形如 $\{X_{(1)},X_{(2)},\cdots,X_{(l)}\}$ $(3\leqslant l\leqslant 11)$的类,如欲分成两类都以 $G_1=\{X_{(1)}\}$, $G_2=\{X_{(2)},\cdots,X_{(l)}\}$的分法为最优,它使分类损失函数达到极小.

其次,计算 $\{L[P(l,3)],4\leqslant l\leqslant 11\}$,如 $l=4$ 时由递推公式(6.6.3)第二式、表 6.9 及表 6.10 第一列可得:

$$L[P(4,3)]=\min\{L[P(2,2)]+D(3,4),L[P(3,2)]+D(4,4)\}$$
$$=\min\{0+0.020,0.005+0\}=0.005.$$

表 6.10 中第二列($k=3$)的其余各项的计算方法也类似. 同样,对表 6.10 中 $k=4,5,\cdots,10$ 的其余各列的计算方法也类似. 各列括弧内的数字含义同上.

(3) 求最优分类. 假如我们希望分成三类,即 $k=3$,由表 6.10 最后一行查得 $L[P(11,3)]=0.368$,括号中数字为 8,这说明最优解的分类的损失函数是 0.368,分类时首先分出第三类 $G_3=\{X_{(8)}\sim X_{(11)}\}$. 再对其余的 7 个样品考虑分为两类的最优分法,查表 6.10 中 $l=7,k=2$ 的位置得 $L[P(7,2)]=0.280$,括号中数字为 2,故

$$G_2=\{X_{(2)}\sim X_{(7)}\},\quad G_1=\{X_{(1)}\}.$$

从而求得最优分类

$$P(11,3):\{X_{(1)}\},\{X_{(2)}\sim X_{(7)}\},\{X_{(8)}\sim X_{(11)}\}.$$

当 k 取其余值时，分类情况列于表 6.11.

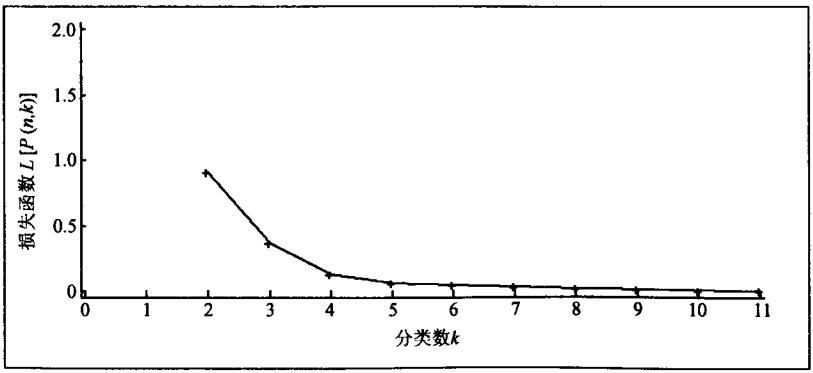
表 6.11　分类情况

k	$L[P(11,k)]$	分　类
1	52.182	{9.3,1.8,1.9,1.7,1.5,1.3,1.4,2.0,1.9,2.3,2.1}
2	0.909	{9.3},{1.8,1.9,1.7,1.5,1.3,1.4,2.0,1.9,2.3,2.1}
3	0.368	{9.3},{1.8,1.9,1.7,1.5,1.3,1.4},{2.0,1.9,2.3,2.1}
4	0.128	{9.3},{1.8,1.9,1.7},{1.5,1.3,1.4},{2.0,1.9,2.3,2.1}
5	0.065	{9.3},{1.8,1.9,1.7},{1.5,1.3,1.4},{2.0,1.9},{2.3,2.1}
6	0.045	{9.3},{1.8,1.9,1.7},{1.5,1.3,1.4},{2.0,1.9},{2.3},{2.1}
7	0.030	{9.3},{1.8,1.9},{1.7},{1.5,1.3,1.4},{2.0,1.9},{2.3},{2.1}
8	0.015	{9.3},{1.8,1.9},{1.7},{1.5},{1.3,1.4},{2.0,1.9},{2.3},{2.1}
9	0.010	{9.3},{1.8},{1.9},{1.7},{1.5},{1.3,1.4},{2.0,1.9},{2.3},{2.1}
10	0.005	{9.3},{1.8},{1.9},{1.7},{1.5},{1.3},{1.4},{2.0,1.9},{2.3},{2.1}
11	0.000	{9.3},{1.8},{1.9},{1.7},{1.5},{1.3},{1.4},{2.0},{1.9},{2.3},{2.1}

(4) 决定 k. 如果从生理角度预先能定出 k，当然最好，这样从表 6.10 即可知道如何分类. 有时事先不能确定 k，这时可作出 $L[P(n,k)]$ 随 k 变化的趋势图，如此例作出的图形见输出6.6.2. 从该输出我们看到曲线在 $k=3,4$ 处拐弯，即分三类或四类为好.

最优分割法虽然计算简单，但是当 n 很大时，由于要存贮直径 $\{D(i,j)\}$，对容量不大的计算机还是有困难的.

输出 6.6.2　损失函数 $L[P(n,k)]$ 随 k 变化的趋势图



三、分类个数的确定

分类数 k 的确定对许多问题都很重要，上面给出的方法是通过 $L[P(n,k)]$ 对 k 作图，在曲线拐弯处来确定 k. 当曲线拐弯很平缓时，可以选取的 k 很多，这时需要有其他的办法来确定，比如均方比和特征根法(省略).

§6.7 变量聚类方法

在本章§6.1已提到，聚类分析根据分类对象的不同可分为Q型(对样品)和R型(对变量). 前面几节讨论的内容是Q型聚类问题，即对样品进行聚类的问题. 在实际工作中，对所考察的一些变量进行分类也是十分重要的. 在统计分析中，为了避免遗漏重要因素，人们往往初始选取所考察的变量时，总是尽可能多地考虑所有相关的因素. 而这样做的结果则是需要考察的变量过多，变量间的相关性也较大，给统计分析带来很大的不便. 因此人们常常希望研究变量间的相似关系，按照变量的相关关系把它们聚合为若干类，然后观察和说明影响系统特性的主要特征.

一、变量分类的系统聚类法

使用类似于Q型聚类分析中最常用的系统聚类法的思路和基本步骤对变量进行聚类. 具体操作时可以采用以下两种方法：

(1) 可以把观测数据阵中样品和变量的地位调换一下，也就是把数据阵作一转置，然后仍使用CLUSTER过程对变量进行聚类. 此方法使用虽简单，但没有考虑到使用相似系数或相关系数作为变量相关性的度量比使用距离要更合适些，故聚类的效果也许不尽如人意.

(2) 计算变量间的相关(或相似)系数矩阵 R，然后按§6.2中小节三“变量间的相似系数和距离”中的定义把相关(或相似)系数矩阵转化为距离矩阵. 用此距离矩阵作为CLUSTER过程的输入数据

阵，实现对变量进行分类的目的.

例 6.7.1　对 305 名女中学生测量 8 项体型指标：X_1 为身高，X_2 为手臂长，X_3 为手肘长，X_4 为小腿长，X_5 为体重，X_6 为颈围，X_7 为胸围，X_8 为胸宽. 表 6.12 是由 305 名中学生的观测数据计算得到的相关系数阵. 试对 8 个体型变量进行分类.

表 6.12　8 个体型变量的相关系数阵

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	0.846	0.805	0.859	0.473	0.398	0.301	0.382
X_2	0.846	1.000	0.881	0.826	0.376	0.326	0.277	0.415
X_3	0.805	0.881	1.000	0.801	0.380	0.319	0.237	0.345
X_4	0.859	0.826	0.801	1.000	0.436	0.329	0.327	0.365
X_5	0.473	0.376	0.380	0.436	1.000	0.762	0.730	0.629
X_6	0.398	0.326	0.319	0.329	0.762	1.000	0.583	0.577
X_7	0.301	0.277	0.237	0.327	0.730	0.583	1.000	0.539
X_8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

解　表 6.12 给出了 8 个变量的相关系数阵，首先把相关系数转换为距离，令 $d_{ij}=1-|r_{ij}|$；然后从变量间的距离矩阵出发，调用 CLUSTER 过程使用类平均法对变量进行聚类，输出 6.7.1 给出其并类过程. 从输出 6.7.1 中可以看出：8 个变量若分为两类，则

$$G_1^{(2)}=\{X_1,X_2,X_3,X_4\},\quad G_2^{(2)}=\{X_5,X_6,X_7,X_8\};$$

若分为三类，则

$$G_1^{(3)}=G_1^{(2)},\quad G_2^{(3)}=\{X_5,X_6,X_7\},\quad G_3^{(3)}=\{X_8\}.$$

输出 6.7.1　变量系统聚类的并类过程

Average Linkage Cluster Analysis

Root-Mean-Square Distance Between Observations = 0.524448

Number of Clusters	--Clusters Joined--		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized RMS Distance	Tie
7	x2	x3	2	22.49	.	0.226905	
6	x1	x4	2	22.22	.	0.268054	
5	CL6	CL7	4	16.68	2.87	0.345879	
4	x5	x6	2	17.07	.	0.453810	
3	CL4	x7	3	14.38	2.57	0.669800	
2	CL3	x8	4	17.15	2.12	0.800759	
1	CL5	CL2	8	.	17.15	1.234171	

二、VARCLUS(变量聚类)过程简介

在SAS/STAT软件中所提供的VARCLUS过程是专门用于对变量进行分类的,它根据相关阵或协方差阵对变量进行分裂聚类或谱系聚类.类的选择原则根据主成分分析和因子分析的思想(参见第七章和第八章),使每一类的第一主成分或重心分量所解释的方差为最大.

1. 变量聚类的步骤

如果没有为过程提供初始分类的情况,VARCLUS过程开始把所有变量看成一个类,然后它重复以下步骤:

步骤1:首先挑选一个将被分裂的类.

步骤2:把选中的类分裂成两个类.首先计算前两个主成分,再进行斜交旋转,并把每个变量分配到旋转分量对应的类里,分配的原则是使变量与这个主成分的相关系数为最大.

步骤3:变量重新归类.通过多次反复循环,变量被重新分配到这些类里,使得由这些类分量所解释的方差为最大.重新分配可能要求保持谱系结构.

当每一类满足用户规定的准则时,过程停止以上分裂类的三个步骤.该准则或是每个类分量所解释的方差的百分比,或是每一类的第二个特征根.如果没有规定准则,则当每个类只有一个特征根大于1时,VARCLUS过程停止.

例6.7.2 对例6.7.1中给出的8项体型指标的相关系数阵,试用VARCLUS过程进行分类.

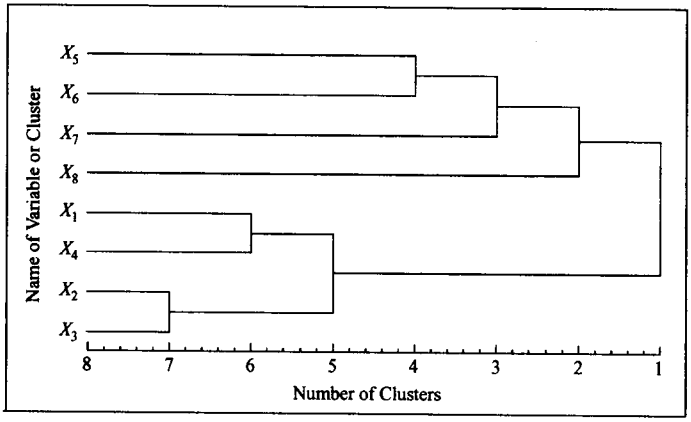
解 调用VARCLUS过程对8个体型变量进行分类.输出6.7.2为给出分类个数由1至8类的分类结果的总结表.该表的每一行对应于一次分裂.

从输出6.7.2和输出6.7.3的结果可看出,把8个体型指标变量分为两类比较合适.第一类包括X_1(身高)、X_2(手臂长)、X_3(手肘长)和X_4(小腿长),它们称为体型的高矮变量类;第二类包括X_5(体重)、X_6(颈围)、X_7(胸围)和X_8(胸宽),它们是体型的肥瘦变量类.

输出 6.7.2　8 个体能变量的谱系聚类结果(规定选项 SUMMARY)

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.672880	0.5841	0.5841	1.770983	0.3810	.
2	6.426502	0.8033	0.7293	0.476418	0.6329	0.4380
3	6.895347	0.8619	0.7954	0.418369	0.7421	0.3634
4	7.271218	0.9089	0.8773	0.238000	0.8652	0.2548
5	7.509218	0.9387	0.8773	0.236135	0.8652	0.1665
6	7.740000	0.9675	0.9295	0.141000	0.9295	0.2560
7	7.881000	0.9851	0.9405	0.119000	0.9405	0.2093
8	8.000000	1.0000	1.0000	0.000000	1.0000	0.0000

输出 6.7.3　8 个体能变量的谱系聚类图



从输出 6.7.3 的谱系聚类图中可以看出,若要求分为三类,则 X_8(胸宽)独立为一类. 该结果与用重心类分量聚类分析所得到的结果是一致的.

习　题　六

6-1　试证明下列结论:

(1) 由两个距离的和所组成的函数仍为距离;

(2) 由一个正常数乘上一个距离所组成的函数仍为距离;

(3) 设 d 为一个距离,$c>0$ 为常数,则 $d^*=d/(d+c)$ 仍是一个距离;

(4) 由两个距离的乘积所组成的函数不一定是距离.

6-2　试证明二值变量的相关系数为(6.2.2)式，其夹角余弦为(6.2.3)式.

6-3　下面是5个样品两两间的距离矩阵

$$D^{(0)}=\begin{bmatrix}0&&&&\\4&0&&&\\6&9&0&&\\1&7&10&0&\\6&3&5&8&0\end{bmatrix},$$

试用最长距离法、类平均法作系统聚类，并画出谱系聚类图.

6-4　利用距离平方的递推公式

$$D_{kr}^2=\alpha_pD_{pk}^2+\alpha_qD_{qk}^2+\beta D_{pq}^2+\gamma|D_{kp}^2-D_{qk}^2|$$

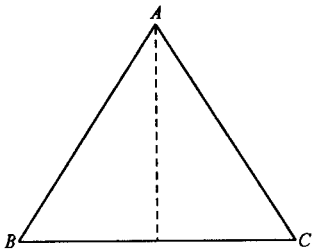
来证明：当$\gamma=0$，$\alpha_p\geqslant 0$，$\alpha_q\geqslant 0$，$\alpha_p+\alpha_q+\beta\geqslant 1$时，系统聚类中的类平均法、可变类平均法、可变法，以及Ward法的单调性.

6-5　试从定义直接证明最长和最短距离法的单调性.

6-6　设A,B,C为平面上三个点(如图所示)，它们之间的距离为

$$d_{AB}^2=d_{AC}^2=1.1,\quad d_{BC}^2=1.0.$$

将该三个点看成三个二元样品，试用此例说明中间距离法和重心法不具有单调性.



题6-6图

6-7　试推导重心法的距离递推公式(6.3.2).

6-8　试推导Ward法的距离递推公式(6.3.3).

6-9　设有5个样品，对每个样品考察一个指标，得数据为1,2,5,7,10.试用离差平方和法求将5个样品分为k类($k=5,4,3,2,1$)的分类法b_k及其相应的总离差平方和$W(k)$.

6-10　今有6个铅弹头，用“中子活化”方法测得7种微量元素的含量数据(见表6.13).

(1) 试用多种系统聚类法对6个弹头进行分类；并比较分类结果；

(2) 试用多种方法对7种微量元素进行分类.

表 6.13　微量元素含量数据

元素 / 样品号	Ag(银) (X_1)	Al(铝) (X_2)	Cu(铜) (X_3)	Ca(钙) (X_4)	Sb(锑) (X_5)	Bi(铋) (X_6)	Sn(锡) (X_7)
1	0.05798	5.5150	347.10	21.910	8586	1742	61.69
2	0.08441	3.9700	347.20	19.710	7947	2000	2440
3	0.07217	1.1530	54.85	3.052	3860	1445	9497
4	0.15010	1.7020	307.50	15.030	12290	1461	6380
5	5.74400	2.8540	229.60	9.657	8099	1266	12520
6	0.21300	0.7058	240.30	13.910	8980	2820	4135

6-11　设在某地区抽取了 14 块岩石标本，其中 7 块含矿，7 块不含矿. 对每块岩石测定了 Cu，Ag，Bi 三种化学成分的含量，得到的数据见第五章表 5.3. 试用几种系统聚类方法进行聚类分析，给出综合的分析结果，并与实际情况进行比较.

6-12　某城市的环保监测站于 1982 年在全市均匀地布置了 16 个监测点，每日三次定时抽取大气样品，测量大气中二氧化硫、氮氧化物和飘尘的含量. 前后 5 天，每个取样点(监测点)对每种污染元素实测 15 次，取 15 次实测值的平均作为该取样点大气污染元素的含量(数据见第五章表 5.5). 试用几种系统聚类方法进行聚类分析，并给出综合的分析结果.

第七章　主成分分析

主成分分析是将多指标化为少数几个综合指标的一种统计分析方法. 在实际问题中，研究多指标的问题是经常遇到的问题. 多元统计分析处理的是多变量(多指标)问题. 由于变量个数太多，并且彼此之间存在着一定的相关性，因而使得所观测到的数据在一定程度上反映的信息有所重叠. 而且当变量较多时，在高维空间中研究样本的分布规律比较复杂，势必增加分析问题的复杂性. 人们自然希望用较少的综合变量来代替原来较多的变量；而这几个综合变量又能够尽可能多地反映原来变量的信息，并且彼此之间互不相关. 利用这种降维的思想，产生了主成分分析、因子分析、典型相关分析、偏最小二乘回归等统计方法. 本章介绍主成分分析方法. 主成分分析又称**主分量分析**或**主轴分析**.

§7.1　总体的主成分

一、主成分的定义

设 $X=(X_1,\cdots,X_p)'$ 是 p 维随机向量，均值 $\mathrm{E}(X)=\mu$，协方差阵 $\mathrm{D}(X)=\Sigma$. 考虑它的线性变换：

$$\begin{cases} Z_1 = a_1'X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p, \\ Z_2 = a_2'X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_p = a_p'X = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p, \end{cases} \tag{7.1.1}$$

易见

$$\begin{aligned} &\mathrm{Var}(Z_i) = a_i'\Sigma a_i, \qquad (i=1,2,\cdots,p), \\ &\mathrm{Cov}(Z_i,Z_j) = a_i'\Sigma a_j, \quad (i,j=1,\cdots,p). \end{aligned} \tag{7.1.2}$$

假如我们希望用 Z_1 来代替原来的 p 个变量 $X_1,\cdots,X_p$，这就要求 Z_1 尽可能多地反映原来 p 个变量的信息，这里所说的“信息”用什么来表达呢？最经典的方法是用 Z_1 的方差来表达. $\mathrm{Var}(Z_1)$越大，表示 Z_1 包含的信息越多. 由(7.1.2)式看出，对 a_1 必须有某种限制，否则可使 $\mathrm{Var}(Z_1)\to\infty$. 常用的限制是：$a_1'a_1=1$. 若存在满足以上约束的 a_1，使 $\mathrm{Var}(Z_1)$达最大，Z_1 就称为第一主成分(或主分量). 如果第一主成分不足以代表原来 p 个变量的绝大部分信息，考虑 X 的第二个线性组合 Z_2. 为了有效地代表原始变量的信息，Z_1 已体现(反映)的信息不希望在 Z_2 中出现，用统计语言来讲，就是要求

$$\mathrm{Cov}(Z_2,Z_1)=a_2'\Sigma a_1=0. \tag{7.1.3}$$

于是求 Z_2，就是在约束 $a_2'a_2=1$ 和(7.1.3)式下，求 a_2 使 $\mathrm{Var}(Z_2)$达最大，所求之 Z_2 称为第二主成分，类似地可求得第三主成分，第四主成分等等.

定义 7.1.1　设 $X=(X_1,X_2,\cdots,X_p)'$ 为 p 维随机向量. 称 $Z_i=a_i'X$ 为 X 的第 i 主成分($i=1,2,\cdots,p$)，如果：

(1) $a_i'a_i=1$ ($i=1,2,\cdots,p$)；

(2) 当 $i>1$ 时，$a_i'\Sigma a_j=0$ ($j=1,\cdots,i-1$)；

(3) $\mathrm{Var}(Z_i)=\max\limits_{a'a=1,a'\Sigma a_j=0\ (j=1,\cdots,i-1)}\mathrm{Var}(a'X)$.

从代数学观点看，主成分就是 p 个原始变量的一些特殊的线性组合；而从几何上看，这些线性组合正是把由 $X_1,\cdots,X_p$ 构成的坐标系经旋转而产生的新坐标系，新坐标轴使之通过样本变差最大的方向(或者说具有最大的样本方差).

考虑 $p=2$，此时原始变量为 X_1，X_2. 设(X_1,X_2)服从二元正态分布，则样品点 $X_{(i)}=(x_{i1},x_{i2})$ ($i=1,2,\cdots n$)的散布图(见图 7.1)在一个椭圆内散布着.

由上可知，对于二维正态随机向量，n 个点散布在一个椭圆内(当原始变量 X_1,X_2 相关性越强，这个椭圆就越扁). 若取椭圆的长轴为坐标轴 Z_1，椭圆的短轴为 Z_2，这相当于在平面上作一个坐标变换，即按逆时针方向旋转一个角度 θ，根据旋转变换公式，新老坐标

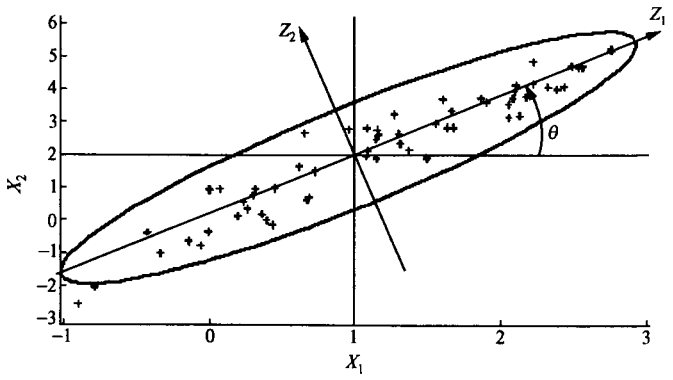


图 7.1 主成分的几何意义

之间有关系：

$$\begin{cases} Z_1 = \cos\theta X_1 + \sin\theta X_2, \\ Z_2 = -\sin\theta X_1 + \cos\theta X_2. \end{cases}$$

Z_1, Z_2 是原始变量 X_1 和 X_2 的特殊线性组合.

从图上可以看出二维平面上 n 个点的波动(用两个变量的方差和表示)大部分可以归结为在 Z_1 方向的波动，而在 Z_2 方向上的波动很小，可以忽略. 这样一来，二维问题即可以降为一维，只取第一个综合变量 Z_1 即可，而 Z_1 是椭圆的长轴.

一般情况，p 个变量组成 p 维空间，n 个样品点就是 p 维空间的 n 个点. 对于 p 维正态随机向量来说，找主成分的问题就是找 p 维空间中椭球的主轴问题.

二、主成分的求法

设 p 维随机向量 X 的均值 $\mathrm{E}(X)=0$，协方差阵 $\mathrm{D}(X)=\Sigma>0$.

由定义 7.1.1，求第一主成分 $Z_1=a_1'X$ 的问题，即为求 $a_1=(a_{11}, a_{21}, \cdots, a_{p1})'$，使得在 $a_1'a_1=1$ 下，$\mathrm{Var}(Z_1)$ 达最大. 这是条件极值问题，用拉格朗日乘子法求解. 令

$$\varphi(a_1) = \mathrm{Var}(a_1'X) - \lambda(a_1'a_1 - 1) = a_1'\Sigma a_1 - \lambda(a_1'a_1 - 1),$$

考虑

$$\begin{cases}\dfrac{\partial \varphi}{\partial a_1}=2(\Sigma-\lambda I)a_1=0,\\ \dfrac{\partial \varphi}{\partial \lambda}=a_1'a_1-1=0.\end{cases}\tag{7.1.4}$$

因 $a_1\neq 0$，故 $|\Sigma-\lambda I|=0$，求解方程组(7.1.4)，其实就是求 Σ 的特征值和特征向量问题. 设 $\lambda=\lambda_1$ 是 Σ 的最大特征值，则相应的单位特征向量 a_1 即为所求. 一般地，求 X 的第 i 主成分可通过求 Σ 的第 i 大特征值所对应的单位特征向量得到.

定理 7.1.1　设 $X=(X_1,\cdots,X_p)'$ 是 p 维随机向量，且 $\mathrm{D}(X)=\Sigma$，Σ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$，$a_1,a_2,\cdots,a_p$ 为相应的单位正交特征向量，则 X 的第 i 主成分为

$$Z_i=a_i'X\quad(i=1,2,\cdots,p).$$

证明　因 Σ 为对称矩阵，利用附录中定理 7.2 的结论(1)可知，对任意非零向量 a，有

$$\lambda_p\leqslant\frac{a'\Sigma a}{a'a}\leqslant\lambda_1,$$

且最大值在 $a=a_1$ 时达到. 故在 $a_1'a_1=1$ 的约束条件下，使得

$$\mathrm{Var}(Z_1)=\mathrm{Var}(a_1'X)=a_1'\Sigma a_1=\lambda_1$$

达到极大值. 根据主成分的定义 7.1.1，$Z_1=a_1'X$ 为 X 的第一主成分.

对 $r=2,3,\cdots,p$，记 $\mathscr{L}_r=\mathscr{L}(a_r,\cdots,a_p)$，利用附录中定理 7.2 的结论(2)，即得

$$\max_{a\neq 0,a\in\mathscr{L}_r}\frac{a'\Sigma a}{a'a}=\lambda_r,$$

且最大值在 $a=a_r$ 时达到. 故在 $a_r'a_r=1$ 的约束条件下，a_r 满足

$$a_r'\Sigma a_j=a_r'\lambda_j a_j=\lambda_j a_r'a_j=0\quad(j=1,\cdots,r-1);$$

且使得

$$\mathrm{Var}(Z_r)=\mathrm{Var}(a_r'X)=a_r'\Sigma a_r=\lambda_r$$

达极大值. 根据主成分的定义 7.1.1，$Z_r=a_r'X$ 为 X 的第 r 主成分.

(证毕)

推论　设 $Z=(Z_1,Z_2,\cdots,Z_p)'$ 为 p 维随机向量，则其分量 Z_i

$(i=1,2,\cdots,p)$依次是 X 的第 i 主成分的充分必要条件是：

(1) $Z=A'X$，A 为正交矩阵；

(2) $\mathrm{D}(Z)=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$，即随机向量 Z 的协方差阵为对角矩阵；

(3) $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$.

三、主成分的性质

记 $\Sigma=(\sigma_{ij})$，$\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$，其中 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 为 Σ 的特征值，$a_1,a_2,\cdots,a_p$ 是相应的单位正交特征向量，记正交矩阵 $A=(a_1,a_2,\cdots,a_p)$. 主成分 $Z=(Z_1,\cdots,Z_p)'$，其中 $Z_i=a_i'$ $(i=1,2,\cdots,p)$.

总体主成分有如下性质：

性质 1 $\mathrm{D}(Z)=\Lambda$，即 p 个主成分的方差为：$\mathrm{Var}(Z_i)=\lambda_i$ $(i=1,2,\cdots,p)$，且它们是互不相关的.

性质 2 $\sum\limits_{i=1}^{p}\sigma_{ii}=\sum\limits_{i=1}^{p}\lambda_i$，通常称 $\sum\limits_{i=1}^{p}\sigma_{ii}$ 为**原总体 X 的总方差(或称总惯量)**.

此性质说明原总体 X 的总方差可分解为不相关的主成分的方差和，且存在 m $(m<p)$，使 $\sum\limits_{i=1}^{p}\sigma_{ii}\approx\sum\limits_{i=1}^{m}\lambda_i$，即 p 个原始变量所提供的总信息(总方差)的绝大部分只需用前 m 个主成分来代替.

性质 3 主成分 Z_k 与原始变量 X_i 的相关系数 $\rho(Z_k,X_i)$ 为

$$\rho(Z_k,X_i)=\sqrt{\lambda_k}a_{ik}/\sqrt{\sigma_{ii}}\quad(k,i=1,2,\cdots p),\qquad(7.1.5)$$

并把主成分 Z_k 与原始变量 X_i 的相关系数称为**因子负荷量(或因子载荷量)**.

证明 事实上，因为

$$\rho(Z_k,X_i)=\frac{\mathrm{Cov}(Z_k,X_i)}{\sqrt{\mathrm{Var}(Z_k)\cdot\mathrm{Var}(X_i)}}=\frac{\mathrm{Cov}(a_k'X,e_k'X)}{\sqrt{\lambda_k\cdot\sigma_{ii}}},$$

其中 $e_i=(0,\cdots,0,1,0,\cdots,0)'$，它是除第 i 个元素为 1 外其余元素均

为0的单位向量.再利用协方差的性质及 a_k 是 Σ 对应于 λ_k 的特征向量可知

$$\mathrm{Cov}(a_k'X,e_i'X)=a_k'\mathrm{D}(X)e_i=a_k'\Sigma e_i=e_i'\Sigma a_k=\lambda_k e_i'a_k=\lambda_k a_{ik}.$$

即得

$$\rho(Z_k,X_i)=\sqrt{\lambda_k}a_{ik}/\sqrt{\sigma_{ii}}\quad(k,i=1,2,\cdots p).\qquad\text{(证毕)}$$

如果把主成分与原始变量的相关系数列成表7.1的形式,则由相关系数的公式(7.1.5),还可得出下列性质4和5.

表7.1　主成分和原始变量的相关系数

	Z_1	$\cdots$	Z_k	$\cdots$	Z_p
X_1	$\rho(Z_1,X_1)$	$\cdots$	$\rho(Z_k,X_1)$	$\cdots$	$\rho(Z_p,X_1)$
X_2	$\rho(Z_1,X_2)$	$\cdots$	$\rho(Z_k,X_2)$	$\cdots$	$\rho(Z_p,X_2)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
X_p	$\rho(Z_1,X_p)$	$\cdots$	$\rho(Z_k,X_p)$	$\cdots$	$\rho(Z_p,X_p)$

性质4　$\sum_{k=1}^{p}\rho^2(Z_k,X_i)=\sum_{k=1}^{p}\frac{\lambda_k a_{ik}^2}{\sigma_{ii}}=1\ (i=1,2,\cdots,p).$

证明　因为由 $A'\Sigma A=\Lambda$,可得 $\Sigma=A\Lambda A'$,故

$$\sigma_{ii}=(a_{i1},\cdots,a_{ip})\Lambda\begin{bmatrix}a_{i1}\\ \vdots\\ a_{ip}\end{bmatrix}=\sum_{k=1}^{p}\lambda_k a_{ik}^2,$$

因此,$\sum_{k=1}^{p}\rho^2(Z_k,X_i)=\sum_{k=1}^{p}\lambda_k a_{ik}^2/\sigma_{ii}=1\ (i=1,2,\cdots,p)$.事实上,主成分 $Z_k\ (k=1,\cdots,p)$是变量 $X_1,X_2,\cdots,X_p$ 的线性组合;反过来,X_i 也可表成 $Z_1,\cdots,Z_p$ 的线性组合.又因 $Z_1,\cdots,Z_p$ 互不相关,由回归分析的知识可知,X_i 与 $Z_1,\cdots,Z_p$ 的全相关系数的平方等于1,即表7.1中每一行的平方和均为1,即得性质4.　(证毕)

性质5　$\sum_{i=1}^{p}\sigma_{ii}\rho^2(Z_k,X_i)=\lambda_k\ (k=1,\cdots,p).$

证明　这只须把 $\rho(Z_k,X_i)$ 的计算公式(7.1.5)代入性质5表达式左端,经整理即得.另一方面,因 Z_k 可表成 $X_1,\cdots,X_p$ 的线性组合,但 $X_i(i=1,\cdots,p)$一般有相关性,由 Z_k 与 X_i 相关系数的公式(7.1.5),也可得出表7.1中 Z_k 所对应的每一列关于各变量相关系

数的加权平方和为 λ_k(即 $\mathrm{Var}(Z_k)$). (证毕)

主成分分析的目的之一是为了简化数据结构,故在实际应用中一般绝不用 p 个主成分,而选用 m ($m<p$)个主成分. m 取多大,这是一个很实际的问题. 为此,我们引进贡献率的概念.

定义 7.1.2 我们称 $\lambda_k \Big/ \sum_{i=1}^{p}\lambda_i$ 为主成分 Z_k 的**贡献率**;又称 $\sum_{k=1}^{m}\lambda_k \Big/ \sum_{i=1}^{p}\lambda_i$ 为主成分 $Z_1,\cdots,Z_m$ ($m<p$)的**累计贡献率**.

通常取 m,使累计贡献率达到 70%以上,累计贡献率的大小仅表达 m 个主成分提取了 $X_1,\cdots,X_p$ 的多少信息,但它没有表达某个变量被提取了多少信息,为此又引入另一个概念.

定义 7.1.3 将前 m 个主成分 $Z_1,\cdots,Z_m$ 对原始变量 X_i 的贡献率 $\nu_i^{(m)}$ 定义为 X_i 与 $Z_1,\cdots,Z_m$ 的相关系数的平方,它等于

$$\nu_i^{(m)} = \sum_{k=1}^{m}\lambda_k a_{ik}^2/\sigma_{ii}. \tag{7.1.6}$$

例 7.1.1 设随机向量 $X=(X_1,X_2,X_3)'$的协方差阵为

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

试求 X 的主成分及主成分对变量 X_i 的贡献率 ν_i ($i=1,2,3$).

解 Σ 的特征值为 $\lambda_1=3+\sqrt{8}$, $\lambda_2=2$, $\lambda_3=3-\sqrt{8}$. 相应单位正交特征向量为

$$a_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad a_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}.$$

故主成分为

$$Z_1 = 0.383X_1 - 0.924X_2,$$

$$Z_2 = X_3 \quad (X_3 \text{本身就是一个主成分,与 } X_1,X_2 \text{ 不相关}),$$

$$Z_3 = 0.924X_1 + 0.383X_2.$$

取 $m=1$ 时,Z_1 对 X 的贡献率可达

$$\frac{3+\sqrt{8}}{\lambda_1+\lambda_2+\lambda_3} = \frac{3+\sqrt{8}}{8} = 72.8\%;$$

取 $m=2$ 时，Z_1,Z_2 的累计贡献率可达 97.85%. 下表列出 m 个主成分对变量 X_i 的贡献率 $\nu_i^{(m)}$.

i	$\rho(Z_1,X_i)$	$\rho(Z_2,X_i)$	$\nu_i^{(1)}(m=1)$	$\nu_i^{(2)}(m=2)$
1	0.925	0	0.856	0.856
2	−0.998	0	0.996	0.996
3	0.000	1	0.000	1.000

由上可见，当 $m=1$ 时，Z_1 的贡献率已达 72.8%，比较理想了. 但 Z_1 对 X_3 的贡献率 $\nu_3^{(1)}=0$，这是因为在 Z_1 中没有包含 X_3 的任何信息，这时仅取 $m=1$ 就不够了，故取 $m=2$，这时 Z_1,Z_2 的累计贡献率为 97.85%，且 Z_1,Z_2 对 X_i 的贡献率 $\nu_i^{(2)}$ $(i=1,2,3)$也较高.

四、标准化变量的主成分及性质

在实际问题中，不同的变量往往有不同的量纲，而通过 Σ 来求主成分总是优先考虑方差大的变量，有时会造成很不合理的结果，为了消除由于量纲的不同可能带来的一些不合理的影响，常采用将变量标准化的方法. 若记 $\mathrm{E}(X_i)=\mu_i$，$\mathrm{Var}(X_i)=\sigma_i^2$，即令

$$X_i^* = \frac{X_i - \mathrm{E}(X_i)}{\sqrt{\mathrm{Var}(X_i)}} = \frac{X_i - \mu_i}{\sigma_i} \quad (i = 1,2,\cdots,p).$$

这时标准化后的随机向量 $X^*=(X_1^*,X_2^*,\cdots,X_p^*)'$ 的协方差阵 Σ^* 就是原随机向量 X 的相关阵 R. 从相关阵 R 出发求主成分，记主成分向量为 $Z^*=(Z_1^*,\cdots,Z_p^*)'$，则 Z^* 有与总体主成分相应的性质：

性质 1　$\mathrm{D}(Z^*)=\Lambda^*=\mathrm{diag}(\lambda_1^*,\lambda_2^*,\cdots,\lambda_p^*)$，其中 $\lambda_1^*\geqslant\lambda_2^*\geqslant\cdots\geqslant\lambda_p^*$ 为相关阵 R 的特征值.

性质 2　$\sum_{i=1}^{p}\lambda_i^*=p.$

性质 3　主成分 Z_k^* 与标准化变量 X_i^* 的相关系数 $\rho(Z_k^*,X_i^*)$ 为

$$\rho(Z_k^*,X_i^*)=\sqrt{\lambda_k^*}a_{ik}^* \quad (k,i=1,2,\cdots p),$$

其中 $a_k^*=(a_{1k}^*,\cdots,a_{pk}^*)'$ 是 R 对应于 λ_k^* 的单位正交特征向量.

性质 4　$\sum_{k=1}^{p}\rho^2(Z_k^*,X_i^*)=\sum_{k=1}^{p}\lambda_k^*(a_{ik}^*)^2=1$ $(i=1,2,\cdots,p).$

性质 5 $\sum_{i=1}^{p}\rho^2(Z_k^*,X_i^*)=\sum_{i=1}^{p}\lambda_k^*(a_{ik}^*)^2=\lambda_k^*\ (k=1,2,\cdots,p)$.

现将主成分 $Z_k^*\ (k=1,\cdots,p)$对标准化变量 X_i^* 的因子负荷量 $\rho_{ik}=\rho(Z_k^*,X_i^*)$列成表 7.2.

表 7.2 变量标准化后的因子负荷量

	Z_1^*	$\cdots$	Z_k^*	$\cdots$	Z_p^*	$\sum_{k=1}^{p}\rho_{ik}^2$
X_1^*	$\sqrt{\lambda_1^*}a_{11}^*$	$\cdots$	$\sqrt{\lambda_k^*}a_{1k}^*$	$\cdots$	$\sqrt{\lambda_p^*}a_{1p}^*$	1
X_2^*	$\sqrt{\lambda_1^*}a_{21}^*$	$\cdots$	$\sqrt{\lambda_k^*}a_{2k}^*$	$\cdots$	$\sqrt{\lambda_p^*}a_{2p}^*$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_p^*	$\sqrt{\lambda_1^*}a_{p1}^*$	$\cdots$	$\sqrt{\lambda_k^*}a_{pk}^*$	$\cdots$	$\sqrt{\lambda_p^*}a_{pp}^*$	1
$\sum_{i=1}^{p}\rho_{ik}^2$	λ_1^*	$\cdots$	λ_k^*	$\cdots$	λ_p^*	$\sum_{k=1}^{p}\sum_{i=1}^{p}\rho_{ik}^2=p$

§7.2 样本的主成分

上节讨论了总体的主成分，在实际问题中，一般协方差阵 Σ 未知，需要通过样本来估计. 设 $X_{(t)}=(x_{t1},\cdots,x_{tp})'\ (t=1,\cdots,n)$为来自总体 X 的样本，记样本数据阵为

$$X=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}=\begin{bmatrix}X_{(1)}'\\X_{(2)}'\\\vdots\\X_{(n)}'\end{bmatrix},$$

则样本协方差阵 S 及样本相关阵 R 分别为

$$S=\frac{1}{n-1}\sum_{t=1}^{n}(X_{(t)}-\overline{X})(X_{(t)}-\overline{X})'\xlongequal{\text{def}}(s_{ij})_{p\times p},$$

其中

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{n}X_{(t)}=(\overline{x}_1,\cdots,\overline{x}_p)',$$

$$s_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}(x_{ti}-\overline{x}_i)(x_{tj}-\overline{x}_j);$$

$$R=(r_{ij})_{p\times p},\quad \text{其中 } r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}\,s_{jj}}}\ (i,j=1,2,\cdots,p).$$

我们可以用样本协方差阵 S 作为 Σ 的估计或用 R 作为总体相关阵的估计，然后按上节的方法即可获得样本的主成分.

一、样本主成分及其性质

假定每个变量的观测数据都已标准化(标准化后的数据阵仍记为 X，它为 $n\times p$ 矩阵). 这时样本协方差阵就是样本相关阵 R，且

$$R=\frac{1}{n-1}X'X. \tag{7.2.1}$$

仍记相关阵 R 的 p 个主成分为 $Z_1,\cdots,Z_p$，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$ 为 R 的特征值，$a_1,a_2,\cdots,a_p$ 为相应的单位正交特征向量，记 $A=(a_1,a_2,\cdots,a_p)$ 为正交矩阵. 显然第 i 个样本主成分 $Z_i=a_i'X$ $(i=1,2,\cdots,p)$. 将第 t 个样品 $X_{(t)}=(x_{t1},\cdots,x_{tp})'$ 的值代入 Z_i 的表达式，经计算得到的值称为第 t 个样品在第 i 个主成分的得分，记为 z_{ti}(显然 $z_{ti}=a_i'X_{(t)}$，$i=1,\cdots,p$). 设 $Z_{(t)}=(z_{t1},z_{t2},\cdots,z_{tp})'=A'X_{(t)}$ $(t=1,\cdots,n)$(见表 7.3). 令

$$Z=\begin{bmatrix}z_{11}&z_{12}&\cdots&z_{1p}\\ z_{21}&z_{22}&\cdots&z_{2p}\\ \vdots&\vdots&&\vdots\\ z_{n1}&z_{n2}&\cdots&z_{np}\end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix}Z'_{(1)}\\ Z'_{(2)}\\ \vdots\\ Z'_{(n)}\end{bmatrix}$$

$$\overset{\text{或 def}}{=\!=}(Z_1,Z_2,\cdots,Z_p).$$

表 7.3　原始数据和样本主成分得分

样品号	原始变量				样本主成分			
	X_1	X_2	$\cdots$	X_p	Z_1	Z_2	$\cdots$	Z_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}	z_{11}	z_{12}	$\cdots$	z_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}	z_{21}	z_{22}	$\cdots$	z_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}	z_{n1}	z_{n2}	$\cdots$	z_{np}

显然，主成分得分阵 Z 和标准化后的原始数据阵 X 满足：

$$Z = XA \quad 或 \quad X = ZA',$$

其中 A 为 p 阶矩阵.

类似总体主成分，样本主成分也具有如下性质：

性质 1 $\overline{Z}=\frac{1}{n}\sum_{t=1}^{n}Z_{(t)}=(\overline{z}_1,\cdots,\overline{z}_p)'=0$ (因为 $\overline{X}=0$)，而

$$\boldsymbol{Z}_i'\boldsymbol{Z}_j=\begin{cases}0, & 当 i\neq j,\\ (n-1)\lambda_i, & 当 i=j.\end{cases}$$

上式说明当 $i\neq j$ 时，第 i 个主成分得分向量 $\boldsymbol{Z}_i$ 与第 j 个主成分得分向量 $\boldsymbol{Z}_j$ 是相互正交的.

性质 2 $\sum_{i=1}^{p}\lambda_i=p$. 称 $\frac{\lambda_k}{p}$ 为样本主成分 Z_k 的**贡献率**；又称 $\frac{\lambda_1+\cdots+\lambda_m}{p}$ 为样本主成分 $Z_1,\cdots,Z_m$ $(m<p)$ 的**累计贡献率**.

性质 3 样本主成分具有使残差平方和最小的优良性.

设 $Z_1,\cdots,Z_p$ 是样本主成分，即

$$\begin{cases}Z_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p=a_1'X,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ Z_p=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p=a_p'X,\end{cases}$$

它等价于 p 维随机向量 Z 可表为

$$Z=\begin{bmatrix}Z_1\\ \vdots\\ Z_p\end{bmatrix}=A'\begin{bmatrix}X_1\\ \vdots\\ X_p\end{bmatrix}=A'X=\begin{bmatrix}a_{11}X_1+\cdots+a_{p1}X_p\\ \vdots\\ a_{1p}X_1+\cdots+a_{pp}X_p\end{bmatrix},$$

于是原 p 维随机向量 X 也类似可表为

$$X=\begin{bmatrix}X_1\\ \vdots\\ X_p\end{bmatrix}=A\begin{bmatrix}Z_1\\ \vdots\\ Z_p\end{bmatrix}=AZ=\begin{bmatrix}a_{11}Z_1+\cdots+a_{1p}Z_p\\ \vdots\\ a_{p1}Z_1+\cdots+a_{pp}Z_p\end{bmatrix},$$

即原始变量 X_j $(j=1,2,\cdots,p)$ 也可以用主成分 $Z_1,\cdots,Z_p$ 的线性组合表示.

如果我们只取前 m 个主成分 $(m<p)$，并考虑用前 m 个主成分 $Z_1,\cdots,Z_m$ 的线性组合来表示 X_j $(j=1,2,\cdots,p)$，即假定 $X_1,\cdots,X_p$ 和 $Z_1,\cdots,Z_m$ 满足以下多因变量的多元线性回归模型：

$$\begin{cases} X_1 = b_{11}Z_1 + b_{12}Z_2 + \cdots + b_{1m}Z_m + \varepsilon_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_p = b_{p1}Z_1 + b_{p2}Z_2 + \cdots + b_{pm}Z_m + \varepsilon_p, \end{cases} \tag{7.2.2}$$

把表 7.3 给出的原始变量的观测数据和前 m 个主成分的得分数据作为以上模型中因变量 $X_1,\cdots,X_p$ 和自变量 $Z_1,\cdots,Z_m$ 的观测数据. 因而使残差平方和最小的问题化为：按最小二乘准则求参数矩阵

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{p1} & \cdots & b_{pm} \end{bmatrix},$$

使得残差平方和 $Q(B)$达最小. 记

$$Z^* = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix},$$

则观测数据满足：$X=Z^*B'+E$,其中 E 为 $n\times p$ 矩阵.

由多因变量的回归分析(参见第四章)的理论知：参数矩阵的最小二乘估计为

$$B' = [(Z^*)'Z^*]^{-1}(Z^*)'X.$$

令 $A^*=(a_1,a_2,\cdots,a_m)$ (A^* 为 $p\times m$ 矩阵),则

$$Z^* = XA^*.$$

且因$(A^*)'RA^*=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m)$,于是

$$\begin{aligned}(Z^*)'Z^* &= (A^*)'X'XA^* = (n-1)(A^*)'RA^* \\ &= (n-1)\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m),\end{aligned}$$

故

$$\begin{aligned}B' &= \frac{1}{n-1}\mathrm{diag}(\lambda_1^{-1},\lambda_2^{-1},\cdots,\lambda_m^{-1})\cdot(A^*)'X'X \\ &= \mathrm{diag}(\lambda_1^{-1},\lambda_2^{-1},\cdots,\lambda_m^{-1})(RA^*)' \\ &= \mathrm{diag}(\lambda_1^{-1},\lambda_2^{-1},\cdots,\lambda_m^{-1})(\lambda_1a_1,\lambda_2a_2,\cdots,\lambda_ma_m)' \\ &= (a_1,a_2,\cdots,a_m)' = (A^*)'.\end{aligned}$$

这表明当 $b_{jk}=a_{jk}$ ($j=1,2,\cdots,p;k=1,\cdots,m$)时,可使回归模型(7.2.2)的残差平方和达最小值,而且回归方程中 X_j 的决定系数

$R^2(j)$有：$R^2(j)=\nu_j^{(m)}$(见本章习题第 7-8 题). 前面我们把 $\nu_j^{(m)}$ 称为 m 个主成分对原始变量 X_j 的贡献率，$\nu_j^{(m)}$ 的大小反映了 m 个主成分能够反映 X_j 变差的比例大小.

二、主成分的个数及解释

主成分分析的目的之一是简化数据结构，用尽可能少的主成分 $Z_1,\cdots,Z_m$ $(m<p)$代替原来的 p 个变量，这样就把 p 个变量的 n 次观测数据简化为 m 个主成分的得分数据. 这要求：

(1) m 个主成分所反映的信息与原来 p 个变量提供的信息差不多；

(2) m 个主成分又能够对数据所具有的意义进行解释.

主成分的个数 m 如何选取是实际工作者关心的问题. 关于主成分的个数如何确定，常用的标准有两个：一个是按累计贡献率达到一定程度(如 70%或 80%以上)来确定 m；另一个是先计算 S 或 R 的 p 个特征值的均值 $\bar{\lambda}$，取大于 $\bar{\lambda}$ 的特征值个数 m. 当变量个数 $p<20$ 时，大量实践表明，第一个标准容易取太多的主成分，而第二个标准容易取太少的主成分，故最好将两者结合起来使用，同时还要考虑 m 个主成分对 X_i 的贡献率 $\nu_i^{(m)}$ $(i=1,2,\cdots,p)$.

例 7.2.1(中学生身体四项指标的主成分分析) 在某中学随机抽取某年级 30 名学生，测量其身高(X_1)、体重(X_2)、胸围(X_3)和坐高(X_4)，数据见表 7.4. 试对这 30 名中学生身体四项指标数据做主成分分析.

表 7.4 30 名中学生身体四项指标数据

序号	X_1	X_2	X_3	X_4	序号	X_1	X_2	X_3	X_4
1	148	41	72	78	2	139	34	71	76
3	160	49	77	86	4	149	36	67	79
5	159	45	80	86	6	142	31	66	76
7	153	43	76	83	8	150	43	77	79
9	151	42	77	80	10	139	31	68	74
11	140	29	64	74	12	161	47	78	84
13	158	49	78	83	14	140	33	67	77

（续表）

序号	X_1	X_2	X_3	X_4	序号	X_1	X_2	X_3	X_4
15	137	31	66	73	16	152	35	73	79
17	149	47	82	79	18	145	35	70	77
19	160	47	74	87	20	156	44	78	85
21	151	42	73	82	22	147	38	73	78
23	157	39	68	80	24	147	30	65	75
25	157	48	80	88	26	151	36	74	80
27	144	36	68	76	28	141	30	67	76
29	139	32	68	73	30	148	38	70	78

解　此例 $p=4, n=30$. 调用 SAS/STAT 软件中 PRINCOMP 过程，由相关阵出发进行主成分分析.

输出 7.2.1　相关阵的特征值和特征向量

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.54109800	3.22771484	0.8853	0.8853
2	0.31338316	0.23397420	0.0783	0.9636
3	0.07940895	0.01329906	0.0199	0.9835
4	0.06610989		0.0165	1.0000

Eigenvectors

	z1	z2	z3	z4
x1	0.496966	-.543213	-.449627	0.505747
x2	0.514571	0.210246	-.462330	-.690844
x3	0.480901	0.724621	0.175177	0.461488
x4	0.506928	-.368294	0.743908	-.232343

由输出 7.2.1 中相关阵的特征值可以看出，第一主成分的贡献率已高达 88.53%；且前两个主成分的累计贡献率已达 96.36%. 因此只需用两个主成分就能很好地概括这组数据. 另由第四个特征值近似为 0，可以得出这 4 个标准化后的身体指标变量 X_i^* $(i=1,2,3,4)$ 有近似的线性关系（即所谓共线性）：

$$0.5057X_1^* - 0.6908X_2^* + 0.4615X_3^* - 0.2323X_4^* \approx 0.$$

由最大的两个特征值对应的特征向量可以写出第一和第二主成

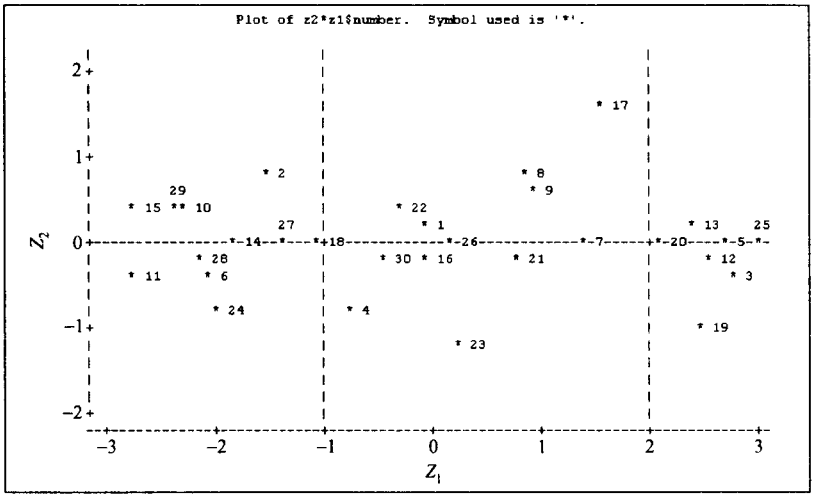
分：

$$Z_1 = 0.4970X_1^* + 0.5146X_2^* + 0.4809X_3^* + 0.5069X_4^*,$$

$$Z_2 = -0.5432X_1^* + 0.2102X_2^* + 0.7246X_3^* - 0.3683X_4^*.$$

由上可知，第一和第二主成分都是标准化后变量 X_i^* $(i=1,2,3,4)$ 的线性组合，且组合系数就是特征向量的分量.

输出 7.2.2 第二主成分得分对第一主成分得分的散布图



利用特征向量各分量的值可以对各个主成分进行解释. 第一大特征值对应的第一个特征向量的各个分量值均在 0.5 附近，且都是正值，它反映中学生身材的魁梧程度：身体高大的学生，他的 4 个部位的尺寸都比较大；而身体矮小的学生，他的 4 个部位的尺寸都比较小，因此我们称第一主成分为大小因子. 第二大特征值对应的特征向量中第一个分量(即身高 X_1 的系数)和第四个分量(即坐高 X_4 的系数)皆为负值，而第二个分量(即体重 X_2 的系数)和第三个分量(即胸围 X_3 的系数)皆为正值，它反映中学生的胖瘦情况，故称第二主成分为体形因子(或胖瘦因子).

输出 7.2.2 是第二主成分得分对第一主成分得分的散布图，从图中可以直观地看出，按学生的身体指标尺寸，这 30 名学生大约应分成三组(以第一主成分得分值为 -1 和 2 为分界点). 每一组包括

哪几名学生由每个散点旁边的序号可以得知.

§7.3　主成分分析的应用

设变量 $X_1,\cdots,X_p$ 的 n 次观测数据阵 X 已标准化,这时样本协方差阵就是样本相关阵 R,且

$$R=\frac{1}{n-1}X'X\xlongequal{\text{def}}(r_{ij})_{p\times p}.$$

R 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$,其相应标准化特征向量为 $a_1,a_2,\cdots,a_p$,样本主成分为

$$Z_j=a_j'X\quad(j=1,2,\cdots,p).$$

设 m 为满足累计贡献率$>P_0$(一般取 $1>P_0\geqslant0.70$)的最小正整数,取前 m 个主成分,由样本观测数据可求得 m 个主成分的得分值:

$$z_{ij}=a_j'X_{(i)}=a_{1j}x_{i1}+a_{2j}x_{i2}+\cdots+a_{pj}x_{ip}$$
$$(i=1,2,\cdots,n;j=1,2,\cdots,m).$$

记 $Z_j=(z_{1j},z_{2j},\cdots,z_{nj})'$ 为第 j 个主成分的 n 次得分值($j=1,2,\cdots,m$). 利用样本主成分的性质 3,下面可得出由主成分得分值估计变量 X_k 的得分.

记

$$X_k^*=(x_{1k}^*,x_{2k}^*,\cdots,x_{nk}^*)'\quad(k=1,\cdots,p),\tag{7.3.1}$$

其中

$$x_{ik}^*=a_{k1}z_{i1}+\cdots+a_{km}z_{im}\quad(i=1,\cdots,n),$$

$$X^*=\begin{bmatrix}x_{11}^* & \cdots & x_{1p}^*\\ \vdots & & \vdots\\ x_{n1}^* & \cdots & x_{np}^*\end{bmatrix}\xlongequal{\text{def}}(X_1^*,\cdots,X_p^*).\tag{7.3.2}$$

可以证明:

$$\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x_{ij}^*)^2=(n-1)\sum_{k=m+1}^{p}\lambda_k.$$

故有

$$X \cong X^*, \quad \text{且} \quad \boldsymbol{X}_k^* = Z^* \begin{bmatrix} a_{k1} \\ \vdots \\ a_{km} \end{bmatrix} = \sum_{t=1}^{m} a_{kt} \boldsymbol{Z}_t,$$

其中

$$Z^* = \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} \overset{\text{def}}{=\!=} (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_m).$$

一、指标分类(变量分类)

如果第 i 个变量和第 j 个变量的相关系数 $r_{ij} \approx 1$,显然这两个变量应归为一类.

仍用 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 表示这两个变量的 n 次观测向量,在 n 维空间中即为两个点.考虑 n 维空间中这两点间的距离:

$$\begin{aligned} \| \boldsymbol{X}_i - \boldsymbol{X}_j \|^2 &= (\boldsymbol{X}_i - \boldsymbol{X}_j)'(\boldsymbol{X}_i - \boldsymbol{X}_j) = \boldsymbol{X}_i'\boldsymbol{X}_i - 2\boldsymbol{X}_i'\boldsymbol{X}_j + \boldsymbol{X}_j'\boldsymbol{X}_j \\ &= (n-1)(r_{ii} - 2r_{ij} + r_{jj}) = 2(n-1)(1 - r_{ij}). \end{aligned}$$

又

$$\begin{aligned} 2(1 - r_{ij}) &= \left\| \frac{1}{\sqrt{n-1}} \boldsymbol{X}_i - \frac{1}{\sqrt{n-1}} \boldsymbol{X}_j \right\|^2 \\ &\approx \left\| \frac{1}{\sqrt{n-1}} \boldsymbol{X}_i^* - \frac{1}{\sqrt{n-1}} \boldsymbol{X}_j^* \right\|^2 \\ &= \left\| \frac{1}{\sqrt{n-1}} [(a_{i1} - a_{j1})\boldsymbol{Z}_1 + \cdots + (a_{im} - a_{jm})\boldsymbol{Z}_m] \right\|^2 \\ &= \lambda_1 (a_{i1} - a_{j1})^2 + \cdots + \lambda_m (a_{im} - a_{jm})^2. \end{aligned}$$

因第 k 个主成分 Z_k 与 X_i 的相关系数

$$\rho(Z_k, X_i) = \sqrt{\lambda_k} a_{ik} \overset{\text{def}}{=\!=} \rho_{ik},$$

ρ_{ik} 也称为第 k 个主成分在 X_i 上的因子负荷量.这时

$$2(1 - r_{ij}) = (\rho_{i1} - \rho_{j1})^2 + \cdots + (\rho_{im} - \rho_{jm})^2.$$

若 $r_{ij} \approx 1$,则有

$$(\rho_{i1} - \rho_{j1})^2 + \cdots + (\rho_{im} - \rho_{jm})^2 \approx 0,$$

亦即 $\| X_i - X_j \| \approx 0$，即第 i 个变量和第 j 个变量应归为一类.

考察 m 维空间的 p 个点 Q_i，其坐标为

$$Q_i = (\rho_{i1}, \rho_{i2}, \cdots, \rho_{im}) \quad (i = 1, 2, \cdots, p),$$

按距离最近准则对 p 个点进行分类.

当 $m=2$ 时，p 个点可在平面上表示出来，利用散布图可直观地对指标进行分类.

例 7.3.1　服装定型分类问题. 对 128 个成年男子的身材进行测量，每人各测得 16 项指标：身高(X_1)、坐高(X_2)、胸围(X_3)、头高(X_4)、裤长(X_5)、下档(X_6)、手长(X_7)、领围(X_8)、前胸(X_9)、后背(X_{10})、肩厚(X_{11})、肩宽(X_{12})、袖长(X_{13})、肋围(X_{14})、腰围(X_{15})和腿肚(X_{16}). 16 项指标的相关阵 R 见表 7.5(因相关阵为对称矩阵，只给出相关阵的上三角部分). 试从相关阵 R 出发进行主成分分析，并对 16 项指标进行分类. (此例选自参考文献[10].)

表 7.5　16 项身体指标数据的相关阵

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
X_1	1.0	0.79	0.36	0.96	0.89	0.79	0.76	0.26	0.21	0.26	0.07	0.52	0.77	0.25	0.51	0.21
X_2		1.0	0.31	0.74	0.58	0.58	0.55	0.19	0.07	0.16	0.21	0.41	0.47	0.17	0.35	0.16
X_3			1.0	0.38	0.31	0.30	0.35	0.58	0.28	0.33	0.38	0.35	0.41	0.64	0.58	0.51
X_4				1.0	0.90	0.78	0.75	0.25	0.20	0.22	0.08	0.53	0.79	0.27	0.57	0.26
X_5					1.0	0.79	0.74	0.25	0.18	0.23	−0.02	0.48	0.79	0.27	0.51	0.23
X_6						1.0	0.73	0.18	0.18	0.23	0.00	0.38	0.69	0.14	0.26	0.00
X_7							1.0	0.24	0.29	0.25	0.10	0.44	0.67	0.16	0.38	0.12
X_8								1.0	−0.04	0.49	0.44	0.30	0.32	0.51	0.51	0.38
X_9									1.0	−0.34	−0.16	−0.05	0.23	0.21	0.15	0.18
X_{10}										1.0	0.23	0.50	0.31	0.15	0.29	0.14
X_{11}											1.0	0.24	0.10	0.31	0.28	0.31
X_{12}												1.0	0.62	0.17	0.41	0.18
X_{13}													1.0	0.26	0.50	0.24
X_{14}														1.0	0.63	0.50
X_{15}															1.0	0.65
X_{16}																1.0

解　此例 $p=16, n=128$. 使用 SAS/STAT 软件中 PRINCOMP 过程，由相关阵出发进行主成分分析，所得主要的计算结果见输出 7.3.1.

输出 7.3.1 给出当取 $m=3$ 时，相关阵 R 的最大特征值为

$$\lambda_1 = 7.0365, \quad \lambda_2 = 2.6140, \quad \lambda_3 = 1.6321;$$

由与此 3 个特征值对应的特征向量 a_1,a_2,a_3(见输出 7.3.1 第 2 个表的表头为 Prin1,Prin2,Prin3 所对应的列)可以计算因子负荷向量 $f_i=\sqrt{\lambda_i}a_i$ $(i=1,2,3)$. 输出 7.3.2 给出前两个因子负荷向量在平面上 16 个点的位置.

输出 7.3.1 相关阵的特征值和特征向量

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7.03647744	4.42244473	0.4398	0.4398
2	2.61403272	0.98192786	0.1634	0.6032
3	1.63210486		0.1020	0.7052

Eigenvectors

	Prin1	Prin2	Prin3
x1	0.341771	-.200400	0.005720
x2	0.264992	-.143202	-.056565
x3	0.234152	0.328625	0.139937
x4	0.344233	-.181124	0.032229
x5	0.326118	-.199650	0.032945
x6	0.285914	-.269807	-.029540
x7	0.295261	-.192150	0.019608
x8	0.189273	0.370267	-.150284
x9	0.084793	-.067472	0.625563
x10	0.154295	0.174246	-.527507
x11	0.098355	0.347850	-.202115
x12	0.242546	0.017665	-.314796
x13	0.317158	-.111914	-.018841
x14	0.180113	0.371353	0.252416
x15	0.266359	0.271225	0.135449
x16	0.158333	0.362824	0.243441

从输出 7.3.2 可以看出,16 个指标可分为三类:

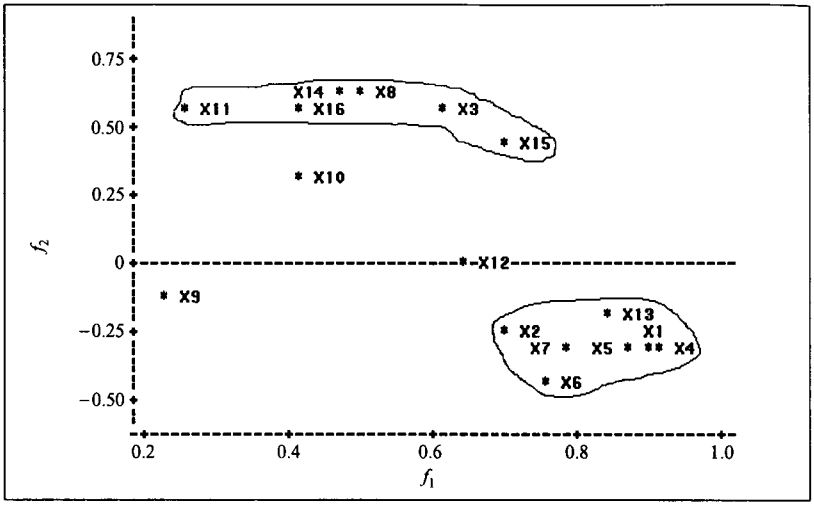
第一类为“长”的指标:身长(X_1),坐高(X_2),头高(X_4),裤长(X_5),下裆(X_6),手长(X_7),袖长(X_{13});

第二类为“围”的指标:胸围(X_3),领围(X_8),肩厚(X_{11}),肋围(X_{14}),腰围(X_{15}),腿肚(X_{16});

第三类为体形特征指标:前胸(X_9),后背(X_{10}),肩宽(X_{12}).

在第六章聚类分析中介绍的 VARCLUS 过程也是利用主成分分析的原理对指标进行分类.

输出 7.3.2　第二因子负荷向量对第一因子负荷向量的散布图



二、样品分类

对 p 个变量(指标)观测 n 次,得 n 个样品,记 $X_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})'$ 为第 i 个样品,看成 p 维空间的点,可按距离相近的程度进行分类(参见第六章),即若

$$\| X_{(i)} - X_{(j)} \| \approx 0,$$

把第 i 个样品和第 j 个样品归为一类.

因原始数据阵 $X\approx X^*$,故

$$\| X_{(i)} - X_{(j)} \| \approx \| X^*_{(i)} - X^*_{(j)} \|,$$

由(7.3.1)及(7.3.2)式知

$$X^*_{(i)}=\begin{bmatrix} x^*_{i1} \\ \vdots \\ x^*_{ip} \end{bmatrix}=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pm} \end{bmatrix}\begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}$$

$$=(a_1,\cdots,a_m)\begin{bmatrix} z_{i1} \\ \vdots \\ z_{im} \end{bmatrix}.$$

因

$$\| X^*_{(i)} - X^*_{(j)} \|^2 = \| a_1(z_{i1} - z_{j1}) + \cdots + a_m(z_{im} - z_{jm}) \|^2$$
$$= (z_{i1} - z_{j1})^2 + \cdots + (z_{im} - z_{jm})^2.$$

这样就把考察两个 p 维空间点的靠近程度转化为考察两个 m ($m<p$)维空间点的靠近程度. 若取 $m=2$,n 个样品点可在平面上表示出,利用点的分布规律对样品进行分类.

例 7.3.2 服装定型分类问题(续例 7.3.1). 仍然利用 128 人 16 项指标的观测数据,试对 128 人的服装尺寸进行分类(即样品分类问题:把 128 人分为几类,每类找出典型代表,以该代表的服装尺寸作为这一类的尺寸).

解 取 $m=2$,求出两个主成分,并计算样本主成分得分值 $Z_{(i)} = (z_{i1}, z_{i2})'$ ($i=1,2,\cdots,128$). 把这 128 个点全部表示在平面上,利用平面散布图(图见参考文献[10]),把 128 个点分为七类:

第一类共有 25 个点,聚集中心是 $Z_{(25)}$;

第二类有 14 个点,聚集中心是 $Z_{(114)}$;

第三类有 9 个点,聚集中心是 $Z_{(89)}$;

第四类有 7 个点,聚集中心是 $Z_{(112)}$;

第五类有 12 个点,聚集中心是 $Z_{(9)}$;

第六类有 20 个点,聚集中心是 $Z_{(47)}$;

第七类有 8 个点,聚集中心是 $Z_{(118)}$.

这七个类的典型代表分别是第 25 号,114 号,89 号,112 号,9 号,47 号和 118 号样品,以它们的服装尺寸作为一个型号的标准尺寸. 如型号 I(第一类)的标准尺寸就是第 25 号样品的尺寸等等. 各种型号服装的生产数量也按 25∶14∶9∶7∶12∶20∶8 这样的比例来生产.

注意:这七类并没有把 128 个点全部包括在内,还有 33 个样品不能归入这七个类,可认为它们是一些特殊体形的样品.

三、样品排序或系统评估

对 p 元总体 X 的样本进行主成分分析往往不是最终的目的,而常常是完成某个实际问题的一种手段. 如例 7.2.1 中由第一主成分得分对 30 名中学生的身体魁梧程度进行排序.

在实际工作中常会遇到多指标系统的排序评估问题，比如对某类企业的经济效益进行评估比较，影响企业经济效益的指标有很多，如何更科学、更客观地将一个多指标问题综合为单个指数的形式. 主成分分析方法为样品排序或多指标系统评估提供可行的方法.

对多指标系统进行排序评估的主要方法是加权评估法，比如专家评估方法、综合评分法、层次分析法等. 随着多元统计方法的普及与应用，主成分分析方法也成为构造系统排序评估指数的常用方法之一.

设 Z_1 是标准化随机向量 $X=(X_1,\cdots,X_p)'$ 的第一主成分. 由主成分的性质可知，Z_1 与原始标准化变量 $X_1,X_2,\cdots,X_p$ 的综合相关程度最强，即

$$\sum_{k=1}^{p}\rho^2(Z_1,X_k)=\lambda_1 \text{ 达最大},$$

其中 λ_1 为 X 的相关阵 R 的最大特征值. 如果只选一个综合变量来代表原来所有的原始变量，最佳的选择就是 Z_1.

另一方面，由于第一主成分 Z_1 对应于数据变异最大的方向，这说明 Z_1 是使数据信息损失最小、精度最高的一维综合变量，因此它可用于构造系统排序评估指数.

几点说明：

(1) 第一主成分 Z_1 并不是总能够被用来作为排序评估指数. 如果

$$Z_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p$$

中的系数 a_{i1} $(i=1,\cdots,p)$ 既有正又有负或近似为零，说明 Z_1 与原始变量 $X_1,X_2,\cdots,X_p$ 中有一部分为正相关，而另一部分为负相关或不相关，这时 Z_1 有可能是无序指数，不能用 Z_1 作为排序评估指数.

(2) 一般情况下，$Z_2,Z_3,\cdots,Z_p$ 不适合用来构造排序评估指数. 因 Z_k $(k=2,\cdots,p)$ 一般与原始变量 $X_1,X_2,\cdots,X_p$ 中有一部分为正相关，而另一部分为负相关或不相关. 或者说 $Z_2,Z_3,\cdots,Z_p$ 一般是无序指数.

(3) 传统的专家评估和第一主成分评估法的结合. 把传统的专家调查研究的信息用来对主成分评估法进行修正. 具体作法可按以

下步骤来进行：

① 把原始数据阵标准化后仍记为 X，标准化的随机向量仍记为 $X=(X_1,\cdots,X_p)'$.

② 把专家调查研究后得到的信息对变量的重要程度分别赋予不同的权重. 设 X_j 为第 j 个变量的 n 次观测向量，令

$$X_j^* = (1+\alpha_j)X_j = (x_{1j}^*, x_{2j}^*, \cdots, x_{nj}^*)' \quad (\alpha_j > 0, j = 1, \cdots, p).$$

记 $X^* = (x_{ij}^*)_{n\times p}$.

③ 由标准化且重新加权后的数据阵 X^* 出发来计算样本协方差阵 Σ^*，并求 Σ^* 的最大特征值 λ 和相应的特征向量 l.

④ 令 $Y=X^*l=(y_1,y_2,\cdots,y_n)'$（其中 l 为 p 维向量），然后按 $y_1,y_2,\cdots,y_n$ 的大小排序后进行评估.

四、主成分回归

在考虑因变量 Y 与 p 个自变量 $X_1,\cdots,X_p$ 的回归模型中，当自变量间有较强的线性相关（多重共线性）时，利用经典的回归方法求回归系数的最小二乘估计，一般效果较差. 利用 p 个变量的主成分 $Z_1,\cdots,Z_p$ 所具有的性质，如它们是互不相关的，$\mathrm{Var}(Z_i)=\lambda_i$ 为第 i 大特征值等，可由前 m 个主成分 $Z_1,\cdots,Z_m$ 来建立主成分回归模型：

$$Y = b_0 + b_1Z_1 + \cdots + b_mZ_m \quad (m \leqslant p).$$

由原始变量的观测数据计算前 m 个主成分的得分值，将其作为主成分 $Z_1,\cdots,Z_m$ 的观测值，建立 Y 与 $Z_1,\cdots,Z_m$ 的回归模型即得主成分回归方程. 这时就把 p 元数据降为 m 元. 这样既简化了回归方程的结构，且消除了变量间相关性带来的影响；但另一方面，主成分回归也给回归模型的解释带来一定的复杂性，因为主成分是原始变量的线性组合，不是直接观测的变量，其含义有时不明确. 在求得主成分回归方程后，经常又使用逆变换将其变为原始变量的回归方程.

当原始变量间有较强的多重共线性，其主成分又有特殊的含义时，往往采用主成分回归，其效果比较好. 下面通过具体例子来说明主成分回归方法.

例 7.3.3　经济分析数据的主成分回归. 考察进口总额 Y 与三个自变量：国内总产值 x_1，存储量 x_2，总消费量 x_3(单位均为 10 亿法郎)之间的关系. 现收集了 1949 至 1959 年共 11 年的数据. 对表 7.6 的数据试用主成分回归分析方法求进口总额与总产值、存储量和总消费量的定量关系式.

表 7.6　经济分析数据

序号	x_1	x_2	x_3	Y
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239.0	0.7	167.6	26.3

解　首先把各变量的观测数据标准化，再调用 SAS/STAT 软件中 PRINCOMP 过程对 3 个自变量做主成分分析；然后用主成分得分数据进行主成分回归.

由输出结果可知，相关阵的三个特征值分别为

$$\lambda_1 = 1.999,\quad \lambda_2 = 0.998,\quad \lambda_3 = 0.003;$$

前两个主成分的累计贡献率在 99%以上. 取两个主成分(用 x_i^* 表示 x_i 的标准化变量)：

$$\hat{Z}_1 = 0.7063x_1^* + 0.0435x_2^* + 0.7065x_3^*,$$

$$\hat{Z}_2 = -0.0357x_1^* + 0.9990x_2^* - 0.0258x_3^*.$$

由主成分回归得到的标准化回归方程为(Y^* 表示 Y 的标准化变量)：

$$\begin{aligned}\hat{Y}^* &= 0.68998\hat{Z}_1 + 0.1913\hat{Z}_2\\ &= 0.4804x_1^* + 0.2211x_2^* + 0.4825x_3^*,\end{aligned}$$

用原始变量可将 $\hat{Y}$ 的回归式表示为

$$\hat{Y} = -9.130 + 0.0727x_1 + 0.6091x_2 + 0.1062x_3. \tag{7.3.3}$$

在 SAS 系统 6.11 以上的版本中，我们还可以使用 REG 过程的选项由原始数据来完成主成分回归. 输出 7.3.3 给出删去第三个主成分(取选项 PCOMIT=1)后的主成分回归方程(同(7.3.3)式，见输出 7.3.3 中 Obs 为 2 的那一行). 这个主成分回归方程中回归系数的符号都是有意义的；主成分回归方程的均方根误差(_RMSE_=0.55)，虽比普通回归方程的均方根误差(_RMSE_=0.48887)有所增大，但增加并不多.

输出 7.3.3 经济分析数据主成分回归的结果

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	x1	x2	x3	y
1	MODEL1	PARMS	y	.	.	0.48887	-10.1280	-0.051396	0.58695	0.28685	-1
2	MODEL1	IPC	y	.	1	0.55001	-9.1301	0.072780	0.60922	0.10626	-1
3	MODEL1	IPC	y	.	2	1.05206	-7.7458	0.073814	0.08269	0.10735	-1

五、主成分检验法

设

$$X_{(i)} = (x_{i1}, \cdots, x_{ip})' \quad (i = 1, \cdots, n)$$

为来自 p 元总体 X 的样本，要检验总体 X 是否为 p 元正态总体，这是第三章 §3.6 中讨论的问题.

设 $D(X)=\Sigma$，如果 Σ 是对角矩阵，即 p 维向量的分量间不相关，这时把 p 元正态性检验问题可转化为 p 个一元正态性检验问题. 但一般 Σ 不是对角矩阵，即分量间是相关的. 利用主成分分析方法，求得 X 的 p 个主成分 $Z_1, \cdots, Z_p$(不相关)，并由原样本值计算 p 个主成分的得分值，作为 p 个不相关的综合变量的样本值. 这时就把 p 元正态性检验问题化为 p 个一元综合变量(主成分)的正态性检验. 这就是多元正态性检验的主成分检验法. 实际检验时，利用主成分的性质，只须对前 m ($m<p$)个主成分得分数据逐个做正态性检验.

习　题　七

7-1　设 $X=(X_1,X_2)'$ 的协方差阵 $\Sigma=\begin{bmatrix}1 & 4\\ 4 & 100\end{bmatrix}$，试从协方差阵 Σ 和相关阵 R 出发求出总体主成分，并加以比较.

7-2　设 $X=(X_1,X_2)'\sim N_2(0,\Sigma)$，协方差阵 $\Sigma=\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$，其中 ρ 为 X_1 和 X_2 的相关系数($\rho>0$).

(1) 试从 Σ 出发求 X 的两个总体主成分；

(2) 求 X 的等概密度椭圆的主轴方向；

(3) 试问当 ρ 取多大时才能使第一主成分的贡献率达 95%以上.

7-3　设 p 元总体 X 的协方差阵为

$$\Sigma=\sigma^2\begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & & \vdots\\ \rho & \rho & \cdots & 1\end{bmatrix}\quad(0<\rho\leqslant 1).$$

(1) 试证明总体的第一主成分 $Z_1=\dfrac{1}{\sqrt{p}}(X_1+X_2+\cdots+X_p)$；

(2) 试求第一主成分的贡献率.

7-4　设总体 $X=(X_1,\cdots,X_p)'\sim N_p(\mu,\Sigma)$ ($\Sigma>0$)，等概率密度椭球为

$$(X-\mu)'\Sigma^{-1}(X-\mu)=C^2\quad(C\text{ 为常数}).$$

试问椭球的主轴方向是什么?

7-5　设三元总体 X 的协方差阵为 $\Sigma=\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 2\end{bmatrix}$，试求总体主成分.

7-6　设三元总体 X 的协方差阵为 $\Sigma=\begin{bmatrix}\sigma^2 & \rho\sigma^2 & 0\\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2\\ 0 & \rho\sigma^2 & \sigma^2\end{bmatrix}$，试求

总体主成分，并计算每个主成分解释的方差比例（$|\rho|\leqslant 1/\sqrt{2}$）.

7-7　设 4 维随机向量 X 的协方差阵是

$$\Sigma=\begin{bmatrix}\sigma^2 & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{12} & \sigma^2 & \sigma_{14} & \sigma_{13}\\ \sigma_{13} & \sigma_{14} & \sigma^2 & \sigma_{12}\\ \sigma_{14} & \sigma_{13} & \sigma_{12} & \sigma^2\end{bmatrix},$$

其中 $\sigma_{12}\geqslant\sigma_{13}\geqslant\sigma_{14}\geqslant 0$，$\sigma^2+\sigma_{14}\geqslant\sigma_{12}+\sigma_{13}$. 试求 X 的主成分.

7-8　已知总体 $X=(X_1,\cdots,X_p)'$ 的 n 次观测数据阵为 $X=(x_{ij})_{n\times p}$；$Z_i=a_i'X$ $(i=1,\cdots,m;m<p)$ 是 X 的前 m 个样本主成分. 设变量 X_j 与 $Z_1,\cdots,Z_m$ 的回归模型为

$$X_j=b_{j1}Z_1+\cdots+b_{jm}Z_m+\varepsilon_j\overset{\text{def}}{=\!=}b_j'Z+\varepsilon_j\quad(j=1,\cdots,p).$$

(1) 试求参数 b_j 的最小二乘估计 $\hat{b}_j$ $(j=1,2,\cdots,p)$；

(2) 求 X_j 回归方程的回归平方和 U_j、残差平方和 Q_j，以及决定系数 R_j^2 $(j=1,2,\cdots,p)$.

7-9　设 $X=(X_1,\cdots,X_p)'\sim N_p(\mu,\Sigma)$，$\Sigma$ 有一个 p 重特征值 λ_1（即 $\Sigma=\lambda_1 I_p$），

(1) 在观测值 $X_{(\alpha)}=(x_{\alpha 1},\cdots,x_{\alpha p})'$ $(\alpha=1,\cdots,n)$ 的基础上，证明 λ_1 的最大似然估计是

$$\hat{\lambda}_1=\frac{1}{pn}\sum_{i=1}^{p}\sum_{\alpha=1}^{n}(x_{\alpha i}-\bar{x}_i)^2,$$

其中 $\bar{x}_i=\dfrac{1}{n}\sum\limits_{\alpha=1}^{n}x_{\alpha i}$；

(2) 证明 X 的主成分由 $B'X$ 给出，其中 B 是任何 p 阶正交矩阵.

7-10　若随机向量 $X=(X_1,\cdots,X_p)'$ 的协方差阵是非负定矩阵 Σ，随机向量 $Y=(Y_1,\cdots,Y_p)'$ 的协方差阵是 $\Sigma+\sigma^2 I$，则 $L'X$ 是 X 的主成分的充要条件是，$L'Y$ 是 Y 的主成分，其中 L 是正交矩阵.

7-11　用主成分分析方法探讨城市工业主体结构. 表 7.7 是某市工业部门 13 个行业 8 项指标的数据.

表 7.7　某市工业部门 13 个行业 8 项指标的数据

指标 / 行业	年末固定资产净值 X_1（万元）	职工人数 X_2（人）	工业总产值 X_3（万元）	全员劳动生产率 X_4（元/人年）	百元固定原资产值实现产值 X_5（元）	资金利税率 X_5（%）	标准燃料消费量 X_7（吨）	能源利用效果 X_8（万元/吨）
1(冶金)	90342	52455	101091	19272	82.000	16.100	197435	0.172
2(电力)	4903	1973	2035	10313	34.200	7.100	592077	0.003
3(煤炭)	6735	21139	3767	1780	36.100	8.200	726396	0.003
4(化学)	49454	36241	81557	22504	98.100	25.900	348226	0.985
5(机械)	139190	203505	215898	10609	93.200	12.600	139572	0.628
6(建材)	12215	16219	10351	6382	62.500	8.700	145818	0.066
7(森工)	2372	6572	8103	12329	184.400	22.200	20921	0.152
8(食品)	11062	23078	54935	23804	370.400	41.000	65486	0.263
9(纺织)	17111	23907	52108	21796	221.500	21.500	63806	0.276
10(缝纫)	1206	3930	6126	15586	330.400	29.500	1840	0.437
11(皮革)	2150	5704	6200	10870	184.200	12.000	8913	0.274
12(造纸)	5251	6155	10383	16875	146.400	27.500	78796	0.151
13(文教艺术用品)	14341	13203	19396	14691	94.600	17.800	6354	1.574

(1) 试用主成分分析方法确定 8 项指标的样本主成分(综合变量)；若要求损失信息不超过 15%，应取几个主成分；并对这几个主成分进行解释；

(2) 利用主成分得分对 13 个行业进行排序和分类.

7-12　试对第六章表 6.7 中 16 个地区农民生活水平的调查数据进行主成分分析，并利用前两个主成分对 16 个地区的农民生活水平进行分类(请与第六章的分类计算结果进行比较).

第八章　因子分析

§8.1　引　　言

因子分析是主成分分析的推广和发展，它也是多元统计分析中降维的一种方法. 因子分析是研究相关阵或协方差阵的内部依赖关系，它将多个变量综合为少数几个因子，以再现原始变量与因子之间的相关关系.

因子分析的形成和早期发展，一般认为是从 Charles Spearman 在 1904 年发表的文章开始. 他提出这种方法用来解决智力测验得分的统计分析. 目前因子分析在心理学、社会学、经济学等学科都取得成功的应用.

下面我们列举几个实际问题以说明如何应用因子分析来构造因子模型.

实例 1　为了了解学生的学习能力，观测了 n 个学生 p 个科目的成绩(分数)，用 $X_1,\cdots,X_p$ 表示 p 个科目(例如代数、几何、语文、英语、政治，……)，$X_{(t)}=(x_{t1},\cdots,x_{tp})'$ $(t=1,\cdots n)$表示第 t 个学生的 p 个科目的成绩. 我们对这些资料进行归纳分析，可以看出各个科目(即变量)由两部分组成：

$$X_i = a_iF + \varepsilon_i \quad (i = 1,\cdots,p), \tag{8.1.1}$$

其中 F 是对所有 $X_i(i=1,\cdots,p)$都起作用的公共因子，它表示智能高低的因子；系数 a_i 称为因子载荷，表示第 i 个科目在智能高低因子上的体现；ε_i 是科目(变量)X_i 特有的特殊因子. 这就是一个最简单的因子模型.

进一步地可把这个简单因子模型推广到多个因子的情况，即全部科目 X 所共有的因子有 m 个，如数学推导因子、记忆因子、计算因子等，分别记为 $F_1,\cdots,F_m$，即

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \quad (i = 1, \cdots, p). \tag{8.1.2}$$

用这 m 个不可观测的互不相关的公共因子 $F_1, \cdots, F_m$(也称为潜因子)和一个特殊因子 ε_i 来描述原始可测的相关变量(科目)$X_1, \cdots, X_p$,并解释分析学生的学习能力. 它们的系数 $a_{i1}, \cdots, a_{im}$ 称为因子载荷,表示第 i 个科目在 m 个方面的表现. 这就是一个因子分析模型.

实例 2　调查青年对婚姻家庭的态度,抽取 n 个青年回答了 $p=50$ 个问题的答卷,这些问题可归纳为如下几个方面:对相貌的重视、对孩子的观点、对老人的态度等,这也是一个因子分析的模型,每一个方面就是一个因子.

实例 3　考察人体的五项生理指标:收缩压(X_1)、舒张压(X_2)、心跳间隔(X_3)、呼吸间隔(X_4)和舌下温度(X_5). 从生理学的知识可知,这五项指标是受植物神经支配的,植物神经又分为交感神经和副交感神经,因此这五项指标至少受到两个公共因子的影响,也可用因子分析的模型去处理它.

通过以上几个实例,我们可以看到,因子分析的主要应用有两方面,一是寻求基本结构,简化观测系统,将具有错综复杂关系的对象(变量或样品)综合为少数几个因子(不可观测的随机变量),以再现因子与原始变量之间的内在联系;二是用于分类,对 p 个变量或 n 个样品进行分类.

因子分析根据研究对象的不同可以分为 R 型和 Q 型因子分析. R 型因子分析研究变量(指标)之间的相关关系,通过对变量的相关阵或协方差阵内部结构的研究,找出控制所有变量的几个公共因子(或称主因子、潜因子),用以对变量或样品进行分类. Q 型因子分析研究样品之间的相关关系,通过对样品的相似矩阵内部结构的研究找出控制所有样品的几个主要因素(或称主因子). 这两种因子分析的处理方法一样,只是出发点不同. R 型从变量的相关阵出发,Q 型从样品的相似矩阵出发. 对一批观测数据,可以根据实际问题的需要来决定采用哪一种类型的因子分析. 本章主要介绍 R 型因子分析.

因子分析与主成分分析有区别. 主成分分析不能作为一个模型来描述,它只是通常的变量变换,而因子分析需要构造因子模型;主

成分分析中主成分的个数和变量个数 p 相同，它是将一组具有相关关系的变量变换为一组互不相关的变量(注意应用主成分分析解决实际问题时，一般只选取前 m ($m<p$)个主成分)，而因子分析的目的是要用尽可能少的公因子，以便构造一个结构简单的因子模型；主成分分析是将主成分表示为原始变量的线性组合，而因子分析是将原始变量表示为公因子和特殊因子的线性组合，用假设的公因子来"解释"相关阵的内部依赖关系. 另一方面这两种分析方法之间在某些情况下也有一定联系，读者将从下面的介绍中看到.

§8.2 因子模型

一、正交因子模型

设 $X=(X_1,\cdots,X_p)'$ 是可观测的随机向量，$\mathrm{E}(X)=\mu$，$\mathrm{D}(X)=\Sigma$，且设 $F=(F_1,\cdots,F_m)'$ ($m<p$)是不可观测的随机向量，$\mathrm{E}(F)=0$，$\mathrm{D}(F)=I_m$(即 F 的各分量方差为 1，且互不相关). 又设 $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_p)'$ 与 F 互不相关，且

$$\mathrm{E}(\varepsilon)=0,\quad \mathrm{D}(\varepsilon)=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_p^2)\overset{\text{def}}{=\!=}D\ (\text{对角矩阵}).$$

假定随机向量 X 满足以下的模型：

$$\begin{cases} X_1-\mu_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1,\\ X_2-\mu_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\varepsilon_2,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ X_p-\mu_p=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\varepsilon_p, \end{cases} \tag{8.2.1}$$

则称模型(8.2.1)为**正交因子模型**. 用矩阵表示为

$$X=\mu+AF+\varepsilon, \tag{8.2.2}$$

其中 $F=(F_1,\cdots,F_m)'$，$F_1,\cdots,F_m$ 称为 X 的**公共因子**；$\varepsilon=(\varepsilon_1,\cdots,\varepsilon_p)'$，$\varepsilon_1,\cdots,\varepsilon_p$ 称为 X 的**特殊因子**；公共因子 $F_1,\cdots,F_m$ 一般对 X 的每一个分量 X_i 都有作用，而 ε_i 只对 X_i 起作用，而且各特殊因子之间以及特殊因子与所有公共因子之间都是互不相关的.

模型中的矩阵 $A=(a_{ij})_{p\times m}$ 是待估的系数矩阵，称为**因子载荷**

矩阵. $a_{ij}(i=1,\cdots,p;j=1,\cdots,m)$ 称为第 i 个变量在第 j 个因子上的载荷(简称为**因子载荷**).

正交因子模型(8.2.1)中用 $m+p$ 个不可观测的随机变量 $F_1,\cdots,F_m,\varepsilon_1,\cdots,\varepsilon_p$ 来表示 p 个原始变量 $X_1,\cdots,X_p$,这是 R 型正交因子模型与回归模型的区别所在. 试图用回归方法确定因子载荷矩阵 A 是不可行的. 上述(8.2.1)的模型中对 F 和 ε 作了一系列的假定,使得模型具有特定的且能验证的协方差结构. 在以上一系列假定中有两个关键性的假设:

(1) 特殊因子互不相关,且 $\mathrm{D}(\varepsilon)=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_p^2)\overset{\text{def}}{=\!=}D$;

(2) 特殊因子同公共因子互不相关,即 $\mathrm{COV}(\varepsilon,F)=O_{p\times m}$.

在主成分分析中,回归模型(7.2.2)中的残差通常是彼此相关的. 在因子分析中,特殊因子起着残差的作用,但被定义为彼此不相关且与公共因子也不相关;而且每个公共因子假定至少对两个变量有贡献,否则它将是一个特殊因子.

在正交因子模型中,假定公共因子彼此不相关且具有单位方差,即 $\mathrm{D}(F)=I_m$. 在这种情况下,由

$$\begin{aligned}\Sigma=\mathrm{D}(X)&=\mathrm{E}[(X-\mu)(X-\mu)']\\&=\mathrm{E}[(AF+\varepsilon)(AF+\varepsilon)']=A\mathrm{D}(F)A'+\mathrm{D}(\varepsilon)=AA'+D,\end{aligned}$$

即由

$$\Sigma-D=AA' \tag{8.2.3}$$

可知,正交因子模型意味着第 j 个变量和第 k 个变量的协方差 σ_{jk} 由下式给出

$$\sigma_{jk}=a_{j1}a_{k1}+a_{j2}a_{k2}+\cdots+a_{jm}a_{km}\quad(j\neq k),$$

$$\sigma_{jj}=a_{j1}^2+a_{j2}^2+\cdots+a_{jm}^2+\sigma_j^2\quad(j=k).$$

如果原始变量已被标准化为单位方差,在(8.2.3)式中将用相关阵代替协方差阵. 在此意义上,公共因子解释了观测变量间的相关性. 用正交因子模型预测的相关与实际的相关之间的差异就是剩余相关. 评估正交因子模型拟合优度的好方法就是考察剩余相关的大小.

因子分析的目的首先是由样本协方差阵 $\hat{\Sigma}$ 估计 Σ,然后由分解

式(8.2.3)求得 A 和 D. 也就是从可以观测的变量 $X_1,\cdots,X_p$ 给出的样本资料中,求出载荷矩阵 A,然后预测公共因子 $F_1,\cdots,F_m$. 又因

$$\begin{aligned}\mathrm{COV}(X,F)&=\mathrm{E}(X-\mathrm{E}(X))(F-\mathrm{E}(F))'=\mathrm{E}[(X-\mu)F']\\&=\mathrm{E}[(AF+\varepsilon)F']=A\mathrm{E}(FF')+\mathrm{E}(\varepsilon F')\\&=A,\end{aligned}\tag{8.2.4}$$

其中 A 为 $p\times m$ 矩阵. 可见 A 中元素 a_{ij} 刻画变量 X_i 与 F_j 之间的相关性,称为 X_i 在 F_j 上的因子载荷.

关系式(8.2.3)和(8.2.4)称为正交因子模型的**协方差结构**.

二、正交因子模型中各个量的统计意义

1. 因子载荷的统计意义

由因子模型(8.2.1)及(8.2.4)可知,X_i 与 F_j 的协方差

$$\mathrm{Cov}(X_i,F_j)=a_{ij}.$$

如果变量 X_i 是标准化变量(即 $\mathrm{E}(X_i)=0$, $\mathrm{Var}(X_i)=1$),则

$$\rho_{ij}=\frac{\mathrm{Cov}(X_i,F_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(F_j)}}=\mathrm{Cov}(X_i,F_j)=a_{ij}.$$

这时因子载荷 a_{ij} 就是第 i 个变量与第 j 个公共因子的相关系数. 由模型(8.2.1), X_i 是 $F_1,\cdots,F_m$ 的线性组合,系数 $a_{i1},\cdots,a_{im}$ 是用来度量 X_i 可由 $F_1,\cdots,F_m$ 线性组合表示之程度的. 模型中 $F_1,\cdots,F_m$ 的系数 $a_{i1},\cdots,a_{im}$ 用统计学的术语叫做"**权重**",它表示 X_i 依赖 F_j 的分量(比重). 由于历史的原因,在心理学中将模型(8.2.1)的系数 a_{ij} 叫做"载荷", 即第 i 个变量在第 j 个因子上的载荷(或负荷),反映了第 i 个变量在第 j 个公共因子上的相对重要性.

2. 变量共同度的统计意义

因子载荷矩阵 A 中各行元素的平方和记为 h_i^2,

$$h_i^2=\sum_{j=1}^{m}a_{ij}^2\quad(i=1,2,\cdots,p)$$

称为变量 X_i 的**共同度**.

为了给出 h_i^2 的统计意义,下面来计算 X_i 方差:

$$\mathrm{Var}(X_i)=\mathrm{Var}\Big(\sum_{t=1}^{m}a_{it}F_t+\varepsilon_i\Big)=\sum_{t=1}^{m}a_{it}^2\mathrm{Var}(F_t)+\mathrm{Var}(\varepsilon_i)=h_i^2+\sigma_i^2.$$

上式表明 X_i 的方差由两部分组成，第一部分 h_i^2 是全部公共因子对变量 X_i 的总方差所作出的贡献，称为**公因子方差**；第二部分 σ_i^2 是由特定因子 ε_i 产生的方差，它仅与变量 X_i 有关，也称为**剩余方差**.

显然，若 h_i^2 大，σ_i^2 必小. 而 h_i^2 大表明 X_i 对公共因子 $F_1,\cdots,F_m$ 的共同依赖程度大. 当 $h_i^2=1$（设 $\mathrm{Var}(X_i)=1$）时，$\sigma_i^2=0$，即 X_i 完全能够由公共因子的线性组合表示；当 $h_i^2\approx 0$ 时，表明公共因子对 X_i 影响很小，X_i 主要由特殊因子 ε_i 来描述. 可见 h_i^2 反映了变量 X_i 对公因子 F 依赖的程度，故也称公因子方差 h_i^2 为变量 X_i 的**共同度**.

3. 公共因子 F_j 的方差贡献的统计意义

在因子载荷矩阵 A 中，求 A 的各列的平方和，记为 q_j^2，即

$$q_j^2=\sum_{i=1}^{p}a_{ij}^2\quad(j=1,2,\cdots,m).$$

q_j^2 的统计意义与 X_i 的共同度 h_i^2 恰好相反，q_j^2 表示第 j 个公共因子 F_j 对 X 的所有分量 $X_1,\cdots,X_p$ 的总影响，称为第 j 个公共因子 F_j 对 X 的贡献，它是衡量第 j 个公共因子相对重要性的指标.

显然，q_j^2 愈大，表明 F_j 对 X 的贡献愈大. 如果我们把载荷矩阵 A 的各列平方和都计算出来，使相应的贡献有顺序：$q_1^2\geqslant q_2^2\geqslant\cdots\geqslant q_m^2$，我们就能够以此为依据，找出最有影响的公共因子. 要解决此问题，关键是求载荷矩阵 A 的估计.

关于因子模型有下列两点需要指出，以便引起读者的注意：

（1）模型不受量纲影响. 变量 X 量纲的变化等价于作变换：$X^*=CX$，其中 C 为对角矩阵，则

$$\mathrm{D}(X^*)=C\mathrm{D}(X)C'=C(AA'+D)C'=CAA'C'+CDC',$$

$$\mathrm{E}(X^*)=C\mu.$$

记 $\mu^*=C\mu$，$A^*=CA$，$C\Sigma C'=\Sigma^*$，$CDC'=D^*$，$\varepsilon^*=C\varepsilon$，则仍有因子模型

$$\begin{cases}X^*=\mu^*+A^*F+\varepsilon^*,\\ \Sigma^*=A^*(A^*)'+D^*.\end{cases}$$

(2) 因子载荷矩阵 A 是不唯一的. 若 Γ 是任一 $m\times m$ 正交矩阵,则模型(8.2.2)可表为

$$X=\mu+(A\Gamma)(\Gamma'F)+\varepsilon.$$

因

$$\mathrm{E}(\Gamma'F)=0,\quad \mathrm{D}(\Gamma'F)=I,$$

$$\mathrm{COV}(\Gamma'F,\varepsilon)=\Gamma'\mathrm{COV}(F,\varepsilon)=O,$$

故将 $\Gamma'F$ 看成公共因子, $A\Gamma$ 看成相应的因子载荷矩阵. 这时

$$\Sigma=AA'+D=(A\Gamma)(A\Gamma)'+D,$$

可见,因子载荷矩阵 A 不唯一.

例 8.2.1 已知 $X=(X_1,\cdots,X_4)'$的协方差阵 Σ 为:

$$\Sigma=\begin{bmatrix}19 & 30 & 2 & 12\\ 30 & 57 & 5 & 23\\ 2 & 5 & 38 & 47\\ 12 & 23 & 47 & 68\end{bmatrix},$$

试求满足(8.2.3)式的因子载荷矩阵 A 和特殊因子的协方差阵 D, 并计算 X_1 的共同度.

解 容易验证,

$$\Sigma=\begin{bmatrix}4 & 1\\ 7 & 2\\ -1 & 6\\ 1 & 8\end{bmatrix}\begin{bmatrix}4 & 7 & -1 & 1\\ 1 & 2 & 6 & 8\end{bmatrix}+\begin{bmatrix}2 & 0 & 0 & 0\\ 0 & 4 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 3\end{bmatrix}$$

$$=AA'+D.$$

因而因子载荷矩阵 A 和特殊因子协方差阵 D 分别为:

$$A=\begin{bmatrix}4 & 1\\ 7 & 2\\ -1 & 6\\ 1 & 8\end{bmatrix},\quad D=\begin{bmatrix}2 & 0 & 0 & 0\\ 0 & 4 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 3\end{bmatrix}.$$

即 X 的协方差阵 Σ 具有 $m=2$ 的正交因子模型结构,且 X_1 的共同度为

$$h_1^2=4^2+1^2=17.$$

第一个特殊因子 ε_1 的方差 $\sigma_1^2=2$, X_1 的方差可分解为:

$$19=17+2,$$

即

$$\text{方差}=\text{共同度}+\text{特殊方差}.$$

对 $X_i(i=2,3,4)$也有类似地分解.

三、正交因子模型的几何解释

在因子分析中,我们可以把互不相关的,即两两之间的相关系数(夹角余弦)为 0,各自方差为 1 的 m 个公共因子和 p 个特殊因子想象成 $m+p$ 个相互正交的单位向量,以它们为坐标轴构成 $m+p$ 维空间的一个直角坐标系,并称为因子空间. 于是根据模型(8.2.1),变量 X_i 可以用因子空间中向量 $P_i=(a_{i1},a_{i2},\cdots,a_{im},0,\cdots,\sigma_i,\cdots,0)'$表示,其中 σ_i 是 X_i 在对应于自己的特殊因子轴上的载荷. 由于 X_i 标准化,显然 P_i 的长度等于 1,即

$$\| P_i \| = \sqrt{a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \sigma_i^2} = \sqrt{\mathrm{Var}(X_i)} = 1,$$

此时 P_i 与各个因子轴 F_j 的夹角余弦为

$$\cos\langle P_i,F_j\rangle = \| P_i \| \cos\langle P_i,F_j\rangle = a_{ij} = r_{P_iF_j},$$

这表明了 P_i 与各公共因子的夹角余弦就等于其相应的坐标,也就是等于变量 X_i 与各公共因子的相关系数. 此外,对于因子空间中分别表示变量 X_i 和 X_j 的向量 P_i 与 P_j 的夹角余弦即为它们的内积

$$\cos\langle P_i,P_j\rangle = \frac{P_i'P_j}{\| P_i \| \| P_j \|} = P_i'P_j = \sum_{t=1}^{m} a_{it}a_{jt} = r_{X_iX_j},$$

它恰好等于变量 X_i 与 X_j 的相关系数.

§8.3　参数估计方法

已知 p 个相关变量的 n 次观测值 $X_{(i)}=(x_{i1},\cdots,x_{ip})'$ $(i=1,\cdots,n)$. 因子分析的目的是用少数几个公共因子(设为 m 个)来描述 p 个相关变量间的协方差结构:

$$\Sigma = AA' + D,$$

其中 $A=(a_{ij})$为 $p\times m$ 的因子载荷矩阵; $D=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_p^2)$为 p 阶对角矩阵. 也就是估计公共因子的个数 m、因子载荷矩阵 A 及特殊因子方差 $\sigma_i^2(i=1,\cdots,p)$,使得满足 $\Sigma=AA'+D$.

由 p 个相关变量的观测数据计算样本协方差阵 S，作为协方差阵的估计. 为了建立公因子模型，首先要估计因子载荷 a_{ij} 和特殊因子方差 σ_i^2. 常用的参数估计方法有以下三种：主成分法，主因子解和极大似然法.

一、主成分法

设样本协方差阵 S 的特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，相应单位正交特征向量为 $l_1, l_2, \cdots, l_p$，则 S 有谱分解式：

$$S = \sum_{i=1}^{p} \lambda_i l_i l_i'.$$

当最后 $p-m$ 个特征值较小时，S 可近似地分解为

$$\begin{aligned} S &\approx \lambda_1 l_1 l_1' + \cdots + \lambda_m l_m l_m' + D \\ &= (\sqrt{\lambda_1} l_1, \cdots, \sqrt{\lambda_m} l_m) \begin{bmatrix} \sqrt{\lambda_1} l_1' \\ \vdots \\ \sqrt{\lambda_m} l_m' \end{bmatrix} + \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_p^2 \end{bmatrix} \\ &= AA' + D, \end{aligned} \tag{8.3.1}$$

其中

$$\begin{cases} A = (\sqrt{\lambda_1} l_1, \cdots, \sqrt{\lambda_m} l_m) \xlongequal{\text{def}} (a_{ij})_{p \times m}, \\ \sigma_i^2 = s_{ii} - \sum_{t=1}^{m} a_{it}^2 \quad (i = 1, 2, \cdots, p). \end{cases} \tag{8.3.2}$$

(8.3.2)式给出的 A 和 D 就是因子模型的一个解. 载荷矩阵 A 中的第 j 列(即第 j 个公共因子 F_j 在 X 上的载荷)和 X 的第 j 个主成分的系数相差一个倍数 $\sqrt{\lambda_j}$ $(j=1,2,\cdots,m)$. 故(8.3.2)式给出的这个解常称为因子模型的**主成分解**.

若记 $\varepsilon = S - (AA' + D) \xlongequal{\text{def}} (\varepsilon_{ij})_{p \times p}$，可以证明(见本章习题第8-4题)：

$$Q(m) = \sum_{i=1}^{p} \sum_{j=1}^{p} \varepsilon_{ij}^2 \leqslant \lambda_{m+1}^2 + \cdots + \lambda_p^2, \tag{8.3.3}$$

当 m 选择适当，则近似式(8.3.1)的误差平方和 $Q(m)$ 很小.

公因子个数 m 的确定方法一般有两种，一是根据实际问题的意

义或专业理论知识来确定；二是用确定主成分个数的原则，选 m 为满足：

$$\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\cdots+\lambda_m+\cdots+\lambda_p}\geqslant P_0$$

的最小整数(比如取 $P_0\geqslant 0.70$ 且 $P_0<1$).

当相关变量所取单位不同时，我们常常先对变量标准化. 标准化变量的样本协方差阵就是原始变量的样本相关阵 R，再用 R 代替 S，与上类似，即可得主成分解.

二、主因子解

从 R 出发，下面来介绍主成分法的一种修正. 设 $R=AA'+D$，则

$$R-D=AA'\xlongequal{\text{def}}R^*$$

称为**约相关阵**. 如果我们已知特殊方差的初始估计 $(\hat{\sigma}_i^*)^2$，也就是已知初始公因子方差(即共同度)的估计为

$$(h_i^*)^2=1-(\hat{\sigma}_i^*)^2,$$

则约相关阵 $R^*=R-D$ 为

$$R^*=\begin{bmatrix}(h_1^*)^2 & r_{12} & \cdots & r_{1p}\\ r_{21} & (h_2^*)^2 & \cdots & r_{2p}\\ \vdots & \vdots & & \vdots\\ r_{p1} & r_{p2} & \cdots & (h_p^*)^2\end{bmatrix}.$$

计算 R^* 的特征值和单位正交特征向量，可取前 m 个正特征值 $\lambda_1^*\geqslant\lambda_2^*\geqslant\cdots\geqslant\lambda_m^*>0$，其相应的单位正交特征向量为 $l_1^*,l_2^*,\cdots,l_m^*$，则有近似分解式：

$$R^*=AA',$$

其中

$$A=(\sqrt{\lambda_1^*}\,l_1^*,\cdots,\sqrt{\lambda_m^*}\,l_m^*).$$

令

$$\hat{\sigma}_i^2=1-\sum_{t=1}^{m}a_{it}^2\quad(i=1,\cdots,p),$$

则 A 和 $D^*=\mathrm{diag}(\hat{\sigma}_1^2,\cdots,\hat{\sigma}_p^2)$ 为因子模型的一个解，这个解就称为**主因子解**.

在实际应用中特殊因子方差 σ_i^2 或公因子方差(也称为共同度) h_i^2 是未知的. 以上得到的解是近似解. 为了得到近似程度更好的解, 常常采用迭代主因子法, 即利用上面得到的 $D^* = \mathrm{diag}(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2)$ 作为特殊方差的初始估计, 重复上述步骤, 直到解稳定为止.

因特殊因子方差 $\sigma_i^2 = 1 - h_i^2$, 故求特殊因子方差的初始估计等价于求公因子方差 h_i^2 的初始估计. 下面介绍公因子方差常用的初始估计的几种方法:

(1) h_i^2 取为第 i 个变量与其他所有变量的多重相关系数的平方(或者取 $\sigma_i^2 = 1/r^{ii}$, 其中 r^{ii} 是 R^{-1} 的对角元素, 则 $h_i^2 = 1 - \sigma_i^2$);

(2) h_i^2 取为第 i 个变量与其他变量相关系数绝对值的最大值;

(3) 取 $h_i^2 = 1$, 它等价于主成分解.

三、极大似然法

假定公因子 F 和特殊因子 ε 服从正态分布, 那么我们可得到因子载荷矩阵和特殊方差的极大似然估计. 设 p 维观测向量 $X_{(1)}, \cdots, X_{(n)}$ 为来自正态总体 $N_p(\mu, \Sigma)$ 的随机样本, 则样本似然函数为 μ, Σ 的函数 $L(\mu, \Sigma)$.

设 $\Sigma = AA' + D$, 取 $\mu = \overline{X}$, 则似然函数 $L(\overline{X}, AA' + D)$ 的对数为 A, D 的函数, 记为 $\varphi(A, D)$, 求 A, D 使 φ 达最大. 可以证明使 $\varphi(A, D)$ 达极大的解 $\hat{A}$ 和 $\hat{D}$ 满足如下方程组:

$$\begin{cases} S\hat{D}^{-1}\hat{A} = \hat{A}(I + \hat{A}'\hat{D}^{-1}\hat{A}), \\ \hat{D} = \mathrm{diag}(S - \hat{A}\hat{A}'), \end{cases} \tag{8.3.4}$$

其中 $S = \dfrac{1}{n}\sum_{i=1}^{n}(X_{(i)} - \overline{X})(X_{(i)} - \overline{X})'$.

为保证方程组(8.3.4)得到唯一解, 可附加计算上方便的唯一性条件:

$$A'D^{-1}A = \text{对角矩阵}.$$

这个建议基于以下的贝叶斯观点: 设 $F \sim N_m(0, I_m)$, 当给定 $F = f$ 时, $X - \mu$ 的条件分布为 $N_p(Af, D)$; 当给定 X 时, F 的条件分布为 $N_m(A'\Sigma^{-1}(x - \mu), (A'D^{-1}A + I_m)^{-1})$, 若 $A'D^{-1}A$ 为对角矩阵, 则 F

的分量相互独立.

方程组(8.3.4)一般用迭代方法来求得极大似然估计 $\hat{A}$ 和 $\hat{D}$ (见参考文献[1]或[5]).

四、主成分估计法的具体步骤

以上三种估计方法中,主成分解应用较广泛.设样本数据阵为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p},$$

则应用主成分估计法的具体步骤如下:

(1) 由样本数据阵 X 计算样本均值、样本离差阵及样本相关阵.样本均值 $\overline{X}$ 为

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{n}X_{(t)}=(\overline{x}_1,\cdots,\overline{x}_p)';$$

样本离差阵 E 为

$$E=\sum_{t=1}^{n}(X_{(t)}-\overline{X})(X_{(t)}-\overline{X})'\xlongequal{\text{def}}(e_{ij}),$$

其中

$$e_{ij}=\sum_{t=1}^{n}(x_{ti}-\overline{x}_i)(x_{tj}-\overline{x}_j);$$

样本相关阵 $R=(r_{ij})$,其中 $r_{ij}=\dfrac{e_{ij}}{\sqrt{e_{ii}e_{jj}}}$ $(i,j=1,2,\cdots,p)$.

(2) 求 R 的特征值和标准化特征向量.记 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$ 为 R 的特征值,其相应的单位正交特征向量为 $l_1,l_2,\cdots,l_p$.

(3) 求因子模型的因子载荷矩阵 A:

① 确定公共因子的个数 m.比如取 m 为满足 $(\lambda_1+\lambda_2+\cdots+\lambda_m)/p\geqslant 0.80$ (或 0.70 或 0.90)的最小正整数;

② 令 $a_i=\sqrt{\lambda_i}\,l_i$ $(i=1,2,\cdots,m)$,则 $A=(a_1,\cdots,a_m)$ 为因子载荷矩阵.

(4) 求特殊因子方差 $\hat{\sigma}_i^2=1-\sum_{t=1}^{m}a_{it}^2$ $(i=1,\cdots,p)$, X_i 的共同

度 h_i^2 的估计为

$$\hat{h}_i^2 = \sum_{t=1}^{m} a_{it}^2 \quad (i = 1, \cdots, p).$$

(5) 对 m 个公共因子(或称潜因子,主因子)作解释. 求出因子载荷矩阵 A 后,即得可测变量 $X_1, \cdots, X_p$ 由 m 个不可测的公共因子及各自特殊因子的表示式,但这 m 个公共因子表示什么? 则要结合专业知识给出解释.

例 8.3.1 (盐泉水化学分析资料的因子分析) 今有 20 个盐泉,盐泉的水化学特征系数值见表 8.1. 试对盐泉水化学分析资料作因子分析(摘自参考文献[14]).

表 8.1 盐泉水化学特征系数的数据

序号	矿化度 (g/L) (X_1)	$\frac{Br \cdot 10^3}{Cl}$ (X_2)	$\frac{K \cdot 10^3}{\Sigma 盐}$ (X_3)	$\frac{K \cdot 10^3}{Cl}$ (X_4)	$\frac{Na}{K}$ (X_5)	$\frac{Mg \cdot 10^2}{Cl}$ (X_6)	$\frac{\epsilon Na}{\epsilon Cl}$ (X_7)
1	11.835	0.480	14.360	25.210	25.21	0.810	0.98
2	45.596	0.526	13.850	24.040	26.01	0.910	0.96
3	3.525	0.086	24.400	49.300	11.30	6.820	0.85
4	3.681	0.370	13.570	25.120	26.00	0.820	1.01
5	48.287	0.386	14.500	25.900	23.32	2.180	0.93
6	17.956	0.280	9.750	17.050	37.20	0.464	0.98
7	7.370	0.506	13.600	34.280	10.69	8.800	0.56
8	4.223	0.340	3.800	7.100	88.20	1.110	0.97
9	6.442	0.190	4.700	9.100	73.20	0.740	1.03
10	16.234	0.390	3.100	5.400	121.50	0.420	1.00
11	10.585	0.420	2.400	4.700	135.60	0.870	0.98
12	23.535	0.230	2.600	4.600	151.80	0.310	1.02
13	5.398	0.120	2.800	6.200	111.20	1.140	1.07
14	283.149	0.148	1.763	2.968	215.86	0.140	0.98
15	316.604	0.317	1.453	2.432	263.41	0.249	0.98
16	307.310	0.173	1.627	2.729	235.70	0.214	0.99
17	322.515	0.312	1.382	2.320	282.21	0.024	1.00
18	254.580	0.297	0.899	1.476	410.30	0.239	0.93
19	304.092	0.283	0.789	1.357	438.36	0.193	1.01
20	202.446	0.042	0.741	1.266	309.77	0.290	0.99

解 此例,$n=20$, $p=7$. 使用 SAS/STAT 软件中 FACTOR 过

程由相关阵出发进行因子分析.

规定 $p_0=0.80$ 来选取公共因子个数 m，使 m 为满足 $(\lambda_1+\lambda_2+\cdots+\lambda_m)/p\geqslant 0.80$ 的最小正整数.

输出 8.3.1　相关阵的特征值、相邻特征值之差、贡献率和累计贡献率

The SAS System

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1

	1	2	3	4
Eigenvalue	4.2444	1.2513	0.9202	0.4384
Difference	2.9931	0.3312	0.4818	0.3248
Proportion	0.6063	0.1788	0.1315	0.0626
Cumulative	0.6063	0.7851	0.9166	0.9792

	5	6	7
Eigenvalue	0.1136	0.0314	0.0007
Difference	0.0822	0.0307	
Proportion	0.0162	0.0045	0.0001
Cumulative	0.9954	0.9999	1.0000

从输出 8.3.1 可以看出，取公共因子的个数 $m=3$，3 个公共因子的累计贡献率在 0.90 以上(0.9166)，即说明前三个公共因子反映原始变量的信息已占总信息的 90%以上.

查看输出 8.3.2 给出的因子载荷矩阵 A（它是由前三个特征值及其相应的特征向量计算得到的），比如查看 FACTOR3 对应的列（第三个公共因子的载荷向量），除 $a_{23}=0.89278$ 较大外，其余各数值都较小，这表示可用 X_2 来解释公因子 F_3（或者说 F_3 主要反映 X_2 的信息）. 对某个公共因子进行解释的基本想法是：首先看哪些变量在这个因子中的负荷量大，然后利用专业知识讨论这些变量组合的实际含意. 由输出 8.3.2 可见，第一、第二公共因子的载荷中有一些数值在 0.5 附近的中等负荷，虽也能按以上想法进行解释，但容易使公共因子的意义含糊不清. 在 §8.4 中将介绍因子旋转后的因子载荷矩阵，其实际含意将更明显.

由输出 8.3.2，查看每一行，如第一行，可以得出：

$$X_1=-0.7156F_1+0.5645F_2+0.0456F_3+\varepsilon_1,$$

它给出了变量 X_1 与公共因子 F_1,F_2,F_3 及特殊因子 ε_1 的关系（即因子模型），其他各行类似.

输出 8.3.2 因子载荷矩阵 $A(m=3)$及每个公共因子解释的方差

```
3 factors will be retained by the PROPORTION criterion.

                    Factor Pattern

                FACTOR1    FACTOR2    FACTOR3

     X1        -0.71560    0.56452    0.04559
     X2         0.41233   -0.13191    0.89278
     X3         0.90960   -0.06429   -0.17215
     X4         0.94490    0.04843   -0.17493
     X5        -0.83458    0.46939    0.04773
     X6         0.82555    0.49675   -0.13428
     X7        -0.68122   -0.66459   -0.20123

          Variance explained by each factor

          FACTOR1    FACTOR2    FACTOR3
         4.244417   1.251331   0.920181
```

输出 8.3.3 最终公因子方差(即 $m=3$ 时各变量的共同度)的估计

```
        Final Communality Estimates: Total = 6.415930

     X1          X2          X3          X4          X5          X6          X7
0.832839    0.984480    0.861139    0.925789    0.919127    0.946322    0.946232
```

从输出 8.3.3 可以得出各变量共同度的估计.从而给出了特殊方差的估计,如

$$\hat{\sigma}_1^2=1-\hat{h}_1^2=1-0.8328=0.1672.$$

§8.4 方差最大的正交旋转

因子分析的目的不仅是求出公共因子,更主要的是应该知道每个公共因子的实际意义.但由于在§8.3中介绍的估计方法所求出的公因子解,其初始因子载荷矩阵并不满足"简单结构准则",即各个公共因子的典型代表变量不很突出,因而容易使公共因子的实际意义含糊不清,不利于对因子进行解释.为此必须对因子载荷矩阵施行旋转变换,使得各因子载荷矩阵的每一列各元素的平方按列向 0 或 1 两极转化,达到其结构简化的目的.

一、理论依据

设因子模型：$X=AF+\varepsilon$，$F=(F_1,\cdots,F_m)'$为公因子向量，对 F 施行正交变换：令 $Z=\Gamma'F$（Γ 为任一 m 阶正交矩阵），则

$$X = A\Gamma Z + \varepsilon, \tag{8.4.1}$$

且

$$\left.\begin{aligned} &\mathrm{D}(Z) = \mathrm{D}(\Gamma'F) = \Gamma'\mathrm{D}(F)\Gamma = I_m,\\ &\mathrm{COV}(Z,\varepsilon) = \mathrm{COV}(\Gamma'F,\varepsilon) = \Gamma'\mathrm{COV}(F,\varepsilon) = O,\\ &\mathrm{D}(X) = \mathrm{D}(A\Gamma Z) + \mathrm{D}(\varepsilon) = A\Gamma\mathrm{D}(Z)\Gamma'A' + D\\ &\qquad = AA' + D. \end{aligned}\right\} \tag{8.4.2}$$

(8.4.1)和(8.4.2)式说明，若 F 是正交因子模型的公因子向量，则对任一正交矩阵 Γ，$\Gamma'F \xrightarrow{\text{def}} Z$ 也是公因子向量. 相应的 $A\Gamma$ 是公因子 Z 的因子载荷矩阵. 利用此性质，在因子分析的实际计算中，当求得初始因子载荷矩阵 A 后，反复右乘正交矩阵 Γ，使 $A\Gamma$ 具有更明显的实际意义. 这种变换载荷矩阵的方法，称为因子轴的**正交旋转**.

二、因子载荷方差

设因子模型 $X=AF+\varepsilon$，$A=(a_{ij})_{p\times m}$为公因子向量 F 的因子载荷矩阵，

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1m}\\ \vdots & & \vdots & & \vdots\\ a_{i1} & \cdots & a_{ij} & \cdots & a_{im}\\ \vdots & & \vdots & & \vdots\\ a_{p1} & \cdots & a_{pj} & \cdots & a_{pm} \end{bmatrix},$$

而

$$h_i^2 = \sum_{t=1}^{m} a_{it}^2 \quad (i=1,\cdots,p)$$

为变量 X_i 的共同度.

如果 A 的每一列(即因子载荷向量)数值越分散，相应的因子载荷向量的方差越大. 为消除由于 a_{ij} 符号不同的影响及各变量对公共因子依赖程度不同的影响，令

$$d_{ij}^2=\frac{a_{ij}^2}{h_i^2}\quad(i=1,\cdots,p;j=1,\cdots,m),$$

将第 j 列的 p 个数据 $d_{1j}^2,\cdots,d_{pj}^2$ 的方差定义为

$$V_j=\sum_{i=1}^{p}(d_{ij}^2-\bar{d}_j)^2/p=\frac{1}{p^2}\left[p\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4}-\left(\sum_{i=1}^{p}\frac{a_{ij}^2}{h_i^2}\right)^2\right],$$

其中 $\bar{d}_j=\frac{1}{p}\sum_{i=1}^{p}d_{ij}^2$ $(j=1,\cdots,m)$. 则因子载荷矩阵 A 的方差为

$$V=\sum_{j=1}^{m}V_j=\frac{1}{p^2}\left\{\sum_{j=1}^{m}\left[p\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4}-\left(\sum_{i=1}^{p}\frac{a_{ij}^2}{h_i^2}\right)^2\right]\right\}.$$

若 V_j 值越大，A 的第 j 个因子载荷向量数值越分散，如果载荷值或是趋于 1 或是趋于 0，这时相应的公共因子 F_j 即具有简化结构，因而我们希望因子载荷矩阵 A 的方差尽可能大.

三、方差最大的正交旋转

设 $m=2$，因子载荷矩阵为

$$A=\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\\\vdots&\vdots\\a_{p1}&a_{p2}\end{bmatrix}.$$

取正交矩阵 $\Gamma=\begin{bmatrix}\cos\varphi&-\sin\varphi\\\sin\varphi&\cos\varphi\end{bmatrix}$，则

$$B=A\Gamma=\begin{bmatrix}a_{11}\cos\varphi+a_{12}\sin\varphi&-a_{11}\sin\varphi+a_{12}\cos\varphi\\a_{21}\cos\varphi+a_{22}\sin\varphi&-a_{21}\sin\varphi+a_{22}\cos\varphi\\\vdots&\vdots\\a_{p1}\cos\varphi+a_{p2}\sin\varphi&-a_{p1}\sin\varphi+a_{p2}\cos\varphi\end{bmatrix}$$

$$\xlongequal{\text{def}}\begin{bmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\\\vdots&\vdots\\b_{p1}&b_{p2}\end{bmatrix}$$

是 $Z=\Gamma' F$ 的因子载荷矩阵，这相当于将由 F_1,F_2 确定的因子平面旋转一个角度 φ. 此时

$$V_t=\frac{1}{p^2}\left[p\sum_{i=1}^{p}\frac{b_{it}^4}{h_i^4}-\left(\sum_{i=1}^{p}\frac{b_{it}^2}{h_i^2}\right)^2\right]\quad(t=1,2),$$

令　$$\frac{\partial V}{\partial \varphi}=\frac{\partial}{\partial \varphi}(V_1+V_2)=0.$$

经整理后可知，为使 $\frac{\partial V}{\partial \varphi}=0$，$\varphi$ 应满足以下关系式：

$$\tan 4\varphi=\frac{d-2\alpha\beta/p}{c-(\alpha^2-\beta^2)/p}. \tag{8.4.3}$$

若记

$$\mu_j=\left(\frac{a_{j1}}{h_j}\right)^2-\left(\frac{a_{j2}}{h_j}\right)^2,\quad \nu_j=2\frac{a_{j1}a_{j2}}{h_j^2}\quad (j=1,2,\cdots,p),$$

则

$$\alpha=\sum_{j=1}^{p}\mu_j,\quad \beta=\sum_{j=1}^{p}\nu_j,$$

$$c=\sum_{j=1}^{p}(\mu_j^2-\nu_j^2),\quad d=2\sum_{j=1}^{p}\mu_j\nu_j.$$

当 $m>2$ 时，可逐次对每两个因子 F_i，$F_j(i\neq j)$ 进行以上旋转. 选择正交旋转的角度 φ_{ij} 使满足(8.4.3)式，即使这两个因子的方差之和达最大. m 个因子的全部配对旋转，共需旋转 C_m^2 次，全部旋转完毕即算一次循环(或一轮). 经第一轮旋转后计算所得的因子载荷方差 $V_{(1)}$，此时不能认为 $V_{(1)}$ 就是最大方差，还需从旋转后的载荷矩阵出发，再进行第二轮、第三轮旋转，直到 V 不能再增大为止.

例 8.4.1（例 8.3.1 的继续）　在例 8.3.1 中，考虑对因子载荷矩阵作方差最大的正交旋转，并由旋转后的因子载荷矩阵解释公共因子的含义.

解　使用 SAS/STAT 软件中 FACTOR 过程的选项要求对因子载荷矩阵进行方差最大的正交旋转，并指定公共因子个数 $m=3$.

在此例中，因子载荷矩阵 A 经五轮正交旋转后，每轮因子载荷矩阵的方差 $V_{(i)}$ 不断增加，即

$$V_{(1)}=0.173828\leqslant 0.390722\leqslant 0.391048$$
$$\leqslant 0.391049\leqslant 0.391049=V_{(5)};$$

且当 $|V_{(5)}-V_{(4)}|\leqslant 0.000001$ 时停止旋转，此时

$$V_{(5)}=0.391049$$

为 V 的最大值. 这时经方差最大正交旋转后的因子载荷矩阵 A^* 见

输出 8.4.1.

输出 8.4.1 方差最大正交旋转后的因子载荷矩阵

Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	-0.73239	0.65030	0.20179
2	0.64257	0.75814	-0.11102
3	0.22519	-0.04835	0.97311

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
X1	0.89711	-0.03957	-0.16271
X2	-0.18571	0.12496	0.96663
X3	-0.74626	0.55109	0.02317
X4	-0.70031	0.65964	0.01507
X5	0.92360	-0.18917	-0.17408
X6	-0.31567	0.91995	-0.01923
X7	0.02655	-0.93712	-0.25950

Variance explained by each factor

FACTOR1	FACTOR2	FACTOR3
2.840008	2.516292	1.059629

旋转后的因子载荷矩阵的总方差 $V_{(5)}$ 约为初始因子载荷矩阵总方差 $V_{(1)}$ 的 2.25 倍，A^* 各列的载荷明显向 0 或 1 两极方向分化，这就大大有利于对公共因子进行解释.

(1) 方差最大正交旋转后第一公共因子中各变量的因子载荷有正有负，正载荷主要是 X_5(Na/K 系数)和 X_1(矿化度)，它们是钠盐形成的显示. 负载荷主要是 X_3($K \cdot 10^3/\Sigma$ 盐)和 X_4($K \cdot 10^3/Cl$)，它们表示了钾盐形成的必要物质来源.

(2) 方差最大正交旋转后第二公共因子中各变量的因子载荷最大值为 X_6($Mg \cdot 10^2/Cl$)，其载荷为 0.9199，而 X_3，X_4 降为次要地位. 这说明第二公共因子是钾盐形成条件的显示——$Mg \cdot 10^2/Cl$，正如我们通常所知道的，它反映了钾盐沉淀环境和钾矿物中的成分之一的存在，而 $K \cdot 10^3/\Sigma$ 盐(X_3)和 $K \cdot 10^3/Cl$(X_4)，则是钾盐形成的不可缺少的物质条件.

(3) 在第三公共因子中，主要起作用的变量是 X_2($Br \cdot 10^3/Cl$)，它是钾盐或钾矿化的一个环境标志.

由上面的分析看出，影响钾盐成矿的主要特征系数是 $K \cdot 10^3/\Sigma$ 盐、$K \cdot 10^3/Cl$ 和 $Mg \cdot 10^2/Cl$，以及较次要的 $Br \cdot 10^3/Cl$，而影响决定钠盐形成的是矿化度与 Na/K 系数.

前面讨论的因子模型(8.2.1)中公共因子是互不相关的(或正交的)，进行因子旋转时也是正交旋转，在正交旋转过程中始终保持公共因子之间互不相关的特点. 因此经正交旋转后所得到的解仍是正交因子解. 如果我们把 m 个公共因子看作坐标轴，有 m 个坐标的点 $(a_{i1}, a_{i2}, \cdots, a_{ip})$ 代表因子空间中第 i 个点的位置，并假设可把 p 个变量分为几个没有重叠的类，旋转到一个简单结构的正交旋转，就与坐标轴的刚性旋转是对应的，它使得这些坐标轴在旋转之后尽可能地通过这几个类.

在实际问题中对变量(或样品)产生影响的公共因子，它们之间往往有相互的关系，即在因子模型(8.2.1)中 $D(F)=V \neq I_m$(公共因子之间是相关的)，我们称这种因子相关的因子模型为**斜交因子模型**，而相关的公共因子称为**斜交公因子**(简称**斜交因子**). 在大量的实际问题中，一组相关变量 $X_1, X_2, \cdots, X_p$ 满足的因子模型一般都是斜交因子模型，而正交因子模型只是一种特例，或者作为斜交因子模型的一种近似. 而旋转到一个简单结构的斜交旋转，则对应于一个非刚性的旋转坐标系，它使得旋转后的坐标轴不再垂直，但(近似地)通过这几个类. 关于斜交因子的概念及求解方法请见参考文献[14].

§8.5　因子得分

以上几节我们已经讨论了如何从样本协方差阵 S 或相关阵 R 出发，来获得公共因子和因子载荷矩阵(或经正交旋转后的因子载荷矩阵)，并给出公共因子的解释. 但有时要求把公共因子表示成变量的线性组合，或反过来对每一个样品计算公共因子的估计值，即所谓的**因子得分**. 因子得分可用于模型的诊断，也可作为进一步分析的原始数据. 但请注意，因子得分的计算并不是通常意义下的参数估计，而是对不可观测的随机向量 F 取值的估计.

下面介绍估计因子得分的几种常用方法.

一、加权最小二乘法

设 X 满足正交因子模型(不妨设 $\mu=0$)

$$X = AF + \varepsilon.$$

假定因子载荷矩阵 A 和特殊因子方差已知,而把特殊因子 ε 看作误差. 因 $\mathrm{Var}(\varepsilon_i)=\sigma_i^2$ $(i=1,\cdots,p)$一般不相等. 于是我们用加权最小二乘法估计公因子 F 的值.

用误差方差的倒数作为权重的误差平方和:

$$\sum_{i=1}^{p}\frac{\varepsilon_i^2}{\sigma_i^2} = \varepsilon' D^{-1}\varepsilon = (X-AF)'D^{-1}(X-AF) \xlongequal{\text{def}} \varphi(F). \tag{8.5.1}$$

在(8.5.1)式中, A,D 已知, X 为可观测的,其值也是已知的,求 F 的估计值 $\hat{F}$,使 $\varphi(\hat{F})=\min\varphi(F)$.

由 $\dfrac{\partial\varphi(F)}{\partial F}=0$,可得到 F 的估计值:

$$\hat{F} = (A'D^{-1}A)^{-1}A'D^{-1}X, \tag{8.5.2}$$

这就是因子得分的加权最小二乘估计.

若假定 $X\sim N_p(AF,D)$, X 的似然函数的对数为

$$L(F) = -\frac{1}{2}(X-AF)'D^{-1}(X-AF) - \frac{1}{2}\ln|2\pi D|,$$

由此得到 F 的最大似然估计仍为(8.5.2)式,这个估计也称为**巴特莱特因子得分**.

实际问题中, A,D 未知,自然的作法是将它们的某个估计代入(8.5.2)式. 对于样品 $X_{(i)}$,其因子得分为

$$\hat{F}_{(i)} = (A'D^{-1}A)^{-1}A'D^{-1}X_{(i)} \quad (i=1,\cdots,n).$$

如果我们用主成分法估计因子载荷矩阵,那么在计算因子得分的估计时,通常用不加权的最小二乘法,即用极小化

$$\sum_{i=1}^{p}\varepsilon_i^2 = \varepsilon'\varepsilon = (X-AF)'(X-AF) \xlongequal{\text{def}} \phi(F)$$

来估计 F. 由

$$\frac{\partial\phi(F)}{\partial F} = 2A'(X-AF) = 0,$$

可得到 F 的估计值：$\hat{F}=(A'A)^{-1}A'X$，这就是因子得分的最小二乘估计. 对于样品 $X_{(i)}$，其因子得分为

$$\hat{F}_{(i)} = (A'A)^{-1}A'X_{(i)} \quad (i = 1,\cdots,n).$$

对照第七章介绍的样本主成分，可以看到，因子得分 $\hat{F}_{(i)}=(\widetilde{f}_{i1},\cdots,\widetilde{f}_{im})'$ 和样本主成分得分 $Z_{(i)}=(z_{i1},\cdots,z_{im})'$ 仅相差一个常数：

$$\widetilde{f}_{ij} = z_{ij}/\sqrt{\lambda_j} \quad (i = 1,\cdots,n; j = 1,\cdots,m),$$

其中 λ_j 是相关阵 R 的第 j 大特征值.

二、回归法

在因子模型中，我们也可以反过来将公因子表示为变量的线性组合，即用

$$F_i = \beta_{i1}X_1 + \cdots + \beta_{ip}X_p \quad (i = 1,\cdots,m) \qquad (8.5.3)$$

来计算每个样品的公因子得分. (8.5.3)式称为因子得分函数. 以下用回归法给出(8.5.3)式中组合系数 β_{ij} 的估计.

假设变量 X 为标准化变量，公因子 F 也已标准化. 在最小二乘法意义下对因子得分函数进行估计，并记建立的公因子 F 对变量 X 满足的回归方程为

$$F_j = b_{j1}X_1 + \cdots + b_{jp}X_p + \varepsilon_j \quad (j = 1,\cdots,m). \qquad (8.5.4)$$

下面来估计回归方程(8.5.4)中的回归系数 $b_{j1},b_{j2},\cdots,b_{jp}$.

上述问题虽是多对多的回归问题，但 $F_1,\cdots,F_m$ 为不可观测的，即因子得分 F_j 的值是待估的. 现我们仅知道由样本可得到因子载荷矩阵 $A=(a_{ij})_{p\times m}$，对因子 F_j，由因子载荷的意义，有

$$\begin{aligned} a_{ij} &= \mathrm{E}(X_iF_j) = \mathrm{E}[X_i(b_{j1}X_1 + \cdots + b_{jp}X_p)] \\ &= b_{j1}r_{i1} + \cdots + b_{jp}r_{ip} \quad (i = 1,2,\cdots,p), \end{aligned}$$

即

$$\begin{cases} b_{j1}r_{11} + \cdots + b_{jp}r_{1p} = a_{1j}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{j1}r_{p1} + \cdots + b_{jp}r_{pp} = a_{pj}, \end{cases} \qquad (8.5.5)$$

以及

$$(8.5.5)\text{ 式} \Longleftrightarrow Rb_{(j)} = a_j \quad (R\text{ 为相关阵}),$$

其中 $b_{(j)}=(b_{j1},b_{j2},\cdots,b_{jp})'$，$a_j=(a_{1j},a_{2j},\cdots,a_{pj})'$，故

$$b_{(j)}=R^{-1}a_j \quad (j=1,2,\cdots,m).$$

记

$$B=\begin{bmatrix} b'_{(1)} \\ \vdots \\ b'_{(m)} \end{bmatrix}=\begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mp} \end{bmatrix},$$

则有

$$B=\begin{bmatrix} (R^{-1}a_1)' \\ \vdots \\ (R^{-1}a_m)' \end{bmatrix}=\begin{bmatrix} a'_1 \\ \vdots \\ a'_m \end{bmatrix}R^{-1}=A'R^{-1},$$

于是利用回归方法所建立的公因子 F 对变量 X 的回归方程为

$$\hat{F}=\begin{bmatrix} \hat{F}_1 \\ \vdots \\ \hat{F}_m \end{bmatrix}=\begin{bmatrix} b'_{(1)}X \\ \vdots \\ b'_{(m)}X \end{bmatrix}=BX=A'R^{-1}X, \qquad (8.5.6)$$

其中 R 为 $p\times p$ 样本相关阵. 由样本值计算相关阵 R，并估计因子载荷矩阵 A(为 $p\times m$ 矩阵)，代入(8.5.6)式，即得因子得分函数 $\hat{F}$ 的计算公式.

此方法是由汤普森(Thompson)(1939)提出来的，所得因子得分在文献上常称为**汤普森因子得分**.

此估计也可以从贝叶斯统计的思想来求得. 在因子模型中，假设 X，F 和 ε 服从正态分布. 若 F 有一先验分布为 $N_m(0,I_m)$，当给定 F 时，X 的条件分布为 $N_p(AF,D)$，用贝叶斯统计的典型手法可以求得，当给定 X 时 F 的条件分布(即后验分布)仍为正态分布，其均值为

$$\mathrm{E}(F|X)=A'(AA'+D)^{-1}X.$$

记 $B=A'(AA'+D)^{-1}$为因子 F 对 X 的回归系数. 因子得分函数有表达式：

$$\hat{F}=A'(AA'+D)^{-1}X=A'\Sigma^{-1}X.$$

当 $X=X_{(j)}$ $(j=1,2,\cdots,n)$时得第 j 个观测的因子得分

$$\hat{F}_{(j)}=BX_{(j)} \quad (j=1,2,\cdots,n).$$

用样本值可以计算样本协方差阵

$$S=\frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)}-\overline{X})(X_{(i)}-\overline{X})',$$

并以它作为 Σ 的估计，因子载荷矩阵的估计仍记为 A. 于是因子得分函数的计算公式为 $\hat{F}=A'S^{-1}X$，当变量 X 为标准化变量时，样本协方差阵 S 就是样本相关阵 R，故有 $\hat{F}=A'R^{-1}X$.

三、两种估计的比较

分 3 种情形进行比较.

(1) 记

$$\hat{F}(1)=(A'D^{-1}A)^{-1}A'D^{-1}X,$$
$$\hat{F}(2)=A'(AA'+D)^{-1}X,$$

利用矩阵间的关系式(见本章习题第 8-3 题)：

$$A'(AA'+D)^{-1}=(I_m+A'D^{-1}A)^{-1}A'D^{-1},$$

故有

$$\hat{F}(2)=(I_m+A'D^{-1}A)^{-1}A'D^{-1}X,$$

显然

$$\begin{aligned}\hat{F}(1)&=(A'D^{-1}A)^{-1}(I_m+A'D^{-1}A)\hat{F}(2)\\&=[I_m+(A'D^{-1}A)^{-1}]\hat{F}(2),\end{aligned}$$

式中 A,D 满足约束条件：$A'D^{-1}A=$对角矩阵，当对角元素近似等于 0 时，两种估计得出的因子得分几乎相等.

(2) 因

$$\mathrm{E}(\hat{F}(1)|F)=F,$$
$$\mathrm{E}(\hat{F}(2)|F)=(I_m+A'D^{-1}A)^{-1}A'D^{-1}AF,$$

这表明第一种估计是无偏的，而汤普森因子得分(回归估计)是有偏的.

(3) 因

$$\mathrm{E}[(\hat{F}(1)-F)(\hat{F}(1)-F)']=(A'D^{-1}A)^{-1},$$
$$\mathrm{E}[(\hat{F}(2)-F)(\hat{F}(2)-F)']=(I_m+A'D^{-1}A)^{-1},$$

这表示第二种估计(汤普森因子得分)有较小的平均预报误差.

两种估计到底哪一种好，长期以来一直有争论，至今尚未有定

论.

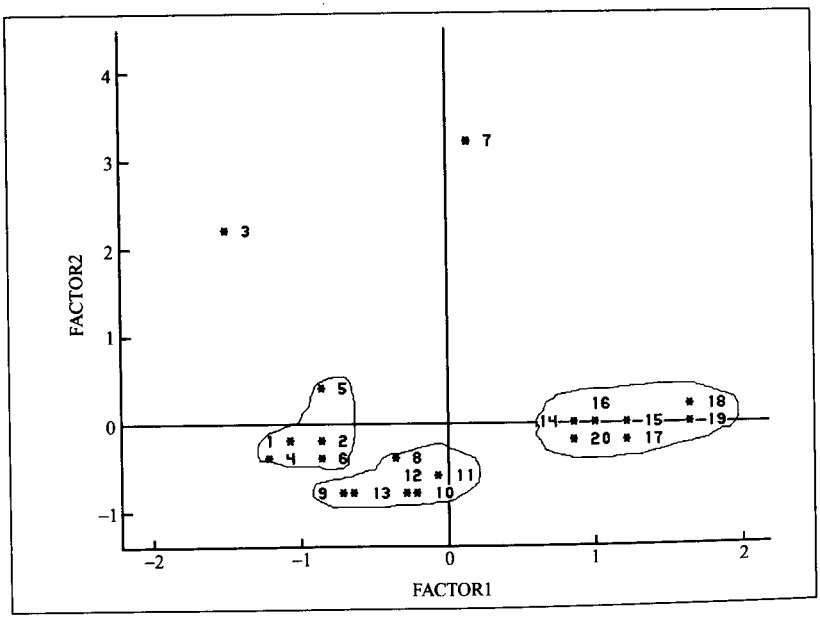
例 8.5.1（例 8.4.1 的继续） 在例 8.4.1 中，用回归法求因子得分函数，计算 20 个样品的因子得分，并绘制第一和第二因子得分的散布图.

解 使用 SAS/STAT 软件中 FACTOR 过程及选项进行因子分析并输出因子得分系数. 由此得出因子得分函数：

$$\begin{aligned}F_1 &= 0.42452X_1 + 0.07959X_2 - 0.23210X_3 - 0.18099X_4\\ &\quad + 0.39673X_5 + 0.07977X_6 - 0.27297X_7,\\ F_2 &= 0.22999X_1 - 0.06366X_2 + 0.10946X_3 + 0.18330X_4\\ &\quad + 0.15401X_5 + 0.43450X_6 - 0.49645X_7,\\ F_3 &= -0.03589X_1 + 0.97545X_2 - 0.13310X_3 - 0.14437X_4\\ &\quad - 0.03085X_5 - 0.14683X_6 - 0.18622X_7.\end{aligned}$$

把 20 个样品的观测值逐个代入以上因子得分函数，即得样品的因子得分值. 第一和第二因子得分的散布图见输出 8.5.1.

输出 8.5.1 第一、第二因子得分的散布图



输出 8.5.1 是根据样品的因子得分，取 F_1 和 F_2 两个公共因子为坐标轴而绘制的因子得分图. 可见 20 个盐泉除 3 号和 7 号外可分为三类，第一类为第 14～20 号盐泉，它们是以第一公共因子轴 F_1 上得分高，第二公共因子轴 F_2 上得分低为特征；第二类为 8～13 号盐泉，它们是以 F_1 上得分小，F_2 上得分为较大的负值为特征；第三类为 1、2、4～6 号盐泉，它们是以 F_1 上得分为较大负值为特征. 这三类表示三种不同的盐泉.

§8.6　Q 型因子分析

根据研究对象的不同，因子分析可分为 R 型和 Q 型两种. 当研究对象是变量时，属于 R 型因子分析. 以上的讨论都是以变量作为研究对象，在样品的基础上研究变量之间的相关关系. 而变量之间的相互关系表现在原始数据矩阵的列与列之间，由相关阵或协方差阵出发，研究变量间的相关关系.

当研究对象是样品时，属于 Q 型因子分析，它是在变量的基础上研究样品之间的相互关系. 而样品之间的相互关系则表现在原始数据矩阵的行与行之间. 因此进行 Q 型因子分析时，只需把在 R 型因子分析中的变量和样品的作用互相调换，其余处理方法是一致的.

在进行 R 型因子分析时，变量间的相互关系我们常用相关系数来描述. 在进行 Q 型分析时，应当选择样品间合适的相似性度量，一般用相似系数(即夹角余弦)作为样品间相似性的度量.

设 $X_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})'$，$X_{(j)}=(x_{j1},x_{j2},\cdots,x_{jp})'$ 是两个样品向量，它们间夹角的余弦

$$\cos\langle X_{(i)},X_{(j)}\rangle=\sum_{t=1}^{p}x_{it}x_{jt}\Big/\sqrt{\sum_{t=1}^{p}x_{it}^2\sum_{t=1}^{p}x_{jt}^2}\quad(i,j=1,2,\cdots,n),\tag{8.6.1}$$

两样品向量间的夹角余弦反映了这两个样品中各变量的观测值之间的比例关系，称为**相似系数**. 相似系数矩阵为

$$Q=(q_{ij})_{n\times n},$$

其中 $q_{ij}=\cos\langle X_{(i)},X_{(j)}\rangle$. Q 是 n 阶方阵，进行Q型因子分析时，需要计算 Q 的特征值和特征向量. 一般 n 较大(n 比起变量个数 p 总是大很多)时，直接计算 Q 的特征值和特征向量是比较困难的(如超出计算机内存，或花费太多的机时等). 解决的方法是利用线性代数的结论：设 Z 为 $n\times p$ 矩阵，则矩阵 $Z'Z$ 与 ZZ' 有相同的非零特征值：$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_q\geqslant 0$ $(q\leqslant p<n)$，其对应的特征向量也有一定的关系. 由此得出一种双重型的因子分析方法——对应分析方法，这是第九章将介绍的统计方法.

例 8.6.1 试对表 8.1 中 20 个盐泉的水化学特征系数作Q型因子分析.

解 此例中 $n=20$，$p=7$. 首先计算样品相似矩阵 Q，然后调用 SAS/STAT 软件中 FACTOR 过程进行Q型因子分析，并对Q型因子载荷矩阵进行方差最大的正交旋转.

部分结果见输出 8.6.1 至输出 8.6.3. 相似矩阵 Q 的特征值及相应贡献率见输出 8.6.1，由此输出可见，选三个主因子，它们提供的信息已占总信息量的 99.8%，几乎完全反映了 20 个样品的信息.

输出 8.6.1 相似矩阵 Q 的前 5 个特征值和相应贡献率

Eigenvalues of the Correlation Matrix: Total = 20 Average = 1

	1	2	3	4	5
Eigenvalue	15.0897	2.8032	2.0741	0.0325	0.0003
Difference	12.2865	0.7291	2.0416	0.0322	0.0001
Proportion	0.7545	0.1402	0.1037	0.0016	0.0000
Cumulative	0.7545	0.8946	0.9983	1.0000	1.0000

因子载荷矩阵见输出 8.6.2. 显见各正交因子的载荷不满足“结构简化”的要求，需进一步对载荷矩阵进行方差最大的正交旋转. 经此旋转后的因子载荷矩阵见输出 8.6.3.

由输出 8.6.3 可得以下几个结论：

(1) 方差最大的正交旋转后第一主因子中 8 至 13 号样品的因子载荷都在 0.9 以上；其余样品的载荷都较小，故这 6 个样品为一类，这一类中载荷最大值 0.93691 所对应的 13 号样品可作为典型代表.

(2) 第二主因子中 14 至 20 号样品的因子载荷都为较大的值，

其余样品除个别的以外载荷都较小，故这 7 个样品为一类. 这一类中载荷最大值 0.91908 所对应的 14 号样品可作为典型代表.

输出 8.6.2　因子载荷矩阵 $A(m=3)$

Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
Q1	0.84851	0.51566	0.09659
Q2	0.84499	0.31041	-0.43379
Q3	0.44151	0.88695	0.13488
Q4	0.79656	0.52415	0.29434
Q5	0.80955	0.34396	-0.47524
Q6	0.98015	0.18344	0.06484
Q7	0.54353	0.82619	0.05429
Q8	0.89034	-0.17885	0.41860
Q9	0.91332	-0.13279	0.38496
Q10	0.91546	-0.21930	0.33735
Q11	0.89260	-0.23056	0.38731
Q12	0.92306	-0.23256	0.30632
Q13	0.88514	-0.20829	0.41594
Q14	0.88691	-0.17384	-0.42781
Q15	0.90222	-0.18472	-0.38951
Q16	0.88770	-0.17583	-0.42535
Q17	0.91067	-0.18997	-0.36667
Q18	0.96790	-0.23772	-0.08070
Q19	0.96373	-0.23196	-0.13135
Q20	0.96647	-0.23445	-0.10395

输出 8.6.3　方差最大的正交旋转后的因子载荷矩阵 $A^*(m=3)$

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
Q1	0.44716	0.38337	0.80516
Q2	0.16963	0.81275	0.55605
Q3	0.06946	0.02914	0.99706
Q4	0.53916	0.20087	0.81538
Q5	0.10695	0.81514	0.56889
Q6	0.63163	0.54915	0.54592
Q7	0.10616	0.16535	0.97075
Q8	0.93172	0.29004	0.21839
Q9	0.90871	0.32195	0.26567
Q10	0.90951	0.37386	0.18165
Q11	0.93099	0.32382	0.16823
Q12	0.89892	0.40421	0.16887
Q13	0.93691	0.29379	0.18907
Q14	0.37271	0.91908	0.12739
Q15	0.41189	0.90219	0.12744
Q16	0.37556	0.91809	0.12614
Q17	0.43435	0.89147	0.12835
Q18	0.67691	0.72292	0.13793
Q19	0.63889	0.75708	0.13601
Q20	0.65956	0.73880	0.13781

(3) 第三主因子中 1 至 7 号样品的因子载荷都在 0.545 以上，其余样品的载荷都很小，故这 7 个样品为一类. 这一类中载荷最大值 0.99706 所对应的 3 号样品可作为典型代表.

(4) 由表 8.1 的原始数据，比较 3,13,14 号这三个典型盐泉的特征系数，3 号盐泉位于钾盐矿区，其特征系数中 $X_3(K\cdot 10^3/Cl)$，$X_4(K\cdot 10^3/\Sigma 盐)$的系数高，而 $X_5(Na/K)$系数低，属地下水溶滤钾盐层后而形成的盐泉. 14 号盐泉属钠盐矿区，钠盐中不具钾矿化，其特征系数中 $X_3(K\cdot 10^3/Cl)$，$X_4(K\cdot 10^3/\Sigma 盐)$的系数甚低，而 $X_5(Na/K)$的系数则很高，属于溶滤钠盐而形成的盐泉. 13 号盐泉的特征系数在 3 与 14 号之间，属过渡类型的盐泉.

为了更直观地对 20 个盐泉样品分类，可用第一、第二主因子的载荷值在平面上作图，然后进行分类.

习 题 八

8-1 设标准化变量 X_1,X_2,X_3 的协方差阵(即相关阵)为

$$R=\begin{bmatrix}1.00 & 0.63 & 0.45\\ 0.63 & 1.00 & 0.35\\ 0.45 & 0.35 & 1.00\end{bmatrix},$$

试求 $m=1$ 的正交因子模型.

8-2 已知题 8-1 中 R 的特征值和特征向量分别为

$\lambda_1=1.9633$，$l_1=(0.6250,0.5932,0.5075)'$，

$\lambda_2=0.6795$，$l_2=(-0.2186,-0.4911,0.8432)'$，

$\lambda_3=0.3672$，$l_3=(0.7494,-0.6379,-0.1772)'$.

(1) 取公共因子个数 $m=1$ 时，求因子模型的主成分解，并计算误差平方和 $Q(1)$；

(2) 取公共因子个数 $m=2$ 时，求因子模型的主成分解，并计算误差平方和 $Q(2)$；

(3) 试求误差平方和 $Q(m)<0.1$ 的主成分解.

8-3 验证下列矩阵关系式(A 为 $p\times m$ 矩阵)

(1) $(I+A'D^{-1}A)^{-1}A'D^{-1}A=I-(I+A'D^{-1}A)^{-1}$；

(2) $(AA'+D)^{-1}=D^{-1}-D^{-1}A(I+A'D^{-1}A)^{-1}A^{-1}D^{-1}$；

(3) $A'(AA'+D)^{-1}=(I_m+A'D^{-1}A)^{-1}A'D^{-1}$.

提示：考虑分块矩阵 $\begin{bmatrix}D & -A\\ A' & I_m\end{bmatrix}\begin{matrix}p\\ m\end{matrix}$ 的逆.

8-4　证明公共因子个数为 m 的主成分解，其误差平方和 $Q(m)$ 满足以下不等式

$$Q(m)=\sum_{i=1}^{p}\sum_{j=1}^{p}\varepsilon_{ij}^2\leqslant\sum_{j=m+1}^{p}\lambda_j^2,$$

其中 $\varepsilon=S-(AA'+D)=(\varepsilon_{ij})$，$A,D$ 是因子模型的主成分估计.

8-5　试比较主成分分析和因子分析的相同之处与不同点.

8-6　我国山区某大型化工厂，在厂区及邻近地区挑选有代表性的 8 个大气取样点. 每日 4 次同时抽取大气样品，测定其中包含的 6 种气体的浓度，前后共 4 天，每个样品每种气体实测 16 次，并计算出每个取样点每种气体的平均浓度(数据如表 8.2). 试用因子分析和主成分方法分析处理表 8.2 的数据.

表 8.2　大气污染数据

取样点 \ 气体名 平均浓度	氯	硫化氢	SO_2	C_4	环氧氯丙烷	环己烷
1	0.056	0.084	0.031	0.038	0.0081	0.0220
2	0.049	0.055	0.100	0.110	0.0220	0.0073
3	0.038	0.130	0.079	0.170	0.0580	0.0430
4	0.034	0.095	0.058	0.160	0.2000	0.0290
5	0.084	0.066	0.029	0.320	0.0120	0.0410
6	0.064	0.072	0.100	0.210	0.0280	1.3800
7	0.048	0.089	0.062	0.260	0.0380	0.0360
8	0.069	0.087	0.027	0.050	0.0890	0.0210

8-7　下表列出邓阜仙岩体的部分化学成分，试用此组数据作因子分析.

	SiO_2	TiO_2	FeO	CaO	K_2O
1	75.20	0.140	1.86	0.91	5.21
2	75.15	0.160	2.11	0.74	4.93
3	72.19	0.130	1.52	0.69	4.65
4	72.35	0.130	1.37	0.83	4.87
5	72.74	0.100	1.41	0.72	4.99
6	73.29	0.033	1.07	0.17	3.15
7	73.72	0.033	0.77	0.28	2.78

8-8　在某年级 44 名学生的期末考试中，有的课程采用闭卷，有的课程采用开卷(考试成绩见表 8.3)．试用因子分析方法分析这组数据．

表 8.3　44 名学生闭卷与开卷考试的成绩表

力学(闭) X_1	物理(闭) X_2	代数(开) X_3	分析(开) X_4	统计(开) X_5	力学(闭) X_1	物理(闭) X_2	代数(开) X_3	分析(开) X_4	统计(开) X_5
77	82	67	67	81	63	78	80	70	81
75	73	71	66	81	55	72	63	70	68
63	63	65	70	63	53	61	72	64	73
51	67	65	65	68	59	70	68	62	56
62	60	58	62	70	64	72	60	62	45
52	64	60	63	54	55	67	59	62	44
50	50	64	55	63	65	63	58	56	37
31	55	60	57	73	60	64	56	54	40
44	69	53	53	53	42	69	61	55	45
62	46	61	57	45	31	49	62	63	62
44	61	52	62	46	49	41	61	49	64
12	58	61	63	67	49	53	49	62	47
54	49	56	47	53	54	53	46	59	44
44	56	55	61	36	18	44	50	57	81
46	52	65	50	35	32	45	49	57	64
30	69	50	52	45	46	49	53	59	37
40	27	54	61	61	31	42	48	54	68
36	59	51	45	51	56	40	56	54	35
46	56	57	49	32	45	42	55	56	40
42	60	54	49	33	40	63	53	54	25
23	55	59	53	44	48	48	49	51	37
41	63	49	46	34	46	52	53	41	40

第九章　对应分析方法

§9.1　什么是对应分析方法

对应分析方法是在R型和Q型因子分析基础上发展起来的多元统计分析方法，又称为**R-Q型因子分析**.

因子分析方法是用少数几个公共因子去提取研究对象的绝大部分信息，既减少了因子的数目，又把握住了研究对象的相互关系. 在因子分析中根据研究对象的不同，分为R型和Q型，如果研究变量间的相互关系时则采用R型因子分析；如果研究样品间相互关系时则采用Q型因子分析. 但无论是R型或Q型都未能很好地揭示变量和样品间的双重关系. 另方面当样品容量 n 很大(如 $n>1000$)，进行Q型因子分析时，计算 n 阶方阵的特征值和特征向量对于微型计算机而言，其容量和速度都是难以胜任的. 还有进行数据处理时，为了将数量级相差很大的变量进行比较，常常先对变量作标准化处理，然而这种标准化处理对样品就不好进行了，换言之，这种标准化处理对于变量和样品是非对等的，这给寻找R型和Q型之间的联系带来一定的困难.

针对上述问题，在20世纪70年代初，由法国统计学家Benzecri提出了对应分析方法，这个方法是在因子分析的基础上发展起来的，它对原始数据采用适当的标度方法. 把R型和Q型分析结合起来，同时得到两方面的结果——在同一因子平面上对变量和样品一块进行分类，从而揭示所研究的样品和变量间的内在联系.

对应分析由R型因子分析的结果，可以很容易地得到Q型因子分析的结果，这不仅克服样品量大时作Q型因子分析所带来计算上的困难，且把R型和Q型因子分析统一起来，把样品点和变量点同

时反映到相同的因子轴上，这就便于我们对研究的对象进行解释和推断.

基本思想：由于 R 型因子分析和 Q 型分析都是反映一个整体的不同侧面，因而它们之间一定存在内在的联系. 对应分析就是通过对应变换后的标准化矩阵 Z 将两者有机地结合起来.

具体地说，首先给出变量间的协方差阵 $S_R=Z'Z$ 和样品间的协方差阵 $S_Q=ZZ'$，由于 $Z'Z$ 和 ZZ' 有相同的非零特征值，记为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$，如果 S_R 的特征值 λ_i 对应的标准化特征向量为 v_i，则 S_Q 的特征值 λ_i 对应的标准化特征向量

$$u_i=\frac{1}{\sqrt{\lambda_i}}Zv_i.$$

由此可以很方便地由 R 型因子分析而得到 Q 型因子分析的结果.

由 S_R 的特征值和特征向量即可写出 R 型因子分析的因子载荷矩阵(记为 A_R)和 Q 型因子分析的因子载荷矩阵(记为 A_Q)：

$$A_R=\begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ v_{p1}\sqrt{\lambda_1} & v_{p2}\sqrt{\lambda_2} & \cdots & v_{pm}\sqrt{\lambda_m} \end{bmatrix}$$

$$=(\sqrt{\lambda_1}v_1,\cdots,\sqrt{\lambda_m}v_m),$$

$$A_Q=\begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{n1}\sqrt{\lambda_1} & u_{n2}\sqrt{\lambda_2} & \cdots & u_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

$$=(\sqrt{\lambda_1}u_1,\cdots,\sqrt{\lambda_m}u_m).$$

由于 S_R 和 S_Q 具有相同的非零特征值，而这些特征值又正是各个公共因子的方差，因此可以用相同的因子轴同时表示变量点和样品点，即把变量点和样品点同时反映在具有相同坐标轴的因子平面上，以便对变量点和样品点一起考虑进行分类.

§9.2　对应分析方法的原理

一、对应分析的数据变换方法

设有 n 个样品，每个样品观测 p 个指标，原始数据阵为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

为了消除量纲或数量级的差异，经常对变量进行标准化处理，如标准化变换、极差标准化变换等，这些变换对变量和样品是不对称的. 这种不对称性是导致变量和样品之间关系复杂化的主要原因. 在对应分析中，采用数据的变换方法即可克服这种不对称性(假设所有数据 $x_{ij}>0$，否则对所有数据同加一适当常数，便会满足以上要求). 数据变换方法的具体步骤如下：

(1) 对数据阵先分别按行和列求和，再求总和：

x_{11}	x_{12}	$\cdots$	x_{1p}	$\sum_{k=1}^{p} x_{1k}=X_{1.}$
x_{21}	x_{22}	$\cdots$	x_{2p}	$\sum_{k=1}^{p} x_{2k}=X_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	$\sum_{k=1}^{p} x_{nk}=X_{n.}$
$X_{.1}$	$X_{.2}$	$\cdots$	$X_{.p}$	$\sum_{l=1}^{n}\sum_{k=1}^{p} x_{lk}=X_{..}\overset{\text{def}}{=\!=} T$

其中 $X_{.j}=\sum_{i=1}^{n} x_{ij}\ (j=1,2,\cdots,p)$.

(2) 化数据阵 X 为规格化的“概率”矩阵 P，令

$$P=\frac{1}{T}X\overset{\text{def}}{=\!=}(p_{ij})_{n\times p}, \tag{9.2.1}$$

其中 $p_{ij}=\frac{1}{T}x_{ij}$ $(i=1,\cdots,n;j=1,\cdots,p)$. 不难看出 $0\leqslant p_{ij}\leqslant 1$,且

$$\sum_{i=1}^{n}\sum_{j=1}^{p}p_{ij}=1.$$

因而 p_{ij} 可理解为数据 x_{ij} 出现的"概率",并称 P 为**对应阵**.

类似地可以写出对应阵 P 的行和与列和,并把 P 表示成如下一张列联表:

p_{11}	p_{12}	$\cdots$	p_{1p}	$P_{1\cdot}$
p_{21}	p_{22}	$\cdots$	p_{2p}	$P_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
p_{n1}	p_{n2}	$\cdots$	p_{np}	$P_{n\cdot}$
$P_{\cdot 1}$	$P_{\cdot 2}$	$\cdots$	$P_{\cdot p}$	1

其中 $P_{\cdot j}=\sum_{i=1}^{n}p_{ij}$ 可理解为第 j 个变量的边缘概率$(j=1,\cdots,p)$;

$P_{i\cdot}=\sum_{j=1}^{p}p_{ij}$ 可理解为第 i 个样品的边缘概率$(i=1,2,\cdots,n)$.

记

$$r=\begin{bmatrix}P_{1\cdot}\\ \vdots\\ P_{n\cdot}\end{bmatrix},\quad c=\begin{bmatrix}P_{\cdot 1}\\ \vdots\\ P_{\cdot p}\end{bmatrix},$$

则

$$r=P\mathbf{1}_p,\quad c=P'\mathbf{1}_n, \tag{9.2.2}$$

其中 $\mathbf{1}_p=(1,1,\cdots,1)'$ 为元素全为 1 的 p 维常向量.

(3) 从对应阵 P 出发计算变量的协方差阵(考虑 R 型因子分析),我们把 P 矩阵中的 n 个行作为 p 维空间中 n 个样品点.

① 消除各样品点出现概率大小的影响,称

$$R_i'=\left(\frac{p_{i1}}{P_{i\cdot}},\frac{p_{i2}}{P_{i\cdot}},\cdots,\frac{p_{ip}}{P_{i\cdot}}\right)\quad(i=1,\cdots,n)$$

为样品 i 的形象,或 p 个变量在第 i 个样品上的分布轮廓(row profiles),显然有

$$R_i'=\left(\frac{p_{i1}}{P_{i\cdot}},\frac{p_{i2}}{P_{i\cdot}},\cdots,\frac{p_{ip}}{P_{i\cdot}}\right)=\left(\frac{x_{i1}}{X_{i\cdot}},\frac{x_{i2}}{X_{i\cdot}},\cdots,\frac{x_{ip}}{X_{i\cdot}}\right).$$

研究样品点的相互关系一般用两个样品点的欧氏距离来表示，为消除各变量量纲不同的影响，引入第 k 个和 l 个样品间的加权平方距离公式(或称卡方距离)：

$$D^2(k,l)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{P_{k\cdot}}-\frac{p_{lj}}{P_{l\cdot}}\right)^2\Big/P_{\cdot j}$$

$$=\sum_{j=1}^{p}\left(\frac{p_{kj}}{P_{k\cdot}\sqrt{P_{\cdot j}}}-\frac{p_{lj}}{P_{l\cdot}\sqrt{P_{\cdot j}}}\right)^2. \qquad (9.2.3)$$

② 消除各变量量纲不同的影响，把第 i 个样品点的坐标化为

$$\left(\frac{p_{i1}}{P_{i\cdot}\sqrt{P_{\cdot 1}}},\frac{p_{i2}}{P_{i\cdot}\sqrt{P_{\cdot 2}}},\cdots,\frac{p_{ip}}{P_{i\cdot}\sqrt{P_{\cdot p}}}\right)\quad(i=1,\cdots,n).$$

③ 计算第 j 个变量(即第 j 列)的加权平均值. 以第 i 个样品点的概率 $P_{i\cdot}$ 作为权重来计算第 j 个变量的加权平均值：

$$\sum_{i=1}^{n}\frac{p_{ij}}{P_{i\cdot}\sqrt{P_{\cdot j}}}\cdot P_{i\cdot}=\sqrt{P_{\cdot j}}\quad(j=1,2,\cdots,p).$$

④ 用加权方法计算第 i 个变量与第 j 个变量的协方差

$$a_{ij}=\sum_{\alpha=1}^{n}\left(\frac{p_{\alpha i}}{P_{\alpha\cdot}\sqrt{P_{\cdot i}}}-\sqrt{P_{\cdot i}}\right)\left(\frac{p_{\alpha j}}{P_{\alpha\cdot}\sqrt{P_{\cdot j}}}-\sqrt{P_{\cdot j}}\right)\cdot P_{\alpha\cdot}$$

$$=\sum_{\alpha=1}^{n}\left(\frac{p_{\alpha i}}{\sqrt{P_{\alpha\cdot}P_{\cdot i}}}-\sqrt{P_{\alpha\cdot}P_{\cdot i}}\right)\left(\frac{p_{\alpha j}}{\sqrt{P_{\alpha\cdot}P_{\cdot j}}}-\sqrt{P_{\alpha\cdot}P_{\cdot j}}\right)$$

$$=\sum_{\alpha=1}^{n}\frac{p_{\alpha i}-P_{\alpha\cdot}P_{\cdot i}}{\sqrt{P_{\alpha\cdot}P_{\cdot i}}}\cdot\frac{p_{\alpha j}-P_{\alpha\cdot}P_{\cdot j}}{\sqrt{P_{\alpha\cdot}P_{\cdot j}}}$$

$$\overset{\text{def}}{=\!=}\sum_{\alpha=1}^{n}z_{\alpha i}z_{\alpha j}, \qquad (9.2.4)$$

其中

$$z_{\alpha i}=\frac{p_{\alpha i}-P_{\alpha\cdot}P_{\cdot i}}{\sqrt{P_{\alpha\cdot}P_{\cdot i}}}=\frac{x_{\alpha i}-X_{\alpha\cdot}X_{\cdot i}/T}{\sqrt{X_{\alpha\cdot}X_{\cdot i}}}.$$

令 $Z=(z_{ij})$ 为 $n\times p$ 矩阵，则变量间的协方差阵为

$$S_R=Z'Z=(a_{ij})_{p\times p}.$$

(4) 从 P 出发计算样品间的协方差阵(考虑 Q 型因子分析). 用类似地方法可以得出 n 个样品间的协方差阵 S_Q 为

$$S_Q = ZZ' = (b_{ij})_{n\times n}.$$

(5) 进行数据的对应变换, 令

$$Z = (z_{ij})_{n\times p},$$

其中

$$z_{ij} = \frac{p_{ij} - P_{i\cdot}\, P_{\cdot j}}{\sqrt{P_{i\cdot}\, P_{\cdot j}}} = \frac{x_{ij} - X_{i\cdot}\, X_{\cdot j}/T}{\sqrt{X_{i\cdot}\, X_{\cdot j}}} \quad (i = 1,\cdots,n; j = 1,\cdots,p). \tag{9.2.5}$$

公式(9.2.5)即是我们从同时研究 R 型和 Q 型因子分析的角度导出的数据对应变换公式.

如果把所研究的 p 个变量看成一个属性变量的 p 个类目; 而把 n 个样品看成另一个属性变量的 n 个类目, 这时原始数据阵 X 就可以看成一张由观测得到的频数表或计数表. 首先由双向频数表 X 矩阵得到对应阵 P:

$$P = (p_{ij}), \quad p_{ij} = \frac{1}{T}x_{ij} \quad (i = 1,\cdots,n; j = 1,\cdots,p).$$

设 $n>p$, 且 $\mathrm{rank}(P)=p$. 下面我们从代数学角度由对应阵 P 来导出数据对应变换的公式:

(1) 对 P 中心化, 令

$$\tilde{p}_{ij} = p_{ij} - P_{i\cdot}P_{\cdot j} = p_{ij} - m_{ij}/T,$$

其中 $m_{ij}=\dfrac{X_{i\cdot}X_{\cdot j}}{T}=T\cdot P_{i\cdot}P_{\cdot j}$, 它是假定行与列两个属性变量不相关时在第$(i,j)$单元上的期望频数值.

记 $\tilde{P}=(\tilde{p}_{ij})_{n\times p}$. 由(9.2.2)式可得

$$\tilde{P} = P - rc', \tag{9.2.6}$$

因 $\tilde{P}\mathbf{1}_p=P\mathbf{1}_p-rc'\mathbf{1}_p=r-r=0$, 所以 $\mathrm{rank}(\tilde{P})\leqslant p-1$. 令

$$D_r = \mathrm{diag}(P_{1\cdot},\cdots,P_{n\cdot}), \quad D_c = \mathrm{diag}(P_{\cdot 1},\cdots,P_{\cdot p}). \tag{9.2.7}$$

(2) 对 P 标准化得 Z, 令

$$Z = D_r^{-1/2}\tilde{P}D_c^{-1/2} \xlongequal{\text{def}} (z_{ij})_{n\times p}, \tag{9.2.8}$$

其中

$$z_{ij}=\frac{p_{ij}-P_{i\cdot}P_{\cdot j}}{\sqrt{P_{i\cdot}P_{\cdot j}}}=\frac{x_{ij}-X_{i\cdot}X_{\cdot j}/T}{\sqrt{X_{i\cdot}X_{\cdot j}}}.$$

故经对应变换后所得到的新数据矩阵 Z，可以看成是由对应阵 P 经中心化和标准化后所得到的矩阵.

设用于检验行与列两个属性变量是否不相关的 χ^2 统计量为

$$\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{ij}-m_{ij})^2}{m_{ij}}=\sum_{i=1}^{n}\sum_{j=1}^{p}\chi_{ij}^2, \tag{9.2.9}$$

其中 χ_{ij}^2 表示第(i,j)单元在检验行与列两个属性变量是否不相关时对总 χ^2 统计量的贡献(cellchi2)：

$$\chi_{ij}^2=\frac{(x_{ij}-m_{ij})^2}{m_{ij}}=Tz_{ij}^2.$$

故 $\chi^2=T\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=T\,\mathrm{tr}(Z'Z)=T\,\mathrm{tr}(S_R)\overset{或}{=\!=}T\,\mathrm{tr}(S_Q)$.

二、对应分析的原理与依据

将原始数据阵 X 变换为 Z 矩阵后，则变量点和样品点的协方差阵分别为 $S_R=Z'Z$ 和 $S_Q=ZZ'$. S_R 和 S_Q 这两个矩阵存在明显的简单的对应关系，而且将原始数据 x_{ij} 变换为 z_{ij} 后，z_{ij} 关于 i,j 是对等的，即 z_{ij} 对变量和样品是对等的.

为了进一步研究 R 型与 Q 型因子分析，我们利用矩阵代数的一些结论.

引理 9.2.1　设 $S_R=Z'Z$，$S_Q=ZZ'$，则 S_R 和 S_Q 的非零特征值相同.

引理 9.2.2　若 v 是 $Z'Z$ 相应于特征值 λ 的特征向量，则 $u=Zv$ 是 ZZ' 相应于特征值 λ 的特征向量.

定义 9.2.1（矩阵的奇异值分解）　设 Z 为 $n\times p$ 矩阵，

$$\mathrm{rank}(Z)=m\leqslant\min(n-1,p-1),$$

$Z'Z$ 的非零特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$，令 $d_i=\sqrt{\lambda_i}$ $(i=1,\cdots,m)$，则称 d_i 为 Z 的**奇异值**. 如果存在分解式：

$$Z=U\Lambda V', \tag{9.2.10}$$

其中 U 为 $n\times n$ 正交矩阵，V 为 $p\times p$ 正交矩阵，Λ 为 $n\times p$ 对角矩阵(前 m 个对角元为 $d_1,\cdots,d_m$，其余元素均为 0)，则称分解式 $Z=U\Lambda V'$ 为矩阵 Z 的**奇异值分解**.

记

$$U=(U_1\vdots U_2),\quad V=(V_1\vdots V_2),\quad \Lambda_m=\mathrm{diag}(d_1,\cdots,d_m),$$

其中 U_1 为 $n\times m$ 的列正交矩阵，V_1 为 $p\times m$ 的列正交矩阵，则奇异值分解式(9.2.10)等价于

$$Z=U_1\Lambda_m V_1'. \tag{9.2.11}$$

引理 9.2.3 任意非零矩阵 Z 的奇异值分解必存在.

引理 9.2.3 的证明就是具体求出矩阵 Z 的奇异值分解式(见参考文献[7]). 从证明中可以看出：列正交矩阵 V_1 的 m 个列向量分别是 $Z'Z$ 的非零特征值 $\lambda_1,\cdots,\lambda_m$ 对应的特征向量；而列正交矩阵 U_1 的 m 个列向量分别是 ZZ' 的非零特征值 $\lambda_1,\cdots,\lambda_m$ 对应的特征向量，且 $U_1=ZV_1\Lambda_m^{-1}$.

矩阵代数的这几个结论为我们建立了因子分析中 R 型与 Q 型的关系. 借助以上引理 9.2.1 和引理 9.2.2，我们从 R 型因子出发可以直接得到 Q 型因子分析的结果.

由于 S_R 与 S_Q 有相同的非零特征值，而这些非零特征值又表示各个公共因子所提供的方差，因此变量空间 $\mathbb{R}^p$ 中的第一公共因子、第二公共因子、…，直到第 m 个公共因子，它们与样本空间 $\mathbb{R}^n$ 中对应的各个公共因子在总方差中所占的百分比全部相同.

从几何的意义上看，即 $\mathbb{R}^p$ 中诸样品点与 $\mathbb{R}^p$ 中各因子轴的距离平方和，以及 $\mathbb{R}^n$ 中诸变量点与 $\mathbb{R}^n$ 中相对应的各因子轴的距离平方和是完全相同的. 因此可以把变量点和样品点同时反映在同一因子轴所确定的平面上 (即取同一个坐标系)，根据接近程度，可以对变量点和样品点同时考虑进行分类.

三、对应分析的计算步骤

对应分析的具体计算步骤如下：

(1) 由原始数据阵 X 出发计算对应阵 P 和对应变换后的新数

据阵 Z，计算公式见(9.2.1)和(9.2.5).

(2) 计算行轮廓分布(或行形象分布)，记

$$R=\left(\frac{x_{ij}}{X_{i\cdot}}\right)_{n\times p}=\left(\frac{p_{ij}}{P_{i\cdot}}\right)_{n\times p}=D_r^{-1}P\xlongequal{\text{def}}\begin{bmatrix}R_1'\\ \vdots\\ R_n'\end{bmatrix}.$$

R 矩阵由 X 矩阵(或对应阵 P)的每一行除以行和得到，其目的在于消除行点(即样品点)出现“概率”不同的影响. 比如 i_1,i_2 行表示两篇文章(即两个样品)使用 p 种词汇(即 p 个变量，下表为 $p=6$)的频

	X_1	X_2	X_3	X_4	X_5	X_6	
i_1	80	30	25	15	60	55	$X_{i_1\cdot}=265$
i_2	160	60	50	30	120	110	$X_{i_2\cdot}=530$

数，因第 i_2 篇文章篇幅大，每种词汇出现的频数都是第 i_1 篇的两倍，此时这两个行形象应完全相同：

$$\left(\frac{80}{265},\frac{30}{265},\frac{25}{265},\frac{15}{265},\frac{60}{265},\frac{55}{265}\right)=\left(\frac{160}{530},\frac{60}{530},\frac{50}{530},\frac{30}{530},\frac{120}{530},\frac{110}{530}\right).$$

记 $N(R)=\{R_i,i=1,\cdots,n\}$，$N(R)$表示 n 个行形象组成的 p 维空间的点集，则点集 $N(R)$的重心(每个样品点以 $P_{i\cdot}$ 为权重)为

$$\sum_{i=1}^{n}P_{i\cdot}R_i=\sum_{i=1}^{n}P_{i\cdot}\begin{bmatrix}\frac{p_{i1}}{P_{i\cdot}}\\ \vdots\\ \frac{p_{ip}}{P_{i\cdot}}\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}p_{i1}\\ \vdots\\ \sum_{i=1}^{n}p_{ip}\end{bmatrix}=\begin{bmatrix}P_{\cdot 1}\\ \vdots\\ P_{\cdot p}\end{bmatrix}=c,$$

由(9.2.2)式可知，c 是 p 个列变量的边缘分布.

(3) 计算列轮廓分布(或列形象分布)，记

$$C=\left(\frac{x_{ij}}{X_{\cdot j}}\right)_{n\times p}=\left(\frac{p_{ij}}{P_{\cdot j}}\right)_{n\times p}=PD_c^{-1}\xlongequal{\text{def}}(C_1,\cdots,C_p).$$

C 矩阵由 X 矩阵(或对应阵 P)的每一列除以列和得到，其目的在于消除列点(即变量点)出现“概率”不同的影响.

(4) 计算总惯量和 χ^2 统计量，首先由(9.2.3)的加权平方距离公式可知，第 k 个与第 l 个样品点的 χ^2 距离为

$$D^2(k,l)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{P_{k\cdot}}-\frac{p_{lj}}{P_{l\cdot}}\right)^2\Big/P_{\cdot j}=(R_k-R_l)'D_c^{-1}(R_k-R_l).$$

我们把 n 个样品点(即行点)到重心 c 的加权平方距离的总和定义为行形象点集 $N(R)$的总惯量 Q:

$$\begin{aligned}Q&=\sum_{i=1}^{n}P_{i\cdot}D^2(i,c)=\sum_{i=1}^{n}P_{i\cdot}\sum_{j=1}^{p}\frac{1}{P_{\cdot j}}\left(\frac{p_{ij}}{P_{i\cdot}}-P_{\cdot j}\right)^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{P_{i\cdot}}{P_{\cdot j}}\frac{(p_{ij}-P_{i\cdot}P_{\cdot j})^2}{P_{i\cdot}^2}=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-P_{i\cdot}P_{\cdot j})^2}{P_{i\cdot}P_{\cdot j}}\\&=\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=\chi^2/T,\end{aligned}\tag{9.2.12}$$

其中 χ^2 统计量是检验行点和列点是否互不相关的检验统计量，χ^2 的计算公式见(9.2.9).

(5) 对标准化后的新数据阵 Z 作奇异值分解，由(9.2.11)式知：

$$Z=U_1\Lambda_mV_1',\quad m=\operatorname{rank}(Z)\leqslant\min(n-1,p-1),$$

其中

$$\Lambda_m=\operatorname{diag}(d_1,\cdots,d_m),\quad V_1'V_1=I_m,\quad U_1'U_1=I_m$$

(即 V_1,U_1 分别为 $p\times m$ 和 $n\times m$ 列正交矩阵). 求 Z 的奇异值分解式其实是通过求 $S_R=Z'Z$ 矩阵的特征值和标准化特征向量来得到. 设特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$，相应标准化特征向量为 $v_1,v_2,\cdots,v_m$. 在实际应用中常按累计贡献率

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_l}{\lambda_1+\cdots+\lambda_l+\cdots+\lambda_m}\geqslant 0.80\quad(\text{或 }0.70,\text{或 }0.85)$$

确定所取公共因子个数 $l(l\leqslant m)$，Z 的奇异值 $d_j=\sqrt{\lambda_j}$ $(j=1,\cdots,m)$. 以下我们仍用 m 表示选定的因子个数.

(6) 计算行轮廓的坐标 G 和列轮廓的坐标 F. 令

$$a_i=D_c^{-1/2}v_i,$$

则 $a_i'D_ca_i=1$ $(i=1,\cdots,m)$. R 型因子分析的“因子载荷矩阵”(或列轮廓坐标)为

$$F=(d_1a_1,d_2a_2,\cdots,d_ma_m)=D_c^{-1/2}V_1\Lambda_m$$

$$=\begin{bmatrix} \frac{d_1}{\sqrt{P_{\cdot 1}}}v_{11} & \frac{d_2}{\sqrt{P_{\cdot 1}}}v_{12} & \cdots & \frac{d_m}{\sqrt{P_{\cdot 1}}}v_{1m} \\ \frac{d_1}{\sqrt{P_{\cdot 2}}}v_{21} & \frac{d_2}{\sqrt{P_{\cdot 2}}}v_{22} & \cdots & \frac{d_m}{\sqrt{P_{\cdot 2}}}v_{2m} \\ \vdots & \vdots & & \vdots \\ \frac{d_1}{\sqrt{P_{\cdot p}}}v_{p1} & \frac{d_2}{\sqrt{P_{\cdot p}}}v_{p2} & \cdots & \frac{d_m}{\sqrt{P_{\cdot p}}}v_{pm} \end{bmatrix},$$

其中 $D_c^{-1/2}$ 为 p 阶矩阵，V_1为 $p\times m$ 矩阵.

令 $b_i=D_r^{-1/2}u_i$，则 $b_i' D_r b_i=1$ $(i=1,\cdots,m)$. Q 型因子分析的“因子载荷矩阵”(或行轮廓坐标)为

$$G=(d_1b_1,d_2b_2,\cdots,d_mb_m)=D_r^{-1/2}U_1\Lambda_m$$

$$=\begin{bmatrix} \frac{d_1}{\sqrt{P_{1\cdot}}}u_{11} & \frac{d_2}{\sqrt{P_{1\cdot}}}u_{12} & \cdots & \frac{d_m}{\sqrt{P_{1\cdot}}}u_{1m} \\ \frac{d_1}{\sqrt{P_{2\cdot}}}u_{21} & \frac{d_2}{\sqrt{P_{2\cdot}}}u_{22} & \cdots & \frac{d_m}{\sqrt{P_{2\cdot}}}u_{2m} \\ \vdots & \vdots & & \vdots \\ \frac{d_1}{\sqrt{P_{n\cdot}}}u_{n1} & \frac{d_2}{\sqrt{P_{n\cdot}}}u_{n2} & \cdots & \frac{d_m}{\sqrt{P_{n\cdot}}}u_{nm} \end{bmatrix},$$

其中 $D_r^{-1/2}$ 为 n 阶矩阵，U_1为 $n\times m$ 矩阵. 我们常把 a_i 或 b_i $(i=1,\cdots,m)$称为加权意义下有单位长度的特征向量.

注意　行轮廓的坐标 G 和列轮廓的坐标 F 的定义与 Q 型和 R 型因子载荷矩阵稍有差别. 关于行和列坐标的定义在 SAS/STAT 软件的 CORRESP 过程中给出几种不同的定义供使用者选择，以上给出的是最常用的行和列坐标的定义. G 的前两列包含了数据最优二维表示中的各对行点(样品点)的坐标，而 F 的前两列则包含了数据最优二维表示中的各对列点(变量点)的坐标.

(7) 在相同二维平面上用行轮廓的坐标 G 和列轮廓的坐标 F (取 $m=2$)绘制出点的平面图. 也就是把 n 个行点(样品点)和 p 个列点(变量点)在同一个平面坐标系中点图，对一组行点或一组列点，二维图中的欧氏距离与原始数据中各行(或列)轮廓之间的加权距离是

相对应的.但请注意,对应行轮廓的点与对应列轮廓的点之间没有直接的距离关系.

(8) 求总惯量 Q 和 χ^2 统计量的分解式. 由(9.2.12)和(9.2.9)式可知

$$Q=\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=\mathrm{tr}(Z'Z)=\sum_{i=1}^{m}\lambda_i=\sum_{i=1}^{m}d_i^2, \quad (9.2.13)$$

其中 λ_i $(i=1,\cdots,m)$是 $Z'Z$ 的特征值, $d_i=\sqrt{\lambda_i}$ $(i=1,\cdots,m)$是 Z 的奇异值. (9.2.13)式就给出 Q 的分解式,第 i 个因子$(i=1,\cdots,m)$轴末端的惯量 $Q_i=d_i^2$. 相应的

$$\chi^2=TQ=T\sum_{i=1}^{m}d_i^2 \quad (9.2.14)$$

给出总 χ^2 统计量的分解式.

(9) 对样品点和变量点进行分类,并结合专业知识进行成因解释.

§9.3 应用例子

在有些多元统计分析的教科书或文献上,介绍对应分析方法时分析处理的数据是二维频数表(或称双向列联表). SAS/STAT 软件中的 CORRESP(对应分析)过程就是讨论二维频数表中行和列之间各种联系的低维图表示法.该过程可进行简单的和多重的对应分析.分析处理的数据可以是双向列联表,或者是两个或多个属性变量的原始类目响应数据.

对应分析是列联表的一类加权主成分分析,它用于寻求列联表的行和列之间联系的低维图形表示法.每一行或每一列用单元频数确定的欧氏空间中的一个点表示.

例 9.3.1 表 9.1 中的数据是美国在 1973 到 1978 年间授予哲学博士学位的数目(美国人口调查局,1979).试用对应分析方法分析该组数据(摘自参考文献[19]).

表 9.1　美国于 1973 到 1978 年间授予哲学博士学位的数目

学科＼年	1973	1974	1975	1976	1977	1978
L(生命科学)	4489	4303	4402	4350	4266	4361
P(物理学)	4101	3800	3749	3572	3410	3234
S(社会学)	3354	3286	3344	3278	3137	3008
B(行为科学)	2444	2587	2749	2878	2960	3049
E(工程学)	3338	3144	2959	2791	2641	2432
M(数学)	1222	1196	1149	1003	959	959

解　如果把年度和学科作为两个属性变量，年度考虑 1973 至 1978 年这 6 年的情况(6 个类目)，学科也考虑 6 种学科，那么表 9.1 就是一张两个属性变量的列联表. 这张列联表也可以从所调查的 107904 个学生的原始数据得到.

使用 SAS/STAT 软件中 CORRESP 过程对表 9.1 的数据进行对应分析，可得出行形象(或称行剖面)、惯量(inertia)和 χ^2(Chi-Square，有时中文用“卡方”)分解，以及行和列的坐标等.

输出结果见输出 9.3.1 至输出 9.3.5. 输出 9.3.1 给出行轮廓分布(或行形象)就是列联表的每一行除以该行的行总和得到的. 比如 L(生命科学)这一行在“1973”这一列的行形象分量值为

0.171526＝4489/26171　(其中 26171＝4489＋4303＋…＋4361).

输出 9.3.2 给出的总 χ^2 统计量等于 383.856，该值在这个中心化后的列联表($\tilde{P}$)的全部 5 维中是行和列之间相关性的度量，它的最大的维数 5(或坐标轴)是行数和列数的最小值减 1. 在总 χ^2 或总惯量的 96%以上可用第一维说明，也就是说，行和列的类目之间的联系实质上可用一维表示.

输出 9.3.1　行轮廓分布阵 *R*

Row Profiles

	1973	1974	1975	1976	1977	1978
L(生命科学)	0.171526	0.164419	0.168201	0.166215	0.163005	0.166635
P(物理学)	0.187551	0.173786	0.171453	0.163359	0.155950	0.147901
S(社会学)	0.172824	0.169320	0.172309	0.168908	0.161643	0.154996
B(行为科学)	0.146637	0.155217	0.164937	0.172677	0.177596	0.182936
E(工程学)	0.192892	0.181682	0.170991	0.161283	0.152615	0.140537
M(数学)	0.188348	0.184340	0.177096	0.154593	0.147811	0.147811

输出 9.3.2 惯量和 χ^2(卡方)分解

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.05845	0.00342	368.653	96.04	96.04	**********
0.00861	0.00007	7.995	2.08	98.12	*
0.00694	0.00005	5.197	1.35	99.48	
0.00414	0.00002	1.852	0.48	99.96	
0.00122	0.00000	0.160	0.04	100.00	
Total	0.00356	383.856	100.00		

Degrees of Freedom = 25

总 χ^2 统计量就是检验两个属性变量是否互不相关时的检验统计量. 行(或列)形象点集间的总惯量 Q 定义为行点 R_i $(i=1,\cdots,n)$ 到重心 c 的 χ^2 距离的加权和. 设 $A=Z'Z$ 的特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m > 0,$$

则
$$Q = \sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij}^2 = \mathrm{tr}(A) = \sum_{i=1}^{m}\lambda_i, \quad \chi^2 = T\sum_{i=1}^{m}\lambda_i.$$

故总惯量 Q 和总 χ^2 统计量在输出结果中也给出分解的形式. 由输出 9.3.2 可看出,总 χ^2 或总惯量的 96%以上可用第一维说明,它表示行点和列点之间的关系用一维表示就足够了.

由输出 9.3.3 可以看出,第一维(Dim1)显示 6 门学科(样品)授予博士学位数目的变化方向;同时也可看出:在第一维中坐标最大的样品点(0.110006)所对应的学科是"行为科学",该学科授予博士学位的数目是随年度的变化而上升的(见输出 9.3.1);"生命科学"和"社会学"变化不大;而另外三个学科授予博士学位的数目是随年度的变化而下降的.

输出 9.3.3 行坐标

Row Coordinates

	Dim1	Dim2
L(生命科学)	0.025813	0.008097
P(物理学)	-.041273	-.002420
S(社会学)	0.001352	-.011413
B(行为科学)	0.110006	-.001299
E(工程学)	-.070379	-.003671
M(数学)	-.063942	0.022762

输出 9.3.4　列坐标

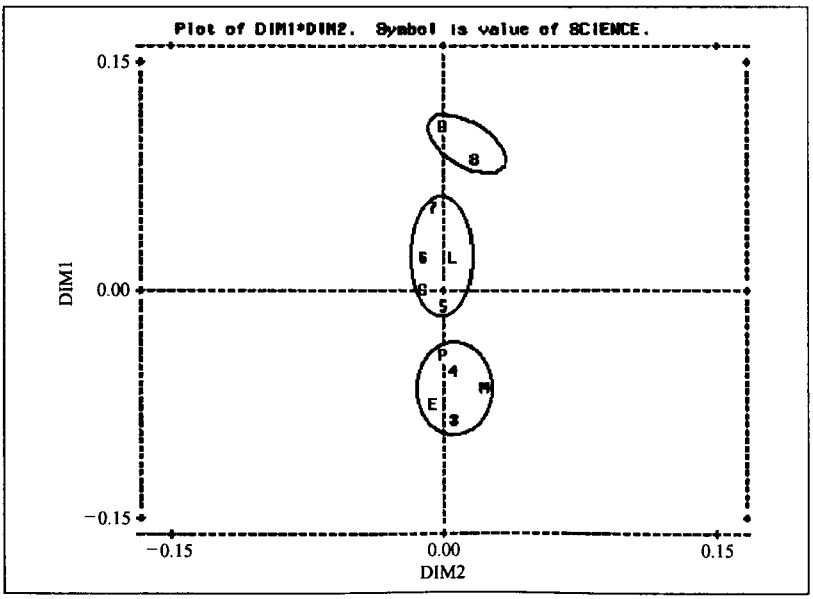
Column Coordinates

	Dim1	Dim2
1973	-.084027	0.003252
1974	-.050893	0.002939
1975	-.014823	0.000793
1976	0.024241	-.012926
1977	0.051249	-.008190
1978	0.086413	0.014276

由输出 9.3.4 可以看出，第一维显示出 6 个年度(变量)授予博士学位的数目随年份的增加而递增的变化方向.

输出 9.3.5 给出行、列坐标的散布图. 该图上的单个数字，如 3，4，…，8 是用年份减去 1970 而得到的，它们是表示各年度的列点；该图上的字符，如 L，P，…，M 表示各学科，它们是用行标签(学科)的第一个字符作为行点的符号. 由该散布图可以看出，表示从 1973 至 1978 年这些年的 3，4，…，8 在第一坐标轴(纵轴)方向上是按顺序排成一排，不过这并不是这次分析所要求的.

输出 9.3.5　行点和列点的散布图



从该散布图又可看出，由表示学科的行点沿纵轴——第一维方向上的排列显示出，随年度变化授予的博士学位数目从最大(如表示“行为学科”的 B)减少到最小(如表示“工程学”的 E)的学科排列次序. 这幅图给出了授予的博士学位数目依赖于学科变化的变化率. 行轮廓分布表(见输出 9.3.1)支持这些解释.

结合输出 9.3.1 给出的行轮廓分布，由输出 9.3.5 可看出，6 个行点和 6 个列点可以分为三类：第一类包括“行为学科(B)”，它在 1978 年授予的博士学位数目的比例最大；第二类包括“社会学(S)”和“生命科学(L)”，它们在 1975 至 1977 年授予的博士学位数目的比例都是随年度下降；第三类包括“物理学(P)”、“工程学(E)”和“数学(M)”，它们在 1973 和 1974 年这两年授予的博士学位数目的比例最大.

例 9.3.2 试用对应分析研究我国部分省份的农村居民家庭人均消费支出结构. 选取 7 个变量：A 为食品支出比重，B 为衣着支出比重，C 为居住支出比重，D 为家庭设备及服务支出比重，E 为医疗保健支出比重，F 为交通和通讯支出比重，G 为文教娱乐、日用品及服务支出比重. 考察的地区(即样品)有 10 个：山西、内蒙古、吉林、辽宁、黑龙江、海南、四川、贵州、甘肃、青海(原始数据见表 9.2).

表 9.2 中国 10 个省份农村居民家庭人均消费支出数据

	地区	A	B	C	D	E	F	G
1	山 西	0.583910	0.111480	0.092473	0.050073	0.038193	0.018803	0.079946
2	内蒙古	0.581218	0.081315	0.112380	0.042396	0.043280	0.040004	0.083339
3	辽 宁	0.565036	0.100121	0.123970	0.041121	0.043429	0.031328	0.078919
4	吉 林	0.530918	0.105360	0.116952	0.045064	0.043735	0.038508	0.095256
5	黑龙江	0.555201	0.096500	0.143498	0.037566	0.052111	0.026267	0.072829
6	海 南	0.654952	0.047852	0.095238	0.047945	0.022134	0.018519	0.096844
7	四 川	0.640012	0.061680	0.116677	0.048471	0.033529	0.017439	0.072043
8	贵 州	0.725239	0.056362	0.073262	0.044388	0.016366	0.015720	0.057261
9	甘 肃	0.678630	0.058043	0.088316	0.038100	0.039794	0.015167	0.067999
0	青 海	0.665913	0.088508	0.096899	0.038191	0.039275	0.019243	0.033801

解 表 9.2 给出的数据是一般的 $n\times p$ 原始数据阵，显然不是列联表，但我们要使用 SAS/STAT 软件提供的 CORRESP 过程进

行对应分析计算. 数据表中列变量(A,B,C,D,E,F,G)是消费支出的几个指标,可以理解为属性变量"消费支出"的几个水平(或类目). 数据表中的样品(行变量)是几个不同的地区,可理解为属性变量"地区"的几个不同水平(或类目). 数据表 9.2 可以理解为双向频率表,使用 CORRESP 过程进行对应分析计算.

把 7 个变量用字母 A,B,C,D,E,F,G 表示的目的是为下面作散布图时可用这几个字母来表示不同的变量.

输出 9.3.6 和输出 9.3.7 只给出最主要的结果.

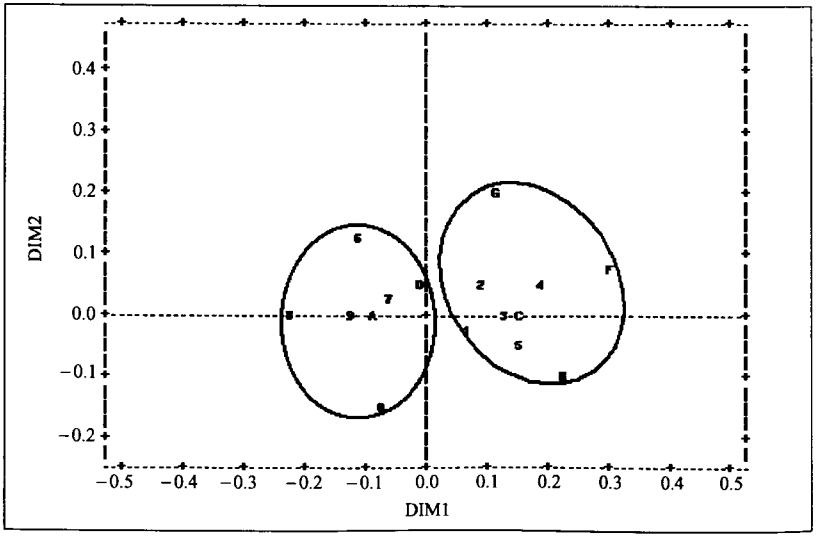
输出 9.3.6　惯量和 χ^2 分解

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	13 26 39 52 65
0.13161	0.01732	0.17031	65.59	65.59	*************************
0.06968	0.00486	0.04774	18.39	83.98	*******
0.04817	0.00232	0.02281	8.79	92.77	***
0.03582	0.00128	0.01261	4.86	97.63	**
0.02294	0.00053	0.00517	1.99	99.62	*
0.01002	0.00010	0.00099	0.38	100.00	
Total	0.02641	0.25963	100.00		

Degrees of Freedom = 54

输出 9.3.7　行点和列点的散布图



输出 9.3.6 给出惯量和 χ^2 统计量的分解，总 $\chi^2=0.25963$，总惯量$=0.02641$，总 χ^2 或总惯量的 83.98%(=65.59%+18.39%)可用前二维即可说明，它表示行点和列点之间的关系用二维表示就足够了.

在输出 9.3.7 中，给出 10 个样品点(用 1,2,…,9,0 表示)和 7 个变量点(用 A,B,…,G 表示)在相同坐标系上绘制的散布图. 从图中可以看出，样品点和变量点可以分为两类：第一类包括变量点 B,C,E,F,G 和样品点 1,2,3,4,5；第二类包括变量点 A,D 和样品点 6,7,8,9,10.

在第一类中，变量为衣着(B)，居住(C)，医疗保健(E)，交通和通讯(F)，文教娱乐、日用品及服务(G)的支出分别占总支出的比重；地区有：山西(1)，内蒙古(2)，辽宁(3)，吉林(4)，黑龙江(5)，它们位于我国的东部和北部地区，说明这 5 个地区的消费支出结构相似. 在第二类中，变量为食品(A)，家庭设备及服务(D)的支出分别占总支出的比重；地区有：海南(6)，四川(7)，贵州(8)，甘肃(9)，青海(0)，它们位于我国的南部和西部地区，说明这 5 个地区的消费支出结构相似.

习 题 九

9-1 我国山区某大型化工厂，在厂区及邻近地区挑选有代表性的 8 个大气取样点. 每日四次同时抽取大气样品，测定其中包含的 6 种气体的浓度，前后共四天，每个取样点每种气体实测 16 次，计算每个取样点每种气体的平均浓度，数据见第八章表 8.2.

(1) 试用对应分析方法对取样点及大气污染气体进行分类.

(2) 用 R 型因子分析方法(参数估计方法用主成分法)分析该组数据；并与(1)的结果比较之；

(3) 用 Q 因子分析方法分析该组数据；与(1)，(2)的结果比较之.

9-2 第六章表 6.7 是我国 16 个地区农民 1982 年支出情况的

抽样调查的汇总资料，每个地区都调查了反映每人平均生活消费支出情况的 6 个指标.

(1) 试用对应分析方法对所考察的 6 项指标和 16 个地区进行分类.

(2) 用 R 型因子分析方法(参数估计方法用主成分法)分析该组数据；并与(1)的结果比较之；

(3) 用聚类分析方法分析该组数据；与(1)，(2)的结果比较之.

9-3　试证明对应分析中行点坐标矩阵 G 和列点坐标矩阵 F 之间有以下关系：

$$G = D_r^{-1}\tilde{P}F\Lambda_m^{-1}, \quad F = D_c^{-1}\tilde{P}'G\Lambda_m^{-1}.$$

第十章　典型相关分析

典型相关分析是研究两组变量之间相关关系的一种统计方法.

在实际问题中,经常遇到要研究一部分变量和另一部分变量之间的相关关系.例如:在工厂里,考察原料的主要质量指标($X_1,\cdots,X_p$)与产品的主要质量指标($Y_1,\cdots,Y_q$)间的相关性;在经济学中,研究主要肉食品的价格与销售量之间的相关性;在地质学中,为研究岩石形成的成因关系,考察岩石的化学成分与其周围围岩化学成分的相关性;在气象学中为可靠地分析预报 24 小时后的天气,研究当天和前一天气象因子间的相关关系;在教育学中,研究学生在高考的各科成绩与高二年级各主科成绩间的相关关系;在婚姻的研究中,考察小伙子对追求姑娘的主要指标与姑娘向往的小伙子的主要尺度之间的相关关系等等.

一般地,假设有一组变量 $X_1,\cdots,X_p$ 与另一组变量 $Y_1,\cdots,Y_q$,我们要研究这两组变量的相关关系,如何给两组变量之间的相关性以数量的描述?

当 $p=q=1$ 时,就是研究两个变量 X 与 Y 之间的相关关系.相关系数是最常见的度量,其定义为

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

当 $p\geqslant 1,q=1$(或 $q\geqslant 1,p=1$)时,p 维随机向量 $X=(X_1,X_2,\cdots,X_p)'$,设 $\begin{bmatrix}X\\Y\end{bmatrix}\sim N_{p+1}(\mu,\Sigma)$,$\Sigma=\begin{bmatrix}\Sigma_{XX}&\Sigma_{XY}\\\Sigma_{YX}&\sigma_{YY}\end{bmatrix}$,则称

$$R=\sqrt{\frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\sigma_{YY}}}$$

为 Y 与 $X_1,X_2,\cdots,X_p$ 的**全相关系数**,全相关系数用于度量一个随机变量 Y 与一组随机变量 $X_1,X_2,\cdots,X_p$ 的相关关系.

当 $p,q>1$ 时，利用主成分分析的思想，可以把多个变量与多个变量之间的相关化为两个新的综合变量之间的相关. 也就是求 $\alpha=(\alpha_1,\cdots,\alpha_p)'$ 和 $\beta=(\beta_1,\cdots,\beta_q)'$，使得新的综合变量

$$V=\alpha_1X_1+\alpha_2X_2+\cdots+\alpha_pX_p=\alpha'X,$$

和

$$W=\beta_1Y_1+\beta_2Y_2+\cdots+\beta_qY_q=\beta'Y$$

之间有最大可能的相关，基于这个思想就产生了典型相关分析(Canonical correlatinal analysis). 这就是本章将介绍的典型相关分析和典型冗余分析，它们是偏最小二乘回归(第十一章)的基础.

§10.1　总体典型相关

一、典型相关和典型相关变量的定义

设 $X=(X_1,\cdots,X_p)'$ 及 $Y=(Y_1,\cdots,Y_q)'$ 为随机向量，我们用 X 和 Y 的线性组合 $\alpha'X$ 和 $\beta'Y$ 之间的相关性来研究两组随机变量 X 和 Y 之间的相关性. 我们希望找到 α 和 β，使 $\rho(\alpha'X,\beta'Y)$ 最大. 由相关系数的定义

$$\rho(\alpha'X,\beta'Y)=\frac{\mathrm{Cov}(\alpha'X,\beta'Y)}{\sqrt{\mathrm{Var}(\alpha'X)}\sqrt{\mathrm{Var}(\beta'Y)}},$$

易得出对任意常数 e,f,c 和 d，均有

$$\rho[e(\alpha'X)+f,c(\beta'Y)+d]=\rho(\alpha'X,\beta'Y).$$

这说明使得相关系数最大的 $\alpha'X$ 和 $\beta'Y$ 并不唯一. 故求综合变量时常限定 $\mathrm{Var}(\alpha'X)=1$，$\mathrm{Var}(\beta'Y)=1$. 于是有以下定义.

定义 10.1.1　设 $X=(X_1,\cdots,X_p)'$，$Y=(Y_1,\cdots,Y_q)'$，$p+q$ 维随机向量 $\begin{bmatrix}X\\Y\end{bmatrix}$ 的均值向量为 0，协方差阵 $\Sigma>0$ (不妨设 $p\leqslant q$). 如果存在 $a_1=(a_{11},\cdots,a_{p1})'$ 和 $b_1=(b_{11},\cdots,b_{q1})'$，使得

$$\rho(a_1'X,b_1'Y)=\max_{\mathrm{Var}(\alpha'X)=1,\ \mathrm{Var}(\beta'Y)=1}\rho(\alpha'X,\beta'Y),$$

则称 $a_1'X$，$b_1'Y$ 是 X,Y 的**第一对(组)典型相关变量**，它们之间的相关系数称为**第一个典型相关系数**；如果存在 $a_k=(a_{1k},\cdots,a_{pk})'$ 和 $b_k=(b_{1k},\cdots,b_{qk})'$，使得

(1) $a_k'X, b_k'Y$ 和前面 $k-1$ 对典型相关变量都不相关；

(2) $\mathrm{Var}(a_k'X)=1$，$\mathrm{Var}(b_k'Y)=1$；

(3) $a_k'X$ 与 $b_k'Y$ 的相关系数最大，

则称 $a_k'X$，$b_k'Y$ 是 X,Y 的**第 k 对(组)典型相关变量**，它们之间的相关系数称为**第 k 个典型相关系数** $(k=2,\cdots,p)$.

二、典型相关变量的解法

设随机向量

$$Z=\begin{bmatrix}X\\Y\end{bmatrix},$$

其中 $X=(X_1,\cdots,X_p)'$，$Y=(Y_1,\cdots,Y_q)'$(不妨设 $p\leqslant q$)；$\mathrm{E}(Z)=0$；以及 $\mathrm{D}(Z)=\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}>0$.

1. 第一对典型相关变量的求法

令 $V=\alpha'X, W=\beta'Y$，则 V,W 的相关系数

$$\rho(V,W)=\frac{\alpha'\Sigma_{12}\beta}{\sqrt{\alpha'\Sigma_{11}\alpha}\sqrt{\beta'\Sigma_{22}\beta}}.$$

求第一对典型相关变量就等价于求 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_p)'$ 和 $\beta=(\beta_1,\beta_2,\cdots,\beta_q)'$，使得在条件 $\mathrm{Var}(\alpha'X)=1$ 和 $\mathrm{Var}(\beta'Y)=1$ 下，

$$\rho(\alpha'X,\beta'Y)=\alpha'\Sigma_{12}\beta$$

达到最大. 这是条件极值问题，用拉格朗日乘子法，令

$$\varphi(\alpha,\beta)=\alpha'\Sigma_{12}\beta-\frac{\lambda_1}{2}(\alpha'\Sigma_{11}\alpha-1)-\frac{\lambda_2}{2}(\beta'\Sigma_{22}\beta-1),$$

其中 λ_1,λ_2 为拉格朗日乘子. 为求 φ 的极大值. 对上式分别关于 α,β 求偏导，并令其为零，得

$$\begin{cases}\dfrac{\partial\varphi}{\partial\alpha}=\Sigma_{12}\beta-\lambda_1\Sigma_{11}\alpha=0,\\[2ex] \dfrac{\partial\varphi}{\partial\beta}=\Sigma_{21}\alpha-\lambda_2\Sigma_{22}\beta=0;\end{cases}\tag{10.1.1}$$

再分别用 α',β' 左乘方程(10.1.1)，得

$$\lambda_1 = \lambda_2 = \alpha' \Sigma_{12} \beta = \rho(V, W) \xlongequal{\text{def}} \lambda,$$

则方程组(10.1.1)等价于

$$\begin{cases} -\lambda \Sigma_{11} \alpha + \Sigma_{12} \beta = 0, \\ \Sigma_{21} \alpha - \lambda \Sigma_{22} \beta = 0. \end{cases} \tag{10.1.2}$$

方程组(10.1.2)有非零解的充要条件是

$$\begin{vmatrix} -\lambda \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda \Sigma_{22} \end{vmatrix} = 0, \tag{10.1.3}$$

该方程左端是 λ 的 $p+q$ 次多项式. 求解 λ 的高次方程(10.1.3),把求得的最大的 λ 代回方程组(10.1.2),再求得 α 和 β,从而得出第一对典型相关变量.

具体计算时,因 λ 的高次方程(10.1.3)不易解;将其代入方程组(10.1.2)后还需求解($p+q$)阶方程组. 为了计算上的简便,常作以下变换:

用 $\Sigma_{12}\Sigma_{22}^{-1}$ 左乘方程组(10.1.2)的第二式,并把

$$\Sigma_{12} \beta = \frac{1}{\lambda} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \alpha$$

代入方程组(10.1.2)的第一式得:

$$\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \alpha - \lambda^2 \Sigma_{11} \alpha = 0;$$

再用 Σ_{11}^{-1} 左乘上式得:

$$(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \lambda^2 I_p) \alpha = 0,$$

然后由 p 阶特征方程求解 λ 和 α. 类似地,可通过求解 q 阶特征方程

$$(\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 I_q) \beta = 0$$

来得到 λ 和 β. 故求解方程(10.1.3)等价于求解方程组(10.1.4):

$$\begin{cases} |\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \lambda^2 I_p| = 0, \\ |\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 I_q| = 0. \end{cases} \tag{10.1.4}$$

由于 $\Sigma_{11} > 0$, $\Sigma_{22} > 0$,故 $\Sigma_{11}^{-1} > 0$, $\Sigma_{22}^{-1} > 0$,所以有

$$\begin{aligned} M_1 &= \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\ &= \Sigma_{11}^{-1/2} (\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2} \Sigma_{22}^{-1/2} \Sigma_{21}) \xlongequal{\text{def}} AB, \end{aligned}$$

其中　　$A = \Sigma_{11}^{-1/2}$,　$B = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2} \Sigma_{22}^{-1/2} \Sigma_{21}$.

但 AB 与 $BA=(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21})\cdot\Sigma_{11}^{-1/2}$有相同的非零特征值，如果记

$$T=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2},$$

则 $BA=TT'$，故 M_1 与 TT' 有相同的非零特征值. 类似地

$$M_2=\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

与 $T'T$ 有相同的非零特征值.

由以上分析可知，M_1 与 M_2 有相同的非零特征值，且非零特征值的个数至多为 p 个(因 $p\leqslant q$).

设 $|TT'-\lambda^2 I_p|=0$ 的 p 个特征值依次为 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2>0$，则 $T'T$ 的 q 个特征值中，除以上 p 个外，其余 $q-p$ 个皆为 0. 故方程(10.1.3)的 $p+q$ 个根依次为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0=\cdots=0>-\lambda_p\geqslant\cdots\geqslant-\lambda_2\geqslant-\lambda_1$($\lambda_i$ 是 λ_i^2 的正平方根，$i=1,\cdots,p$). 取其中最大的 λ_1 代入方程组(10.1.2)，即可求得 $\alpha=a_1$，$\beta=b_1$(设 a_1,b_1 满足 $a_1'\Sigma_{11}a_1=1$，$b_1'\Sigma_{22}b_1=1$). 令 $V_1=a_1'X$，$W_1=b_1'Y$，则 V_1，W_1 为第一对典型相关变量；而 $\rho(V_1,W_1)=a_1'\Sigma_{12}b_1=\lambda_1$ 为第一个典型相关系数.

其实可以证明，a_1 是 TT' 的特征值 λ_1^2 对应的满足 $a_1'\Sigma_{11}a_1=1$ 的特征向量；b_1 是 $T'T$ 的特征值 λ_1^2 对应的满足 $b_1'\Sigma_{22}b_1=1$ 的特征向量. 故求第一对典型相关变量及典型相关系数的问题，就等价于求解 TT' 的最大特征值及相应的特征向量.

2. 典型相关变量的一般求法

从第一对典型相关变量的解法中，我们知道求第一对典型相关变量和第一个典型相关系数的问题，就是求解 TT' 的最大特征值和相应的特征向量. 不仅如此，求解第 k 对典型相关变量和典型相关系数，类似地也是求 TT' 的第 k 大特征值和相应的特征向量.

定理 10.1.1 设 $Z=\begin{bmatrix}X\\Y\end{bmatrix}$，其中 $X=(X_1,\cdots,X_p)'$ 为 p 维随机向量，$Y=(Y_1,\cdots,Y_q)'$ 为 q 维随机向量(不妨设 $p\leqslant q$). 已知

$$\mathrm{E}(Z)=0,\quad \mathrm{D}(Z)=\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}>0.$$

记 $T=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$，并设 p 阶方阵 TT' 的特征值依次为 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots$

$\geqslant\lambda_p^2>0$ ($\lambda_i>0$, $i=1,\cdots,p$)；而 $l_1,l_2,\cdots,l_p$ 为相应的单位正交特征向量. 令

$$a_k=\Sigma_{11}^{-1/2}l_k,\quad b_k=\lambda_k^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_k\quad(k=1,2,\cdots,p).$$

则 $V_k=a_k'X$, $W_k=b_k'Y$ 为 X,Y 的第 k 对典型相关变量，λ_k 为第 k 个典型相关系数.

证明　当 $k=1$ 时，取

$$a_1=\Sigma_{11}^{-1/2}l_1,\quad b_1=\lambda_1^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_1,$$

其中 l_1 是 TT' 的最大特征值 λ_1^2 所对应的单位长度特征向量. 下面来证明：

$$\rho(a_1'X,b_1'Y)=a_1'\Sigma_{12}b_1=\max_{\mathrm{Var}(\alpha'X)=1,\mathrm{Var}(\beta'Y)=1}\rho(\alpha'X,\beta'Y).$$

任给满足约束条件 $\alpha'\Sigma_{11}\alpha=1$ 和 $\beta'\Sigma_{22}\beta=1$ 的 p 维向量 α 和 q 维向量 β，则 $\alpha'X$ 和 $\beta'Y$ 的相关系数为

$$\begin{aligned}\rho(\alpha'X,\beta'Y)&=\alpha'\Sigma_{12}\beta=\beta'\Sigma_{21}\alpha\\&=\beta'\Sigma_{22}^{1/2}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{11}^{1/2}\alpha\\&=\widetilde{\beta}'T'\,\widetilde{\alpha}\quad(\text{其中 }\widetilde{\alpha}=\Sigma_{11}^{1/2}\alpha,\widetilde{\beta}=\Sigma_{22}^{1/2}\beta)\\&\leqslant[\widetilde{\beta}'\,\widetilde{\beta}]^{1/2}[(T'\,\widetilde{\alpha})'(T'\,\widetilde{\alpha})]^{1/2}\\&=[\widetilde{\alpha}'TT'\,\widetilde{\alpha}]^{1/2}.\end{aligned}$$

以上不等式利用附录中引理 7.1. 不妨设

$$\widetilde{\alpha}=\sum_{i=1}^{p}c_il_i,$$

其中 $l_1,\cdots,l_p$ 是 TT' 的 p 个单位正交特征向量. 由 $\widetilde{\alpha}'\widetilde{\alpha}=1$，故 $\sum_{i=1}^{p}c_i^2=1$. 于是

$$\begin{aligned}\widetilde{\alpha}'TT'\,\widetilde{\alpha}&=\Big(\sum_{i=1}^{p}c_il_i\Big)'TT'\Big(\sum_{j=1}^{p}c_jl_j\Big)\\&=\sum_{i=1}^{p}\sum_{j=1}^{p}c_ic_jl_i'TT'l_j=\sum_{i=1}^{p}c_i^2\lambda_i^2\leqslant\lambda_1^2,\end{aligned}$$

从而

$$\rho(\alpha'X,\beta'Y)=\beta'\Sigma_{21}\alpha\leqslant\lambda_1.$$

因 λ_1^2 是 TT' 的最大特征值，l_1 是相应的单位长度特征向量，所以

$$a_1'\Sigma_{12}b_1 = l_1'\Sigma_{11}^{-1/2}\Sigma_{12}\lambda_1^{-1}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}l_1$$
$$= \frac{1}{\lambda_1}l_1'TT'l_1 = \frac{1}{\lambda_1}l_1'\lambda_1^2 l_1 = \lambda_1$$
$$= \max_{\mathrm{Var}(\alpha'X)=1,\mathrm{Var}(\beta'Y)=1}\rho(\alpha'X,\beta'Y),$$

由定义 10.1.1 知，$a_1'X, b_1'Y$ 是 X,Y 的第一对典型相关变量，它们的典型相关系数(即第一典型相关系数)为 λ_1.

假定已经求得 $k-1$ 对典型相关变量

$$V_i = a_i'X, \quad W_i = b_i'Y \quad (i = 1,\cdots,k-1),$$

其中 $a_i=\Sigma_{11}^{-1/2}l_i$, $b_i=\lambda_i^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_i$，而 l_i 是 TT' 的第 i 大特征值 λ_i^2 所对应的单位长度特征向量，而且 V_i $(i=1,\cdots,k-1)$互不相关，W_i $(i=1,\cdots,k-1)$也互不相关. 现在来求第 k 对典型相关变量及第 k 个典型相关系数. 取

$$a_k = \Sigma_{11}^{-1/2}l_k, \quad b_k = \lambda_k^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_k,$$

记 $V_k=a_k'X$, $W_k=b_k'Y$. 显然

(1) V_k 和 V_i, W_k 和 W_i 的相关系数分别为

$$\rho(V_k,V_i) = a_k'\Sigma_{11}a_i = a_k'\Sigma_{11}^{1/2}\cdot\Sigma_{11}^{1/2}a_i = l_k'l_i = 0;$$
$$\rho(W_k,W_i) = b_k'\Sigma_{22}b_i = \lambda_k^{-1}a_k'\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\lambda_i^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_i$$
$$= \frac{1}{\lambda_k\lambda_i}l_k'TT'l_i = 0 \quad (i = 1,\cdots,k-1).$$

(2) $\mathrm{Var}(V_k)=a_k'\Sigma_{11}a_k=1$, $\mathrm{Var}(W_k)=b_k'\Sigma_{22}b_k=1$.

(3) V_i 和 W_j 的相关系数为

$$\rho(V_i,W_j) = a_i'\Sigma_{12}b_j = \lambda_j^{-1}l_i'\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}l_j = \lambda_j^{-1}l_i'TT'l_j$$
$$= \begin{cases}\lambda_j, & \text{当 } i=j \text{ 时},\\ 0, & \text{当 } i\neq j \text{ 时}\end{cases} \quad (i,j=1,2,\cdots,k).$$

下面只须证明对满足下列条件

$$\begin{cases}\mathrm{Var}(\alpha'X)=1, \quad \mathrm{Var}(\beta'Y)=1,\\ \rho(\alpha'X,V_i)=0, \quad \rho(\beta'Y,W_i)=0 \quad (i=1,\cdots,k-1)\end{cases} \tag{10.1.5}$$

的一切 α,β 有

$$\rho(V_k,W_k) = \max_{\alpha,\beta\text{ 满足条件(10.1.5)}}\rho(\alpha'X,\beta'Y).$$

用类似的方法可以证明，对任给满足条件(10.1.5)的 α,β，有

$$\rho(\alpha'X,\beta'Y)=\beta'\Sigma_{21}\alpha\leqslant[\tilde{\alpha}'TT'\tilde{\alpha}]^{1/2},$$

其中 $\tilde{\alpha}=\Sigma_{11}^{1/2}\alpha$，满足 $\tilde{\alpha}'\tilde{\alpha}=1$. 不妨设 $\tilde{\alpha}=\sum_{i=1}^{p}c_il_i$，其中 $l_1,l_2,\cdots,l_p$ 是 TT' 的 p 个单位正交特征向量. 由条件(10.1.5)，对 $i=1,\cdots,k-1$ 有

$$\begin{aligned}0&=\rho(\alpha'X,V_i)=\alpha'\Sigma_{11}a_i\\&=\tilde{\alpha}'\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}l_i=\left(\sum_{j=1}^{p}c_jl_j\right)'l_i=c_i,\end{aligned}$$

故

$$\tilde{\alpha}=\sum_{i=k}^{p}c_il_i,\quad\text{且}\quad\sum_{i=k}^{p}c_i^2=1,$$

于是

$$\tilde{\alpha}'TT'\tilde{\alpha}=\left(\sum_{i=k}^{p}c_il_i\right)'TT'\left(\sum_{i=k}^{p}c_il_i\right)=\sum_{i=k}^{p}c_i^2\lambda_i^2\leqslant\lambda_k^2.$$

当 $\alpha=a_k$，$\beta=b_k$ 时，$a_k'\Sigma_{12}b_k=\lambda_k$，所以

$$\rho(a_k'X,b_k'Y)=\max_{\alpha,\beta\text{ 满足条件(10.1.5)}}\rho(\alpha'X,\beta'Y).$$

由定义 10.1.1 知，$a_k'X,b_k'Y$ 是 X,Y 的第 k 对典型相关变量，它们的第 k 个典型相关系数为 λ_k $(k=2,\cdots,p)$. **(证毕)**

以上定理 10.1.1 中，我们假定 $\Sigma>0$，一般情况协方差阵非负定，从而 Σ_{11}^{-1}，Σ_{22}^{-1}不一定存在. 但注意到方程(10.1.2)总有非零解，因此我们可用广义逆矩阵来求解.

定义 10.1.2　给定一个矩阵 A，如果有矩阵 D 满足

$$ADA=A,\quad DAD=D,\quad(AD)'=AD,\quad(DA)'=DA,$$

则称 D 是 A 的加号逆，记作 A^+.

A^+是存在唯一的，还有一些基本性质(见参考文献[7]). 以下根据加号逆不加证明地给出更一般的结论.

定理 10.1.2　设 $Z=\begin{bmatrix}X\\Y\end{bmatrix}$，其中 $X=(X_1,\cdots,X_p)'$ 为 p 维随机向量，$Y=(Y_1,\cdots,Y_q)'$ 为 q 维随机向量(不妨设 $p\leqslant q$). 已知

$$\mathrm{E}(Z)=0,\quad \mathrm{D}(Z)=\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\geqslant 0,$$

记 $T=(\Sigma_{11}^{1/2})^{+}\Sigma_{12}(\Sigma_{22}^{1/2})^{+}$，$m=\mathrm{rank}(TT')\leqslant\min(p,q)$. 并设 p 阶方阵 TT' 的非零特征值依次为 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_m^2>0$ $(\lambda_i>0,i=1,\cdots,m)$；而 $l_1,l_2,\cdots,l_m$ 为其相应的单位正交特征向量. 令

$$a_k=(\Sigma_{11}^{1/2})^{+}l_k,\quad b_k=\lambda_k^{-1}(\Sigma_{22})^{+}\Sigma_{21}a_k\quad(k=1,2,\cdots,m).$$

则 $V_k=a_k'X$，$W_k=b_k'Y$ 为 X,Y 的第 k 对典型相关变量，λ_k 为第 k 个典型相关系数.

三、典型变量的性质

性质 1 设 $V_k=a_k'X$，$W_k=b_k'Y$ 为 X,Y 的第 k 对典型相关变量 $(k=1,\cdots,p)$；令 $V=(V_1,\cdots,V_p)'$，$W=(W_1,\cdots,W_p)'$. 则

$$D\begin{bmatrix}V\\ W\end{bmatrix}=\begin{bmatrix}I_p & \Lambda\\ \Lambda & I_p\end{bmatrix},$$

其中 $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$.

此性质说明 V_i $(i=1,\cdots,p)$互不相关；W_j $(j=1,\cdots,p)$互不相关；且 V_i 与 W_j $(i\neq j)$也互不相关；而 $\rho(V_i,W_i)=\lambda_i$ $(i=1,\cdots,p)$.

性质 2 原始变量与典型变量之间的相关性.

求出典型变量后，进一步来计算原始变量与典型变量之间的相关系数矩阵，也称为**典型结构**.

记 $A=(a_1,a_2,\cdots,a_p)$ 为 $p\times p$ 矩阵，$B=(b_1,b_2,\cdots,b_p)$ 为 $q\times p$ 矩阵. 设典型随机向量

$$V=(V_1,\cdots,V_p)'=(a_1'X,\cdots,a_p'X)'=A'X,$$
$$W=(W_1,\cdots,W_p)'=(b_1'Y,\cdots,b_p'Y)'=B'Y,$$

$p+q$ 维随机向量 Z 的协方差阵为 $\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}>0$. 则

$$\mathrm{COV}(X,V)=\mathrm{COV}(X,A'X)=\Sigma_{11}A,$$
$$\mathrm{COV}(X,W)=\mathrm{COV}(X,B'Y)=\Sigma_{12}B,$$
$$\mathrm{COV}(Y,V)=\mathrm{COV}(Y,A'X)=\Sigma_{21}A,$$
$$\mathrm{COV}(Y,W)=\mathrm{COV}(Y,B'Y)=\Sigma_{22}B.$$

利用协方差进一步可以计算原始变量与典型变量之间的相关系

数. 若假定原始变量均为标准化变量，则通过以上计算所得到的原始变量与典型变量的协方差阵就是相关系数矩阵.

若计算这四个相关系数矩阵中各列(或各行)相关系数的平方和，还将得出一些有关典型冗余分析的概念(见 § 10.3).

性质 3　设 X 和 Y 分别为 p 维和 q 维随机向量，令 $X^*=C'X+d$，$Y^*=G'Y+h$，其中 C 为 $p\times p$ 非退化矩阵，d 为 p 维常向量，G 为 $q\times q$ 非退化矩阵，h 为 q 维常向量. 则

(1) X^* 和 Y^* 的典型相关变量为 $(a_i^*)'X^*$ 和 $(b_i^*)'Y^*$，其中 $a_i^*=C^{-1}a_i$，$b_i^*=G^{-1}b_i$ $(i=1,2,\cdots,p)$；而 a_i,b_i 是 X 和 Y 的第 i 对典型相关变量的系数.

(2) $\rho[(a_i^*)'X^*,(b_i^*)'Y^*]=\rho(a_i'X,b_i'Y)$，即线性变换不改变相关性.

例 10.1.1　已知 p 维随机向量 X 和 q 维随机向量 Y 的协方差阵分别为 Σ_{11} 和 Σ_{22}. 试从 $Z=\begin{bmatrix}X\\Y\end{bmatrix}$ 的相关阵 R 出发求典型相关变量和典型相关系数.

解　取

$$C=(\mathrm{diag}(\Sigma_{11}))^{-1/2},\quad G=(\mathrm{diag}(\Sigma_{22}))^{-1/2}$$

(记号 $\mathrm{diag}(C)$ 表示由 C 的对角元素组成的对角矩阵).

设 $\begin{bmatrix}X\\Y\end{bmatrix}$ 的相关阵 $R=\begin{bmatrix}R_{11}&R_{12}\\R_{21}&R_{22}\end{bmatrix}$，令

$$X^*=CX,\quad Y^*=GY.$$

这时线性变换后的 $\begin{bmatrix}X^*\\Y^*\end{bmatrix}$ 的协方差阵就是 $\begin{bmatrix}X\\Y\end{bmatrix}$ 的相关阵 R.

如果已知 R，欲求 X,Y 的典型相关变量及典型相关系数时，可从 R 矩阵出发，求 $T^*(T^*)'$ 的特征值和单位正交特征向量(其中 $T^*=R_{11}^{-1/2}R_{12}R_{22}^{-1/2}$)，从而得 X^*,Y^* 的典型相关变量 $(a_i^*)'X^*$ 和 $(b_i^*)'Y^*$ 及典型相关系数 λ_i $(i=1,\cdots,p)$. 令

$$a_i=Ca_i^*,\quad b_i=Gb_i^*\quad (i=1,\cdots,p),$$

则 $a_i'X$ 和 $b_i'Y$ 即为 X,Y 的第 i 对典型相关变量，它们的相关系数为 λ_i $(i=1,\cdots,p)$.

●

例 10.1.2 已知标准化变量 $X=(X_1,X_2)'$ 和 $Y=(Y_1,Y_2)'$ 的相关阵

$$R=\begin{bmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{bmatrix},$$

其中

$$R_{11}=\begin{bmatrix}1 & \alpha\\ \alpha & 1\end{bmatrix},\quad R_{22}=\begin{bmatrix}1 & \nu\\ \nu & 1\end{bmatrix},$$

$$R_{12}=R_{21}=\begin{bmatrix}\beta & \beta\\ \beta & \beta\end{bmatrix}\quad (0<\beta<1).$$

试求 X,Y 的典型相关变量和典型相关系数.

解 由已知的相关阵 R 即可求出

$$R_{11}^{-1}=\frac{1}{1-\alpha^2}\begin{bmatrix}1 & -\alpha\\ -\alpha & 1\end{bmatrix},\quad R_{22}^{-1}=\frac{1}{1-\nu^2}\begin{bmatrix}1 & -\nu\\ -\nu & 1\end{bmatrix},$$

$$M_1^*=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}=\frac{2\beta^2}{(1+\alpha)(1+\nu)}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.$$

由于 $J=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$ 的特征值为 2 和 0，故 M_1^* 的特征值为

$$\lambda_1^2=\frac{4\beta^2}{(1+\alpha)(1+\nu)},\quad \lambda_2^2=0.$$

M_1^* 对应 λ_1^2 的特征向量为 $(1/\sqrt{2},1/\sqrt{2})'$，故满足 $a'R_{11}a=1$ 的向量 a 为

$$a=\frac{1}{\sqrt{2(1+\alpha)}}\begin{bmatrix}1\\ 1\end{bmatrix};$$

类似可得

$$b=\frac{1}{\sqrt{2(1+\nu)}}\begin{bmatrix}1\\ 1\end{bmatrix}.$$

所以第一对典型相关变量为

$$V_1=a'X=\frac{1}{\sqrt{2(1+\alpha)}}(X_1+X_2),$$

$$W_1=b'Y=\frac{1}{\sqrt{2(1+\nu)}}(Y_1+Y_2).$$

而第一个典型相关系数为

$$\rho_1=\frac{2\beta}{\sqrt{(1+\alpha)(1+\nu)}}\quad (0<\rho_1<1).$$

因 $|\alpha|<1$，$|\nu|<1$，显然有 $\rho_1>\beta$，这表明第一个典型相关系数一般

大于两组原变量之间的相关系数.

§10.2　样本典型相关

设总体 $Z=(X_1,\cdots,X_p,Y_1,\cdots,Y_q)'$，在实际问题中，总体的均值向量 $E(Z)=\mu$ 和协方差阵 $D(Z)=\Sigma$ 通常是未知的，因而无法求得总体的典型相关变量和典型相关系数. 首先需要根据观测到的样本数据阵对 Σ 进行估计.

已知总体 Z 的 n 次观测数据为：

$$Z_{(t)}=\begin{bmatrix}X_{(t)}\\ Y_{(t)}\end{bmatrix}_{(p+q)\times 1}\qquad (t=1,2,\cdots,n),$$

于是样本数据阵为

$$\left[\begin{array}{cccc:cccc}x_{11}&x_{12}&\cdots&x_{1p}&y_{11}&y_{12}&\cdots&y_{1q}\\ x_{21}&x_{22}&\cdots&x_{2p}&y_{21}&y_{22}&\cdots&y_{2q}\\ \vdots&\vdots&&\vdots&\vdots&\vdots&&\vdots\\ x_{n1}&x_{n2}&\cdots&x_{np}&y_{n1}&y_{n2}&\cdots&y_{nq}\end{array}\right]_{n\times(p+q)}.$$

若假定 $Z\sim N_{p+q}(\mu,\Sigma)$，则协方差阵 Σ 的最大似然估计为

$$\hat{\Sigma}=\frac{1}{n}\sum_{t=1}^{n}(Z_{(t)}-\bar{Z})(Z_{(t)}-\bar{Z})',$$

其中 $\bar{Z}=\dfrac{1}{n}\sum\limits_{t=1}^{n}Z_{(t)}$. 令 $S=\dfrac{n}{n-1}\hat{\Sigma}$，$S$ 矩阵也称为样本协方差阵.

设 $\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{bmatrix}$，$\Sigma_{11}$ 为 p 阶矩阵. 将 S 相应剖分为

$$S=\begin{bmatrix}S_{11}&S_{12}\\ S_{21}&S_{22}\end{bmatrix}.$$

显然，$S_{ij}\ (i,j=1,2)$ 是 Σ_{ij} 的无偏估计量. 下面我们将从样本协方差阵 S 出发，来讨论两组变量间的相关关系.

一、样本典型相关变量和典型相关系数

不妨设 $S>0$. 令 $\hat{T}=S_{11}^{-1/2}S_{12}S_{22}^{-1/2}$，并设 $\hat{T}\hat{T}'$ 的特征值依次为

$\hat{\lambda}_1^2 \geqslant \hat{\lambda}_2^2 \geqslant \cdots \geqslant \hat{\lambda}_p^2 > 0$ ($\hat{\lambda}_i > 0, i=1,\cdots,p$)，$\hat{l}_k$ ($k=1,\cdots,p$)为 $\hat{T}\hat{T}'$的特征根 $\hat{\lambda}_k^2$ 所对应的单位正交特征向量. 令

$$\begin{cases} \hat{a}_k = S_{11}^{-1/2}\hat{l}_k, \\ \hat{b}_k = \hat{\lambda}_k^{-1}S_{22}^{-1}S_{21}\hat{a}_k, \end{cases} \tag{10.2.1}$$

则 $\hat{V}_k = \hat{a}_k'X$，$\hat{W}_k = \hat{b}_k'Y$ 为 X,Y 的第 k 对样本典型相关变量；而 $\hat{\lambda}_k$ 为 X,Y 的第 k 个样本典型相关系数.

以上我们从样本协方差阵 S 出发，导出了样本典型相关变量和样本典型相关系数. 另外也可以从样本相关阵 R 出发来导出样本典型相关变量和样本典型相关系数.

设样本相关阵 $R=(r_{ij})$，其中 $r_{ij}=s_{ij}/\sqrt{s_{ii}s_{jj}}$，$s_{ij}$为样本协方差阵 S 的元素. 把 R 相应剖分为

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

记

$$D_1 = \begin{bmatrix} \sqrt{s_{11}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{s_{pp}} \end{bmatrix}, \quad D_2 = \begin{bmatrix} \sqrt{s_{p+1,p+1}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{s_{p+q,p+q}} \end{bmatrix},$$

则

$$S_{11} = D_1R_{11}D_1, \quad S_{22} = D_2R_{22}D_2,$$
$$S_{12} = D_1R_{12}D_2, \quad S_{21} = D_2R_{21}D_1.$$

记 $\widetilde{T} = R_{11}^{-1/2}R_{12}R_{22}^{-1/2}$，求 $\widetilde{T}\widetilde{T}'$的特征值. 依次记 $\widetilde{T}\widetilde{T}'$的特征值为 $\tilde{\lambda}_1^2 \geqslant \tilde{\lambda}_2^2 \geqslant \cdots \geqslant \tilde{\lambda}_p^2 > 0$ ($\tilde{\lambda}_i > 0, i=1,\cdots,p$)，$\tilde{l}_k$ ($k=1,\cdots,p$)为 $\widetilde{T}\widetilde{T}'$的特征值 $\tilde{\lambda}_k^2$ 所对应的单位正交特征向量，则

$$\begin{cases} \tilde{V}_k = (D_1^{-1}R_{11}^{-1/2}\tilde{l}_k)'X = \tilde{a}_k'X, \\ \tilde{W}_k = (D_2^{-1}\tilde{\lambda}_k^{-1}R_{22}^{-1}R_{21}R_{11}^{-1/2}\tilde{l}_k)'Y = \tilde{b}_k'Y \end{cases}$$

为 X,Y 的第 k 对样本典型相关变量；而 $\tilde{\lambda}_k$ 为 X,Y 的第 k 个样本典型相关系数($k=1,2,\cdots,p$).

二、典型相关系数的显著性检验

1. 检验 H_0：$\Sigma_{12}=O$

总体 Z 的两组变量 $X=(X_1,\cdots,X_p)'$ 和 $Y=(Y_1,\cdots,Y_q)'$ 如果不相关，即 $\mathrm{COV}(X,Y)=\Sigma_{12}=O$，则以上有关两组变量典型相关的讨论就毫无意义. 故在讨论两组变量间的相关关系之前，应首先对假设 H_0 作统计检验.

设总体 $Z\sim N_{p+q}(\mu,\Sigma)$，用似然比方法可导出检验 H_0 的似然比统计量

$$\Lambda=\frac{|S|}{|S_{11}||S_{22}|},\tag{10.2.2}$$

其中 $p+q$ 阶矩阵 S 是 Σ 的最大似然估计，S_{11},S_{22} 分别是 Σ_{11},Σ_{22} 的最大似然估计.

似然比统计量 Λ 的精确分布已由霍特林(1936)，Girshik(1939)和 Anderson(1958)给出，但表达式很复杂. 以下我们给出 Λ 的近似分布.

利用矩阵行列式及其分块行列式的关系，可得出

$$\begin{aligned}|S|&=|S_{22}|\cdot|S_{11}-S_{12}S_{22}^{-1}S_{21}|\\&=|S_{22}|\cdot|S_{11}|\cdot|I_p-S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}|.\end{aligned}$$

故

$$\Lambda=|I_p-S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}|=\prod_{i=1}^{p}(1-\hat{\lambda}_i^2),$$

其中 $\hat{\lambda}_i^2$ 是 $\hat{T}\hat{T}'$ 的特征值($\hat{T}=S_{11}^{-1/2}S_{12}S_{22}^{-1/2}$). 由 Λ 统计量出发导出检验 H_0 的近似检验方法，例如威尔克斯的 λ 统计量、Pillai 的迹、Hotelling-Lawley 迹和 Roy 的极大根等. 现采用 Box (1949)给出的检验方法：当 $n\to\infty$ 时，在 H_0 成立时有

$$P\{-m\ln\Lambda\leqslant C\}=P\{V\leqslant C\},$$

其中 $m=n-1-\frac{1}{2}(p+q+1)$，$f=pq$，V 为服从 $\chi^2(f)$ 的统计量. 当样本容量 n 足够大时，由样本值计算样本典型相关系数 $\hat{\lambda}_i^2$ $(i=1,2,\cdots,p)$ 及

$$Q_1=-m\ln\prod_{i=1}^{p}(1-\hat{\lambda}_i^2)=-m\sum_{i=1}^{p}\ln(1-\hat{\lambda}_i^2),$$

$$p = P\{V > Q_1\} \quad (V \sim \chi^2(f)).$$

如果 $p<\alpha$(比如取 $\alpha=0.05$),则否定 H_0,即两组变量 X 和 Y 相关;否则 H_0 相容.

2. 检验 $H_0^{(k)}$: $\lambda_k=0$ $(k=2,\cdots,p)$

当否定 H_0 时,表明 X,Y 相关,进而可得出至少第一个典型相关系数 $\lambda_1\neq 0$,相应的第一对典型相关变量 V_1,W_1 可能已经提取了两组变量相关关系的绝大部分信息.两组变量余下的部分可认为不相关,这时 $\lambda_i\approx 0$ $(k=2,\cdots,p)$. 故在否定 H_0 后,有必要再检验 $H_0^{(k)}$ $(k=2,\cdots,p)$,即第 k 个及以后的所有典型相关系数均为 0 $(k=2,3,\cdots,p)$.

这时采用 Bartlett 提出的大样本 χ^2 检验,取检验统计量为

$$Q_k = -\left[n-k-\frac{1}{2}(p+q+1)\right]\sum_{i=k}^{p}\ln(1-\hat{\lambda}_i^2),$$

由样本值计算样本典型相关系数 $\hat{\lambda}_i^2$ $(i=1,2,\cdots,p)$及 Q_k 值,并计算显著性概率 $p_k=P\{V>Q_k\}$,其中 $V\sim\chi^2(f_k)$, $f_k=(p-k+1)(q-k+1)$. 如果 $p_k<\alpha$(比如取 $\alpha=0.05$),则否定 $H_0^{(k)}$,即第 k 个典型相关系数显著的不等于 0. 否则认为 $\lambda_k=0$. 对 $H_0^{(k)}$ 从 $k=2$ 开始逐个检验,直到某个 k_0,使 $H_0^{(k_0)}$ 相容时为止. 这时说明第 k_0 个及以后的所有典型相关系数均为 0.

三、样本典型变量的得分值

假设经检验,有 r $(r\leqslant p)$个典型相关系数显著不等于 0,这时可得 r 对典型相关变量(V_i,W_i) $(i=1,\cdots,r)$. 将样品

$$Z_{(t)} = \begin{bmatrix} X_{(t)} \\ Y_{(t)} \end{bmatrix}$$

代入第 i 对典型变量中,令

$$v_{it} = a_i'(X_{(t)} - \overline{X}) \quad (i=1,2,\cdots,r),$$

$$w_{it} = b_i'(Y_{(t)} - \overline{Y}) \quad (t=1,2,\cdots,n).$$

称(v_{it},w_{it})为第 t 个样品 $Z_{(t)}$的第 i 对**样本典型变量的得分值**.

对每个 i，可用 (v_{ti}, w_{ti}) $(t=1,2,\cdots,n)$ 来绘制散点的散布图，散点应近似在一条直线上，若有异常点，则应分析原因.

例 10.2.1　为了了解某矿区下部矿 Pt(铂)、Pd(钯)与 Cu(铜)、Ni(镍)的共生组合规律. 我们从其钻孔中取出 27 个样品(数据见表 10.1). 试用典型相关分析研究 Pt、Pd 与 Cu、Ni 的相关关系.

解　在此例中 $p=q=2$，$n=27$. 我们使用 SAS/STAT 软件中 CANCORR 过程进行典型相关分析. 部分输出结果见输出 10.2.1.

输出 10.2.1 中包含三个表格，位于最上面的表格给出第一典型相关系数为 $\lambda_1=0.8946$，它比两组变量间最大的相关系数(即 $\rho(X_2, Y_2)=0.8691$)都大.

表 10.1　矿区下部矿 Pt、Pd 与 Cu、Ni 的数据

序号	Pt X_1	Pd X_2	Cu Y_1	Ni Y_2	序号	Pt X_1	Pd X_2	Cu Y_1	Ni Y_2
1	0.14	0.30	0.03	0.14	15	0.43	0.90	0.13	0.22
2	0.20	0.50	0.14	0.22	16	0.47	0.97	0.26	0.22
3	0.06	0.11	0.03	0.02	17	0.49	0.79	0.21	0.20
4	0.07	0.11	0.04	0.13	18	0.47	0.77	0.51	0.22
5	0.12	0.22	0.06	0.12	19	0.40	0.88	0.33	0.19
6	0.52	0.87	0.19	0.20	20	0.66	1.30	0.21	0.30
7	0.23	0.47	0.14	0.10	21	0.63	1.30	0.45	0.28
8	1.19	0.38	0.09	0.11	22	0.52	1.43	0.31	0.23
9	0.37	0.66	0.14	0.15	23	0.44	0.87	0.17	0.25
10	0.36	0.60	0.12	0.14	24	0.03	0.07	0.05	0.08
11	0.42	0.77	0.17	0.10	25	0.20	0.28	0.04	0.08
12	0.35	0.85	0.30	0.19	26	0.04	0.10	0.11	0.07
13	0.50	0.87	0.23	0.22	27	0.17	0.28	0.15	0.09
14	0.56	1.15	0.29	0.28					

位于输出 10.2.1 中间的表格给出检验假设 $H_0^{(i)}$：$\lambda_i=0$ $(i=1,2)$ 的结果. 由该表格中“1”所在的这一行，可得到似然比的值为 0.19964，近似 F 统计量为 14.2379，显著性概率(p 值)为 0.0001(即表格中最右列 Pr>F 的值)，故在 $\alpha=0.01$ 的显著性水平下，否定所有典型相关为 0 的假设，也就是至少有一个典型相关是显著的. 由“2”所在的这一行，得到的结果说明第二个典型相关是不显著的 ($p=0.9633>\alpha$).

输出 10.2.1 典型相关系数及显著性检验

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.894618	0.889469	0.039156	0.800341
2	0.009496	.	0.196098	0.000090

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.19964137	14.2379	4	46	0.0001
2	0.99990983	0.0022	1	24	0.9633

Multivariate Statistics and F Approximations

S=2 M=-0.5 N=10.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.19964137	14.2379	4	46	0.0001
Pillai's Trace	0.80043080	8.0072	4	48	0.0001
Hotelling-Lawley Trace	4.00862032	22.0474	4	44	0.0001
Roy's Greatest Root	4.00853014	48.1024	2	24	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

由 CANCORR 过程还可以生成许多输出结果(省略),如由这两组标准化变量可得到的典型相关变量,第一对标准化典型变量(X_i^* 和 Y_j^* 表示 X_i 和 Y_j 的标准化变量)为

$$V_1 = -0.0263X_1^* + 1.0161X_2^*,$$

$$W_1 = 0.3231Y_1^* + 0.7514Y_2^*.$$

来自第一组变量的第一个典型变量 V_1 主要代表 X_2^*(即元素 Pd);而来自第二组变量的典型变量 W_1 代表 Y_2^*(Ni)和 Y_1^*(Cu).

§10.3 典型冗余分析

由样本观测数据阵 Z 计算样本协方差阵

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

由 S 矩阵求出样本典型变量后,进一步可以来计算原始变量与 r 对典型变量之间的相关系数矩阵(或称典型结构).

假定两组原始变量均为标准化变量(即 $S=R$). 若记原始变量 X (或 Y)与典型变量 V(或 W)的相关阵为

$$R(X,V)=R_{11}A=(R_{11}a_1,\cdots,R_{11}a_p)$$
$$\xrightarrow{\text{def}}\begin{bmatrix} r(X_1,V_1) & \cdots & r(X_1,V_p)\\ \vdots & & \vdots\\ r(X_p,V_1) & \cdots & r(X_p,V_p)\end{bmatrix}_{p\times p},$$
$$R(Y,W)=R_{22}B=(R_{22}b_1,\cdots,R_{22}b_p)$$
$$\xrightarrow{\text{def}}\begin{bmatrix} r(Y_1,W_1) & \cdots & r(Y_1,W_p)\\ \vdots & & \vdots\\ r(Y_q,W_1) & \cdots & r(Y_q,W_p)\end{bmatrix}_{q\times p},$$

分别计算两组原始变量 X,Y 与典型变量 V,W 之间的相关系数阵中各列相关系数的平方和,还将得出下面一些有关的概念.

一、几个概念

设 $\operatorname{rank}(\Sigma_{12})=r\leqslant\min(p,q)$. 类似于主成分分析,把 V_k 看成是由第一组标准化变量 X 提取的成分, W_k 看成是由第二组标准化变量 Y 提取的成分,由相关阵 $R(X,V)=R_{11}A=[r(X_j,V_k)]_{p\times r}$ 和 $R(Y,W)=R_{22}B=[r(Y_j,W_k)]_{q\times r}$,分别计算第 k 列平方和除以原变量组变差总和 p 或 q. 记

$$R_d(X;V_k)=\frac{1}{p}\sum_{j=1}^{p}r^2(X_j,V_k),$$

$$R_d(Y;W_k)=\frac{1}{q}\sum_{j=1}^{q}r^2(Y_j,W_k)\quad(k=1,\cdots,r),$$

并称 $R_d(X;V_k)$(或 $R_d(Y;W_k)$)为第 k 个典型变量 V_k(或 W_k)解释本组变量 X(或 Y)**总变差的百分比**. 记

$$R_d(X;V_1,\cdots,V_m)=\frac{1}{p}\sum_{k=1}^{m}\sum_{j=1}^{p}r^2(X_j,V_k),$$

$$R_d(Y;W_1,\cdots,W_m)=\frac{1}{q}\sum_{k=1}^{m}\sum_{j=1}^{q}r^2(Y_j,W_k),$$

并称 $R_d(X;V_1,\cdots,V_m)$(或 $R_d(Y;W_1,\cdots,W_m)$)为前 m $(m\leqslant r)$个典型变量 $V_1,\cdots,V_m$(或 $W_1,\cdots,W_m$)解释本组变量 X(或 Y)**总变差的累计百分比**.

在典型相关分析中，从两组变量分别提取的两个典型成分首先要求相关程度最大，同时也希望每个典型成分解释各组变差的百分比也尽可能的大. 百分比的多少反映由每组变量提取的用于典型相关分析的变差的多少.

类似于主成分分析，还可以引入前 m 个典型变量对本组第 j 个变量 X_j(或 Y_j)的贡献等概念(见参考文献[11]).

二、典型冗余分析

我们进一步来讨论典型变量解释另一组变量总变差百分比的问题. 在典型相关分析中，因所提取的每对典型成分保证其相关程度达最大，故每个典型成分不仅解释了本组变量的信息，还解释了另一组变量的信息. 典型相关系数越大，典型成分解释对方变量组变差的信息也将越多.

类似可以定义 $R_d(X;W_k)$(或 $R_d(Y;V_k)$)为 W_k(或 V_k)解释另一组总变差的百分比. 以下给出利用典型变量解释本组变差的百分比来计算解释另一组变差百分比的公式：

$$\begin{aligned} R_d(X;W_k) &= \frac{1}{p}\sum_{j=1}^{p} r^2(X_j,W_k) \\ &= \lambda_k^2 R_d(X;V_k) \quad (k=1,\cdots,r), \\ R_d(Y;V_k) &= \frac{1}{q}\sum_{j=1}^{q} r^2(Y_j,V_k) \\ &= \lambda_k^2 R_d(Y;W_k) \quad (k=1,\cdots,r). \end{aligned}$$

事实上，由样本典型变量的系数 a_k 与 b_k 的公式(10.2.1)还可以得出 a_k 与 b_k 之间的关系

$$\begin{aligned} b_k = \frac{1}{\lambda_k} S_{22}^{-1} S_{21} a_k &\iff \lambda_k b_k = S_{22}^{-1} S_{21} a_k \\ &\iff \lambda_k S_{22} b_k = S_{22} S_{22}^{-1} S_{21} a_k = S_{21} a_k, \end{aligned}$$

以及典型变量与原始变量(假定已标准化)的相关阵，即得

$$r(Y_j,V_k) = \lambda_k r(Y_j;W_k),$$

故有

$$R_d(Y;V_k) = \lambda_k^2 R_d(Y;W_k).$$

类似还可证明另一式.

$R_d(X;W_k)$表示第一组中典型变量解释原变量组的变差被第二组中典型变量重复解释的百分比,简称为第一组典型变量的**冗余测度**; $R_d(Y;V_k)$表示第二组中典型变量解释原变量组的变差被第一组中典型变量重复解释的百分比,简称为第二组典型变量的冗余测度.冗余测度体现了两组变量之间的相关程度.

冗余测度的大小表示这对典型变量能够对另一组变差相互解释的程度大小,它将为进一步讨论多对多建模提供一些有用的信息. SAS/STAT 软件中的 CANCORR 过程可完成典型冗余分析.

例 10.3.1 (康复俱乐部 20 名成员测试数据的典型相关分析) 康复俱乐部对 20 名中年人测量了三个生理指标: WEIGHT(体重)、WAIST(腰围)、PULSE(脉搏),以及三个训练指标: CHINS(单杠)、SITUPS(仰卧起坐)和 JUMPS(跳高). 数据见表 10.2. 试分析生理指标和训练指标这两组变量间的相关性.

解　在此例中 $p=q=3, n=20$. 我们使用 SAS/STAT 软件中的 CANCORR 过程对表 10.2 的数据进行典型相关分析.

表 10.2　康复俱乐部 20 名成员的测试数据

体重	腰围	脉搏	单杠	仰卧起坐	跳高	体重	腰围	脉搏	单杠	仰卧起坐	跳高
191	36	50	5	162	60	189	37	52	2	110	60
193	38	58	12	101	101	162	35	62	12	105	37
189	35	46	13	155	58	182	36	56	4	101	42
211	38	56	8	101	38	167	34	60	6	125	40
176	31	74	15	200	40	154	33	56	17	251	250
169	34	50	17	120	38	166	33	52	13	210	115
154	34	64	14	215	105	247	46	50	1	50	50
193	36	46	6	70	31	202	37	62	12	210	120
176	37	54	4	60	25	157	32	52	11	230	80
156	33	54	15	225	73	138	33	68	2	110	43

计算结果首先输出 6 个变量的均值、标准差及生理指标和训练指标之间的相关系数阵(省略了). 生理指标和训练指标之间的相关性都为中等,其中 WAIST(腰围)和 SITUPS(仰卧起坐)的相关系数最大,为 -0.6456.

输出 10.3.1 典型相关系数及显著性检验

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.795608	0.754056	0.084197	0.632992
2	0.200556	-.076399	0.220188	0.040223
3	0.072570	.	0.228208	0.005266

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.35039053	2.0482	9	34.22293	0.0635
2	0.95472266	0.1758	4	30	0.9491
3	0.99473355	0.0847	1	16	0.7748

输出 10.3.1 给出典型相关分析的一般结果. 第一典型相关系数为 0.7956,它比生理指标和训练指标两组间的任一个相关系数都大. 检验总体中所有典型相关系数均为 0 的零假设时显著性概率为 0.0635 (即 Pr>F 的值),故在 $\alpha=0.10$ 的显著性水平下,否定所有典型相关为 0 的假设,也就是至少有一个典型相关系数是显著的. 从后面的检验结果可知,只有第一典型相关系数是显著不等于 0 的. 因此,两组变量相关性的研究可转化为研究第一对典型相关变量的相关性.

输出结果中还给出原始变量和标准化变量的典型相关变量的系数. 因 6 个变量没有使用相同单位进行测量,因此我们来分析标准化后的系数. 来自生理指标的第一典型变量 V_1 为(原始变量的右上角带"*"表示为标准化变量):

$$V_1=-0.7754\ \text{WEIGHT}^*+1.5793\ \text{WAIST}^*-0.0591\ \text{PULSE}^*,$$

它近似地是 WAIST*(腰围)和 WEIGHT*(体重)的加权差,在 WAIST* 上的权重更大些,V_1 在 PULSE*(脉搏)上系数近似为 0. 来自训练指标的第一典型变量 W_1 为

$$W_1=-0.3495\ \text{CHINS}^*-1.0540\ \text{SITUPS}^*+0.7164\ \text{JUMPS}^*,$$

它在 SITUPS*(仰卧起坐)上的系数最大. 这一对典型变量主要是反映 WAIST*(腰围)和 SITUPS*(仰卧起坐)的负相关关系.

由输出 10.3.2 又可看出,来自生理指标的第一典型变量 V_1 与

WAIST(腰围),以及 V_1 与 WEIGHT(体重)的相关系数分别为 0.9254,0.6206,都是正的.但在典型变量 V_1 的表示式中 WEIGHT* 的系数为负的(−0.7754),即 WEIGHT 在 V_1 表示式中的系数和它与 V_1 的相关系数反号.来自训练指标的第一典型变量 W_1 与三个训练指标的相关系数都是负值,其中 JUMPS(跳高)在 W_1 表示式中的系数(0.7164)和它与 W_1 的相关系数(−0.1622)也是反号的.因此,WEIGHT 和 JUMPS 在这两组变量中是一个校正(或抑制)变量.

输出 10.3.2　典型结构——原始变量和典型变量的相关系数阵

Canonical Structure

Correlations Between the 生理指标 and Their Canonical Variables

	V1	V2	V3	
WEIGHT	0.6206	-0.7724	-0.1350	体重
WAIST	0.9254	-0.3777	-0.0310	腰围
PULSE	-0.3328	0.0415	0.9421	脉搏

Correlations Between the 训练指标 and Their Canonical Variables

	W1	W2	W3	
CHINS	-0.7276	0.2370	-0.6438	单杠
SITUPS	-0.8177	0.5730	0.0544	仰卧起坐
JUMPS	-0.1622	0.9586	-0.2339	跳高

Correlations Between the 生理指标 and the Canonical Variables of the 训练指标

	W1	W2	W3	
WEIGHT	0.4938	-0.1549	-0.0098	体重
WAIST	0.7363	-0.0757	-0.0022	腰围
PULSE	-0.2648	0.0083	0.0684	脉搏

Correlations Between the 训练指标 and the Canonical Variables of the 生理指标

	V1	V2	V3	
CHINS	-0.5789	0.0475	-0.0467	单杠
SITUPS	-0.6506	0.1149	0.0040	仰卧起坐
JUMPS	-0.1290	0.1923	-0.0170	跳高

输出 10.3.3 给出典型冗余分析的结果.我们来分析标准化的方差,第一典型变量 V_1 可以解释 45.08%组内变差,并解释 25.84%的另一组(训练指标)的变差;而典型变量 W_1 可以解释 40.81%组内变差,并解释 28.54%的另一组(生理指标)的变差.可见第一对典型变量 V_1 和 W_1 都不能很好地全面地预测另一组变量.第二和第三对典型变量实际上都没有给出什么信息,三个典型变量解释另一组总变差的累计百分比分别为 0.2969 和 0.2767.

位于输出 10.3.3 中的第 4 个表格，给出训练指标组中各个变量被生理指标变量组提取的前 M 个($M=1,2,3$)典型变量 $V_1,\cdots,V_M$ 解释变差的累计百分比，即多重相关的平方和为：

$$\sum_{k=1}^{M} r^2(Y_j,V_k),$$

可以看出，只有 CHINS(单杠) (0.3351)和 SITUPS(仰卧起坐) (0.4233)可被对方变量组的第一典型变量 V_1 预测，V_1 对 JUMPS (跳高) (0.0167)几乎没有预测能力. 从该输出中的第 3 个表格，可以类似地得出，来自训练指标的第一典型变量 W_1 对 WAIST(腰围) (0.5421)有相当好的预测能力，对 WEIGHT(体重) (0.2438)较差，而对 PULSE(脉搏) (0.0701)几乎没有预测能力.

输出 10.3.3 CANCORR 过程产生的典型冗余分析结果

Standardized Variance of the 生理指标 Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.4508	0.4508	0.6330	0.2854	0.2854
2	0.2470	0.6978	0.0402	0.0099	0.2953
3	0.3022	1.0000	0.0053	0.0016	0.2969

Standardized Variance of the 训练指标 Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.4081	0.4081	0.6330	0.2584	0.2584
2	0.4345	0.8426	0.0402	0.0175	0.2758
3	0.1574	1.0000	0.0053	0.0008	0.2767

Squared Multiple Correlations Between the 生理指标 and the First 'M' Canonical Variables of the 训练指标

M	1	2	3	
WEIGHT	0.2438	0.2678	0.2679	体重
WAIST	0.5421	0.5478	0.5478	腰围
PULSE	0.0701	0.0702	0.0749	脉搏

Squared Multiple Correlations Between the 训练指标 and the First 'M' Canonical Variables of the 生理指标

M	1	2	3	
CHINS	0.3351	0.3374	0.3396	单杠
SITUPS	0.4233	0.4365	0.4365	仰卧起坐
JUMPS	0.0167	0.0536	0.0539	跳高

习　题　十

10-1　设标准化变量 $X=(X_1,X_2)'$，$Y=(Y_1,Y_2)'$. 已知 $Z=\begin{bmatrix}X\\Y\end{bmatrix}$ 的相关阵为

$$R=\left[\begin{array}{cc:cc}1.0 & 0.5 & 0.7 & 0.7\\0.5 & 1.0 & 0.7 & 0.7\\\hdashline 0.7 & 0.7 & 1.0 & 0.6\\0.7 & 0.7 & 0.6 & 1.0\end{array}\right]\overset{\text{def}}{=\!=}\begin{bmatrix}R_{11} & R_{12}\\R_{21} & R_{22}\end{bmatrix}.$$

试求 X,Y 的典型相关变量和典型相关系数.

10-2　在 140 个学生中进行了阅读速度 X_1，阅读能力 X_2，运算速度 X_3 和运算能力 X_4 共 4 种测验，由所得测验成绩算出相关系数阵为

$$R=\begin{bmatrix}1.00 & 0.63 & 0.24 & 0.59\\0.63 & 1.00 & -0.06 & 0.07\\0.24 & -0.06 & 1.00 & 0.42\\0.59 & 0.07 & 0.42 & 1.00\end{bmatrix},$$

试分析学生的阅读能力和运算能力之间的相关程度.

10-3　在某年级 44 名学生的期末考试中，有的课程用闭卷，有的课程用开卷(考试成绩见表 8.3). 试对闭卷(X_1,X_2)和开卷(X_3，X_4,X_5)两组变量进行典型相关分析.

10-4　表 10.3 中是从 25 个家庭中测到的成年长子和次子的头宽、头长的数据. 试用典型相关分析方法分析长子和次子头宽、头长的相关情况.

表 10.3　成年长子和次子的头宽、头长的数据

样品号	长子头长 (X_1)	长子头宽 (X_2)	次子头长 (X_3)	次子头宽 (X_4)	样品号	长子头长 (X_1)	长子头宽 (X_2)	次子头长 (X_3)	次子头宽 (X_4)
1	191	155	179	145	14	190	159	195	157
2	195	149	201	152	15	188	151	187	158

（续表）

样品号	长子头长 (X_1)	长子头宽 (X_2)	次子头长 (X_3)	次子头宽 (X_4)	样品号	长子头长 (X_1)	长子头宽 (X_2)	次子头长 (X_3)	次子头宽 (X_4)
3	181	148	185	149	16	163	137	161	130
4	183	153	188	149	17	195	155	183	158
5	176	144	171	142	18	186	153	173	148
6	208	157	192	152	19	181	145	182	146
7	189	150	190	149	20	175	140	165	137
8	197	159	189	152	21	192	154	185	152
9	188	152	197	159	22	174	143	178	147
10	192	150	187	151	23	176	139	176	143
11	179	158	186	148	24	197	167	200	158
12	183	147	174	147	25	190	163	187	150
13	174	150	185	152					

10-5 某学校为研究学生的体质与运动能力的关系，对 38 名学生的体质情况，每人测试了 7 项指标：X_1（反复横荡的次数）、X_2（纵跳高度）、X_3（背力）、X_4（握力）、X_5（踏台升降指数）、X_6（立姿体前屈）、X_7（卧姿上体后仰）；对运动能力情况每人测试了 5 项指标：X_8（50 米跑）、X_9（1000 米长跑）、X_{10}（投掷）、X_{11}（悬垂次数）、X_{12}（持久走）. 7 项体质数据和 5 项运动数据见表 10.4. 试对这两组数据进行典型相关分析.

表 10.4 学生体质与运动能力数据

学生	体质情况							运动能力				
序号	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
1	46	55	126	51	75.0	25	72	6.8	489	27	8	360
2	52	55	95	42	81.2	18	50	7.2	464	30	5	348
3	46	69	107	38	98.0	18	74	6.8	430	32	9	386
4	49	50	105	48	97.6	16	60	6.8	362	26	6	331
5	42	55	90	46	66.5	2	68	7.2	453	23	11	391
6	48	61	106	43	78.0	25	58	7.0	405	29	7	389
7	49	60	100	49	90.6	15	60	7.0	420	21	10	379
8	48	63	122	52	56.0	17	68	7.0	466	28	2	362
9	45	55	105	48	76.0	15	61	6.8	415	24	6	386
10	48	64	120	38	60.2	20	62	7.0	413	28	7	398
11	49	52	100	42	53.4	6	42	7.4	404	23	6	400
12	47	62	100	34	61.2	10	62	7.2	427	25	7	407

（续表）

学生序号	体质情况							运动能力				
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
13	41	51	101	53	62.4	5	60	8.0	372	25	3	409
14	52	55	125	43	86.3	5	62	6.8	496	30	10	350
15	45	52	94	50	51.4	20	65	7.6	394	24	3	399
16	49	57	110	47	72.3	19	45	7.0	446	30	11	337
17	53	65	112	47	90.4	15	75	6.6	420	30	12	357
18	47	57	95	47	72.3	9	64	6.6	447	25	4	447
19	48	60	120	47	86.4	12	62	6.8	398	28	11	381
20	49	55	113	41	84.1	15	60	7.0	398	27	4	387
21	48	69	128	42	47.9	20	63	7.0	485	30	7	350
22	42	57	122	46	54.2	15	63	7.2	400	28	6	388
23	54	64	155	51	71.4	19	61	6.9	511	33	12	298
24	53	63	120	42	56.6	8	53	7.5	430	29	4	353
25	42	71	138	44	65.2	17	55	7.0	487	29	9	370
26	46	66	120	45	62.2	22	68	7.4	470	28	7	360
27	45	56	91	29	66.2	18	51	7.9	380	26	5	358
28	50	60	120	42	56.6	8	57	6.8	460	32	5	348
29	42	51	126	50	50.0	13	57	7.7	398	27	2	383
30	48	50	115	41	52.9	6	39	7.4	415	28	6	314
31	42	52	140	48	56.3	15	60	6.9	470	27	11	348
32	48	67	105	39	69.2	23	60	7.6	450	28	10	326
33	49	74	151	49	54.2	20	58	7.0	500	30	12	330
34	47	55	113	40	71.4	19	64	7.6	410	29	7	331
35	49	74	120	53	54.5	22	59	6.9	500	33	21	348
36	44	52	110	37	54.9	14	57	7.5	400	29	2	421
37	52	66	130	47	45.9	14	45	6.8	505	28	11	355
38	48	68	100	45	53.6	23	70	7.2	522	28	9	352

第十一章　偏最小二乘回归分析

在实际问题中，经常遇到需要研究两组多重相关变量间的相互依赖关系，并研究用一组变量（常称为自变量或预测变量）去预测另一组变量（常称为因变量或响应变量），除了使用最小二乘准则下的经典多元线性回归分析（MLR），提取自变量组主成分的主成分回归分析（PCR）等方法外，还有近年发展起来的偏最小二乘回归分析（PLS）方法.

偏最小二乘回归分析提供一种多对多线性回归建模的方法，特别当两组变量的个数很多，且都存在多重相关性，而观测数据的数量（样本量）又较少时，用偏最小二乘回归分析建立的模型具有传统的经典回归分析等方法所没有的优点.

偏最小二乘回归分析在建模过程中集中了主成分分析、典型相关分析和线性回归分析方法的特点，因此在分析结果中，除了可以提供一个更为合理的回归模型以外，还可以同时完成一些类似于主成分分析和典型相关分析的研究内容，提供更加丰富、并深入的一些信息.

本章结合 SAS/STAT 软件中用于完成偏最小二乘回归分析的 PLS 过程，介绍偏最小二乘回归分析的建模方法；并通过例子从预测角度对所建立的回归模型进行比较.

§11.1　偏最小二乘回归分析方法

考虑 p 个因变量 $Y_1,\cdots,Y_p$ 与 m 个自变量 $X_1,\cdots,X_m$ 的建模问题. 偏最小二乘回归分析的基本作法是，首先在自变量集中提取第一成分 T_1（T_1 是 $X_1,\cdots,X_m$ 的线性组合，且尽可能多地提取原自变量集中的变异信息）；同时在因变量集中也提取第一成分 U_1，并要求

T_1 与 U_1 相关程度达最大；然后建立因变量 $Y_1,\cdots,Y_p$ 与 T_1 的回归方程. 如果回归方程已达到满意的精度，则算法终止；否则继续第二对成分的提取，直到能达到满意的精度为止. 若最终对自变量集提取 r 个成分 $T_1,T_2,\cdots,T_r$，偏最小二乘回归分析将通过建立 $Y_1,\cdots,Y_p$ 与 $T_1,T_2,\cdots,T_r$ 的回归方程，然后再表示为 $Y_1,\cdots,Y_p$ 与原自变量的回归方程，即偏最小二乘回归方程.

为了比较方便，假定 p 个因变量 $Y_1,\cdots,Y_p$ 与 m 个自变量 $X_1,\cdots,X_m$ 均为标准化变量. 因变量组和自变量组的 n 次标准化观测数据阵分别记为：

$$Y_0=\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix},\quad X_0=\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}.$$

现在介绍偏最小二乘回归分析建模的具体步骤：

(1) 分别提取两变量组的第一对成分，并使之相关性达最大. 假设从两组变量分别提取第一对成分为 T_1 和 U_1，T_1 是自变量集 $X=(X_1,\cdots,X_m)'$ 的线性组合：

$$T_1=w_{11}X_1+\cdots+w_{1m}X_m=w_1'X,$$

U_1 是因变量集 $Y=(Y_1,\cdots,Y_p)'$ 的线性组合：

$$U_1=v_{11}Y_1+\cdots+v_{1p}Y_p=v_1'Y.$$

为了回归分析的需要，要求：

① T_1 和 U_1 各自尽可能多地提取所在变量组的变异信息；

② T_1 和 U_1 的相关程度达到最大.

由两组变量集的标准化观测数据阵 X_0 和 Y_0，可以计算第一对成分的得分向量，分别记为 t_1 和 u_1：

$$t_1=X_0w_1=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}\begin{bmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1m} \end{bmatrix}=\begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix},$$

$$u_1 = Y_0 v_1 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1p} \end{bmatrix} = \begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{bmatrix}.$$

第一对成分 T_1 和 U_1 的协方差 $\mathrm{Cov}(T_1, U_1)$可用第一对成分的得分向量 t_1 和 u_1 的内积来计算. 故而上述①和②两个要求可化为数学分析中的条件极值问题：

$$\begin{cases} \langle t_1, u_1 \rangle = \langle X_0 w_1, Y_0 v_1 \rangle = w_1' X_0' Y_0 v_1 \to \text{最大}, \\ w_1' w_1 = \| w_1 \|^2 = 1, \quad v_1' v_1 = \| v_1 \|^2 = 1. \end{cases}$$

利用拉格朗日乘子法，上述问题可化为求单位向量 w_1 和 v_1，使 $\theta_1 = w_1' X_0' Y_0 v_1 \to$ 最大. 问题的求解只须通过计算 m 阶矩阵

$$M = X_0' Y_0 Y_0' X_0$$

的特征值和特征向量，且 M 的最大特征值为 θ_1^2，相应的单位特征向量就是所求的解 w_1，而 v_1 可由 w_1 计算得到：

$$v_1 = \frac{1}{\theta_1} Y_0' X_0 w_1.$$

在 SAS/STAT 软件的 PLS 过程中称 $w_1 = (w_{11}, \cdots, w_{1m})'$ 为**模型效应权重**(Model Effect Weights)；称 $v_1 = (v_{11}, \cdots, v_{1p})'$ 为**因变量权重**(Dependent Variable Weights).

(2) 建立 $Y_1, \cdots, Y_p$ 对 T_1 的，以及 $X_1, \cdots, X_m$ 对 T_1 的回归方程. 假定回归模型为

$$\begin{cases} X_0 = t_1 \alpha_1' + E_1, \\ Y_0 = t_1 \beta_1' + F_1, \end{cases}$$

其中 t_1 为 n 维得分向量，$\alpha_1' = (\alpha_{11}, \cdots, \alpha_{1m})$，$\beta_1' = (\beta_{11}, \cdots, \beta_{1p})$分别是多因变量而只有一个自变量的回归模型中的参数向量，E_1 和 F_1 分别为 $n \times m$ 和 $n \times p$ 残差阵. 回归系数向量 α_1, β_1 的最小二乘估计为

$$\begin{cases} \alpha_1' = (t_1' t_1)^{-1} t_1' X_0, \\ \beta_1' = (t_1' t_1)^{-1} t_1' Y_0, \end{cases} \quad \text{或} \quad \alpha_1 = \frac{X_0' t_1}{\| t_1 \|^2}, \ \beta_1 = \frac{Y_0' t_1}{\| t_1 \|^2}.$$

在 PLS 过程中我们称 $\alpha_1=(\alpha_{11},\cdots,\alpha_{1m})'$ 为模型效应载荷量(Model Effect Loadings).

(3) 用残差阵 E_1 和 F_1 代替 X_0 和 Y_0，然后再重复以上步骤. 记 $\hat{X}_0=t_1\alpha_1'$，$\hat{Y}_0=t_1\beta_1'$，则残差阵 $E_1=X_0-\hat{X}_0$，$F_1=Y_0-\hat{Y}_0$. 如果残差阵 F_1 中元素的绝对值近似为 0，则认为用第一个成分建立的回归方程精度已满足需要了，可以停止抽取成分. 否则用残差阵 E_1 和 F_1 代替 X_0 和 Y_0，重复以上步骤，即得：

$$w_2=(w_{21},\cdots,w_{2m})';\quad v_2=(v_{21},\cdots,v_{2p})',$$

分别为第二对成分的权重. 而 $t_2=E_1w_2$，$u_2=F_1v_2$ 为第二对成分的得分向量，

$$\alpha_2=\frac{E_1't_2}{\|t_2\|^2},\quad \beta_2=\frac{F_1't_2}{\|t_2\|^2}$$

分别为 X,Y 的第二对成分的载荷量. 这时有

$$\begin{cases}X_0=t_1\alpha_1'+t_2\alpha_2'+E_2,\\ Y_0=t_1\beta_1'+t_2\beta_2'+F_2.\end{cases}$$

(4) 设 $n\times m$ 数据阵 X_0 的秩为 $r\leqslant\min(n-1,m)$，则存在 r 个成分 $t_1,t_2,\cdots,t_r$，使得

$$\begin{cases}X_0=t_1\alpha_1'+\cdots+t_r\alpha_r'+E_r,\\ Y_0=t_1\beta_1'+\cdots+t_r\beta_r'+F_r.\end{cases}$$

设 X_i^* $(i=1,\cdots,m)$，Y_j^* $(j=1,\cdots,p)$表示标准化变量，把

$$t_k=w_{k1}X_1^*+\cdots+w_{km}X_m^*\quad(k=1,\cdots,r)$$

代入

$$Y_j^*=t_1\beta_{1j}+t_2\beta_{2j}+\cdots+t_r\beta_{rj}\quad(j=1,2,\cdots,p),$$

即得 p 个标准化因变量的偏最小二乘回归方程：

$$\hat{Y}_j^*=a_{j1}^*X_1^*+\cdots+a_{jm}^*X_m^*\quad(j=1,\cdots,p).$$

然后再还原为原始变量的偏最小二乘回归方程：

$$\hat{Y}_j=a_{j0}+a_{j1}X_1+\cdots+a_{jm}X_m\quad(j=1,\cdots,p).$$

(5) 确定抽取成分的个数 l. 一般情况下，偏最小二乘回归分析

并不需要选用存在的所有 r 个成分 $t_1,t_2,\cdots,t_r$ 来建立回归方程，而像主成分分析一样，只选用前 l 个成分($l\leqslant r$)，即可得到预测能力较好的回归模型. 下面讨论确定抽取成分个数 l 的几种方法：

① “舍一交叉验证方法”：每次舍去第 i 个观测($i=1,\cdots,n$)，用余下的 $n-1$ 个观测按偏最小二乘回归方法建模，并考虑抽取 k 个成分后拟合的回归方程，然后把舍去的第 i 个观测点代入所拟合的回归方程，得到 Y_j ($j=1,\cdots,p$)在第 i 个观测点上的预测值 $\hat{y}_{j(i)}(k)$. 对 $i=1,2,\cdots,n$ 重复以上的验证，即得抽取 k 个成分时第 j 个因变量 Y_j ($j=1,\cdots,p$)的预测残差平方和(PRESS)为

$$\mathrm{PRESS}_j(k)=\sum_{i=1}^{n}(y_{ij}-\hat{y}_{j(i)}(k))^2 \quad (j=1,\cdots,p).$$

$Y=(Y_1,\cdots,Y_p)'$ 的预测残差平方和为

$$\mathrm{PRESS}(k)=\sum_{j=1}^{p}\mathrm{PRESS}_j(k).$$

先对抽取成分的个数 k 从 1 至 r 个逐个计算 Y 的预测残差平方和 $\mathrm{PRESS}(k)$，然后选取使 Y 的预测残差平方和达到最小值的 k，让 $l=k$.

② “分批交叉验证方法”：每次扣留连续的 q 个观测作为检验数据集，$q=1$ 时就是“舍一交叉验证方法”，类似地，按使预测残差平方和达到最小的准则确定抽取成分的个数.

③ “分裂样本(Split-Sample)交叉验证方法”：此方法中所扣留的用以作为检验数据集的观测不必是连续的，而是按一定宽度抽取而构成的. 例如第一次扣留的观测为{1,11,21,…}，然后是{2,12,22,…}等等.

④ “随机样本交叉验证方法”：此方法中所扣留的用以作为检验数据集的观测可以是随机抽取.

以上方法是 SAS/STAT 软件中 PLS 过程提供的，此外还有其它的方法(见参考文献[11]). 在实际应用中这些方法所确定的成分个数也不完全一致，最后确定成分的个数可综合各种验证的结果及从理论上给出的检验方法.

§11.2　应用例子

例 11.2.1　康复俱乐部对 20 名中年人测量了三个生理指标：weight(体重)、waist(腰围)、pulse(脉搏)，并测量了三个训练指标：chins(单杠)、situps(仰卧起坐)、jumps(跳高)(测量数据见第十章表 10.2). 试用偏最小二乘回归方法建立由三个生理指标分别预测三个训练指标的回归模型，并对所得的计算结果进行分析.

解　使用 SAS/STAT 软件中 PLS 过程来完成偏最小二乘回归分析，并对标准化数据进行分析计算. 输出的数据集中包括成分的得分向量：$t_1,\cdots,t_l,u_1,\cdots,u_l$，以及偏最小二乘回归方程对 p 个因变量的预测结果，其部分结果见输出 11.2.1 至输出 11.2.3.

从输出 11.2.1 可以看出，由自变量组抽取的第 k 个成分 T_k ($k=1,2,3$)可解释 $X=(X_1,\cdots,X_m)'$(或称模型效应)变差的百分比分别为 69.4781%，22.6694%和 7.8525%. 而 T_k ($k=1,2,3$)可解释因变量组 $Y=(Y_1,\cdots,Y_p)'$变差的百分比分别为 20.9447%，2.9491%和 3.7718%. 由此可见 T_2,T_3 对 Y 的解释能力已非常微弱了. 从这里也可以初步直观地看出，只需抽取一个成分 T_1 就已足够了.

输出 11.2.1　被偏最小二乘因子解释的变差的百分比

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	69.4781	69.4781	20.9447	20.9447
2	22.6694	92.1475	2.9491	23.8938
3	7.8525	100.0000	3.7718	27.6656

输出 11.2.2 给出模型效应的权重 w_k ($k=1,2,3$)和因变量的权重 v_k ($k=1,2,3$). 比如 $k=1$ 时可得出 T_1 和 U_1 为

$$T_1=w_1'=-0.5985\text{weight}^*-0.7826\text{waist}^*+0.2423\text{pulse}^*,$$

$$U_1=v_1'Y=0.6133\text{chins}^*+0.7470\text{situps}^*+0.2567\text{jumps}^*.$$

输出 11.2.2 模型效应和因变量的权重

Model Effect Weights

Number of Extracted Factors	weight	waist	pulse	Inner Regression Coefficients
1	-0.598464	-0.782550	0.242348	0.549051
2	0.584054	-0.707666	-0.842793	0.360681
3	0.657469	-0.287058	0.696658	0.693060

Dependent Variable Weights

Number of Extracted Factors	chins	situps	jumps
1	0.613307	0.746972	0.256685
2	0.748517	0.647049	0.145086
3	0.688603	0.657104	-0.306659

输出 11.2.3 模型效应的负荷量

Model Effect Loadings

Number of Extracted Factors	weight	waist	pulse
1	-0.656377	-0.666340	0.353780
2	-0.015877	-0.284692	-0.958488
3	0.657469	-0.287058	0.696658

输出 11.2.3 给出模型效应的负荷量 α_k $(k=1,2,3)$. 例如

$$\alpha_1=(-0.656377,-0.666340,0.353780)'$$

是生理指标变量 weight*，waist*和 pulse*关于 T_1 的回归系数.

因没有指定抽取成分(即偏最小二乘因子)的个数,以上只给出所有可能成分 $r=3$ 的结果. 这时所得到的偏最小二乘回归方程就是最小二乘准则下多对多线性回归方程(省略了). 如果指定抽取成分的个数为 2,所得到的偏最小二乘回归方程见输出 11.2.4.

由输出 11.2.4 可得出因变量 situps 的标准化回归方程(右上角带“*”的变量表示标准化变量)及还原为原始变量的回归方程分别为:

$$\text{situps}^*=-0.1385\text{weight}^*-0.5244\text{waist}^*-0.0854\text{pulse}^*,$$

$$\text{situps}=612.5671-0.3509\text{weight}-10.2477\text{waist}-0.7412\text{pulse}.$$

经进一步计算可知，回归方程是显著的($p=0.0155<0.05$)，决定系数 $R^2=0.3876$. 回归方程中 weight 和 waist 系数的符号皆为负，这是合乎实际意义的. 我们还可以用 REG 过程(在 MODEL 语句后加选项 R)得出由“舍一交叉验证方法”而得到的预测残差平方和：PRESS=17.14692(这里没有显示).

输出 11.2.4　抽取成分个数为 2 时偏最小二乘回归方程的参数估计

Parameter Estimates for Centered and Scaled Data

	chins	situps	jumps
Intercept	0.0000000000	0.0000000000	0.0000000000
weight	-.0777704452	-.1384668393	-.0603559018
waist	-.4989281575	-.5244458437	-.1559181910
pulse	-.1321877334	-.0854203022	-.0072854215

Parameter Estimates

	chins	situps	jumps
Intercept	47.0197329	612.5671220	183.9848928
weight	-0.0166508	-0.3508797	-0.1253477
waist	-0.8237024	-10.2476753	-2.4969262
pulse	-0.0969133	-0.7412177	-0.0518112

输出 11.2.5　用交叉验证法确定抽取成分的个数为 1

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS
0	1.052632
1	0.996829
2	1.046888
3	1.075091

Minimum root mean PRESS	0.9968
Minimizing number of factors	1

如果在 PLS 过程中要求使用“舍一交叉验证方法”进行交叉验证，并确定抽取成分(即偏最小二乘因子)的个数. 计算结果是当抽取一个偏最小二乘因子时，得到的预测残差平方和的均方达最小(见输出 11.2.5)，其最小值为 0.9968.

当使用其他交叉验证方法时确定的因子个数也均为 1. 当抽取因子的个数为 1 时，得到的偏最小二乘回归方程的参数估计见输出

11.2.6.

输出 11.2.6 因子个数为 1 时偏最小二乘回归方程的参数估计

Parameter Estimates for Centered and Scaled Data

	chins	situps	jumps
Intercept	0.0000000000	0.0000000000	0.0000000000
weight	-.2015249745	-.2454453490	-.0843434721
waist	-.2635136754	-.3209438740	-.1102873645
pulse	0.0816076463	0.0993932254	0.0341549342

Parameter Estimates

	chins	situps	jumps
Intercept	29.2001686	430.2510765	150.4807263
weight	-0.0431468	-0.6219668	-0.1751653
waist	-0.4350463	-6.2712454	-1.7661788
pulse	0.0598306	0.8624650	0.2428971

由输出 11.2.6 可得出因变量 situps 的标准化回归方程(右上角带“*”的变量表示标准化变量)及还原为原始变量的回归方程分别为:

$$\text{situps}^* = -0.2454\text{weight}^* - 0.3209\text{waist}^* + 0.0994\text{pulse}^*,$$

$$\text{situps} = 430.2511 - 0.6220\text{weight} - 6.2712\text{waist} + 0.8625\text{pulse}.$$

经进一步计算可知,回归方程是显著的($p=0.0059<0.05$),决定系数 $R^2=0.3506$. 回归方程中 weight 和 waist 系数的符号皆为负的,这合乎实际意义. 使用“舍一交叉验证方法”而得到的预测残差平方和

$$\text{PRESS}=14.99592 \quad \text{(见输出 11.2.7)}.$$

另两个因变量的回归方程这里省略. 输出 11.2.7 给出在 REG

输出 11.2.7 当抽取成分个数为 1 时因变量 SITUPS 的预测残差平方和 PRESS

The REG Procedure
Model: nfar1
Dependent Variable: situps 仰卧起坐

Sum of Residuals	0
Sum of Squared Residuals	12.33873
Predicted Residual SS (PRESS)	14.99592

经进一步计算可知，回归方程是显著的($p=0.0155<0.05$)，决定系数 $R^2=0.3876$. 回归方程中 weight 和 waist 系数的符号皆为负，这是合乎实际意义的. 我们还可以用 REG 过程(在 MODEL 语句后加选项 R)得出由“舍一交叉验证方法”而得到的预测残差平方和：PRESS=17.14692(这里没有显示).

输出 11.2.4　抽取成分个数为 2 时偏最小二乘回归方程的参数估计

Parameter Estimates for Centered and Scaled Data

	chins	situps	jumps
Intercept	0.0000000000	0.0000000000	0.0000000000
weight	-.0777704452	-.1384668393	-.0603559018
waist	-.4989281575	-.5244458437	-.1559181910
pulse	-.1321877334	-.0854203022	-.0072854215

Parameter Estimates

	chins	situps	jumps
Intercept	47.0197329	612.5671220	183.9848928
weight	-0.0166508	-0.3508797	-0.1253477
waist	-0.8237024	-10.2476753	-2.4969262
pulse	-0.0969133	-0.7412177	-0.0518112

输出 11.2.5　用交叉验证法确定抽取成分的个数为 1

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS
0	1.052632
1	0.996829
2	1.046888
3	1.075091

Minimum root mean PRESS	0.9968
Minimizing number of factors	1

如果在 PLS 过程中要求使用“舍一交叉验证方法”进行交叉验证，并确定抽取成分(即偏最小二乘因子)的个数. 计算结果是当抽取一个偏最小二乘因子时，得到的预测残差平方和的均方达最小(见输出 11.2.5)，其最小值为 0.9968.

当使用其他交叉验证方法时确定的因子个数也均为 1. 当抽取因子的个数为 1 时，得到的偏最小二乘回归方程的参数估计见输出

11.2.6.

输出 11.2.6 因子个数为 1 时偏最小二乘回归方程的参数估计

Parameter Estimates for Centered and Scaled Data

	chins	situps	jumps
Intercept	0.0000000000	0.0000000000	0.0000000000
weight	-.2015249745	-.2454453490	-.0843434721
waist	-.2635136754	-.3209438740	-.1102873645
pulse	0.0816076463	0.0993932254	0.0341549342

Parameter Estimates

	chins	situps	jumps
Intercept	29.2001686	430.2510765	150.4807263
weight	-0.0431468	-0.6219668	-0.1751653
waist	-0.4350463	-6.2712454	-1.7661788
pulse	0.0598306	0.8624650	0.2428971

由输出 11.2.6 可得出因变量 situps 的标准化回归方程(右上角带“*”的变量表示标准化变量)及还原为原始变量的回归方程分别为:

$$\text{situps}^* = -0.2454\text{weight}^* - 0.3209\text{waist}^* + 0.0994\text{pulse}^*,$$

$$\text{situps} = 430.2511 - 0.6220\text{weight} - 6.2712\text{waist} + 0.8625\text{pulse}.$$

经进一步计算可知,回归方程是显著的($p=0.0059<0.05$),决定系数 $R^2=0.3506$. 回归方程中 weight 和 waist 系数的符号皆为负的,这合乎实际意义. 使用“舍一交叉验证方法”而得到的预测残差平方和

$$\text{PRESS}=14.99592 \quad \text{(见输出 11.2.7)}.$$

另两个因变量的回归方程这里省略. 输出 11.2.7 给出在 REG

输出 11.2.7 当抽取成分个数为 1 时因变量 SITUPS 的预测残差平方和 PRESS

The REG Procedure
Model: nfar1
Dependent Variable: situps 仰卧起坐

Sum of Residuals	0
Sum of Squared Residuals	12.33873
Predicted Residual SS (PRESS)	14.99592

过程中使用选项 R 得到的关于 PRESS 的结果.

表 11.1 给出了对三个因变量关于几种建模方法的预测残差平方和 PRESS 值的比较,这几种建模方法为:经典的多元线性回归方法(MLR),主成分个数为 1 或 2 时的主成分回归方法(PCR(1)或 PCR(2)),以及偏最小二乘因子个数为 1 或 2 时的偏最小二乘回归方法(PLS(1)或 PLS(2)).

表 11.1　对三个因变量比较几种建模方法的 PRESS 值

	MLR	PCR(1)	PCR(2)	PLS(1)	PLS(2)
chins	19.2108	18.7536	20.1552	17.6540	19.3307
situps	19.4546	15.4981	17.0632	14.9959	17.1469
jumps	27.2164	21.8387	23.6942	21.8714	23.2857

从表 11.1 明显可以看出,当抽取偏最小二乘因子(成分)的个数为 1 时所建立的偏最小二乘回归方程(见输出 11.2.6)的预测残差平方和 PRESS 比其他几个方法都小.此例中两个变量组的个数都不多,多重相关性也不是特别严重,PLS 方法的优点在这里还没有更充分地显示出来.

有关 SAS/STAT 软件中 PLS 过程的用法及其全面的功能请参考 SAS 系统(8 版本)有关 PLS 过程的帮助系统,其中的应用例子更能说明偏最小二乘回归方法的特点.

习题十一

11-1(化工试验例子)　考察的指标(因变量)Y 表示原辛烷值,自变量 x_1 表示直接蒸馏成分,x_2 表示重整汽油,x_3 表示原油热裂化油,x_4 表示原油催化裂化油,x_5 表示聚合物,x_6 表示烷基化物,x_7 表示天然香精. 7 个变量表示 7 个成分含量的比例(满足 $x_1+x_2+\cdots+x_7=1$). 表 11.2 给出 12 种混合物中 7 种成分和 Y 的数据. 试用偏最小二乘方法建立 Y 与 $x_1,x_2,\cdots,x_7$ 的回归方程,用于确定 7 种构成元素 $x_1,x_2,\cdots,x_7$ 对 Y 的影响.

表 11.2 化工试验的原始数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Y
1	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.7
2	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.8
3	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.6
4	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.0
5	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.6
6	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.2
7	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.9
8	0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.1
9	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.4
10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.2
11	0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.4
12	0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.1

11-2 试对第十章表 10.4 的 38 名学生的体质和运动能力数据，用偏最小二乘法建立 5 个运动能力指标与 7 个体质变量的回归方程.

附录　矩阵代数

矩阵和行列式是多元统计分析的重要工具，本附录对书中需要用到的有关矩阵代数知识作一简单的回顾和介绍，熟悉这些内容对阅读本书将带来很大方便. 如果读者希望对这方面知识有更多、更详细的了解，可参考有关的教科书.

§1　向量与长度

一、向量的定义及几何意义

由 n 个实数 $x_1, x_2, \cdots, x_n$ 组成的一个数组称为 n 维向量，记为

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{或} \quad X = (x_1, x_2, \cdots, x_n)'.$$

n 维向量在几何上可表示为一个有方向的线段. 向量可以进行数乘和加法运算. 向量通过乘一个常数 c 来实现伸长或缩短. 如向量 $Y = cX = (cx_1, cx_2, \cdots, cx_n)'$ (c 为常数)，当 $c > 1$ 时，向量 Y 是由 X 沿正方向伸长为原来的 c 倍得到的；当 $0 < c < 1$，向量 Y 是由 X 沿正方向缩短为原来的 c 倍得到的；当 $c < 0$ 时向量 Y 是由 X 沿反方向伸长或缩短为原来的 c 倍得到的.

两向量 $X = (x_1, x_2, \cdots, x_n)'$ 和 $Y = (y_1, y_2, \cdots, y_n)'$ 的和为

$$X + Y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

二、向量的长度和两向量间的夹角

向量 $X = (x_1, x_2, \cdots, x_n)'$ 的长度记为 L_X，其定义为

$$L_X=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}.$$

若令 $Y=cX$，则 $L_Y=L_{cX}=|c|L_X$. 取 $c=1/L_X$，则得到长度为 1 且与 X 同方向的单位向量 $Y=L_X^{-1}X$.

下面我们来考虑两个向量 X 和 Y 之间的夹角 θ. 当 $n=2$ 时，若记向量 $X=(x_1,x_2)'$ 和 $Y=(y_1,y_2)'$，它们与横坐标的夹角分别为 θ_1 和 θ_2，则两向量之间的夹角为 $\theta=\theta_2-\theta_1$，而且

$$\cos\theta=\cos(\theta_2-\theta_1)=\cos(\theta_2)\cos(\theta_1)+\sin(\theta_2)\sin(\theta_1)$$

$$=\frac{y_1}{L_y}\frac{x_1}{L_X}+\frac{y_2}{L_y}\frac{x_2}{L_X}=\frac{x_1y_1+x_2y_2}{L_XL_Y}.$$

推广到 n 维向量也有相似的定义. 如果引入两个 n 维向量 X 和 Y 间的内积 (X,Y)，其定义为

$$(X,Y)=X'Y=Y'X=x_1y_1+x_2y_2+\cdots+x_ny_n,$$

则 n 维向量 X 的长度 L_X 以及两个向量 X 和 Y 的夹角 θ 都可以用内积来表示：

$$L_X=\sqrt{X'X},$$

$$\cos\theta=\frac{x_1y_1+\cdots+x_ny_n}{L_XL_Y}=\frac{X'Y}{\sqrt{X'X}\sqrt{Y'Y}}.$$

当 $X'Y=0$ 时，$\cos\theta=0$，所以我们说，当 $X'Y=0$ 时向量 X 与 Y 相互垂直.

三、向量的线性相关与线性无关

一组 n 维向量 $X_1,X_2,\cdots,X_p$，如果存在不全为零的常数 $c_1,c_2,\cdots,c_p$，使

$$c_1X_1+c_2X_2+\cdots+c_pX_p=0,$$

则称这组向量线性相关. 线性相关意味着这组向量中至少有一个向量能写成其余向量的线性组合. 如果一组 n 维向量不线性相关，就称它们线性无关.

四、向量 X 在向量 Y 上的投影

设 $X=(x_1,x_2,\cdots,x_n)'$，$Y=(y_1,y_2,\cdots,y_n)'$，向量 X 在向量 Y 上的投影为

$$X\text{ 在 }Y\text{ 上的投影}=\frac{(X,Y)}{(Y,Y)}Y=\frac{X'Y}{L_Y}\frac{1}{L_Y}Y,$$

其中单位向量 $\frac{1}{L_Y}Y$ 表示 X 在 Y 上投影的方向. 而向量 X 在向量 Y 上投影的长度即为

$$\frac{|X'Y|}{L_Y}=L_X\left|\frac{X'Y}{L_XL_Y}\right|=L_X|\cos\theta|,$$

其中 θ 是两个向量 X 和 Y 之间的夹角.

§2　矩阵及基本运算

一、矩阵的定义

将 $p\times q$ 个实数 $a_{11},a_{12},\cdots,a_{pq}$ 排列成一个如以下形式的 p 行、q 列的长方形表：

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{bmatrix},$$

称 A 为 $p\times q$ 矩阵，常记作 $A=(a_{ij})_{p\times q}$，其中 a_{ij} 是第 i 行、第 j 列的元素. 本书中 a_{ij} 均为实数.

若 $q=1$，则称 A 为 p 维列向量，记作 a；当 p 维列向量的所有元素均为 1 时，常记为 $\mathbf{1}_p$：

$$a=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix},\quad \mathbf{1}_p=\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

若 $p=1$，则称 A 为 q 维行向量，记作

$$a'=(a_1,a_2,\cdots,a_q).$$

若 A 的所有元素全为零，则称 A 为零矩阵，记作 $A=O_{p\times q}$ 或 $A=O$. 若 A 的所有元素全为 1，常记该矩阵为 J，显然 $J_{p\times q}=\mathbf{1}_p\mathbf{1}_q'$.

若 $p=q$，则称 A 为 p 阶方阵，$a_{11},a_{22},\cdots,a_{pp}$ 称为它的对角线元素，其他元素 a_{ij} $(i\neq j)$ 称为非对角线元素.

若方阵 A 的对角线下方的元素全为零，则称 A 为上三角矩阵. 显然，$a_{ij}=0,i>j$.

若方阵 A 的对角线上方的元素全为零，则称 A 为下三角矩阵. 显然，$a_{ij}=0,i<j$.

若方阵 A 的所有非对角线元素均为零，则称 A 为对角矩阵，简记为 $A=\mathrm{diag}(a_{11},a_{22},\cdots,a_{pp})$.

若 p 阶对角矩阵 A 的所有 p 个对角线元素均为 1，则称 A 为 p 阶单位矩

阵，记作 $A=I_p$ 或 $A=I$.

若将矩阵 A 的行与列互换，则得到的矩阵称为 A 的转置，记作 A'，即

$$A'=\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & & \vdots \\ a_{1q} & a_{2q} & \cdots & a_{pq} \end{bmatrix}_{q\times p}.$$

若 A 是方阵，且 $A'=A$，则称 A 为对称矩阵. 显然 $a_{ij}=a_{ji}$.

若 A 是方阵，且 $A'=-A$，则称 A 为斜对称矩阵. 显然 $a_{ii}=0$，$a_{ij}=-a_{ji}$ $(i\neq j)$.

二、矩阵的运算

若 $A=(a_{ij})$ 为 $p\times q$ 矩阵，$B=(b_{ij})$ 为 $p\times q$ 矩阵，则 A 与 B 的和定义为

$$A+B=(a_{ij}+b_{ij})_{p\times q}.$$

若 c 为一常数，则它与 A 的积定义为

$$cA=(ca_{ij})_{p\times q}.$$

若 $A=(a_{ij})$ 为 $p\times q$ 矩阵，$B=(b_{ij})$ 为 $q\times r$ 矩阵，则 A 与 B 的积定义为

$$AB=\left(\sum_{k=1}^{q}a_{ik}b_{kj}\right)_{p\times r}.$$

从上述定义中容易得出如下的运算规律：

(1) $(A+B)'=A'+B'$.

(2) $(AB)'=B'A'$.

(3) $A(B_1+B_2)=AB_1+AB_2$.

(4) $A\left(\sum_{\alpha=1}^{k}B_\alpha\right)=\sum_{\alpha=1}^{k}AB_\alpha$.

(5) $c(A+B)=cA+cB$.

若 p 阶方阵 A 满足 $AA'=I$，则称 A 为正交矩阵. 显然，$\sum_{j=1}^{p}a_{ij}^2=1$ $(i=1,2,\cdots,p)$，称 A 的 p 个行向量为单位向量；$\sum_{j=1}^{p}a_{ij}a_{kj}=0$ $(i\neq k)$，称 A 的 p 个行向量两两相互正交. 又从 $A'A=I$ 得：$\sum_{i=1}^{p}a_{ij}^2=1$ $(j=1,\cdots,p)$，$\sum_{i=1}^{p}a_{ij}a_{ik}=0$ $(j\neq k)$，即 A 的 p 个列向量也是一组相互正交的单位向量. 例如，以下三个矩阵都是正交阵：

$$\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix},\quad \begin{bmatrix}\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{bmatrix},\quad \begin{bmatrix}\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0\\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-2}{\sqrt{3\cdot 2}}\end{bmatrix}.$$

若方阵 A 满足 $A^2=A$，则称 A 为幂等矩阵. 例如，

$$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad \begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix},\quad \begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix}.$$

对称的幂等矩阵称为投影矩阵.

矩阵的分块是在处理阶数较高的矩阵时常用的方法. 有时，我们把一个高阶矩阵看成是由一些低阶矩阵组成的，就像矩阵是由数值组成的一样. 设 $A=(a_{ij})$ 为 $p\times q$ 矩阵，将它剖分成四块，表示成

$$A=\left[\begin{array}{c:c}A_{11} & A_{12}\\ \hdashline A_{21} & A_{22}\end{array}\right],$$

其中 A_{11} 为 $k\times l$ 矩阵，A_{12} 为 $k\times(q-l)$ 矩阵，A_{21} 为 $(p-k)\times l$ 矩阵，A_{22} 为 $(p-k)\times(q-l)$ 矩阵. 分块矩阵也满足一般矩阵的加法、乘法等运算规律.

若 A 和 B 有相同的分块，则

$$A+B=\begin{bmatrix}A_{11}+B_{11} & A_{12}+B_{12}\\ A_{21}+B_{21} & A_{22}+B_{22}\end{bmatrix}.$$

若 C 为 $q\times r$ 矩阵，剖分成

$$C=\begin{bmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{bmatrix},$$

其中 C_{11} 为 $l\times m$ 矩阵，C_{12} 为 $l\times(r-m)$ 矩阵，C_{21} 为 $(q-l)\times m$ 矩矩，C_{22} 为 $(q-l)\times(r-m)$ 矩阵，则有

$$\begin{aligned}AC&=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{bmatrix}\\ &=\begin{bmatrix}A_{11}C_{11}+A_{12}C_{21} & A_{11}C_{12}+A_{12}C_{22}\\ A_{21}C_{11}+A_{22}C_{21} & A_{21}C_{12}+A_{22}C_{22}\end{bmatrix}.\end{aligned}$$

§3　行　列　式

p 阶方阵 $A=(a_{ij})$ 的行列式定义为

$$|A| = \sum_{j_1 j_2 \cdots j_p} (-1)^{\tau(j_1 j_2 \cdots j_p)} a_{1j_1} a_{2j_2} \cdots a_{pj_p},$$

这里 $\sum_{j_1 j_2 \cdots j_p}$ 表示对 $1,2,\cdots,p$ 的所有排列求和，$\tau(j_1 j_2 \cdots j_p)$ 是排列 $j_1, j_2, \cdots, j_p$ 中逆序的总数，称它为这个排列的逆序数. 一个逆序是指在一个排列中一对数的前后位置与大小顺序相反，即前面的数大于后面的数. 例如，$\tau(3,1,4,2)=3$.

由行列式的定义可以得到如下的一些基本性质：

性质 1 若 A 的某行(或列)为零，则 $|A|=0$.

性质 2 $|A'|=|A|$.

性质 3 若将 A 的某一行(或列)乘以常数 c，则所得矩阵的行列式为 $c|A|$.

性质 4 若 A 是一个 p 阶方阵，c 为一常数，则 $|cA|=c^p|A|$.

性质 5 若互换 A 的任意两行(或列)，则行列式符号改变.

性质 6 若 A 的某两行(或列)相同，则行列式为零.

性质 7 若将 A 的某一行(或列)的倍数加到另一行(或列)，则所得行列式不变.

性质 8 若 A 的某一行(或列)是其他一些行(或列)的线性组合，则行列式为零.

性质 9 若 p 阶方阵 A 为上三角矩阵或下三角矩阵或对角矩阵，则 $|A| = \prod_{i=1}^{p} a_{ii}$.

性质 10 若 A 和 B 均为 p 阶方阵，则 $|AB|=|A||B|$.

性质 11 因 AA' 为非负定矩阵，故有 $|AA'|\geqslant 0$.

性质 12 若 A 与 B 都是方阵，则 $\begin{vmatrix} A & C \\ O & B \end{vmatrix} = \begin{vmatrix} A & O \\ C & B \end{vmatrix} = |A||B|$.

性质 13 若 A 为 $p\times q$ 矩阵，B 为 $q\times p$ 矩阵，则 $|I_p+AB|=|I_q+BA|$.

证明 因为

$$\begin{bmatrix} I_p & A \\ O & I_q \end{bmatrix}\begin{bmatrix} I_p & -A \\ B & I_q \end{bmatrix} = \begin{bmatrix} I_p + AB & O \\ B & I_q \end{bmatrix},$$

$$\begin{bmatrix} I_p & O \\ -B & I_q \end{bmatrix}\begin{bmatrix} I_p & -A \\ B & I_q \end{bmatrix} = \begin{bmatrix} I_p & -A \\ O & I_q + BA \end{bmatrix},$$

将上述两个等式的两边各取行列式，故得

$$|I_p + AB| = |I_q + BA|. \qquad \text{(证毕)}$$

例如，若 x, y 为两个 p 维向量，则

$$|I_p + xy'| = 1 + y'x.$$

设 A 为 p 阶方阵，将其元素 a_{ij} 所在的第 i 行与第 j 列划去所得 $(p-1)$ 阶矩阵的行列式，称为元素 a_{ij} 的子式，记为 M_{ij}. $A_{ij}=(-1)^{i+j}M_{ij}$ 称为元素 a_{ij} 的代数余子式. 有以下公式成立

$$|A| = \sum_{j=1}^{p} a_{ij}A_{ij} = \sum_{i=1}^{p} a_{ij}A_{ij},$$

$$\sum_{j=1}^{p} a_{kj}A_{ij} = 0 \ (k \neq i), \quad \sum_{i=1}^{p} a_{ik}A_{ij} = 0 \ (k \neq j).$$

例如，若

$$A = \begin{bmatrix} 6 & 5 & 10 & 1 \\ 7 & 10 & 7 & 6 \\ 9 & 8 & 12 & 2 \\ 4 & 9 & 11 & 3 \end{bmatrix}, \quad a_{32} \text{ 的子式为 } M_{32} = \begin{vmatrix} 6 & 10 & 1 \\ 7 & 7 & 6 \\ 4 & 11 & 3 \end{vmatrix},$$

则其代数余子式为 $A_{32} = (-1)^{3+2}M_{32} = -M_{32}$.

§4　逆矩阵、矩阵的秩及分块求逆

一、逆矩阵

若方阵 A 满足 $|A|\neq 0$，则称 A 为非退化方阵或非奇异方阵；若 $|A|=0$，则称 A 为退化方阵. 若 $A=(a_{ij})$ 是一非退化方阵，令

$$B' = (A_{ij})/|A|,$$

其中 A_{ij} 是 a_{ij} 的代数余子式，则有 $AB=BA=I$，称 B 为 A 的逆，记作 $B=A^{-1}$. 由于 $|B|=1/|A|\neq 0$，所以 B 也是一个非退化方阵. 若方阵 C 满足 $AC=I$，则 $C=BAC=B$；同样，若方阵 D 满足 $DA=I$，则 $D=DAB=B$. 因此，A^{-1} 是唯一的，且 $(A^{-1})^{-1}=A$.

逆矩阵具有如下的基本性质：

性质 1　$AA^{-1}=A^{-1}A=I$.

性质 2　$(A')^{-1}=(A^{-1})'$.

性质 3　若 A 和 C 均为 p 阶非退化方阵，则 $(AC)^{-1}=C^{-1}A^{-1}$.

性质 4　$|A^{-1}|=|A|^{-1}$.

性质 5　若 A 是正交矩阵，则 $A^{-1}=A'$.

性质 6　若 $A=\mathrm{diag}(a_{11},a_{22},\cdots,a_{pp})$ 非退化（即 $a_{ii}\neq 0, i=1,2,\cdots,p$），则

$$A^{-1}=\mathrm{diag}(a_{11}^{-1},a_{22}^{-1},\cdots,a_{pp}^{-1}).$$

性质 7 若 A 和 B 为非退化方阵，则 $\begin{bmatrix} A & O \\ O & B \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & O \\ O & B^{-1} \end{bmatrix}$.

二、矩阵的秩

设 A 为 $p\times q$ 矩阵，若存在 A 的一个 r 阶子方阵的行列式不为零，而 A 的一切 $(r+1)$ 阶子方阵的行列式均为零，则称 A 的秩为 r，记作 $\mathrm{rank}(A)=r$.

矩阵的秩具有下述基本性质：

性质 1 $\mathrm{rank}(A)=0$，当且仅当 $A=O_{p\times q}$.

性质 2 若 A 为 $p\times q$ 矩阵，且 $A\neq O_{p\times q}$，则 $1\leqslant \mathrm{rank}(A)\leqslant \min\{p,q\}$.

性质 3 $\mathrm{rank}(A)=\mathrm{rank}(A')$.

性质 4 若 A 为 $p\times q$ 矩阵，B 为 $p\times r$ 矩阵，则

$$\max\{\mathrm{rank}(A),\mathrm{rank}(B)\}\leqslant \mathrm{rank}(A\vdots B)\leqslant \min\{p,\mathrm{rank}(A)+\mathrm{rank}(B)\}.$$

性质 5 $\mathrm{rank}\begin{bmatrix} A & O \\ O & B \end{bmatrix}=\mathrm{rank}\begin{bmatrix} O & A \\ B & O \end{bmatrix}=\mathrm{rank}(A)+\mathrm{rank}(B)$.

性质 6 $\mathrm{rank}(AB)\leqslant \min\{\mathrm{rank}(A),\mathrm{rank}(B)\}$.

性质 7 $\mathrm{rank}(A+B)\leqslant \mathrm{rank}(A)+\mathrm{rank}(B)$.

性质 8 若 A 和 C 为非退化方阵，则 $\mathrm{rank}(ABC)=\mathrm{rank}(B)$.

性质 9 p 阶方阵 A 是非退化的，当且仅当 $\mathrm{rank}(A)=p$. 当 p 阶方阵的秩为 p 时，称 A 是满秩的矩阵.

三、非奇异矩阵的分块矩阵求逆

设 A 是 p 阶满秩矩阵，它的逆矩阵为 A^{-1}，它们可表示为下列分块矩阵：

$$A=\left[\begin{array}{c:c} A_{11} & A_{12} \\ \hdashline A_{21} & A_{22} \end{array}\right],\quad A^{-1}=\left[\begin{array}{c:c} A^{11} & A^{12} \\ \hdashline A^{21} & A^{22} \end{array}\right],$$

其中 A_{11}，A^{11} 均为 $r\times r$ 矩阵，A_{22} 和 A^{22} 均为 $s\times s$ 矩阵，且 $r+s=p$.

若 A_{11} 满秩，记 $A_{22\cdot 1}=A_{22}-A_{21}A_{11}^{-1}A_{12}$.

若 A_{22} 满秩，记 $A_{11\cdot 2}=A_{11}-A_{12}A_{22}^{-1}A_{21}$.

定理 4.1 如果 A 和 A_{11} 非奇异，则

$$\left[\begin{array}{c:c} A^{11} & A^{12} \\ \hdashline A^{21} & A^{22} \end{array}\right]=\left[\begin{array}{c:c} A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1} \\ \hdashline -A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1} & A_{22\cdot 1}^{-1} \end{array}\right];\quad (4.1)$$

如果 A 和 A_{22} 非奇异，则

$$\left[\begin{array}{c:c} A^{11} & A^{12} \\ \hdashline A^{21} & A^{22} \end{array}\right]=\left[\begin{array}{c:c} A_{11\cdot 2}^{-1} & -A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} \\ \hdashline -A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1} & A_{22}^{-1}+A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} \end{array}\right].\quad (4.2)$$

证明　仅证(4.1)式,(4.2)式可类似证明.

由于 A_{11}^{-1}存在,容易验证下式成立:

$$\left[\begin{array}{c|c} I_r & O \\ \hline -A_{21}A_{11}^{-1} & I_s \end{array}\right]\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]\left[\begin{array}{c|c} I_r & -A_{11}^{-1}A_{12} \\ \hline O & I_s \end{array}\right]$$

$$=\left[\begin{array}{c|c} A_{11} & O \\ \hline O & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{array}\right],$$

由于$|A|=|A_{11}|\cdot|A_{22\cdot 1}|\neq 0$,知$|A_{22\cdot 1}|\neq 0$,于是 $A_{22\cdot 1}^{-1}$存在,上式两边求逆并作适当变换得到:

$$\left[\begin{array}{c|c} A^{11} & A^{12} \\ \hline A^{21} & A^{22} \end{array}\right]=\left[\begin{array}{c|c} I_r & -A_{11}^{-1}A_{12} \\ \hline O & I_s \end{array}\right]\left[\begin{array}{c|c} A_{11}^{-1} & O \\ \hline O & A_{22\cdot 1}^{-1} \end{array}\right]\left[\begin{array}{c|c} I_r & O \\ \hline -A_{21}A_{11}^{-1} & I_s \end{array}\right],$$

将其展开后即得(4.1)式.　**(证毕)**

如果 A_{11},A_{22} 都非奇异,则有

$$A_{11\cdot 2}^{-1}=A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1}, \tag{4.3}$$

$$A_{22\cdot 1}^{-1}=A_{22}^{-1}+A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1}. \tag{4.4}$$

另外还有

$$A^{11}=A_{11\cdot 2}^{-1},\quad A^{22}=A_{22\cdot 1}^{-1},$$

$$A^{12}=-A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1}=-A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1},$$

$$A^{21}=-A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1}=-A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}.$$

关于非奇异矩阵的行列式则有以下两个结论.

推论 4.1　如果 A_{11},A_{22} 非奇异,则

$$|A|=|A_{11}||A_{22\cdot 1}|=|A_{22}||A_{11\cdot 2}|.$$

推论 4.2　设 A 是 $p\times q$ 矩阵,B 是 $q\times p$ 矩阵,$\lambda\neq 0$,那么

$$|\lambda I_p-AB|=\lambda^{p-q}|\lambda I_q-BA|. \tag{4.5}$$

证明　利用推论 4.1 可知

$$\left|\begin{array}{cc} \lambda I_p & A \\ B & I_q \end{array}\right|=|\lambda I_p|\left|I_q-\frac{1}{\lambda}BA\right|=\lambda^{p-q}|\lambda I_q-BA|,$$

$$\left|\begin{array}{cc} \lambda I_p & A \\ B & I_q \end{array}\right|=|I_q||\lambda I_p-AB|=|\lambda I_p-AB|.$$　**(证毕)**

推论 4.3　设 A 是 $p\times q$ 矩阵,B 是 $q\times p$ 矩阵,那么

$$|I_p+AB|=|I_q+BA|.$$

证明　只须在推论 4.2 中令 $\lambda=1$,A 用 $-A$ 代替即可.　**(证毕)**

关于涉及矩阵的和的求逆问题下列定理是十分有用的.

定理 4.2 设 P 为 p 阶非奇异矩阵，U 为 $p\times q$ 矩阵，V 为 $q\times p$ 矩阵，而 p 阶方阵 $Q=P+UV$ 和 q 阶方阵 $I_q+VP^{-1}U$ 也非奇异，则

$$\begin{aligned} Q^{-1} &= (P+UV)^{-1} \\ &= P^{-1}-P^{-1}U(I+VP^{-1}U)^{-1}VP^{-1}. \end{aligned} \tag{4.6}$$

证明 令 $A=\begin{bmatrix} P & -U \\ V & I_q \end{bmatrix}$，利用前面的记号，只须将

$$A_{11\cdot 2}=P+UV,\quad A_{22\cdot 1}=I+VP^{-1}U,$$
$$A_{11}=P,\quad A_{22}=I_q,\quad A_{12}=-U,\quad A_{21}=V$$

代入(4.3)式即可得证. **(证毕)**

例如，若 x 和 y 分别为 p 维向量，P 是 p 阶方阵，且满足定理 4.2，则有

$$(P+xy')^{-1}=P^{-1}-(1+y'P^{-1}x)^{-1}P^{-1}xy'P^{-1}.$$

§5 特征值、特征向量和矩阵的迹

一、特征值和特征向量

设 A 是 p 阶方阵，则方程 $|A-\lambda I_p|=0$ 的左边是 λ 的 p 次多项式. 由多项式理论知道，该方程有 p 个根(可能有重根). 虽然 A 是实数矩阵，但方程的根可能为实数，也可能为复数，记作 $\lambda_1,\lambda_2,\cdots,\lambda_p$，并称为 A 的特征值或特征根.

另一方面，若 λ_i 是方程 $|A-\lambda I_p|=0$ 的一个根，则 $(A-\lambda_i I_p)$ 为退化矩阵，故存在一个 p 维非零向量 x_i，使得

$$(A-\lambda_i I_p)x_i=0,$$

即 λ_i 是 A 的一个特征值，而 x_i 称为相应的特征向量. 今后，一般取 x_i 为单位向量，即满足 $x_i'x_i=1$.

特征值和特征向量具有下述基本性质：

性质 1 A 和 A' 有相同的特征值.

性质 2 若 A 和 B 分别是 $p\times q$ 和 $q\times p$ 矩阵，则 AB 和 BA 有相同的非零特征值.

证明 因为

$$\begin{bmatrix} I_p & -A \\ O & \lambda I_q \end{bmatrix}\begin{bmatrix} \lambda I_p & A \\ B & I_q \end{bmatrix}=\begin{bmatrix} \lambda I_p-AB & O \\ \lambda B & \lambda I_q \end{bmatrix},$$

$$\begin{bmatrix} I_p & O \\ -B & \lambda I_q \end{bmatrix}\begin{bmatrix} \lambda I_p & A \\ B & I_q \end{bmatrix}=\begin{bmatrix} \lambda I_p & A \\ O & \lambda I_q-BA \end{bmatrix}.$$

所以

$$\begin{vmatrix} \lambda I_p - AB & O \\ \lambda B & \lambda I_q \end{vmatrix} = \begin{vmatrix} \lambda I_p & A \\ O & \lambda I_q - BA \end{vmatrix},$$

$$\lambda^q |\lambda I_p - AB| = \lambda^p |\lambda I_q - BA|.$$

可见，两个关于 λ 的方程 $|\lambda I_p - AB| = 0$ 和 $|\lambda I_q - BA| = 0$ 有着完全相同的非零根(若有重根，则它们的重数也相同)，故而 AB 和 BA 有相同的非零特征值.

(证毕)

性质 3　若 A 为实对称矩阵，则 A 的特征值全为实数，p 个特征值按大小依次表示为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$. 若 $\lambda_i \neq \lambda_j$，则相应的特征向量 x_i 和 x_j 必正交，即

$$x_i' x_j = 0.$$

性质 4　若 $A = \mathrm{diag}(a_{11}, a_{22}, \cdots, a_{pp})$，则 $a_{11}, a_{22}, \cdots, a_{pp}$ 为 A 的 p 个特征值，相应的特征向量分别为

$$e_1 = (1,0,\cdots,0)', \quad e_2 = (0,1,0,\cdots,0)', \quad \cdots, \quad e_p = (0,\cdots,0,1)'.$$

性质 5　$|A| = \prod_{i=1}^{p} \lambda_i$，即 A 的行列式等于其特征值的乘积. 因此，A 为非退化矩阵，当且仅当 A 的特征值均不为零；A 为退化矩阵，当且仅当 A 至少有一个特征值为零.

性质 6　若 A 为 p 阶对称矩阵，则存在正交矩阵 Γ 及对角矩阵 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$，将 Γ 按列向量分块，并记作 $\Gamma = (l_1, l_2, \cdots, l_p)$，矩阵 A 即有如下分解形式：$A = \sum_{i=1}^{p} \lambda_i l_i l_i'$，并称此分解为 A 的谱分解.

证明　事实上，由于 A 是对称矩阵，则有

$$A = \Gamma \Lambda \Gamma'. \tag{5.1}$$

将等式(5.1)两边右乘 Γ，得 $A\Gamma = \Gamma\Lambda$，于是有

$$A(l_1, l_2, \cdots l_p) = (l_1, l_2, \cdots, l_p) \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix},$$

$$(Al_1, Al_2, \cdots, Al_p) = (\lambda_1 l_1, \lambda_2 l_2, \cdots, \lambda_p l_p),$$

故

$$Al_i = \lambda_i l_i \quad (i = 1,2,\cdots,p).$$

这表明 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 是 A 的 p 个特征值，而 $l_1, l_2, \cdots, l_p$ 为相应的特征向量. 由于 Γ 是正交矩阵，所以相应的特征向量 $l_1, l_2, \cdots, l_p$ 是正交单位特征向量. 又上述矩阵 A 可作如下分解：

$$A = \Gamma\Lambda\Gamma'$$

$$= (l_1, l_2, \cdots, l_p) \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix} \begin{bmatrix} l_1' \\ l_2' \\ \vdots \\ l_p' \end{bmatrix} = \sum_{i=1}^{p} \lambda_i l_i l_i'. \qquad \text{(证毕)}$$

二、矩阵的迹

设 A 为 p 阶方阵，则它的对角线元素之和称为 A 的迹，记作 $\mathrm{tr}(A)$，即

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{pp}.$$

方阵的迹具有下述基本性质：

性质 1 若 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 为 A 的特征值，则 $\mathrm{tr}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$.

性质 2 $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

性质 3 $\mathrm{tr}(A) = \mathrm{tr}(A')$.

性质 4 $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.

性质 5 $\mathrm{tr}\left(\sum_{\alpha=1}^{k} A_\alpha\right) = \sum_{\alpha=1}^{k} \mathrm{tr}(A_\alpha)$.

性质 6 若 A 为投影矩阵，则 $\mathrm{tr}(A) = \mathrm{rank}(A)$.

证明 由于 $A' = A$，所以存在正交矩阵 Γ 和对角矩阵 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$，使得 $A = \Gamma\Lambda\Gamma'$，由 §4 矩阵秩的性质 8 知

$$\mathrm{rank}(A) = \mathrm{rank}(\Lambda) = \lambda_1, \lambda_2, \cdots, \lambda_p \text{ 中的非零个数.}$$

又由于幂等矩阵的特征值或为 0 或为 1，即

$$\lambda_i = 0 \text{ 或 } 1 \quad (i = 1, 2, \cdots, p),$$

故而

$$\begin{aligned} \mathrm{tr}(A) &= \lambda_1 + \lambda_2 + \cdots + \lambda_p = \lambda_1, \lambda_2, \cdots, \lambda_p \text{ 中 1 的个数} \\ &= \mathrm{rank}(A). \end{aligned} \qquad \text{(证毕)}$$

§6 正定矩阵、非负定矩阵和投影矩阵

设 A 是 p 阶对称矩阵，x 是一个 p 维向量，则 $x'Ax$ 称为 A 的二次型. 若对一切 $x \neq 0$，有 $x'Ax > 0$，则称 A 为正定矩阵，记作 $A > 0$；若对一切 $x \neq 0$，有 $x'Ax \geqslant 0$，则称 A 为非负定矩阵，记作 $A \geqslant 0$. $A > B$ 表示 $A - B > 0$；$A \geqslant B$ 表示 $A - B \geqslant 0$.

一、正定矩阵和非负定矩阵的基本性质

性质 1　设 A 是对称矩阵，则 A 是正定(或非负定)矩阵，当且仅当 A 的所有特征值均为正(或非负).

证明　必要性. 由于 A 是对称矩阵，故存在一正交矩阵 Γ，使得 $\Gamma'A\Gamma=\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$. 令 $e_i=(0,\cdots,0,1,0,\cdots,0)'$，其中除第 i 个位置为 1 外，其余元素均为 0，于是

$$\lambda_i=e_i'\Lambda e_i=(\Gamma e_i)'A(\Gamma e_i).$$

因为 $\Gamma e_i\neq 0$，所以当 $A>0$ 时，$\lambda_i>0$；$A\geqslant 0$ 时，$\lambda_i\geqslant 0$.

充分性. 对任意的 $x\neq 0$，令 $y=\Gamma'x$，则

$$x'Ax=x'\Gamma\Lambda\Gamma'x=y'\Lambda y=\sum_{i=1}^{p}\lambda_i y_i^2.$$

由于 $y\neq 0$，即 $y_1,y_2,\cdots,y_p$ 不全为零，所以当 $\lambda_i>0$ $(i=1,2,\cdots,p)$时，$x'Ax>0$，从而 $A>0$；当 $\lambda_i\geqslant 0$ $(i=1,2,\cdots,p)$时，$x'Ax\geqslant 0$，从而 $A\geqslant 0$.　**(证毕)**

性质 2　若 $A>0$，则 $A^{-1}>0$.

性质 3　设 $A\geqslant 0$，则 $A>0$ 当且仅当 $|A|\neq 0$.

性质 4　$BB'\geqslant 0$，对一切矩阵 B 成立.

性质 5　若 $A>0$(或$\geqslant 0$)，则存在 $A^{1/2}>0$(或$\geqslant 0$)，使得 $A=A^{1/2}A^{1/2}$，$A^{1/2}$ 称为 A 的平方根矩阵.

证明　因为 A 是对称矩阵，所以存在正交矩阵 Γ 和对角矩阵 $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_p)$使得 $A=\Gamma\Lambda\Gamma'$. 由 $A>0$(或$\geqslant 0$)可知，$\lambda_i>0$(或$\geqslant 0$) $(i=1,\cdots,p)$. 令

$$\Lambda^{1/2}=\mathrm{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},\cdots,\sqrt{\lambda_p}),\quad A^{1/2}=\Gamma\Lambda^{1/2}\Gamma',$$

则有

$$A=\Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma'=\Gamma\Lambda^{1/2}\Gamma'\Gamma\Lambda^{1/2}\Gamma'=A^{1/2}A^{1/2}.$$

由于 $A^{1/2}$的特征值$\sqrt{\lambda_i}>0$ (或$\geqslant 0$) $(i=1,2,\cdots,p)$，所以 $A^{1/2}>0$ (或$\geqslant 0$).

(证毕)

性质 6　设 $A\geqslant 0$ 是秩为 r 的 p 阶方阵，则存在一个秩为 r 的 $p\times r$ 矩阵 B，使得

$$A=BB'.$$

证明　因为 $A\geqslant 0$，所以存在正交矩阵 Γ 和对角矩阵

$$\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p),$$

使得 $A=\Gamma\Lambda\Gamma'$，且 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$. 因 $\mathrm{rank}(A)=\mathrm{rank}(\Lambda)=r$，故知

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>\lambda_{r+1}=\cdots=\lambda_p=0.$$

令 $\Lambda_1=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_r)$，记 $\Lambda_1^{1/2}=\mathrm{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},\cdots,\sqrt{\lambda_r})$，将 Γ 的前 r

列组成的 $p\times r$ 矩阵记为 Γ_1，Γ 的后 $p-r$ 列组成的 $p\times(p-r)$ 矩阵记为 Γ_2，即

$$\Gamma=(\Gamma_1\vdots\Gamma_2).$$

又记 $B=\Gamma_1\Lambda_1^{1/2}$. 显然，$B$ 是秩为 r 的 $p\times r$ 矩阵，因此

$$A=\Gamma\Lambda\Gamma'=(\Gamma_1\vdots\Gamma_2)\begin{bmatrix}\Lambda_1 & \vdots & O\\ \cdots & \vdots & \cdots\\ O & \vdots & O\end{bmatrix}\begin{bmatrix}\Gamma_1'\\ \cdots\\ \Gamma_2'\end{bmatrix}$$

$$=\Gamma_1\Lambda_1\Gamma_1'=\Gamma_1\Lambda_1^{1/2}\Lambda_1^{1/2}\Gamma_1'=(\Gamma_1\Lambda_1^{1/2})(\Gamma_1\Lambda_1^{1/2})'=BB'. \quad \text{(证毕)}$$

性质 7 若 $A>0, B>0, A-B>0$，则 $B^{-1}-A^{-1}>0$，且 $|A|>|B|$.

性质 8 若 $A>0$，将 A 剖分为

$$A=\begin{bmatrix}A_{11} & \vdots & A_{12}\\ \cdots & \vdots & \cdots\\ A_{21} & \vdots & A_{22}\end{bmatrix},$$

其中 A_{11} 为方阵，则

$$A_{11}>0,\ A_{22}>0,\ A_{11\cdot 2}=A_{11}-A_{12}A_{22}^{-1}A_{21}>0,\ A_{22\cdot 1}=A_{22}-A_{21}A_{11}^{-1}A_{12}>0.$$

二、投影矩阵及其性质

前面已给出定义：对称幂等矩阵就是投影矩阵. 以下是投影矩阵的一些性质：

性质 1 若 A 是投影矩阵，则 $\mathrm{tr}(A)=\mathrm{rank}(A)$.

性质 2 若 A 是投影矩阵，则 $I-A$ 也是投影矩阵.

性质 3 若 A 是秩为 r 的投影矩阵，则 A 有 r 个特征根为 1，其余为 0. 故满秩的投影矩阵必为 I.

性质 4 若 A 和 B 均为投影矩阵，且 $A+B=I$，则 $AB=BA=O$.

性质 5 若 X 是 $n\times p$ 矩阵，$n\geqslant p$，$\mathrm{rank}(X)=p$，则 $P_x=X(X'X)^{-1}X'$ 是投影矩阵，且 $\mathrm{rank}(P_x)=p$.

§7 特征值的极值问题

本节介绍几个与特征值有关的极值问题.

引理 7.1（柯西-施瓦茨(Cauchy-Schwarz)不等式） 设 b 和 d 是两个 p 维向量，则

$$(b'd)^2\leqslant(b'b)(d'd), \tag{7.1}$$

上述等号当且仅当 $b=cd$（或 $d=cb$）时成立，这里 c 为常数.

证明　当 $b=0$ 或 $d=0$，显然成立. 不妨设 $b\neq 0, d\neq 0$，考虑向量 $b-xd$，这里 x 是一个变数，于是

$$0 \leqslant (b-xd)'(b-xd) = (d'd)x^2 - (2b'd)x + b'b, \tag{7.2}$$

由 x 的二次函数性质知道此时

$$(b'd)^2 - (b'b)(d'd) \leqslant 0.$$

即(7.1)式成立.

若 $b=cd$，立即可知(7.1)式等号成立. 反之若(7.1)式等号成立，则当取 $x=b'd/d'd$ 时(7.2)式为 0，从而存在常数 $c=b'd/d'd$，使得

$$b = cd. \qquad \textbf{(证毕)}$$

引理 7.2（推广的柯西-施瓦茨不等式）　设 b, d 是两个 p 维向量，B 是 p 阶正定矩阵，那么有

$$(b'd)^2 \leqslant (b'Bb)(d'B^{-1}d), \tag{7.3}$$

且等号当且仅当 $b=cB^{-1}d$（或 $d=cBb$）时成立，这里 c 为常数.

证明　由于 B 正定，记 $\tilde{b}=B^{1/2}b$，$\tilde{d}=B^{-1/2}d$，此时 b 与 $\tilde{b}$，d 与 $\tilde{d}$ 同时为零向量或者非零向量，利用引理 7.1 于向量 $\tilde{b}$ 和 $\tilde{d}$，立即可得(7.3)式. **(证毕)**

定理 7.1　设 B 是 p 阶正定矩阵，d 为 p 维向量，对任意 p 维向量 x 下式成立：

$$\max_{x\neq 0} \frac{(x'd)^2}{x'Bx} = d'B^{-1}d, \tag{7.4}$$

且当 $x=cB^{-1}d$ 时达到最大值 $d'B^{-1}d$ ($c\neq 0$ 为常数).

证明　因为 B 是 p 阶正定矩阵，由引理 7.2 知道，对任一 p 维向量 $x\neq 0$，$x'Bx>0$ 成立，以及

$$\frac{(x'd)^2}{x'Bx} \leqslant d'B^{-1}d,$$

且当 $x=cB^{-1}d$ 时等号成立，定理得证. **(证毕)**

定理 7.2　设 B 是 p 阶对称矩阵，$\lambda_i=\mathrm{ch}_i(B)$ 是 B 的第 i 大的特征值，l_i 是相应于 λ_i 的 B 的标准化特征向量($i=1,\cdots,p$)，x 为任一非零 p 维向量，那么有

(1)
$$\lambda_p \leqslant \frac{x'Bx}{x'x} \leqslant \lambda_1, \tag{7.5}$$

上式右边等号当 $x=cl_1$ 时成立，左边等号当 $x=cl_p$ 时成立，这里 c 是非零常数.

(2) 记 $\mathscr{L}_2=\mathscr{L}(l_{r+1},\cdots,l_p)$，即 $\mathscr{L}_2$ 是由 $l_{r+1},\cdots,l_p$ 张成的空间，则

$$\max_{\substack{x\neq 0\\ x\in\mathscr{L}}} \frac{x'Bx}{x'x} = \lambda_{r+1}, \tag{7.6}$$

且当 $x=cl_{r+1}$ 时达到最大值，这里 c 为非零常数.

证明 (1) 记 $P=(l_1,\cdots,l_p)$,它是正交矩阵,由§5中特征值和特征向量的性质6知

$$B = P\mathrm{diag}(\lambda_1,\cdots,\lambda_p)P'.$$

任给 $x\neq 0$,令 $y=P'x/\sqrt{x'x}=(y_1,\cdots,y_p)'$,那么 $y'y=1$,而且

$$\frac{x'Bx}{x'x} = y'\mathrm{diag}(\lambda_1,\cdots,\lambda_p)y = \sum_{i=1}^{p}\lambda_i y_i^2,$$

于是

$$\lambda_p = \lambda_p\sum_{i=1}^{p}y_i^2 \leqslant \sum_{i=1}^{p}\lambda_i y_i^2 \leqslant \lambda_1\sum_{i=1}^{p}y_i^2 = \lambda_1.$$

上式右边等式当 $y=(1,0,\cdots,0)'$ 时成立,从而

$$\frac{x}{\sqrt{x'x}} = Py = (l_1,\cdots,l_p)\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} = l_1,$$

即 $x=cl_1$. 左边等式可类似证明. (7.5)式得证.

(2) 如果 $x\in\mathscr{L}(l_{r+1},\cdots,l_p)$,则 $l_i'x=0$ $(i=1,\cdots,r)$. 因此

$$y = P'x/\sqrt{x'x} = (0,\cdots,0,y_{r+1},\cdots,y_p)',$$

$$y'y = \sum_{i=r+1}^{p}y_i^2 = 1,$$

$$\frac{x'Bx}{x'x} = \sum_{i=r+1}^{p}\lambda_i y_i^2 \leqslant \lambda_{r+1}\sum_{i=r+1}^{p}y_i^2 = \lambda_{r+1},$$

而且等号当 $y=(0,\cdots,0,1,0,\cdots,0)$ 时成立,其中1在第 $r+1$ 位置上. 同上可知此时 $x=cl_{r+1}$, (7.6)式得证. (**证毕**)

推论 7.1 设 A 为 p 阶对称矩阵,B 为 p 阶正定矩阵,μ_i 是 AB^{-1} 的第 i 大特征值,l_i 是相应于 λ_i 的特征向量,那么对任一 p 维非零向量 x,有

$$\mu_p \leqslant \frac{x'Ax}{x'Bx} \leqslant \mu_1.$$

上式右边等号当 $x=cl_1$ 时成立,左边等号当 $x=cl_p$ 成立,这里 c 是非零常数.

§8 矩阵的微商和变换的雅可比行列式

在通常的高等数学教程中,我们曾讨论过对自变量 x 分别为一元变量或 p 维向量时它的函数 $y=f(x)$ 的微分表达形式. 而在本节我们将把上述概念推广到 x 的函数 y(因变量)为一元变量或 q 维向量的各种情况. 同时讨论当自变量

和因变量分别为相同维数的向量或者相同阶数的矩阵时有关变换的雅可比行列式.

一、矩阵的微商

我们将按自变量为一元变量或 p 维向量或 $p\times q$ 矩阵时分别讨论有关函数、向量函数及矩阵函数的微商表达形式.

1. 自变量是一元变量 x

(1) 若 $y=(y_1,\cdots,y_p)'$ 是 x 的向量函数,则记 $\dfrac{\mathrm{d}y}{\mathrm{d}x}=\left(\dfrac{\mathrm{d}y_1}{\mathrm{d}x},\cdots,\dfrac{\mathrm{d}y_p}{\mathrm{d}x}\right)'$ 为 y 对 x 的导数向量.

(2) 若 $Y=F(x)$ 是 x 的矩阵函数,其中 $Y=(y_{ij})$ 是 $p\times q$ 矩阵,那么

$$\frac{\mathrm{d}Y}{\mathrm{d}x}=\left(\frac{\mathrm{d}y_{ij}}{\mathrm{d}x}\right)_{p\times q}$$

是 Y 对 x 的导数矩阵,它仍是 $p\times q$ 矩阵.

2. 自变量是 p 维向量 $x=(x_1,\cdots,x_p)'$

(1) 若 $y=f(x)$ 是 x 的一元函数,通过令 $x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_p$ 为常数对 x_i 求导可求得 y 关于 x_i 的偏导数,记为 $\dfrac{\partial f}{\partial x_i}\left(\text{或}\dfrac{\partial y}{\partial x_i}\right)$,此时 y 关于 x 的偏导数向量记为

$$\frac{\partial f}{\partial x}=\left(\frac{\partial f}{\partial x_1},\cdots,\frac{\partial f}{\partial x_p}\right)'.$$

(2) 若 $y=(y_1,\cdots,y_q)'$ 是向量 x 的 q 维向量函数,即 $y_i=f_i(x)$ $(i=1,\cdots,q)$,简记 $y=f(x)$. 那么规定 y 关于 x 的诸偏导数构成的矩阵为

$$\frac{\partial(y)}{\partial(x)}=\frac{\partial y'}{\partial x}=\left(\frac{\partial y_j}{\partial x_i}\right)_{p\times q}=\begin{bmatrix}\dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_q}{\partial x_1}\\ \vdots & & \vdots\\ \dfrac{\partial y_1}{\partial x_p} & \cdots & \dfrac{\partial y_q}{\partial x_p}\end{bmatrix}. \tag{8.1}$$

例如,若 $y=(y_1,\cdots,y_q)'$, A 为 $q\times p$ 常数矩阵,令 $y=Ax$,则

$$\frac{\partial(y)}{\partial(x)}=\frac{\partial y'}{\partial x}=A'. \tag{8.2}$$

又如,若设 B 为 p 阶方阵, x 是 p 维向量,则

$$\frac{\partial x'Bx}{\partial x}=(B+B')x.$$

特别,当 B 是对称矩阵时有

$$\frac{\partial(x'Bx)}{\partial x}=2Bx. \tag{8.3}$$

二、矩阵变换的雅可比行列式

下面研究变换的雅可比行列式，在多元分布的推导过程中常常在积分号下作变换，这时必须计算它的雅可比行列式.

考虑积分

$$\int_V g(x_1,x_2,\cdots,x_p)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_p \quad (V\subset \mathbf{R}^p),$$

作变换 $y_i=f_i(x_1,x_2,\cdots,x_p)$ $(i=1,\cdots,p)$，记 $y=(y_1,\cdots,y_p)'$，则该变换可简记为 $y=f(x)$. 当逆变换存在时，可记为 $x=f^{-1}(y)$. 由微积分知识知道，以上积分可化为

$$\int_T g(f^{-1}(y))J(x\to y)\mathrm{d}y, \tag{8.4}$$

其中积分区域 $T=\{y:\ y=f(x),x\in V\}$；

$$J(x\to y)=\left|\frac{\partial(x)}{\partial(y)}\right|_+ \quad (|A|_+ \text{ 表示行列式 } |A| \text{ 的绝对值})$$

$J(x\to y)$称为变换的雅可比行列式.

注意，上述变换中自变量 x 和因变量 y 有相同的维数.

若 $y=f(x)$，计算 $J(x\to y)$时常用下列公式：

$$J(x\to y)=[J(y\to x)]^{-1}. \tag{8.5}$$

§9 消去变换

消去变换是通过对矩阵施行一些初等变换来计算矩阵的逆矩阵、广义逆矩阵，以及求解线性方程组的一种很有效的算法. 特别在逐步回归和逐步判别的计算中，消去变换是完成变量筛选的一种非常巧妙的算法. 消去变换在国外的文献上常常称为 Sweep(扫描)变换；也有人称为紧凑或原地求逆变换.

一、消去变换的定义

定义 9.1 设 $A=(a_{ij})_{n\times m}$，$a_{ij}\neq 0$，令

$$b_{\alpha\beta}=\begin{cases}1/a_{ij}, & \alpha=i,\beta=j,\\ -a_{\alpha j}/a_{ij}, & \alpha\neq i,\beta=j,\\ a_{i\beta}/a_{ij}, & \alpha=i,\beta\neq j,\\ a_{\alpha\beta}-a_{\alpha j}a_{i\beta}/a_{ij}, & \alpha\neq i,\beta\neq j.\end{cases} \tag{9.1}$$

记矩阵 $B=(b_{\alpha\beta})=T_{ij}(A)A$，并称 $T_{ij}(A)$为对矩阵 A 施行以(i,j)为主元(或枢轴)的**消去变换**. B 是对 A 施行(i,j)消去变换后得到的矩阵，简记为 $B=T_{ij}A$；当 $i=j$ 时，记 T_{ii}为 T_i.

矩阵 B 的形式如下：

$$B=\begin{bmatrix} * & \cdots & * & -\dfrac{a_{1j}}{a_{ij}} & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ * & \cdots & * & -\dfrac{a_{(i-1)j}}{a_{ij}} & * & \cdots & * \\ \dfrac{a_{i1}}{a_{ij}} & \cdots & \dfrac{a_{i(j-1)}}{a_{ij}} & \dfrac{1}{a_{ij}} & \dfrac{a_{i(j+1)}}{a_{ij}} & \cdots & \dfrac{a_{im}}{a_{ij}} \\ * & \cdots & * & -\dfrac{a_{(i+1)j}}{a_{ij}} & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ * & \cdots & * & -\dfrac{a_{nj}}{a_{ij}} & * & \cdots & * \end{bmatrix},$$

其中 $*$ 表示矩阵 B 中(α,β)位置的元素 $b_{\alpha\beta}=a_{\alpha\beta}-\dfrac{a_{\alpha j}a_{i\beta}}{a_{ij}}$.

二、消去变换 T_{ij} 的基本性质

性质 1　反身性：$T_{ij}T_{ij}A=A$.

设 $A^{(1)}=T_{ij}A\triangleq(a_{ij}^{(1)})$，$A^{(2)}=T_{ij}A^{(1)}\triangleq(a_{ij}^{(2)})$，由(9.1)式可直接验证：

$$a_{ij}^{(2)}=a_{ij}\quad(i=1,2,\cdots,n;j=1,2,\cdots,m).$$

性质 2　可交换性：当 $i\neq k,j\neq l$ 时，$T_{ij}T_{kl}A=T_{kl}T_{ij}A$.

由(9.1)式可直接验证两边元素对应相等.

性质 3　若 $A'=A$(A 为对称矩阵)，记 $B=T_{kk}A\overset{\text{def}}{=\!=}(b_{\alpha\beta})$，则

$$\begin{cases} b_{k\beta}=-b_{\beta k}, & \beta\neq k, \\ b_{\alpha\beta}=b_{\beta\alpha}, & \alpha\neq k,\beta\neq k. \end{cases}$$

对对称矩阵 A 施行(k,k)消去变换后，得矩阵 B. B 除第 k 行 k 列相差一个符号外，其余仍保持对称性，也称 B 为绝对对称矩阵.

性质 4　行列置换与消去变换的次序变化的关系：设 A 为 $n\times m$ 矩阵，P_{ip} 为 i 行和 p 行交换的行置换矩阵，Q_{jq}为 j 列和 q 列交换的列置换矩阵，则

(1) $T_{ij}(AQ_{jq})=(T_{iq}A)Q_{jq}$；

(2) $T_{ij}(P_{ip}A)=P_{ip}(T_{pj}A)$；

(3) $T_{ij}(P_{ip}AQ_{jq})=P_{ip}(T_{pq}A)Q_{jq}$.

性质 5 设 $A=\left[\begin{array}{c:c}A_{11}&A_{12}\\\hdashline A_{21}&A_{22}\end{array}\right]$，$A_{11}$为 r 阶可逆矩阵，则

$$T_rT_{r-1}\cdots T_1A=\left[\begin{array}{c:c}A_{11}^{-1}&A_{11}^{-1}A_{12}\\\hdashline -A_{21}A_{11}^{-1}&A_{22}-A_{21}A_{11}^{-1}A_{12}\end{array}\right].$$

部分习题参考解答或提示

习 题 二

2-1 $Y\sim N_2\left(\begin{bmatrix}2\\1\end{bmatrix},\begin{bmatrix}3&-1\\-1&1\end{bmatrix}\right)$.

2-2 令 $Y_1=X_1+X_2,Y_2=X_1-X_2$.

(1) 因 $\mathrm{Cov}(Y_1,Y_2)=0$,所以 X_1+X_2 与 X_1-X_2 相互独立.

(2) $Y_1=X_1+X_2\sim N(\mu_1+\mu_2,2(1+\rho)\sigma^2)$;

$Y_2=X_1-X_2\sim N(\mu_1-\mu_2,2(1-\rho)\sigma^2)$.

2-3 令 $Y=X^{(1)}+X^{(2)},Z=X^{(1)}-X^{(2)}$.

(1) 因 $\mathrm{COV}(Y,Z)=0_{p\times p}$,所以 $X^{(1)}+X^{(2)}$ 与 $X^{(1)}-X^{(2)}$ 相互独立.

(2) $Y=X^{(1)}+X^{(2)}\sim N_p(\mu^{(1)}+\mu^{(2)},2(\Sigma_1+\Sigma_2))$,

$Z=X^{(1)}-X^{(2)}\sim N_p(\mu^{(1)}-\mu^{(2)},2(\Sigma_1-\Sigma_2))$.

2-4 (1) 条件分布分别为

$$(X_1,X_2\mid X_3)\sim N_2\left(\begin{bmatrix}\mu_1+(x_3-\mu_3)\rho\\ \mu_2+(x_3-\mu_3)\rho\end{bmatrix},\begin{bmatrix}1-\rho^2&\rho-\rho^2\\ \rho-\rho^2&1-\rho^2\end{bmatrix}\right);$$

$$(X_1\mid X_2,X_3)\sim N_1\left(\mu_1+\frac{\rho}{1+\rho}(x_2+x_3-\mu_2-\mu_3),\frac{1+\rho-2\rho^2}{1+\rho}\right).$$

(2) $\sigma_{12\cdot 3}=\rho(1-\rho)$.

2-5 $(X_1\mid X_1+X_2)\sim N_1\left(\frac{1}{2}(x_1+x_2),\frac{1}{2}\right)$.

2-6 (1) $Y=3X_1-2X_2+X_3\sim N_1(13,9)$.

(2) 当二维向量 $a=(a_1,a_2)'$ 满足:$a_1+2a_2=2$ 时,X_3 与 $X_3-a'\begin{bmatrix}X_1\\X_2\end{bmatrix}$ 相互独立.

2-7 利用定理 2.3.1,可得(2),(3)和(4)中这 3 对变量是相互独立的,而另两对不独立.

2-10 如取 $A=\begin{bmatrix}2&0\\1&0\end{bmatrix}$,则 $X\overset{\mathrm{d}}{=\!=}AU\sim N_2(0,\Sigma)$.

2-11 $\mathrm{E}(X)=\begin{bmatrix}4\\3\end{bmatrix},\mathrm{D}(X)=\begin{bmatrix}1&-1\\-1&2\end{bmatrix}$.

2-12 (1) 提示:因 $X_1\sim N(0,1)$,对任意 $x(x>-1)$ 都有:

$$P\{-1<X_1\leqslant x\}=P\{-x\leqslant X_1<1\},$$

由此还可得出，对任给 $x_2\in(-1,1)$，有

$$\begin{aligned}P\{X_2\leqslant x_2\}&=P\{X_2\leqslant -1\}+P\{-1<X_2\leqslant x_2\}\\&=P\{X_1\leqslant -1\}+P\{-x_2\leqslant X_1<1\}=\cdots\\&=P\{X_1\leqslant X_2\}.\end{aligned}$$

(2) 提示：考虑 X_1,X_2 的线性函数 $Y=X_1-X_2$，并说明 Y 不是正态分布的(因 $P\{|Y|=0\}=P\{|X_1|>1\}=0.3174$).

2-15 $\hat{\Sigma}=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_{(i)}-\mu_0)(X_{(i)}-\mu_0)'$.

2-19 样本均值 $\overline{X}=(68.1,46.5,32.29)'$，

样本离差阵 $A=\begin{bmatrix}42.9&&\\-17.5&34.5&\\17.41&5.55&55.709\end{bmatrix}$，

样本协方差阵 $S=\dfrac{1}{9}A=\begin{bmatrix}4.7667&&\\-1.9444&3.8333&\\1.9344&0.6167&6.1899\end{bmatrix}$，

样本相关阵 $R=\begin{bmatrix}1&&\\-0.4549&1&\\0.3561&0.1266&1\end{bmatrix}$.

习　题　三

3-6 $F=\dfrac{n-k}{k(n-1)}T^2\overset{H_0下}{\sim}F(k,n-k)$，其中

$$T^2=(n-1)n(C\overline{X}-r)'[CAC']^{-1}(C\overline{X}-r),$$

又 $A=\sum\limits_{i=1}^{n}(X_{(i)}-\overline{X})(X_{(i)}-\overline{X})'$ 为样本离差阵.

3-7 $F=\dfrac{n-p+1}{(n-1)(p-1)}T^2\overset{H_0下}{\sim}F(p-1,n-p+1)$，其中

$$T^2=(n-1)n(C\overline{X})'[CAC']^{-1}C\overline{X},$$

又 A 为样本离差阵.

3-8 H_0: $C\mu=0_2$，其中 $C=\begin{bmatrix}1&0&-6\\0&1&-4\end{bmatrix}$. 由男婴测量数据($p=3,n=6$)计算可得

$$T^2=47.1434,\quad F=18.8574,\quad p=0.009195<\alpha=0.05,$$

故否定 H_0. 即认为这组数据与人类的一般规律不一致.

3-11 $T^2=5.3117,F=1.4982,p=0.2693>\alpha$，故 H_0 相容.

3-12 (1) $M=24.5240, d=0.4424, \chi^2=13.6755$. 因 $p=0.3219>\alpha=0.05$, 故 H_0 相容.

(2) $T^2=134.8155, F=32.09894$. 因 $p=0.001083<\alpha=0.05$, 故否定 H_0, 即 A 和 B 地区岩石的化学成分有显著差异.

(3) $\Lambda=0.01604, F=18.3903$. 因 $p=2.345\times10^{-6}<\alpha=0.05$, 故否定 H_0, 即 A、B 和 C 三个地区岩石的化学成分有显著差异.

(4) $V=0.7253, \hat{\xi}=-b\ln V=3.2650$. 因 $p=0.3525>\alpha$, 故 H_0 相容, 可认为三种化学成分相互独立.

习　题　四

4-1 (1) $\hat{a}=\frac{1}{6}(y_1+2y_2+y_3), \hat{b}=\frac{1}{5}(-y_2+2y_3)$.

(2) 似然比统计量为

$$\Lambda=\left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{\frac{3}{2}},$$

其中 $\hat{\sigma}^2=\frac{1}{3}[(y_1-\hat{a})^2+(y_2-2\hat{a}+\hat{b})^2+(y_3-\hat{a}-2\hat{b})^2]$; 当 H_0 成立时, 若记 $a=b=a_0$, 则 $\hat{a}_0=\frac{1}{11}(y_1+y_2+3y_3)$, 且

$$\hat{\sigma}_0^2=\frac{1}{3}[(y_1-\hat{a}_0)^2+(y_2-\hat{a}_0)^2+(y_3-3\hat{a}_0)^2],$$

因此与似然比统计量 Λ 等价的 F 统计量及分布为

$$F=\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2-\hat{\sigma}^2}\sim F(1,1)\quad(\text{在 } H_0 \text{ 成立时}).$$

4-2 $\hat{\beta}=(C'C)^{-1}C'Y$, $\hat{\sigma}^2=\frac{1}{n}(Y-C\hat{\beta})'(Y-C\hat{\beta})$.

4-3 (1) $\hat{Y}=-106.7267+3.2518x_1+1.3313x_2-0.6747x_3$;

$R^2=0.9909$, $s^2=(2.4416)^2$.

(2) $\hat{Y}=-30.0098+0.02672x_1^2+0.03130x_1\times x_2$;

$R^2=0.9896$, $s^2=5.43596$.

4-4 (1) OXY = 102.2383 − 0.2199 age − 0.07238 weight − 2.6805 time − 0.0008442 spulse − 0.3732 rpulse + 0.3047 mpulse, $R^2=0.8480$, $s=2.3221$.

(2) 当 $\alpha=0.05$ 时, OXY = 82.4218 − 3.3106 time　($R^2=0.7434$).

当 $\alpha=0.15$ 时,

OXY = 98.1479 − 0.1977 age − 2.7676 time − 0.3481 rpulse

$+0.2705$ mpulse　($R^2=0.8368$).

(3) OXY $=102.2043-0.2196$ age -0.07230 weight -2.6825 time -0.3734 rpulse $+0.3049$ mpulse　(修正 $R^2=0.8176$).

4-10 逐步回归的结果($\alpha_{in}=\alpha_{out}=0.15$)：

$$\hat{Y}_1=-33.8629-1.7261\,x_2+0.9546x_3+194.6098x_7\quad(R^2=0.8621);$$

$$\hat{Y}_2=-6.1372-1.6982\,x_4+0.00429\,x_6\quad(R^2=0.9287).$$

用双重筛选逐步回归，当筛选因变量和筛选自变量的临界值 $F_X=F_Y=3$ 时，两个因变量分为两组，且各组的回归方程同以上逐步回归的结果.

习　题　五

5-1 $P(2|1)=P(1|2)=1-\Phi\left(\dfrac{\mu_1-\mu_2}{\sigma_1+\sigma_2}\right)$.

5-2 (1) 当 $x=2.5$ 时，因 $d_1^2(x)=1$，$d_2^2(x)=1.5625$，$d_3^2(x)=0.25$，故 $x\in G_3$.

(2) 按贝叶斯准则(即广义平方距离准则)，因 $D_1^2(x)=-0.3863$，$D_2^2(x)=2.9488$，$D_3^2(x)=0.25$，故 $x\in G_1$.

5-4 (1) 费希尔线性判别函数为 $u(x)=-\dfrac{1}{\sqrt{89765}}(32\,X_1+33\,X_2)$，判别准则为

$$\begin{cases}\text{判 } X\in G_1, & \text{当 } u(X)>u^*,\\ \text{判 } X\in G_2, & \text{当 } u(X)\leqslant u^*\end{cases}\quad(\text{其中 } u^*=-4.2964).$$

当 $X_{(1)}=(20,20)'$ 时，因 $u(X_{(1)})=-4.3390<u^*$，故判 $X_{(1)}\in G_2$；

当 $X_{(2)}=(15,20)'$ 时，因 $u(X_{(2)})=-3.8050>u^*$，故判 $X_{(1)}\in G_1$.

(2) 按贝叶斯准则，考虑比值(其中 $\bar{\mu}=(15,20)'$)

$$W=\frac{h_1(x)}{h_2(x)}=\frac{75}{10}\frac{f_2(x)}{f_1(x)}=\frac{75}{10}\exp\{-(X-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})\}.$$

当 $X_{(1)}=(20,20)'$ 时，因 $W=7.5\exp(125/54)=75.9229>1$，故判

$$X_{(1)}\in G_2;$$

当 $X_{(2)}=(15,20)'$ 时，因 $W=7.5>1$，故判 $X_{(2)}\in G_2$.

(3) 当 $x=(20,20)'$ 时，$P(1|x)=0.7306$，$P(2|x)=0.2694$.

5-6 $P(1|2)=P(2|1)=1-\Phi\left(\dfrac{d}{2}\right)$，其中 $d^2=(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$.

5-7 距离判别准则：$\begin{cases}\text{判 } X\in G_1, & \text{当 } W(X)>0 \text{ 时},\\ \text{判 } X\in G_2, & \text{当 } W(X)\leqslant 0 \text{ 时},\end{cases}$ 其中

$$W(X)=(X-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),\ \bar{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)}).$$

贝叶斯判别准则：

$$\begin{cases} \text{判 } X \in G_1, & \text{当 } W(X) > d \text{ 时}, \\ \text{判 } X \in G_2, & \text{当 } W(X) \leqslant d \text{ 时}, \end{cases}$$

其中 $d=\ln \dfrac{q_2 L(1|2)}{q_1 L(2|1)}$.

在两个总体且协方差阵相等时两种判别准则形式很相似；当 $q_1=q_2$，$L(1|2)=L(2|1)=1$ 时，$d=0$，这表明在先验概率相等、错判损失均取为 1 时，这两种判别准则是等价的.

5-8 下一步可引入变量 X_4（因 F 统计量的值为 10.6106，$p<0.01$）.

5-9 (1) 两个三元总体均值相等的检验结果：$D=3.1977$，$F=3.1089$，$p=0.0756<0.10$，故在显著性水平 $\alpha=0.10$ 时两总体的均值向量有显著差异（即讨论这两个三元总体的判别问题是有意义的）.

线性判别函数为：

$$Y_1 = -28.7375 + 10.3139\text{Cu} + 8.9904\text{Ag} + 16.8578\text{Bi},$$

$$Y_2 = -31.1105 + 13.7895\text{Cu} + 8.2120\text{Ag} + 11.3311\text{Bi}.$$

判别结果为错判两个观测：含矿的第 6 号错判为不含矿；而不含矿的第 13 号错判为含矿.

(2) 待判的样品被判为不含矿.

5-10 (1) 判别结果为错判了两个观测：第 8 号原属 1 类的观测被错判为 2 类；第 9 号原属 3 类的观测被错判为 1 类. 待判的 3 个观测依次被判归为 1,2,3 类.

(2) 经检验三个总体协方差阵不相等，用二次判别函数，判别结果为 17 个观测全部判对；待判的 3 个观测依次被判归为 1,1,3 类.

用逐步判别法，当引入和剔除的临界值=0.18 时，选出 x_2 和 x_4；用这两个变量建立的线性判别函数判别的结果把第 9 号原属 3 类的观测被错判为 1 类；待判的 3 个观测依次被判归为 1,2,3 类.

5-11 (1) 判别结果：14 个监测点全部判对了.

(2) 待判的两个单位（观测点）分别被判归为 2,3 类.

习　题　六

6-9 $k=5$ 时，$G_i^{(5)}=\{x_{(i)}\}$ $(i=1,2,\cdots,5)$，$W(5)=0$；

$k=4$ 时，$G_1^{(4)}=\{x_{(1)},x_{(2)}\}$，$G_i^{(4)}=\{x_{(i+1)}\}$ $(i=2,3,4)$，$W(4)=0.5$；

$k=3$ 时，$G_1^{(3)}=\{x_{(1)},x_{(2)}\}$，$G_2^{(3)}=\{x_{(3)},x_{(4)}\}$，$G_3^{(3)}=\{x_{(5)}\}$，$W(3)=2.5$；

$k=2$ 时，$G_1^{(2)}=\{x_{(1)},x_{(2)}\}$，$G_2^{(2)}=\{x_{(3)},x_{(4)},x_{(5)}\}$，$W(2)=13.1667$；

$k=1$ 时，$G_1^{(1)}=\{x_{(1)},x_{(2)},x_{(3)},x_{(4)},x_{(5)}\}$，$W(1)=54.2$.

6-10 (1)用类平均法和 Ward 法对 6 个弹头分为三个类的结果相同：第 1,2,6 号为一类，第 4 号为一类，第 3,5 号为一类.

(2) 当变量间的距离定义为 $d_{ij}^2=1-c_{ij}^2$ 时，用系统聚类的类平均法把 7 种微量元素可分为三类：第一类含 X_1,X_6；第二类含 X_2,X_3,X_4,X_7，第三类只有 X_5. 用系统聚类的 Ward 法分类结果稍有不同，把 7 种微量元素也分为三类：第一类含 X_1,X_6,X_7；第二类含 X_2,X_3,X_4，第三类只有 X_5.

用 VARCLUS 过程的分类结果：把 7 种微量元素分为两类：第一类含 X_1,X_6,X_7；第二类含 X_2,X_3,X_4,X_5.

6-11 用系统聚类的类平均法把 14 个岩石标本(观测)可细分为四类：{1,2,4,7,13}，{3,5}，{8,9,10,11,12}，{6,14}. 用 Ward 法也可分为四类：{1,4}，{2,3,5,7,13}，{8,9,10,11,12}，{6,14}，其中前两类属含矿，后两类属不含矿.

系统聚类的分类结果中有少数观测与实际所属类别有差别.

6-12 用系统聚类的类平均法和 Ward 法均把 16 个监测点分为四类：{1,2,7}，{3,10,14,15}，{6,8,9,13}，{4,5,11,12,16}.

习 题 七

7-1 从协方差阵出发得总体主成分为：

$Z_1=0.040\,X_1+0.999\,X_2$ ($\mathrm{Var}(Z_1)=\lambda_1=100.1614$)；

$Z_2=0.999\,X_1-0.040\,X_2$ ($\mathrm{Var}(Z_2)=\lambda_2=0.8386$).

从相关阵出发得总体主成分为：

$Z_1^*=0.707\,X_1^*+0.707X_2^*$ ($\mathrm{Var}(Z_1^*)=\lambda_1^*=1.4$)；

$Z_2^*=0.707\,X_1^*-0.707X_2^*$ ($\mathrm{Var}(Z_2^*)=\lambda_2^*=0.6$).

7-2 (1) $Z_1=\sqrt{2}/2X_1+\sqrt{2}/2X_2$ ($\mathrm{Var}(Z_1)=\lambda_1=1+\rho$)；

$Z_2=\sqrt{2}/2X_1-\sqrt{2}/2X_2$ ($\mathrm{Var}(Z_2)=\lambda_2=1-\rho$).

(2) 长轴的方向为 $e_1=(\sqrt{2}/2,\sqrt{2}/2)'$，

短轴的方向为 $e_2=(\sqrt{2}/2,-\sqrt{2}/2)'$.

(3) $\rho\geqslant 0.9$.

7-3 (2) 因 $\mathrm{Var}(Z_1)=\lambda_1=\sigma^2[1+(p-1)\rho]$，故第一主成分的贡献率为

$$\frac{\lambda_1}{p\sigma^2}=\rho+\frac{1-\rho}{p}.$$

7-4 椭球的主轴方向为主成分的方向.

7-5 总体主成分为 $Z_i=X_i(i=1,2,3)$;三个主成分的方差分别为 4,4,2.

7-6 当 $0<\rho\leqslant\frac{1}{\sqrt{2}}$ 时,总体主成分为

$Z_1=\frac{X_1+\sqrt{2}X_2+X_3}{2}$, $\mathrm{Var}(Z_1)=\lambda_1=\sigma^2(1+\sqrt{2}\rho)$,其解释的方差比例为 $\frac{1+\sqrt{2}\rho}{3}$;

$Z_2=\frac{1}{\sqrt{2}}(X_1-X_3)$, $\mathrm{Var}(Z_2)=\lambda_2=\sigma^2$,其解释的方差比例为 $\frac{1}{3}$;

$Z_3=\frac{X_1-\sqrt{2}X_2+X_3}{2}$, $\mathrm{Var}(Z_3)=\lambda_3=\sigma^2(1-\sqrt{2}\rho)$,其解释的方差比例为 $\frac{1-\sqrt{2}\rho}{3}$.

7-7 总体主成分为

$Z_1=\frac{1}{2}(X_1+X_2+X_3+X_4)$, $\mathrm{Var}(Z_1)=\lambda_1=\sigma^2+\sigma_{12}+\sigma_{13}+\sigma_{14}$,

$Z_2=\frac{1}{2}(X_1+X_2-X_3-X_4)$, $\mathrm{Var}(Z_2)=\lambda_2=\sigma^2+\sigma_{12}-\sigma_{13}-\sigma_{14}$,

$Z_3=\frac{1}{2}(X_1-X_2+X_3-X_4)$, $\mathrm{Var}(Z_3)=\lambda_3=\sigma^2-\sigma_{12}+\sigma_{13}-\sigma_{14}$,

$Z_4=\frac{1}{2}(X_1-X_2-X_3+X_4)$, $\mathrm{Var}(Z_4)=\lambda_4=\sigma^2-\sigma_{12}-\sigma_{13}+\sigma_{14}$.

7-8 (1) $\hat{b}_j=(a_{1j},a_{2j},\cdots,a_{mj})'$;

(2) X_j 的回归平方和

$$U_j=(n-1)\sum_{t=1}^{m}\lambda_t a_{jt}^2=(n-1)\sum_{t=1}^{m}\rho^2(X_j,Z_t)=(n-1)\nu_j;$$

X_j 的残差平方和 $Q_j=(n-1)(1-\nu_j)$; X_j 的决定系数

$$R_j^2=\nu_j=\sum_{t=1}^{m}\rho^2(X_j,Z_t).$$

7-11 (1) 8 项指标若综合为三个主成分,可解释原变量信息的 86.66%;若综合为四个主成分,可解释原变量信息的 94.68%.

(2) 按第一主成分得分由小到大对 13 个行业排的次序为: 8,10,12,7,9,11,13,6,4,3,2,1,5.

7-12 六项指标若综合为两个主成分,可解释原变量信息的 81.24%;若综合为三个主成分,可解释原变量信息的 91.38%.

按第一主成分得分由小到大对 16 个地区农民的生活水平排的次序为: 山西,河北,河南,江西,内蒙,黑龙江,福建,安徽,山东,吉林,江苏,

辽宁，天津，浙江，北京，上海.

习 题 八

8-1 $m=1$ 的正交因子模型为

$$\begin{cases} X_1 = 0.9F_1 + \varepsilon_1, \\ X_2 = 0.7F_1 + \varepsilon_2, \\ X_3 = 0.5F_1 + \varepsilon_3, \end{cases} \quad D = \begin{bmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}.$$

8-2 (1) $m=1$ 时因子模型的主成分解为

$$\begin{cases} X_1 = 0.8757F_1 + \varepsilon_1, \\ X_2 = 0.8312F_1 + \varepsilon_2, \\ X_3 = 0.7111F_1 + \varepsilon_3, \end{cases} \quad D = \begin{bmatrix} 0.2331 & 0 & 0 \\ 0 & 0.3091 & 0 \\ 0 & 0 & 0.4943 \end{bmatrix},$$

$$Q(1) = 0.1951.$$

(2) $m=2$ 时因子模型的主成分解为

$$\begin{cases} X_1 = 0.8757F_1 - 0.1802F_2 + \varepsilon_1, \\ X_2 = 0.8312F_1 - 0.4048F_2 + \varepsilon_2, \\ X_3 = 0.7111F_1 + 0.6951F_2 + \varepsilon_3, \end{cases}$$

$$D = \begin{bmatrix} 0.2006 & 0 & 0 \\ 0 & 0.1453 & 0 \\ 0 & 0 & 0.01122 \end{bmatrix}, \quad Q(2) = 0.06611.$$

(3) 因 $Q(2)=0.06611<0.1$，故 $m=2$ 的主成分解符合要求.

8-4 提示：利用样本协方差阵 S 的谱分解式有：

$$S = \sum_{i=1}^{p} \lambda_i l_i l_i' = AA' + BB',$$

则 $\varepsilon = S-(AA'+D) = BB'-D$，即 $BB' = E+D$，故

$$\sum_{i=m+1}^{p} \lambda_i^2 = \mathrm{tr}(BB' \cdot BB') = \mathrm{tr}[(E+D)(E+D)']$$

$$= Q(m) + \sum_{i=1}^{p} (\sigma_i^2)^2.$$

8-6 因子分析(主成分解)的结果：取前 4 个公共因子可反映原始变量的 89.26%的信息. 由方差最大正交旋转后的载荷矩阵可得出，第一公共因子主要代表 X_1 和 X_2；第二公共因子主要代表 X_3 和 X_6；第三公共因子主要代表 X_4；第四公共因子主要代表 X_5.

8-7 因子分析(主成分解)的结果：取前两个公共因子可反映原始变量的 95.86%的信息($\lambda_1=3.7526$，$\lambda_2=1.0402$). 正交因子模型为

$$\begin{cases} X_1 = 0.35049F_1 + 0.92792F_2 + \varepsilon_1, \\ X_2 = 0.97604F_1 - 0.09686F_2 + \varepsilon_2, \\ X_3 = 0.93971F_1 + 0.23601F_2 + \varepsilon_3, \\ X_4 = 0.93958F_1 - 0.23403F_2 + \varepsilon_4, \\ X_5 = 0.95461F_1 - 0.24363F_2 + \varepsilon_5. \end{cases}$$

变量的共同度为

$$h_1^2 = 0.983876,\quad h_2^2 = 0.962030,\quad h_3^2 = 0.938752,$$

$$h_4^2 = 0.937588,\quad h_5^2 = 0.970640.$$

8-8 取前两个公共因子可反映原始变量的 73.68%的信息. 由因子载荷矩阵可看出,第一公共因子主要代表学生的总成绩;第二公共因子主要代表开、闭卷的成绩比较. 由方差最大正交旋转后的载荷矩阵可得出,第一公共因子主要代表开卷的三门课的成绩;第二公共因子主要代表闭卷的两门课的成绩.

习　题　九

9-1 (1) 因总 χ^2 统计量(为 3.5891)或总惯量(为 0.7238)的 86.87%可用前二维说明,这表示样品点和变量点用二维表示就可以了.

通过在同一坐标系上绘制 8 个样品点和 6 个变量点的散布图,可粗略地看出,变量点和样品点可分为三类:{污染气体环己烷和 6};{污染气体环氧氯丙烷和 4,8};{污染气体氯,硫化氢,SO_2,碳 4 和 1,2,3,5,7}.

9-2 (1) 因总 χ^2 统计量(为 186.561)或总惯量(为 0.04660)的 92.10%可用前二维说明,这表示样品点和变量点用二维表示就可以了.

通过在同一坐标系上绘制 16 个样品点(地区)和 6 个变量点(指标)的散布图,可粗略地看出,地区和指标可分为五类:{上海和 $X4$(住房)};{北京,天津,山东和 $X5$(生活用品及其他)};{江苏,浙江和 $X1$(食品)};{福建,江西,安徽和 $X3$(燃料),$X6$(文化生活服务支出)};{山西,内蒙,辽宁,黑龙江,河北,河南和 $X2$(衣着)}.

习　题　十

10-1 第一典型相关系数 $\rho=\dfrac{2\times 0.7}{\sqrt{(1+0.5)\times(1+0.6)}}=0.903696$;第一对典型变量为:$V_1=\dfrac{1}{\sqrt{3}}(X_1+X_2)$,$W_1=\dfrac{1}{\sqrt{3.2}}(Y_1+Y_2)$.

10-3 第一典型相关系数 $\rho=0.6299$，在 $\alpha=0.01$ 的水平下它是显著地不为零(似然比为 0.5977，$p=0.0022<0.01$)．第一对标准化的典型变量为：

$$V_1 = 0.6265X_1 + 0.5560X_2,$$

$$W_1 = 0.8600X_3 + 0.5018X_4 - 0.5060X_5.$$

10-4 第一典型相关系数 $\rho=0.7885$，在显著性水平 $\alpha=0.01$ 时它是显著地不为零(似然比为 0.3772，$p=0.0003<0.01$)．第一对标准化的典型变量为：

$$V_1 = 0.5522X_1 + 0.5215X_2,$$

$$W_1 = 0.5044X_3 + 0.5383X_4.$$

10-5 第一典型相关系数 $\rho_1=0.8515$．它在显著性水平 $\alpha=0.01$ 时是显著地不为零(似然比为 0.06119，$p<0.0001$)，第一对标准化典型变量为：

$$V_1 = 0.4421X_1 + 0.2669X_2 + 0.5884X_3 + 0.0614X_4 + 0.2217X_5 + 0.0911X_6 + 0.0138X_7,$$

$$W_1 = -0.4266X_8 + 0.2335X_9 + 0.3696X_{10} + 0.0038X_{11} - 0.3560X_{12}.$$

第二典型相关系数 $\rho_2=0.7284$，它在显著性水平 $\alpha=0.01$ 时是显著地不为零(似然比为 0.2225，$p=0.005<0.01$)：

$$V_2 = -0.2087X_1 + 0.7020X_2 - 0.2102X_3 + 0.0148X_4 - 0.7263X_5 - 0.1749X_6 + 0.2399X_7,$$

$$W_2 = 0.8255X_8 + 1.0405X_9 + 0.1982X_{10} + 0.2218X_{11} + 0.8101X_{12}.$$

而后面的几个典型相关系数 $\rho_i(i=3,\cdots,5)$ 在显著性水平 $\alpha=0.01$ 时可认为是 0.

参考文献

[1] 方开泰编著. 实用多元统计分析. 上海：华东师范大学出版社，1989

[2] 张尧庭，方开泰著. 多元统计分析引论. 北京：科学出版社，1982/2003

[3] 高惠璇编著. 实用统计方法与 SAS 系统. 北京：北京大学出版社，2001

[4] 陈家鼎等编著. 数理统计学讲义. 北京：高等教育出版社，1993

[5] [美] Richard A Johnson，Dean W Wichern 著，陆璇译. 实用多元统计分析. 第四版. 北京：清华大学出版社，2001

[6] 王学仁，王松桂编译. 实用多元统计分析. 上海：上海科学技术出版社，1990

[7] 高惠璇编著. 统计计算. 北京：北京大学出版社，1995

[8] 王学民编著. 应用多元分析. 第二版. 上海：上海财经大学出版社，1999

[9] 于秀林等编著. 多元统计分析. 北京：中国统计出版社，1999

[10] [日]奥野忠一等著. 多变量解析法. 日科技连，1971

[11] 王惠文著. 偏最小二乘回归方法及其应用. 北京：国防工业出版社，1999

[12] 韦博成等著. 统计诊断引论. 南京：东南大学出版社，1991

[13] 王吉利，张尧庭主编. SAS 软件与应用统计. 北京：中国统计出版社，2000

[14] 王学仁编著. 地质数据的多变量统计分析. 北京：科学技术出版社，1982

[15] [英]肯德尔著，中国科学院计算中心概率统计组译. 多元分析. 北京：科学技术出版社，1983

[16] 李天生等. 用双重筛选逐步回归法对广西钦州县松毛虫发生进行分析与预测. 林业科学，1985，21(3)：247～251

[17] 高惠璇等编译. SAS 系统·Base SAS 软件使用手册. 北京：中国统计出版社，1997

[18] 高惠璇等编译. SAS 系统·SAS/STAT 软件使用手册. 北京：中国统计出版社，1997

[19] SAS Institute Inc. SAS/STAT (Version 6) User's Guide (Fourth Ed) Edition. 1994(1)

[20] SAS Institute Inc. SAS/STAT (Version 6) User's Guide (Fourth Ed) Edition. 1994(2)

[21] SAS Institute Inc. Introduction to Statistics Using SAS/INSIGHT Software. 1998

[22] Anderson T W. An Interduction to Multivariate Statistical Analysis. 2nd Edition. New York：World Publishing Co；Wiley，1984

[23] Richard A Johnson，Dean W Wichern. Applied Multivariate Statistical Analysis. 4nd Edition. Englewood Cliffs，N J：Prentice-Hill，1998

[24] 董文泉等编著. 数量化理论及其应用. 长春：吉林人民出版社，1979

[25] [印度]劳(Rao C R)著，张燮等译. 线性统计推断及其应用. 北京：科学出版社，1987

[26] 胡国定，张润楚著. 多元数据分析方法——纯代数处理. 天津：南开大学出版社，1989

[27] 孙文爽，陈兰祥编. 多元统计分析. 北京：高等教育出版社，1994

[28] 周光亚等编著. 多元统计方法. 长春：吉林大学出版社，1988

[29] 吴国富等编. 实用数据分析方法. 北京：中国统计出版社，1992

[30] 孙尚拱，潘恩沛编著. 实用判别分析. 北京：科学出版社，1990

[31] 孙尚拱编著. 医学多变量统计与统计软件. 北京：北京医科大学出版社，2000

主要符号说明

$X=(X_1,\cdots,X_p)'$	p 维随机向量
$X=(x_{ij})_{n\times p}$ 或 $X_{n\times p}=(x_{ij})$	$n\times p$ 随机阵或 $n\times p$ 观测数据阵
$X_{(\alpha)}=(x_{\alpha 1},x_{\alpha 2},\cdots,x_{\alpha p})'$	来自 p 元总体的第 α 次观测样品
$X_j=(x_{1j},x_{2j},\cdots,x_{nj})'$	第 j 个变量的 n 维观测向量或得分向量
$\mathbb{R}^p$	p 维实向量空间
$\mathrm{E}(X)$	随机向量 X 的均值向量
$\mathrm{D}(X)$	随机向量 X 的协方差阵
$\mathrm{Var}(X)$	随机变量 X 的方差
$\mathrm{Cov}(X,Y)$	随机变量 X 和 Y 的协方差
$\mathrm{COV}(X,Y)$	随机向量 X 和 Y 的协方差阵，显然 $\mathrm{COV}(X,X)=\mathrm{D}(X)$
$\mathrm{Vec}(X)$	用 $n\times p$ 矩阵 X 的列向量 $X_1,\cdots,X_p$ 依次连结成一个 np 维长向量
$N(\mu,\sigma^2)$ 或 $N_1(\mu,\sigma^2)$	均值为 μ，方差为 σ^2 的一元正态分布
$N_p(\mu,\Sigma)$	均值向量为 μ，协方差阵为 Σ 的 p 元正态分布；若 $X\sim N_p(\mu,\Sigma)$，则 X 为 p 维正态随机向量
$N_{n\times p}(M,I_n\otimes\Sigma)$	$n\times p$ 矩阵正态分布
$\chi^2(n,\delta)$ 或 $\chi^2_n(\delta)$	自由度为 n，非中心参数为 δ 的 χ^2（卡方）分布；当 $\delta=0$ 为中心 χ^2 分布
$t(n,\delta)$	自由度为 n，非中心参数为 δ 的非中心 t 分布；当 $\delta=0$ 为中心 t 分布
$F(m,n,\delta)$	自由度为 m,n，非中心参数为 δ 的 F 分布；当 $\delta=0$ 为中心的 F 分布
$W_p(n,\Sigma)$	参数为 p,n,Σ 的威沙特(Wishart)分布
$T^2(p,n)$	参数为 p,n 的霍特林(Hotelling) T^2 分布
$\Lambda(p,n,m)$	参数为 p,n 和 m 的威尔克斯(Wilks)分布
$O_{p\times q}$	$p\times q$ 零矩阵

$A>0$	A 为对称正定方阵
$A\geqslant 0$	A 为对称非负定方阵
$\|A\|$	方阵 A 的行列式
$\|A\|_+$	方阵 A 行列式的绝对值
$\mathrm{tr}(A)$	方阵 A 的迹
$\mathrm{rank}(A)$	矩阵 A 的秩
$\mathbf{1}_n=(1,1,\cdots,1)'$	元素全部为 1 的 n 维常向量
$0_p=(0,0,\cdots,0)'$ 或 0	元素全部为 0 的 p 维零向量，当零向量的维数显然，简记为 0
$(A \vdots B)$	$n\times p$ 矩阵 A 和 $n\times q$ 矩阵 B 合并为一个 $n\times(p+q)$ 的矩阵
$\mathrm{ch}_i(A)$	方阵 A 的第 i 大特征值
$\mathscr{L}(l_1,\cdots,l_p)$	由向量 $l_1,\cdots,l_p$ 张成的空间
$J(x\to y)$	p 维向量 x 到 p 维向量 y 变换的雅可比行列式

索　引

（按拼音字母顺序）

Q

R

S

T

V