第1章

Chapter 1 多元正态分布

学习目标

1. 掌握多元分布的有关概念；
2. 掌握统计距离的概念；
3. 理解多元正态分布的定义及其有关性质；
4. 了解常用多元分布及其抽样分布的定义和基本性质。

在基础统计学中，随机变量的正态分布在理论和实际应用中都有着重要的地位。同样，在多元统计学中，多元正态分布也占有相当重要的位置。原因是许多实际问题研究中的随机向量确实遵从或近似遵从多元正态分布；对于多元正态分布，已有一整套统计推断方法，并且可以得到许多完整的结果。

多元正态分布是最常用的一种多元概率分布。此外，还有多元对数正态分布、多项式分布、多元超几何分布、多元 β 分布、多元 χ^2 分布、多元指数分布等。本章从多维变量及多元分布的基本概念开始，着重介绍多元正态分布的定义及一些重要性质，以及常用多元分布及其抽样分布的定义和基本性质。

1.1 多元分布的基本概念

在研究社会、经济现象和许多实际问题时，经常遇到的是多指标的问题。例如研究职工工资构成情况时，计时工资、基础工资与职务工资、各种奖金、各种津贴等都是同时需要考察的指标；又如研究公司的运营情况时，要涉及公司的资金周转能力、偿债能力、获利能力及竞争能力等财务指标，这些都是多指标研究的问题。

显然，由于这些指标之间往往不独立，仅研究某个指标或是将这些指标割裂开来分别研究，都不能从整体上把握所研究问题的实质。一般，假设所研究的问题涉及 p 个指标，进行了 n 次独立观测，将得到 np 个数据，我们的目的就是对观测对象进行分组、分类，或分析这 p 个变量之间的相互关联程度，或找出内在规律等。下面简要介绍多元分析中涉及的一些基本概念。

1.1.1 随机向量

假定所讨论的是多个变量的总体，所研究的数据是同时观测 p 个指标（即变量），进行了 n 次观测得到的，我们把这 p 个指标表示为 X_1，X_2，…，X_p，常用向量

$$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$$

表示对同一个体观测的 p 个变量。若观测了 n 个个体，则可得到如表 1—1 所示的数据，称每一个个体的 p 个变量为一个样品，而全体 n 个样品形成一个样本。

表 1—1

变量 / 序号	X_1	X_2	…	X_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	…	x_{np}

横看表 1—1，记

$$\boldsymbol{X}_{(\alpha)}=(x_{\alpha 1},x_{\alpha 2},\cdots,x_{\alpha p})',\quad \alpha=1,2,\cdots,n$$

它表示第 α 个样品的观测值。竖看表 1—1，第 j 列的元素：

$$\boldsymbol{X}_j=(x_{1j},x_{2j},\cdots,x_{nj})',\quad j=1,2,\cdots,p$$

表示对第 j 个变量 X_j 的 n 次观测数值。

因此，样本资料矩阵可用矩阵语言表示为：

$$\boldsymbol{X}=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p)=\begin{bmatrix}\boldsymbol{X}'_{(1)}\\ \boldsymbol{X}'_{(2)}\\ \vdots\\ \boldsymbol{X}'_{(n)}\end{bmatrix}$$

若无特别说明，本书所称向量均指列向量。

定义 1.1 设 X_1，X_2，…，X_p 为 p 个随机变量，由它们组成的向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 称为随机向量。

1.1.2　分布函数与密度函数

描述随机变量的最基本工具是分布函数。类似地，描述随机向量的最基本工具还是分布函数。

定义 1.2　设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 是一随机向量，它的多元分布函数是

$$F(\boldsymbol{x})=F(x_1,x_2,\cdots,x_p)=P(X_1\leqslant x_1, X_2\leqslant x_2,\cdots,X_p\leqslant x_p) \tag{1.1}$$

式中，$\boldsymbol{x}=(x_1, x_2, \cdots, x_p)\in R^p$，并记成 $\boldsymbol{X}\sim F$。

多元分布函数的有关性质此处从略。

定义 1.3　设 $\boldsymbol{X}\sim F(\boldsymbol{x})=F(x_1, x_2, \cdots, x_p)$，若存在一个非负的函数 $f(\cdot)$，使得

$$F(\boldsymbol{x})=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f(t_1,t_2,\cdots,t_p)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_p \tag{1.2}$$

对一切 $\boldsymbol{x}\in R^p$ 成立，则称 $\boldsymbol{X}$（或 $F(\boldsymbol{x})$）有分布密度 $f(\cdot)$，并称 $\boldsymbol{X}$ 为连续型随机向量。

一个 p 维变量的函数 $f(\cdot)$ 能作为 R^p 中某个随机向量的分布密度，当且仅当

(i)　$f(\boldsymbol{x})\geqslant 0, \quad \forall \boldsymbol{x}\in R^p$

(ii)　$\int_{R^p} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$

例 1—1

若随机向量 (X_1, X_2, X_3) 有密度函数

$$f(x_1,x_2,x_3)=x_1^2+6x_3^2+\frac{1}{3}x_1x_2$$

$$0<x_1<1,\quad 0<x_2<2,\quad 0<x_3<\frac{1}{2}$$

容易验证它符合分布密度函数的两个条件 (i) 和 (ii)。

最重要的连续型多元分布——多元正态分布将留在 1.3 节讨论。

1.1.3　多元变量的独立性

定义 1.4　两个随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 称为相互独立的，若

$$P(\boldsymbol{X}\leqslant\boldsymbol{x},\boldsymbol{Y}\leqslant\boldsymbol{y})=P(\boldsymbol{X}\leqslant\boldsymbol{x})P(\boldsymbol{Y}\leqslant\boldsymbol{y}) \tag{1.3}$$

对一切 $\boldsymbol{x}$，$\boldsymbol{y}$ 成立。若 $F(\boldsymbol{x}, \boldsymbol{y})$ 为 $(\boldsymbol{X}, \boldsymbol{Y})$ 的联合分布函数，$G(\boldsymbol{x})$ 和 $H(\boldsymbol{y})$ 分别

为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的分布函数，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立当且仅当

$$F(\boldsymbol{x},\boldsymbol{y})=G(\boldsymbol{x})H(\boldsymbol{y}) \tag{1.4}$$

若 $(\boldsymbol{X},\boldsymbol{Y})$ 有密度 $f(\boldsymbol{x},\boldsymbol{y})$，用 $g(\boldsymbol{x})$ 和 $h(\boldsymbol{y})$ 分别表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的分布密度，则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 独立当且仅当

$$f(\boldsymbol{x},\boldsymbol{y})=g(\boldsymbol{x})h(\boldsymbol{y}) \tag{1.5}$$

注意在上述定义中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的维数一般是不同的。

类似地，若它们的联合分布等于各自分布的乘积，则称 p 个随机向量 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立。由 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立可以推知任何 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j(i\neq j)$ 独立，但是，若已知任何 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j(i\neq j)$ 独立，并不能推出 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立。

1.1.4 随机向量的数字特征

1. 随机向量 $\boldsymbol{X}$ 的均值

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 有 p 个分量。若 $E(X_i)=\mu_i$ $(i=1,2,\cdots,p)$ 存在，定义随机向量 $\boldsymbol{X}$ 的均值为：

$$E(\boldsymbol{X})=\begin{bmatrix}E(X_1)\\E(X_2)\\\vdots\\E(X_p)\end{bmatrix}=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix}=\boldsymbol{\mu} \tag{1.6}$$

$\boldsymbol{\mu}$ 是一个 p 维向量，称为均值向量。

当 $\boldsymbol{A}$，$\boldsymbol{B}$ 为常数矩阵时，由定义可立即推出如下性质：

(1) $E(\boldsymbol{AX})=\boldsymbol{A}E(\boldsymbol{X})$ (1.7)

(2) $E(\boldsymbol{AXB})=\boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$ (1.8)

2. 随机向量 $\boldsymbol{X}$ 的协方差阵

$$\begin{aligned}\boldsymbol{\Sigma}&=\operatorname{cov}(\boldsymbol{X},\boldsymbol{X})=E(\boldsymbol{X}-\boldsymbol{EX})(\boldsymbol{X}-\boldsymbol{EX})'=D(\boldsymbol{X})\\&=\begin{bmatrix}D(X_1)&\operatorname{cov}(X_1,X_2)&\cdots&\operatorname{cov}(X_1,X_p)\\\operatorname{cov}(X_2,X_1)&D(X_2)&\cdots&\operatorname{cov}(X_2,X_p)\\\vdots&\vdots&&\vdots\\\operatorname{cov}(X_p,X_1)&\operatorname{cov}(X_p,X_2)&\cdots&D(X_p)\end{bmatrix}\\&=(\sigma_{ij})\end{aligned} \tag{1.9}$$

称它为 p 维随机向量 $\boldsymbol{X}$ 的协方差阵，简称为 $\boldsymbol{X}$ 的协方差阵。

称 $|\operatorname{cov}(\boldsymbol{X},\boldsymbol{X})|$ 为 $\boldsymbol{X}$ 的广义方差，它是协方差阵的行列式之值。

3. 随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差阵

设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 和 $\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_q)'$ 分别为 p 维和 q 维随机向量，它们之间的协方差阵定义为一个 $p\times q$ 矩阵，其元素是 $\mathrm{cov}(X_i, Y_j)$，即

$$\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})=(\mathrm{cov}(X_i,Y_j)),\quad i=1,2,\cdots,p;j=1,2,\cdots,q \tag{1.10}$$

若 $\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{0}$，称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是不相关的。

当 $\boldsymbol{A}$，$\boldsymbol{B}$ 为常数矩阵时，由定义可推出协方差阵有如下性质：

(1) $D(\boldsymbol{AX})=\boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}'=\boldsymbol{A\Sigma A}'$

(2) $\mathrm{cov}(\boldsymbol{AX}, \boldsymbol{BY})=\boldsymbol{A}\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}'$

(3) 设 $\boldsymbol{X}$ 为 p 维随机向量，期望和协方差存在，记 $\boldsymbol{\mu}=E(\boldsymbol{X})$，$\boldsymbol{\Sigma}=D(\boldsymbol{X})$，$\boldsymbol{A}$ 为 $p\times p$ 常数阵，则

$$E(\boldsymbol{X}'\boldsymbol{AX})=\mathrm{tr}(\boldsymbol{A\Sigma})+\boldsymbol{\mu}'\boldsymbol{A\mu}$$

对于任何随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 来说，其协方差阵 $\boldsymbol{\Sigma}$ 都是对称阵，同时总是非负定（也称半正定）的。大多数情形下是正定的。

4. 随机向量 $\boldsymbol{X}$ 的相关阵

若随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 的协方差阵存在，且每个分量的方差大于零，则 $\boldsymbol{X}$ 的相关阵定义为：

$$\boldsymbol{R}=(\mathrm{corr}(X_i,X_j))=(r_{ij})_{p\times p}$$

$$r_{ij}=\frac{\mathrm{cov}(X_i,X_j)}{\sqrt{D(X_i)}\sqrt{D(X_j)}},\quad i,j=1,2,\cdots,p \tag{1.11}$$

r_{ij} 也称为分量 X_i 与 X_j 之间的（线性）相关系数。

对于两组不同的随机向量 $\boldsymbol{X}$ 及 $\boldsymbol{Y}$，它们之间的相关问题将在典型相关分析的章节中详细讨论。

在数据处理时，为了克服由于指标的量纲不同对统计分析结果带来的影响，往往在使用某种统计分析方法之前，将每个指标“标准化”，即做如下变换：

$$X_j^*=\frac{X_j-E(X_j)}{[\mathrm{var}(X_j)]^{1/2}},\quad j=1,2,\cdots,p \tag{1.12}$$

$$\boldsymbol{X}^*=(X_1^*, X_2^*, \cdots, X_p^*)$$

于是

$$E(\boldsymbol{X}^*)=\boldsymbol{0}$$

$$D(\boldsymbol{X}^*)=\mathrm{corr}(\boldsymbol{X})=\boldsymbol{R}$$

即标准化数据的协方差阵正好是原指标的相关阵：

$$\boldsymbol{R}=\frac{1}{n-1}\boldsymbol{X}^{*\prime}\boldsymbol{X}^* \tag{1.13}$$

1.2 统计距离

在多指标统计分析中，距离的概念十分重要，样品间的不少特征都可用距离来描述。大部分多元方法是建立在简单的距离概念基础上的，即平时人们熟悉的欧氏距离，或称直线距离。如几何平面上的点 $P=(x_1, x_2)$ 到原点 $O=(0, 0)$ 的欧氏距离，依勾股定理有

$$d(O,P)=(x_1^2+x_2^2)^{1/2} \tag{1.14}$$

一般，若点 P 的坐标 $P=(x_1, x_2, \cdots, x_p)$，则它到原点 $O=(0, 0, \cdots, 0)$ 的欧氏距离，依勾股定理有

$$d(O,P)=\sqrt{x_1^2+x_2^2+\cdots+x_p^2} \tag{1.15}$$

所有与原点距离为 C 的点满足方程

$$d^2(O,P)=x_1^2+x_2^2+\cdots+x_p^2=C^2 \tag{1.16}$$

因为这是一个球面方程（$p=2$ 时是圆），所以，与原点等距离的点构成一个球面，任意两个点 $P=(x_1, x_2, \cdots, x_p)$ 与 $Q=(y_1, y_2, \cdots, y_p)$ 之间的欧氏距离为：

$$d(P,Q)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2} \tag{1.17}$$

但就大部分统计问题而言，欧氏距离是不能令人满意的。这是因为每个坐标对欧氏距离的贡献是同等的。当坐标轴表示测量值时，它们往往带有大小不等的随机波动，在这种情况下，合理的办法是对坐标加权，使变化较大的坐标比变化小的坐标有较小的权系数，这就产生了各种距离。

欧氏距离还有一个缺点，那就是当各个分量为不同性质的量时，“距离”的大小竟然与指标的单位有关。例如，横轴 x_1 代表重量（以 kg 为单位），纵轴 x_2 代表长度（以 cm 为单位）。有四个点 A，B，C，D，它们的坐标如图 1—1 所示。

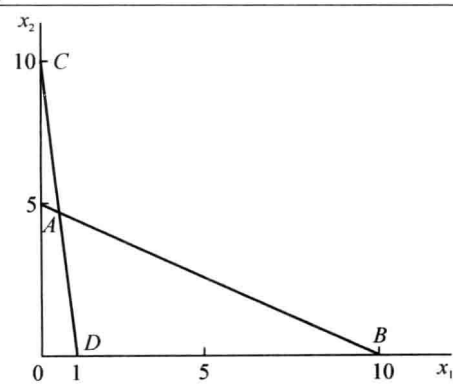


图 1—1

这时

$$AB=\sqrt{5^2+10^2}=\sqrt{125}$$
$$CD=\sqrt{10^2+1^2}=\sqrt{101}$$

显然，AB 要比 CD 长。

现在，如果 x_2 用 mm 作单位，x_1 单位保持不变，此时 A 坐标为 (0，50)，C 坐标为 (0，100)，则

$$AB=\sqrt{50^2+10^2}=\sqrt{2\,600}$$
$$CD=\sqrt{100^2+1^2}=\sqrt{10\,001}$$

结果 CD 反而比 AB 长！这显然是不够合理的。因此，有必要建立一种距离，这种距离应能够体现各个变量在变差大小上的不同，以及有时存在的相关性，还要求距离与各变量所用的单位无关。看来，我们选择的距离要依赖于样本方差和协方差。因此，采用“统计距离”这个术语，以区别通常习惯用的欧氏距离。

下面先介绍统计距离。

设 $P=(x_1，x_2，\cdots，x_p)$，$Q=(y_1，y_2，\cdots，y_p)$，且 Q 的坐标是固定的，点 P 的坐标相互独立地变化。用 S_{11}，S_{22}，…，S_{pp} 表示 p 个变量 x_1，x_2，…，x_p 的 n 次观测的样本方差。为给出坐标的合理权数，用坐标标准离差去除每个坐标，得到标准化坐标，则从 P 到 Q 的统计距离为：

$$d(P,Q)=\sqrt{\frac{(x_1-y_1)^2}{S_{11}}+\frac{(x_2-y_2)^2}{S_{22}}+\cdots+\frac{(x_p-y_p)^2}{S_{pp}}} \tag{1.18}$$

所有与点 Q 的距离平方为常数的点 P 构成一个椭球，其中心在 Q，其长短轴平行于坐标轴。容易看到：

(1) 在式 (1.18) 中，令 $y_1=y_2=\cdots=y_p=0$，得到点 P 到原点 O 的距离。

(2) 如果 $S_{11}=S_{22}=\cdots=S_{pp}$，则用欧氏距离式 (1.17) 是方便可行的。

还可以利用旋转变换的方法得到合理的距离。考虑点 $P=(x_1，x_2，\cdots，x_p)$ 和 $Q=(y_1，y_2，\cdots，y_p)$，这里 Q 为固定点，而 P 的坐标是变化的，且彼此相关，$O=(0，0，\cdots，0)$ 为坐标原点，则 P 到 O 和 Q 的距离分别为：

$$\begin{aligned} d(O,P)&=(a_{11}x_1^2+a_{22}x_2^2+\cdots+a_{pp}x_p^2+2a_{12}x_1x_2+\cdots+2a_{p-1,p}x_{p-1}x_p)^{1/2}\\ &=(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})^{1/2} \end{aligned} \tag{1.19}$$

和

$$\begin{aligned} d(P,Q)=&[a_{11}(x_1-y_1)^2+a_{22}(x_2-y_2)^2+\cdots+a_{pp}(x_p-y_p)^2\\ &+2a_{12}(x_1-y_1)(x_2-y_2)+\cdots\\ &+2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)]^{1/2}\\ =&[(\boldsymbol{X}-\boldsymbol{Y})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{Y})]^{1/2} \end{aligned} \tag{1.20}$$

这里

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix},\quad \boldsymbol{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

且 $\boldsymbol{A}$ 为对称阵，满足条件：对任意的 $\boldsymbol{X}$，恒有 $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\geqslant\boldsymbol{0}$，且等号成立当且仅当 $\boldsymbol{X}=\boldsymbol{0}$，即 $\boldsymbol{A}$ 为正定方阵。

最常用的一种统计距离是印度统计学家马哈拉诺比斯（Mahalanobis）于 1936 年引入的，称为“马氏距离”。下面先用一个一维的例子说明欧氏距离与马氏距离在概率上的差异。设有两个一维正态总体 G_1：$N(\mu_1, \sigma_1^2)$ 和 G_2：$N(\mu_2, \sigma_2^2)$。若有一个样品，其值在点 A 处，点 A 距离哪个总体近些呢？如图 1—2 所示。

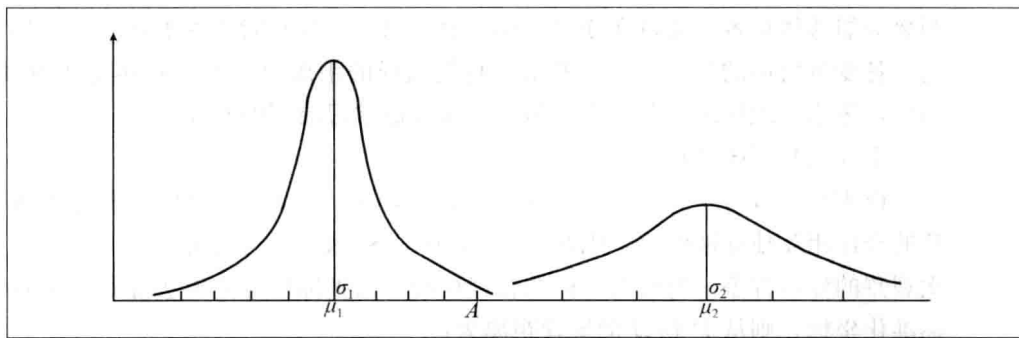


图 1—2

由图 1—2 可看出，从绝对长度来看，点 A 距左面总体 G_1 近些，即点 A 到 μ_1 比点 A 到 μ_2 要近一些（这里用的是欧氏距离，比较的是点 A 坐标与 μ_1 到 μ_2 值之差的绝对值），但从概率观点来看，点 A 在 μ_1 右侧约 $4\sigma_1$ 处，点 A 在 μ_2 的左侧约 $3\sigma_2$ 处，若以标准差的观点来衡量，点 A 离 μ_2 比离 μ_1 要近一些。显然，后者是从概率角度来考虑的，因而更为合理，它是用坐标差平方除以方差（或说乘以方差的倒数），从而转化为无量纲数的，推广到多维就要乘以协方差阵 $\boldsymbol{\Sigma}$ 的逆矩阵 $\boldsymbol{\Sigma}^{-1}$，这就是马氏距离的概念。以后将会看到，这一距离在多元分析中起着十分重要的作用。

有了上面的讨论，现在可以定义马氏距离了。

设 $\boldsymbol{X}$，$\boldsymbol{Y}$ 是从均值向量为 $\boldsymbol{\mu}$，协方差阵为 $\boldsymbol{\Sigma}$ 的总体 G 中抽取的两个样品，定义 $\boldsymbol{X}$，$\boldsymbol{Y}$ 两点之间的马氏距离为：

$$d_m^2(\boldsymbol{X},\boldsymbol{Y})=(\boldsymbol{X}-\boldsymbol{Y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{Y}) \tag{1.21}$$

定义 $\boldsymbol{X}$ 与总体 G 的马氏距离为：

$$d_m^2(\boldsymbol{X},G)=(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \tag{1.22}$$

设 E 表示一个点集，d 表示距离，它是 $E\times E$ 到 $[0, \infty)$ 的函数，可以证明，马氏距离符合如下距离的四条基本公理：

(1) $d(x, y)\geqslant 0$，　$\forall x, y\in E$；

(2) $d(x, y)=0$，　当且仅当 $x=y$；

(3) $d(x, y)=d(y, x)$，　$\forall x, y\in E$；

(4) $d(x, y)\leqslant d(x, z)+d(z, y)$，　$\forall x, y, z\in E$。

1.3　多元正态分布

多元正态分布是一元正态分布的推广。迄今为止，多元分析的主要理论都是建立在多元正态总体基础上的，多元正态分布是多元分析的基础。另一方面，许多实际问题的分布常是多元正态分布或近似正态分布，或虽本身不是正态分布，但它的样本均值近似于多元正态分布。

本节将介绍多元正态分布的定义，并简要给出它的基本性质。

1.3.1　多元正态分布的定义

在概率论中已经讲过，一元正态分布的密度函数为：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad \sigma>0$$

上式可以改写成

$$f(x)=(2\pi)^{-1/2}\sigma^{-1}\exp\left[-\frac{1}{2}(x-\mu)'(\sigma^2)^{-1}(x-\mu)\right] \tag{1.23}$$

式 (1.23) 用 $(x-\mu)'$代表 $(x-\mu)$ 的转置。由于 x，μ 均为一维的数字，转置与否都相同，所以可以这样写。

当遵从一元正态分布的随机变量 X 的概率密度函数改写为式 (1.23) 时，我们就可以将其推广，给出多元正态分布的定义。

定义 1.5　若 p 元随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$的概率密度函数为：

$$f(x_1, x_2, \cdots, x_p)=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\quad \boldsymbol{\Sigma}>\boldsymbol{0} \tag{1.24}$$

则称 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$遵从 p 元正态分布，也称 $\boldsymbol{X}$ 为 p 元正态变量，记为：

$$\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$|\boldsymbol{\Sigma}|$ 为协方差阵 $\boldsymbol{\Sigma}$ 的行列式。

式 (1.24) 实际是在 $|\boldsymbol{\Sigma}|\neq 0$ 时定义的。若 $|\boldsymbol{\Sigma}|=0$，则不存在通常意义下的密度，但可以在形式上给出一个表达式，使有些问题可以利用这一形式对 $|\boldsymbol{\Sigma}|\neq 0$及 $|\boldsymbol{\Sigma}|=0$ 的情况给出统一的处理。

当 $p=2$ 时，可以得到二元正态分布的密度公式。

设 $\boldsymbol{X}=(X_1, X_2)'$遵从二元正态分布，则

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2 r\\ \sigma_2\sigma_1 r & \sigma_2^2\end{bmatrix}, \quad r\neq\pm 1$$

这里 σ_1^2，σ_2^2 分别是 X_1 与 X_2 的方差，r 是 X_1 与 X_2 的相关系数。此时

$$|\boldsymbol{\Sigma}|=\sigma_1^2\sigma_2^2(1-r^2)$$

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-r^2)}\begin{bmatrix}\sigma_2^2 & -\sigma_1\sigma_2 r\\ -\sigma_2\sigma_1 r & \sigma_1^2\end{bmatrix}$$

故 X_1 与 X_2 的密度函数为：

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\ (1-r^2)^{1/2}}$$
$$\exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2r\,\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

这与我们学过的概率统计中的结果是一致的。

如果 $r=0$，那么 X_1 与 X_2 是独立的；若 $r>0$，则 X_1 与 X_2 趋于正相关；若 $r<0$，则 X_1 与 X_2 趋于负相关。

定理 1.1 设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则

$$E(\boldsymbol{X})=\boldsymbol{\mu}, \quad D(\boldsymbol{X})=\boldsymbol{\Sigma}$$

定理 1.1 将正态分布的参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 赋予了明确的统计意义。有关这个定理的证明可参见参考文献［3］。

多元正态分布不止定义 1.5 一种形式，更广泛的可采用特征函数来定义，也可用一切线性组合均为正态的性质来定义等。有关这些定义的方式参见参考文献［3］。

1.3.2 多元正态分布的性质

（1）如果正态随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$的协方差阵 $\boldsymbol{\Sigma}$ 是对角阵，则 $\boldsymbol{X}$ 的各分量是相互独立的随机变量。证明参见参考文献［4］。

（2）多元正态分布随机向量 $\boldsymbol{X}$ 的任何一个分量子集（多变量 $(X_1, X_2, \cdots, X_p)'$ 中的一部分变量构成的集合）的分布（称为 $\boldsymbol{X}$ 的边缘分布）仍然遵从正态分布。反之，若一个随机向量的任何边缘分布均为正态，并不能导出它是多元正态分布。

例如，设 $\boldsymbol{X}=(X_1, X_2)'$有分布密度

$$f(x_1,x_2)=\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}[1+x_1x_2\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}]$$

容易验证，$X_1\sim N(0, 1)$，$X_2\sim N(0, 1)$，但 (X_1, X_2) 显然不是正态分布。

（3）多元正态向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$的任意线性变换仍然遵从多元正态分布。

即设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，而 m 维随机向量 $\boldsymbol{Z}_{m\times 1}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}$，其中 $\boldsymbol{A}=(a_{ij})$ 是 $m\times p$ 阶的常数矩阵，$\boldsymbol{b}$ 是 m 维的常向量，则 m 维随机向量 $\boldsymbol{Z}$ 也是正态的，且 $\boldsymbol{Z}\sim N_m(\boldsymbol{A\mu}+\boldsymbol{b},\ \boldsymbol{A\Sigma A}')$。即 $\boldsymbol{Z}$ 遵从 m 元正态分布，其均值向量为 $\boldsymbol{A\mu}+\boldsymbol{b}$，协方差阵为 $\boldsymbol{A\Sigma A}'$。

(4) 若 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，则

$$d^2=(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\sim\chi^2(p)$$

d^2 若为定值，随着 $\boldsymbol{X}$ 的变化，其轨迹为一椭球面，是 $\boldsymbol{X}$ 的密度函数的等值面。若 $\boldsymbol{X}$ 给定，则 d^2 为 $\boldsymbol{X}$ 到 $\boldsymbol{\mu}$ 的马氏距离。

1.3.3　条件分布和独立性

设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，$p\geqslant 2$，将 $\boldsymbol{X}$，$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 剖分如下：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}^{(1)}\\ \boldsymbol{X}^{(2)}\end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix} \tag{1.25}$$

其中，$\boldsymbol{X}^{(1)}$，$\boldsymbol{\mu}^{(1)}$ 为 $q\times 1$ 维，$\boldsymbol{\Sigma}_{11}$ 为 $q\times q$ 维，我们希望求给定 $\boldsymbol{X}^{(2)}$ 时 $\boldsymbol{X}^{(1)}$ 的条件分布，即 $(\boldsymbol{X}^{(1)}\mid\boldsymbol{X}^{(2)})$ 的分布。下一个定理指出：正态分布的条件分布仍为正态分布。

定理 1.2　设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}>\boldsymbol{0}$，则

$$(\boldsymbol{X}^{(1)}\mid\boldsymbol{X}^{(2)})\sim N_q(\boldsymbol{\mu}_{1\cdot 2},\boldsymbol{\Sigma}_{11\cdot 2})$$

其中

$$\boldsymbol{\mu}_{1\cdot 2}=\boldsymbol{\mu}^{(1)}+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}^{(2)}-\boldsymbol{\mu}^{(2)}) \tag{1.26}$$

$$\boldsymbol{\Sigma}_{11\cdot 2}=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{1.27}$$

证明参见参考文献 [3]。

该定理告诉我们，$\boldsymbol{X}^{(1)}$ 的分布与 $(\boldsymbol{X}^{(1)}\mid\boldsymbol{X}^{(2)})$ 的分布均为正态，它们的协方差阵分别为 $\boldsymbol{\Sigma}_{11}$ 与 $\boldsymbol{\Sigma}_{11\cdot 2}=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$。由于 $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\geqslant\boldsymbol{0}$，故 $\boldsymbol{\Sigma}_{11}\geqslant\boldsymbol{\Sigma}_{11\cdot 2}$，等号成立当且仅当 $\boldsymbol{\Sigma}_{12}=\boldsymbol{0}$。协差阵是用来描述指标之间关系及散布程度的，$\boldsymbol{\Sigma}_{11}\geqslant\boldsymbol{\Sigma}_{11\cdot 2}$，说明了已知 $\boldsymbol{X}^{(2)}$ 的条件下，$\boldsymbol{X}^{(1)}$ 散布的程度比不知道 $\boldsymbol{X}^{(2)}$ 的情况下减小了，只有当 $\boldsymbol{\Sigma}_{12}=\boldsymbol{0}$ 时，两者相同。还可以证明，$\boldsymbol{\Sigma}_{12}=\boldsymbol{0}$，等价于 $\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 独立，这时，即使给出 $\boldsymbol{X}^{(2)}$，对 $\boldsymbol{X}^{(1)}$ 的分布也是没有影响的。

定理 1.3　设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}>\boldsymbol{0}$，将 $\boldsymbol{X}$，$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$ 剖分如下：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}^{(1)}\\ \boldsymbol{X}^{(2)}\\ \boldsymbol{X}^{(3)}\end{bmatrix}\begin{matrix}r\\ s\\ t\end{matrix},\ \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\\ \boldsymbol{\mu}^{(3)}\end{bmatrix}\begin{matrix}r\\ s\\ t\end{matrix},\ \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23}\\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33}\end{bmatrix}\begin{matrix}r\\ s\\ t\end{matrix} \tag{1.28}$$

则 $\boldsymbol{X}^{(1)}$ 有如下的条件均值和条件协方差阵的递推公式：

$$E(\boldsymbol{X}^{(1)}\mid\boldsymbol{X}^{(2)},\boldsymbol{X}^{(3)})=\boldsymbol{\mu}_{1\cdot 3}+\boldsymbol{\Sigma}_{12\cdot 3}\boldsymbol{\Sigma}_{22\cdot 3}^{-1}(\boldsymbol{X}^{(2)}-\boldsymbol{\mu}_{2\cdot 3}) \tag{1.29}$$

$$D(\boldsymbol{X}^{(1)}\mid\boldsymbol{X}^{(2)},\boldsymbol{X}^{(3)})=\boldsymbol{\Sigma}_{11\cdot 3}-\boldsymbol{\Sigma}_{12\cdot 3}\boldsymbol{\Sigma}_{22\cdot 3}^{-1}\boldsymbol{\Sigma}_{21\cdot 3} \tag{1.30}$$

其中 $\boldsymbol{\Sigma}_{ij\cdot k}=\boldsymbol{\Sigma}_{ij}-\boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{kk}^{-1}\boldsymbol{\Sigma}_{kj}$，　$i,j,k=1,2,3$

$\boldsymbol{\mu}_{i\cdot 3}=\boldsymbol{E}(\boldsymbol{X}^{(i)}\mid\boldsymbol{X}^{(3)})$，　$i=1,2$

证明参见参考文献［3］。

定理 1.2 和定理 1.3 在 20 世纪 70 年代中期国家标准部门制定服装标准时有成功的应用，见参考文献［3］。在制定服装标准时需抽样进行人体测量，现从某年龄段女性测量取出部分结果：

X_1：身高，X_2：胸围，X_3：腰围，X_4：上体长，X_5：臀围。已知它们遵从 $N_5(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中

$$\boldsymbol{\mu}=\begin{bmatrix}154.98\\83.39\\70.26\\61.32\\91.52\end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix}29.66&&&&\\6.51&30.53&&&\\1.85&25.54&39.86&&\\9.36&3.54&2.23&7.03&\\10.34&19.53&20.70&5.21&27.36\end{bmatrix}$$

若取 $\boldsymbol{X}^{(1)}=(X_1,X_2,X_3)'$，$\boldsymbol{X}^{(2)}=(X_4)$，$\boldsymbol{X}^{(3)}=(X_5)$，则由式（1.26）和式（1.27）得

$$E\left[\begin{matrix}X_1\\X_2\\X_3\\X_4\end{matrix}\middle|X_5\right]=\begin{bmatrix}154.98\\83.39\\70.26\\61.32\end{bmatrix}+\begin{bmatrix}10.34\\19.53\\20.70\\5.21\end{bmatrix}(27.36)^{-1}(X_5-91.52)$$

$$=\begin{bmatrix}154.98+0.38(X_5-91.52)\\83.39+0.71(X_5-91.52)\\70.26+0.76(X_5-91.52)\\61.32+0.19(X_5-91.52)\end{bmatrix}$$

$$D\left[\begin{matrix}X_1\\X_2\\X_3\\X_4\end{matrix}\middle|X_5\right]=\begin{bmatrix}29.66&6.51&1.85&9.36\\6.51&30.53&25.54&3.54\\1.85&25.54&39.86&2.23\\9.36&3.54&2.23&7.03\end{bmatrix}$$

$$-\begin{bmatrix}10.34\\19.53\\20.70\\5.21\end{bmatrix}(27.36)^{-1}(10.34,19.53,20.70,5.21)$$

$$=\begin{bmatrix}25.76&-0.86&-5.97&7.39\\-0.86&16.59&10.76&-0.18\\-5.97&10.76&24.19&-1.72\\7.39&-0.18&-1.72&6.04\end{bmatrix}$$

再利用式（1.30）得

$$D\begin{bmatrix} X_1 & X_4 \\ X_2 & \\ X_3 & X_5 \end{bmatrix}=\begin{bmatrix} 25.76 & -0.86 & -5.97 \\ -0.86 & 16.59 & 10.76 \\ -5.97 & 10.76 & 24.19 \end{bmatrix}$$
$$-\begin{bmatrix} 7.39 \\ -0.18 \\ -1.72 \end{bmatrix}(6.04)^{-1}(7.39,-0.18,-1.72)$$
$$=\begin{bmatrix} 16.72 & -0.64 & -3.87 \\ -0.64 & 16.58 & 10.71 \\ -3.87 & 10.71 & 23.71 \end{bmatrix}$$

此时看到

$$\text{var}(X_1 \mid X_4, X_5)=16.72<29.66=\text{var}(X_1)$$
$$\text{var}(X_2 \mid X_4, X_5)=16.58<30.53=\text{var}(X_2)$$
$$\text{var}(X_3 \mid X_4, X_5)=23.71<39.86=\text{var}(X_3)$$

这说明，若已知一个人的上体长和臀围，则身高、胸围和腰围的条件方差比原来的方差大大减小。

在定理 1.2 中，我们给出了对 $\boldsymbol{X}$，$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 作形如式（1.25）剖分时条件协方差阵 $\boldsymbol{\Sigma}_{11\cdot 2}$ 的表达式及其与非条件协方差阵的关系。令 $\sigma_{ij\cdot q+1,\cdots,p}$ 表示 $\boldsymbol{\Sigma}_{11\cdot 2}$ 的元素，则可以定义偏相关系数的概念如下：

定义 1.6　当 $\boldsymbol{X}^{(2)}$ 给定时，X_i 与 X_j 的偏相关系数为：

$$\boldsymbol{r}_{ij\cdot q+1,\cdots,p}=\frac{\sigma_{ij\cdot q+1,\cdots,p}}{(\sigma_{ii\cdot q+1,\cdots,p}\sigma_{jj\cdot q+1,\cdots,p})^{1/2}}$$

在上面制定服装标准的例子中，给出 X_4 和 X_5 时，X_1 与 X_2，X_1 与 X_3，X_2 与 X_3 的偏相关系数为：

$$r_{12\cdot 45}=\frac{-0.643}{\sqrt{16.717\times 16.582}}=-0.038\,6$$
$$r_{13\cdot 45}=\frac{-3.873}{\sqrt{16.717\times 23.707}}=-0.195$$
$$r_{23\cdot 45}=\frac{10.707}{\sqrt{16.582\times 23.707}}=0.540$$

定理 1.4　设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，将 $\boldsymbol{X}$，$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$ 按同样方式剖分为：

$$\boldsymbol{X}=\begin{bmatrix} \boldsymbol{X}^{(1)} \\ \vdots \\ \boldsymbol{X}^{(k)} \end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \vdots \\ \boldsymbol{\mu}^{(k)} \end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \vdots & & \vdots \\ \boldsymbol{\Sigma}_{k1} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}$$

其中，$\boldsymbol{X}^{(j)}$：$S_j\times 1$，$\boldsymbol{\mu}^{(j)}$：$S_j\times 1$，$\boldsymbol{\Sigma}_{jj}$：$S_j\times S_j$（$j=1,\ \cdots,\ k$），则 $\boldsymbol{X}^{(1)}$，$\boldsymbol{X}^{(2)}$，$\cdots$，$\boldsymbol{X}^{(k)}$

相互独立当且仅当 $\boldsymbol{\Sigma}_{ij}=\mathbf{0}$，对一切 $i\neq j$。

证明参见参考文献［3］。

因为 $\boldsymbol{\Sigma}_{12}=\mathrm{cov}(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$，该定理同时指出对多元正态分布而言，“$\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 不相关”等价于“$\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 独立”。

1.4 均值向量和协方差阵的估计

上节已经给出了多元正态分布的定义和有关的性质，在实际问题中，通常可以假定研究对象是多元正态分布，但分布中的参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 是未知的，一般的做法是通过样本来估计。

在一般情况下，如果样本资料阵为：

$$\boldsymbol{X}=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p)=\begin{bmatrix}\boldsymbol{X}'_{(1)}\\ \boldsymbol{X}'_{(2)}\\ \vdots\\ \boldsymbol{X}'_{(n)}\end{bmatrix}$$

设样品 $\boldsymbol{X}_{(1)}$，$\boldsymbol{X}_{(2)}$，…，$\boldsymbol{X}_{(n)}$ 相互独立，同遵从于 p 元正态分布 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，而且 $n>p$，$\boldsymbol{\Sigma}>\mathbf{0}$，则总体参数均值 $\boldsymbol{\mu}$ 的估计量为：

$$\hat{\boldsymbol{\mu}}=\overline{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{(i)}=\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}x_{i1}\\ \sum_{i=1}^{n}x_{i2}\\ \vdots\\ \sum_{i=1}^{n}x_{ip}\end{bmatrix}=\begin{bmatrix}\overline{X}_1\\ \overline{X}_2\\ \vdots\\ \overline{X}_p\end{bmatrix}\tag{1.31}$$

即均值向量 $\boldsymbol{\mu}$ 的估计量就是样本均值向量。这可由极大似然法推导出来。很显然，当样本资料选取的是 p 个指标的数据时，当然 $\hat{\boldsymbol{\mu}}=\overline{\boldsymbol{X}}$ 也是 p 维向量。

总体参数协方差阵 $\boldsymbol{\Sigma}$ 的极大似然估计为：

$$\hat{\boldsymbol{\Sigma}}_p=\frac{1}{n}\boldsymbol{L}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})'$$

$$=\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}(x_{i1}-\overline{X}_1)^2 & \cdots & \sum_{i=1}^{n}(x_{i1}-\overline{X}_1)(x_{ip}-\overline{X}_p)\\ & \sum_{i=1}^{n}(x_{i2}-\overline{X}_2)^2 & \cdots & \sum_{i=1}^{n}(x_{i2}-\overline{X}_2)(x_{ip}-\overline{X}_p)\\ & & & \vdots\\ & & & \sum_{i=1}^{n}(x_{ip}-\overline{X}_p)^2\end{bmatrix}$$

其中 $\boldsymbol{L}$ 是离差阵，它是每一个样品（向量）与样本均值（向量）的离差积形成的 n 个 $p\times p$ 阶对称阵的和。同一元相似，$\hat{\boldsymbol{\Sigma}}_p$ 不是 $\boldsymbol{\Sigma}$ 的无偏估计。为了得到无偏估计，我们常用样本协方差阵 $\hat{\boldsymbol{\Sigma}}=\frac{1}{n-1}\boldsymbol{L}$ 作为总体协方差阵的估计。

可以证明，$\bar{\boldsymbol{X}}$ 是 $\boldsymbol{\mu}$ 的无偏估计，是极小极大估计，是强相合估计，$\bar{\boldsymbol{X}}$ 还是 $\boldsymbol{\mu}$ 的充分统计量；$\hat{\boldsymbol{\Sigma}}$ 是 $\boldsymbol{\Sigma}$ 的强相合估计，但用 $\hat{\boldsymbol{\Sigma}}$ 估计 $\boldsymbol{\Sigma}$ 是有偏的，$\frac{1}{n-1}\boldsymbol{L}$ 才是 $\boldsymbol{\Sigma}$ 的无偏估计。在实际应用中，当 n 不是很大时，人们常用 $\frac{1}{n-1}\boldsymbol{L}$ 来估计 $\boldsymbol{\Sigma}$，但当 n 比较大时，用 $\hat{\boldsymbol{\Sigma}}$ 或 $\frac{1}{n-1}\boldsymbol{L}$ 差别不大。关于估计量 $\bar{\boldsymbol{X}}$，$\hat{\boldsymbol{\Sigma}}$ 的统计性质的证明，有兴趣的读者参见参考文献［3］。

1.5　常用分布及抽样分布

多元统计研究的是多指标问题，为了解总体的特征，通过对总体抽样得到代表总体的样本，但因为信息是分散在每个样本上的，就需要对样本进行加工，把样本的信息浓缩到不包含未知量的样本函数中，这个函数称为统计量，如前面介绍的样本均值向量 $\bar{\boldsymbol{X}}$、样本离差阵 $\boldsymbol{L}$ 等都是统计量。统计量的分布称为抽样分布。

在数理统计中常用的抽样分布有 χ^2 分布、t 分布和 F 分布。在多元统计中，与之对应的分布分别为 Wishart 分布、T^2 分布和 Wilks 分布。

1.5.1　χ^2 分布与 Wishart 分布

在数理统计中，若 $X_i\sim N(0,\ 1)(i=1,\ 2,\ \cdots,\ n)$，且相互独立，则 $\sum_{i=1}^{n}X_i^2$ 所遵从的分布为自由度为 n 的 χ^2 分布（chi-squared distribution），记为 $\chi^2(n)$。

$\chi^2(n)$ 分布是刻画正态变量二次型的一个重要分布，在一元统计分析中有着十分重要的地位，在对有关样本均值、样本方差的假设检验或非参数检验中经常用到 χ^2 统计量。

$\chi^2(n)$ 分布的均值和方差分别为：

$$E(\chi^2(n))=n$$
$$D(\chi^2(n))=2n$$

$\chi^2(n)$ 分布有两个重要的性质：

(1) 若 $\chi_i^2\sim\chi^2(n_i)(i=1,\ 2,\ \cdots,\ k)$，且相互独立，则

$$\sum_{i=1}^{k}\chi_i^2 \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

称为相互独立的 χ^2 变量具有可加性。

(2) 设 $X_i \sim N(0, 1)(i=1, 2, \cdots, n)$，且相互独立，$\boldsymbol{A}_j(j=1, 2, \cdots, m)$ 为 n 阶对称阵，且 $\sum_{j=1}^{m}\boldsymbol{A}_j = \boldsymbol{I}_n$（$n$ 阶单位阵），记 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)'$，$Q_j = \boldsymbol{X}'\boldsymbol{A}_j\boldsymbol{X}$，则 $Q_1, Q_2, \cdots, Q_m$ 为相互独立的 χ^2 分布变量的充要条件为 $\sum_{j=1}^{m}\mathrm{rank}(\boldsymbol{A}_j)=n$。此时 $Q_j \sim \chi^2(n_j)$，$n_j=\mathrm{rank}(\boldsymbol{A}_j)$。

这个性质称为 Cochran 定理，在方差分析和回归分析中起着重要作用。

从一元正态总体 $N(\mu, \sigma^2)$ 抽取容量为 n 的随机样本 $X_1, X_2, \cdots, X_n$，其样本均值 $\overline{X}$ 和样本方差 $S^2=\frac{l_{xx}}{n-1}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 的抽样分布有如下结果：

(1) $\overline{X}$ 和 S^2 相互独立。

(2) $\overline{X}\sim N\left(\mu, \frac{\sigma^2}{n}\right)$ 和 $\frac{(n-1)S^2}{\sigma^2}=\frac{l_{xx}}{\sigma^2}\sim\chi^2(n-1)$ 相互独立。

以上两个结论在数理统计中有着重要的应用。

在多元统计中，χ^2 分布发展为 Wishart 分布。Wishart 分布是由统计学家威沙特（Wishart）为研究多元样本离差阵 $\boldsymbol{L}$ 的分布于 1928 年推导出来的，有人就将这个时间作为多元分析诞生的时间。Wishart 分布在多元统计中的作用与 χ^2 分布在一元统计中类似，它可以由遵从多元正态分布的随机向量直接得到，同时它也是构成其他重要分布的基础。

定义 1.7 设 $\boldsymbol{X}_{(\alpha)}(\alpha=1, 2, \cdots, n)$ 相互独立，且 $\boldsymbol{X}_{(\alpha)}\sim N_p(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma})$，记 $\boldsymbol{X}=(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(n)})$，则随机矩阵

$$\boldsymbol{W}=\boldsymbol{X}'\boldsymbol{X}=\sum_{\alpha=1}^{n}\boldsymbol{X}_{(\alpha)}\boldsymbol{X}'_{(\alpha)} \tag{1.32}$$

所遵从的分布称为自由度为 n 的 p 维 Wishart 分布，记为 $W\sim W_p(n, \boldsymbol{\Sigma})$。其中，$n\geqslant p$，$\boldsymbol{\Sigma}>\boldsymbol{0}$。

由 Wishart 分布的定义知，当 $p=1$ 时，$\boldsymbol{\Sigma}$ 退化为 σ^2，此时中心 Wishart 分布就退化为 $\sigma^2\chi^2(n)$，由此可以看出，Wishart 分布实际上是 χ^2 分布在多维正态情形下的推广。

下面不加证明地给出 Wishart 分布的 5 条重要性质：

(1) 若 $\boldsymbol{X}_{(\alpha)}(\alpha=1, 2, \cdots, n)$ 是从 p 维正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 中抽取的 n 个随机样本，$\overline{\boldsymbol{X}}$ 为样本均值，样本离差阵为 $\boldsymbol{L}=\sum_{\alpha=1}^{n}(\boldsymbol{X}_{(\alpha)}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{(\alpha)}-\overline{\boldsymbol{X}})'$，则

1) $\overline{\boldsymbol{X}}$ 和 $\boldsymbol{L}$ 相互独立。

2) $\overline{\boldsymbol{X}}\sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$，$\boldsymbol{L}\sim\boldsymbol{W}_p(n-1, \boldsymbol{\Sigma})$。

(2) 若 $\boldsymbol{W}_i \sim \boldsymbol{W}_p(n_i, \boldsymbol{\Sigma})(i=1, 2, \cdots, k)$ 且相互独立，则

$$\sum_{i=1}^{k} \boldsymbol{W}_i \sim \boldsymbol{W}_p\left(\sum_{i=1}^{k} n_i, \boldsymbol{\Sigma}\right)$$

(3) 若 $\boldsymbol{W} \sim \boldsymbol{W}_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{C}_{q\times p}$ 为非奇异阵，则

$$\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}' \sim \boldsymbol{W}_q(n, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$$

(4) 若 $\boldsymbol{W} \sim \boldsymbol{W}_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{a}$ 为任一 p 元常向量，满足 $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \neq 0$，则 $\dfrac{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} \sim \chi^2(n)$。

(5) 若 $\boldsymbol{W} \sim \boldsymbol{W}_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{a}$ 为任一 p 元非零常向量，比值

$$\frac{\boldsymbol{a}'\boldsymbol{\Sigma}^{-1}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}^{-1}\boldsymbol{a}} \sim \chi^2(n-p+1)$$

特别地，设 w_{ii} 和 σ_{ii} 分别为 $\boldsymbol{W}^{-1}$ 和 $\boldsymbol{\Sigma}^{-1}$ 的第 i 个对角元，则

$$\frac{\sigma_{ii}}{w_{ii}} \sim \chi^2(n-p+1)$$

1.5.2　t 分布与 T^2 分布

在数理统计中，若 $X \sim N(0, 1)$，$Y \sim \chi^2(n)$，且 X 与 Y 相互独立，则称 $T=\dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ 遵从自由度为 n 的 t 分布，又称为学生分布（student distribution），记为 $T \sim t(n)$。如果将 T 平方，即 $T^2=n\dfrac{X^2}{Y}$，则 $T^2 \sim F(1, n)$，即 $t(n)$ 分布变量的平方遵从第一自由度为 1、第二自由度为 n 的中心 F 分布。

将上述 F 分布的定义改写成

$$F=nX'Y^{-1}X$$

式中，用 X' 表示 X 的转置。由于 X 为一维数字，转置与否都相同，所以可以这样写。

在多元统计中，仿照上式推广得到 T^2 分布的定义如下：

定义 1.8　设 $\boldsymbol{W} \sim \boldsymbol{W}_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{X} \sim N_p(\boldsymbol{0}, c\boldsymbol{\Sigma})$，$c>0$，$n \geqslant p$，$\boldsymbol{\Sigma}>\boldsymbol{0}$，$\boldsymbol{W}$ 与 $\boldsymbol{X}$ 相互独立，则称随机变量

$$T^2=\frac{n}{c}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X} \tag{1.33}$$

所遵从的分布为第一自由度为 p、第二自由度为 n 的中心 T^2 分布，记为 $T^2 \sim T^2(p, n)$。

T^2 分布是霍特林（Hotelling）于 1931 年由一元统计推广而来的，故 T^2 分布又称为 Hotelling T^2 分布。其作用相当于数理统计学中的 t 分布。

中心 T^2 分布可化为中心 F 分布，其关系可表示为：

$$\frac{n-p+1}{pn}T^2(p,n)=F(p,n-p+1)$$

显然，当 $p=1$ 时，有 $T^2(1,n)=F(1,n)$。

下面不加证明地给出 T^2 分布的两条重要性质：

(1) 设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{W}\sim W_p(n,\boldsymbol{\Sigma})$，且 $\boldsymbol{X}$ 与 $\boldsymbol{W}$ 相互独立，则

$$n(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{W}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\sim T^2(p,n)$$

推论 设 $\boldsymbol{X}_{(\alpha)}=(X_{\alpha 1},X_{\alpha 2},\cdots,X_{\alpha p})'(\alpha=1,2,\cdots,n)$ 是从 p 维正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 中抽取的 n 个随机样本，$\bar{\boldsymbol{X}}$ 为样本均值，样本离差阵为 $\boldsymbol{L}=\sum_{\alpha=1}^{n}(\boldsymbol{X}_{(\alpha)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(\alpha)}-\bar{\boldsymbol{X}})'$，则

$$n(n-1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{L}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\sim T^2(p,n-1)$$

或

$$n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\sim T^2(p,n-1)$$

其中，$\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{L}$，为样本的方差。

(2) 设 $\boldsymbol{X}_i\sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})(i=1,2)$，从总体 X_1，X_2 中取得容量分别为 n_1，n_2 的两个随机样本，若 $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$，则

$$\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)\sim T^2(p,n_1+n_2-2)$$

或

$$(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)\sim\frac{n_1+n_2}{n_1n_2}T^2(p,n_1+n_2-2)$$

其中，$\bar{\boldsymbol{X}}_1$，$\bar{\boldsymbol{X}}_2$ 为两样本的均值向量；$\boldsymbol{S}_p=\frac{n_1\boldsymbol{S}_1+n_2\boldsymbol{S}_2}{n_1+n_2-2}$；$\boldsymbol{S}_1$，$\boldsymbol{S}_2$ 分别为两样本的方差阵。

这个性质在下章的假设检验中有重要应用。

1.5.3 中心 F 分布与 Wilks 分布

在一元统计学中，若 $X\sim\chi^2(m)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立，则称 $F=\frac{X/m}{Y/n}$ 所遵从的分布为第一自由度为 m、第二自由度为 n 的中心 F 分布，记为 $F\sim F(m,n)$。F 分布本质上是从正态总体 $N(\mu,\sigma^2)$ 中随机抽取的两个样本方差的比。

F 分布能否推广到多元呢？由于 F 分布由两个方差比构成，而多元总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的变异由协方差阵确定，它不是一个数字，这就产生了如何用与协方差阵 $\boldsymbol{\Sigma}$ 有关的一个量来描述总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的变异的问题，它是将 F 分布推广到多元情形的关键。

描述 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$的变异度的统计参数称为广义方差。围绕这一问题产生了许多方法，有的用行列式，有的用迹，主要的方法有以下几种：

(1) 广义方差$\triangleq|\boldsymbol{\Sigma}|$；

(2) 广义方差$\triangleq \operatorname{tr}(\boldsymbol{\Sigma})=\sum_{i=1}^{p}\sigma_i^2=\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2$，其中 $\operatorname{tr}(\boldsymbol{\Sigma})$ 为 $\boldsymbol{\Sigma}$ 的迹，等于 $\boldsymbol{\Sigma}$ 主对角线元素之和；

(3) 广义方差$\triangleq \operatorname{tr}(\boldsymbol{\Sigma})=\prod_{i=1}^{p}\sigma_i^2$；

(4) 广义方差$\triangleq |\boldsymbol{\Sigma}|^{\frac{1}{p}}$；

(5) 广义方差$\triangleq (\operatorname{tr}(\boldsymbol{\Sigma}))^{\frac{1}{2}}=\sqrt{\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2}$；

(6) 广义方差$\triangleq\max\{\lambda_i\}$，其中 λ_i 为 $\boldsymbol{\Sigma}$ 的特征根；

(7) 广义方差$\triangleq\min\{\lambda_i\}$，其中 λ_i 为 $\boldsymbol{\Sigma}$ 的最大特征根。

在以上各种广义方差的定义中，目前使用最多的是第一种，它是 T. W. 安德森 (T. W. Anderson) 于 1958 年提出来的。

下面根据第一种广义方差，仿照 F 分布的定义给出多元统计中两个广义方差之比的统计量，称为 Wilks Λ 分布。

定义 1.9　设 $\boldsymbol{W}_1\sim W_p(n_1, \boldsymbol{\Sigma})$，$\boldsymbol{W}_2\sim W_p(n_2, \boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}>\boldsymbol{0}$，$n_1>p$，且 $\boldsymbol{W}_1$ 与 $\boldsymbol{W}_2$ 相互独立，则

$$\Lambda=\frac{|\boldsymbol{W}_1|}{|\boldsymbol{W}_1+\boldsymbol{W}_2|} \tag{1.34}$$

所遵从的分布称为维数为 p，第一自由度为 n_1，第二自由度为 n_2 的 Wilks $\boldsymbol{\Lambda}$ 分布，记为 $\Lambda\sim\Lambda(p, n_1, n_2)$。

由上述定义，Λ 分布为两个广义方差之比。

由于 Λ 分布在多元统计中的重要性，关于它的近似分布和精确分布不断有学者进行研究。当 p 和 n_2 中的一个比较小时，Λ 分布可化为 F 分布，表 1—2 列举了常见的情况。

表 1—2　　**$\Lambda\sim\Lambda(p, n_1, n_2)$ 与 F 分布的关系，$n_1>p$**

p	n_2	统计量 F	F 的自由度
任意	1	$\dfrac{1-\Lambda}{\Lambda}\dfrac{n_1-p+1}{p}$	p，n_1-p+1
任意	2	$\dfrac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\dfrac{n_1-p}{p}$	$2p$，$2(n_1-p)$
1	任意	$\dfrac{1-\Lambda}{\Lambda}\dfrac{n_1}{n_2}$	n_2，n_1
2	任意	$\dfrac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\dfrac{n_1-1}{n_2}$	$2n_2$，$2(n_1-1)$

当 p，n_2 不属于表 1—2 所列举的情况时，巴特莱特 (Bartlett) 指出可用 χ^2 分布来近似表示，即

$$V=-(n_1+n_2-\frac{p+n_2+1}{2})\ln\Lambda(p,n_1,n_2)$$

近似遵从 $\chi^2(pn_2)$。

拉奥（Rao）后来又研究用 F 分布来近似，即

$$R=\frac{1-\Lambda^{\frac{1}{s}}}{\Lambda^{\frac{1}{s}}}\frac{ts-2\lambda}{pn_2}$$

近似遵从 $F(pn_2,\ ts-2\lambda)$，其中

$$\begin{cases} t=n_1+n_2-\dfrac{p+n_2+1}{2} \\ s=\sqrt{\dfrac{p^2n_2^2-4}{p^2+n_2^2-5}} \\ \lambda=\dfrac{pn_2-2}{4} \end{cases}$$

$ts-2\lambda$ 不一定是整数，用与它最近的整数来作为 F 分布的第二自由度。

若 $n_2<p$，有 $\Lambda(p,\ n_1,\ n_2)=\Lambda(n_2,\ p,\ n_1+n_2-p)$。该结论说明，在使用 Λ 统计量时也可考虑 $n_2>p$ 的情形，有关 Λ 统计量的其他性质参见参考文献［1］。

□ 参考文献

［1］张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982

［2］Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley，1982

［3］方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

［4］王学仁. 地质数据的多变量统计分析. 北京：科学出版社，1986

［5］王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993

［6］G. A. F. Seber. *Multivariate Observations*. John Wiley & Sons，Inc.，1984

□ 思考与练习

1. 在数据处理时，为什么通常要进行标准化处理？

2. 欧氏距离与马氏距离的优缺点是什么？

3. 当变量 X_1 和 X_2 方向上的变差相等，且 X_1 与 X_2 互相独立时，采用欧氏距离与统计距离是否一致？

4. 如果正态随机向量 $\boldsymbol{X}=(x_1, x_2, \cdots, x_p)'$ 的协方差阵 $\boldsymbol{\Sigma}$ 是对角阵，证明 $\boldsymbol{X}$ 的分量是相互独立的随机变量。

5. y_1 与 y_2 是相互独立的随机变量，且 $y_1 \sim N(0, 1)$，$y_2 \sim N(3, 4)$。

(a) 求 y_1^2 的分布。

(b) 如果 $\boldsymbol{y}=\begin{bmatrix} y_1 \\ (y_2-3)/2 \end{bmatrix}$，写出 $\boldsymbol{y}'\boldsymbol{y}$ 关于 y_1 与 y_2 的表达式，并写出 $\boldsymbol{y}'\boldsymbol{y}$ 的分布。

(c) 如果 $\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ 且 $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，写出 $\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ 关于 y_1 与 y_2 的表达式，并写出 $\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ 的分布。

第 2 章

Chapter 2 均值向量和协方差阵的检验

学习目标

1. 掌握均值向量及协方差阵的检验方法；

2. 能够用 SPSS 软件或 R 软件实现均值向量及协方差阵的检验，并正确理解输出结果。

在一元统计中，关于正态总体 $N(\mu, \sigma^2)$ 的均值 μ 和方差 σ^2 的各种检验，已给出了常用的 z 检验、t 检验、F 检验和 χ^2 检验等。对于包含多个指标的正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，各种实际问题同样要求对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 进行统计推断。例如，要考察某工业行业的生产经营状况、今年与去年相比指标的平均水平有无显著差异，以及各生产经营指标间的波动是否有显著差异，需做检验 $H_0: \boldsymbol{\mu}=\boldsymbol{\mu}_0$，$H_1: \boldsymbol{\mu}\neq\boldsymbol{\mu}_0$，或 $H_0: \boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$，$H_1: \boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}_0$ 等。关于 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的各种形式的假设检验构成了本章的内容。本章的很多内容是一元的直接推广，但由于多指标问题的复杂性，本章将只列出检验用的统计量，主要详细介绍如何使用这些统计量做检验，对有关检验问题的理论推证则全部略去。本章最后还将介绍有关检验的上机实现。

2.1 均值向量的检验

2.1.1 一个指标检验的回顾

设从总体 $N(\mu, \sigma^2)$ 中抽取一个样本 $x_1, x_2, \cdots, x_n$，我们要检验假设

$$H_0: \mu=\mu_0, \quad H_1: \mu\neq\mu_0$$

当 σ^2 已知时，用统计量

$$z=\frac{\bar{x}-\mu_0}{\sigma}\sqrt{n} \tag{2.1}$$

式中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 为样本均值。当假设成立时，统计量 z 遵从正态分布，$z\sim N(0,1)$，从而拒绝域为 $|z|>z_{\alpha/2}$，$z_{\alpha/2}$ 为 $N(0,1)$ 的上 $\alpha/2$ 分位点。

当 σ^2 未知时，用

$$S^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{(n-1)}$$

作为 σ^2 的估计，用统计量

$$t=\frac{\bar{x}-\mu_0}{S}\sqrt{n} \tag{2.2}$$

来做检验。当假设成立时，t 统计量遵从自由度为 $n-1$ 的 t 分布，$t\sim t_{n-1}$，拒绝域为 $|t|>t_{n-1}(\alpha/2)$，$t_{n-1}(\alpha/2)$ 为 t_{n-1} 的上 $\alpha/2$ 分位点。统计量（2.2）也可改写成如下形式：

$$t^2=n(\bar{x}-\mu_0)'(S^2)^{-1}(\bar{x}-\mu_0) \tag{2.3}$$

当假设为真时，统计量 t^2 遵从第一自由度为 1、第二自由度为 $n-1$ 的 F 分布，简写成 $t^2\sim F_{1,n-1}$，其拒绝域为：

$$t^2>F_{1,n-1}(\alpha)$$

$F_{1,n-1}(\alpha)$ 为 $F_{1,n-1}$ 的上 α 分位点。

2.1.2　多元均值检验

某工业行业的管理机构想要掌握所属企业的生产经营活动情况，选取了 p 个指标进行考察。根据历史资料的记载，将 p 个指标的历史平均水平记作 $\boldsymbol{\mu}_0$。今年的 p 个指标平均值与历史资料记载的平均值有无显著差异？若有差异，进一步分析差异主要在哪些指标上。类似这样的问题，就是要对下面的假设

$$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0,\quad H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$$

做检验。检验的思想和步骤与一元相似，可归纳如下：

（1）根据问题的要求提出统计假设 H_0 及 H_1；

（2）选取一个合适的统计量，并求出它的抽样分布；

（3）指定 α 风险值（即显著性水平 α 值），并在零假设 H_0 为真的条件下求出能使风险值控制在 α 的临界值 W；

（4）建立判别准则；

（5）由样本观测值计算统计量值，再由准则作统计判断，最后对统计判断做出具体的解释。

设 $\boldsymbol{X}_{(\alpha)}=(X_{\alpha 1}, X_{\alpha 2}, \cdots, X_{\alpha p})'(\alpha=1, 2, \cdots, n)$ 是容量为 n 的一个样本，它们来自均值向量为 $\boldsymbol{\mu}$，协方差阵为 $\boldsymbol{\Sigma}$（$\boldsymbol{\Sigma}>\boldsymbol{0}$ 是正定阵）的 p 元正态总体，对于指定向量 $\boldsymbol{\mu}_0$，要对下面的假设

$$H_0: \boldsymbol{\mu}=\boldsymbol{\mu}_0, \quad H_1: \boldsymbol{\mu}\neq\boldsymbol{\mu}_0 \tag{2.4}$$

做检验。检验的方法与一元相似，将分两种情况讨论。

（1）协方差阵 $\boldsymbol{\Sigma}$ 已知。

类似于式（2.3）的统计量（注意式（2.3）的形式）是

$$\chi_0^2=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0) \tag{2.5}$$

可以证明，在假设 H_0 为真时，统计量 χ_0^2 遵从自由度为 p 的 χ^2 分布；事实上由 1.5 节 $\bar{\boldsymbol{X}}-\boldsymbol{\mu}\sim N_p(\boldsymbol{0}, \frac{1}{n}\boldsymbol{\Sigma})$ 知，当 H_0：$\boldsymbol{\mu}=\boldsymbol{\mu}_0$ 成立时，由多元正态分布的性质（4），有

$$(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'(\frac{1}{n}\boldsymbol{\Sigma})^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\sim\chi^2(p)$$

统计量 χ_0^2 实质上是样本均值 $\bar{\boldsymbol{X}}$ 与已知平均水平 $\boldsymbol{\mu}_0$ 之间的马氏距离的 n 倍，这个值越大，$\boldsymbol{\mu}$ 与 $\boldsymbol{\mu}_0$ 相等的可能性就越小，因而在备择假设 H_1 成立时，χ_0^2 有变大的趋势，所以拒绝域应取 χ_0^2 值较大的右侧部分。式中，$\bar{\boldsymbol{X}}$ 是样本均值；n 是样本容量。

当给定显著性水平 α 后，由样本值可以算出 χ_0^2 的值。当

$$\chi_0^2=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\geqslant\chi_p^2(\alpha)$$

时，便拒绝零假设 H_0，说明均值 $\boldsymbol{\mu}$ 不等于 $\boldsymbol{\mu}_0$，其中 $\chi_p^2(\alpha)$ 是自由度为 p 的 χ^2 分布的上 α 分位点。即

$$P\{\chi_0^2>\chi_p^2(\alpha)\}=\alpha$$

（2）协方差阵 $\boldsymbol{\Sigma}$ 未知。

此时 $\boldsymbol{\Sigma}$ 的无偏估计是 $\hat{\boldsymbol{\Sigma}}=\frac{\boldsymbol{L}}{(n-1)}$，类似于式（2.3）的统计量是

$$\begin{aligned}T^2&=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\\&=n(n-1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{L}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\end{aligned} \tag{2.6}$$

可以证明，统计量 T^2 遵从参数为 p，$n-1$ 的 T^2 分布，即 $T^2\sim T_{p,n-1}^2$。统计量 T^2 实际上也是样本均值 $\bar{\boldsymbol{X}}$ 与已知均值向量 $\boldsymbol{\mu}_0$ 之间的马氏距离再乘以 $n(n-1)$，这个值越大，$\boldsymbol{\mu}$ 与 $\boldsymbol{\mu}_0$ 相等的可能性就越小。因而在备择假设成立时，T^2 的值有变大的趋势，所以拒绝域可取 T^2 值较大的右侧部分。因此，在给定显著性水平 α 后，由样本的数值可以算出 T^2 值。当

$$T^2>T_{p,n-1}^2(\alpha) \tag{2.7}$$

时，便拒绝零假设 H_0。$T_{p,n-1}^2(\alpha)$ 为 $T_{p,n-1}^2$ 的上 α 分位点。

T^2 分布的 5%及 1%的分位点已列成专表，可从网上下载。

由 1.5 节，将 T^2 统计量乘上一个适当的常数后，便成为 F 统计量，也可用 F 分布表获得零假设的拒绝域，即

$$\left\{\frac{n-p}{(n-1)p}T^2>F_{p,n-p}(\alpha)\right\} \tag{2.8}$$

关于 χ_0^2，T^2 的合理性及推证见参考文献 [3]。

在实际工作中，一元检验与多元检验可以联合使用，多元检验具有概括和全面考察的特点，而一元检验容易发现各指标之间的关系和差异，能帮助我们找出存在差异的侧重面，提供了更多的统计分析信息。

2.1.3　两总体均值的比较

在许多实际问题中，往往要比较两个总体的平均水平之间有无差异。例如，两所大学新生录取成绩是否有明显差异。研究职工工资总额的构成情况，若按国民经济行业分组，就是要研究工业与建筑业这两个行业之间是否有明显的不同；同理，可按工业领导关系（中央、省、市、县属工业）分组，也可按工业行业分组。组与组之间的工资总额构成有无显著差异，本质上就是两个总体的均值向量是否相等，这类问题通常也称为两样本问题。两总体均值的比较问题又可分为两总体协方差阵相等与两总体协方差阵不相等两种情形。

1. 协方差阵相等的情形

设 $\boldsymbol{X}_{(\alpha)}=(X_{\alpha 1}, X_{\alpha 2}, \cdots, X_{\alpha p})'(\alpha=1, 2, \cdots, n_1)$ 为来自 p 元正态总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ 的容量为 n_1 的样本，$\boldsymbol{Y}_{(\alpha)}=(Y_{\alpha 1}, Y_{\alpha 2}, \cdots, Y_{\alpha p})'(\alpha=1, 2, \cdots, n_2)$ 为来自 p 元正态总体 $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ 的容量为 n_2 的样本，且两样本相互独立，$n_1>p$，$n_2>p$，假定两总体协方差阵相等但未知，现对假设

$$H_0: \boldsymbol{\mu}=\boldsymbol{\mu}_0, \quad H_1: \boldsymbol{\mu}\neq\boldsymbol{\mu}_0 \tag{2.9}$$

进行检验。与前面类似的统计量的形式为：

$$T^2=\frac{n_1 n_2}{n_1+n_2}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) \tag{2.10}$$

式中，$\bar{\boldsymbol{X}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i$，$\bar{\boldsymbol{Y}}=\frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{y}_i$；$n_1, n_2$ 是样本容量；$\hat{\boldsymbol{\Sigma}}=\frac{(\boldsymbol{L}_x+\boldsymbol{L}_y)}{(n_1+n_2-2)}$，是协方差阵 $\boldsymbol{\Sigma}$ 的估计量；$\boldsymbol{L}_x=\sum_{i=1}^{n_1}(\boldsymbol{x}_i-\bar{\boldsymbol{X}})(\boldsymbol{x}_i-\bar{\boldsymbol{X}})'$，$\boldsymbol{L}_y=\sum_{i=1}^{n_2}(\boldsymbol{y}_i-\bar{\boldsymbol{Y}})(\boldsymbol{y}_i-\bar{\boldsymbol{Y}})'$，是两个总体的样本离差阵。

当假设 H_0：$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 成立时，$T^2\sim T^2_{p,n_1+n_2-2}$，从而

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2\sim F_{p,n_1+n_2-p-1} \tag{2.11}$$

当备择假设 H_1：$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ 成立时，$\dfrac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2=F^*$ 有变大的趋势，因为 T^2 的值与总体均值的马氏距离 $(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})$ 成正比，比值越大，说明两总体的均值很接近的可能性越小，因而拒绝域可以取 F^* 值较大的右侧区域，即当给定显著性水平 α 的值时，若

$$F^* > F_{p,n_1+n_2-p-1}(\alpha) \tag{2.12}$$

拒绝 H_0，否则没有足够理由拒绝 H_0。

2. 协方差阵不相等情形

设从两个总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ 和 $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ 中分别抽取容量为 n_1 和 n_2 的两个样本，$\boldsymbol{X}_{(\alpha)}=(X_{\alpha 1}, X_{\alpha 2}, \cdots, X_{\alpha p})'(\alpha=1, 2, \cdots, n_1)$，$\boldsymbol{Y}_{(\alpha)}=(Y_{\alpha 1}, Y_{\alpha 2}, \cdots, Y_{\alpha p})'(\alpha=1, 2, \cdots, n_2)$，$n_1>p$，$n_2>p$，假定两总体协方差阵不相等，我们考虑对假设式（2.9）做检验。这是著名的 Behrens-Fisher 问题。长期以来，统计学家用许多方法试图解决这个问题。当 $\boldsymbol{\Sigma}_1$ 与 $\boldsymbol{\Sigma}_2$ 相差很大时，T^2 统计量的形式为：

$$\begin{aligned} T^2 &= (\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\left[\frac{\boldsymbol{L}_x}{n_1(n_1-1)}+\frac{\boldsymbol{L}_y}{n_2(n_2-1)}\right]^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) \\ &= (\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\boldsymbol{S}_*^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) \end{aligned} \tag{2.13}$$

式中，$\bar{\boldsymbol{X}}$，$\bar{\boldsymbol{Y}}$，$\boldsymbol{L}_x$ 和 $\boldsymbol{L}_y$ 的统计含义与前相同，$\boldsymbol{S}_*=\dfrac{\boldsymbol{L}_x}{n_1(n_1-1)}+\dfrac{\boldsymbol{L}_y}{n_2(n_2-1)}$。再令

$$\begin{aligned} f^{-1} &= (n_1^3-n_1^2)^{-1}\left[(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\boldsymbol{S}_*^{-1}\left(\frac{\boldsymbol{L}_x}{n_1-1}\right)\boldsymbol{S}_*^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\right]^2 T^{-4} \\ &\quad +(n_2^3-n_2^2)^{-1}\left[(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\boldsymbol{S}_*^{-1}\frac{\boldsymbol{L}_y}{(n_2-1)}\boldsymbol{S}_*^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\right]^2 T^{-4} \end{aligned}$$

当假设式（2.9）的 H_0 成立时，可以证明（见参考文献 [3]）$\left(\dfrac{f-p+1}{fp}\right)T^2$ 近似遵从第一自由度为 p、第二自由度为 $f-p+1$ 的 F 分布，即

$$\left(\frac{f-p+1}{fp}\right)T^2 \sim F_{p,f-p+1} \tag{2.14}$$

当 $\min(n_1, n_2)\to\infty$ 时，T^2 近似于 χ_p^2。

2.1.4 多总体均值的检验

在许多实际问题中，我们要研究的总体往往不止两个。例如，要对全国的工业行业的生产经营状况做比较时，一个行业可以看成一个总体，此时要研究的总体多达几十甚至几百个，就需要运用多元方差分析的知识。多元方差分析是一元方差分析的直接推广。为了便于理解多元方差分析的方法，我们先回顾一元方差分析。

设有 r 个总体 G_1，G_2，…，G_r，它们的分布分别是一元正态 $N(\mu_1, \sigma^2)$，$N(\mu_2, \sigma^2)$，…，$N(\mu_r, \sigma^2)$，从各个总体中抽取的样本如下：

$$X_1^{(1)},X_2^{(1)},\cdots,X_{n_1}^{(1)}\sim N(\mu_1,\sigma^2)$$
$$X_1^{(2)},X_2^{(2)},\cdots,X_{n_2}^{(2)}\sim N(\mu_2,\sigma^2)$$
$$\cdots\cdots\cdots\cdots$$
$$X_1^{(r)},X_2^{(r)},\cdots,X_{n_r}^{(r)}\sim N(\mu_r,\sigma^2)$$

假设 r 个总体的方差相等，要检验的假设就是

$$H_0:\mu_1=\cdots=\mu_r,\quad H_1:\text{至少存在 } i\neq j,\text{使得 } \mu_i\neq\mu_j$$

这个检验的统计量与下列平方和密切相关：

$$\text{组间平方和 } SS(TR)=\sum_{k=1}^{r}n_k(\overline{X}_k-\overline{X})^2$$

$$\text{组内平方和 } SSE=\sum_{k=1}^{r}\sum_{j=1}^{n_k}(X_j^{(k)}-\overline{X}_k)^2$$

$$\text{总平方和 } SST=\sum_{k=1}^{r}\sum_{j=1}^{n_k}(X_j^{(k)}-\overline{X})^2$$

式中，$\overline{X}_k=\dfrac{1}{n_k}\sum_{j=1}^{n_k}X_j^{(k)}$ 是第 k 组的均值；$\overline{X}=\dfrac{1}{n}\sum_{k=1}^{r}\sum_{j=1}^{n_k}X_j^{(k)}$ 是总均值；$n=n_1+n_2+\cdots+n_r$。

一元统计中，构造 F 统计量的方法是

$$\frac{\text{组间平方和/自由度}}{\text{组内平方和/自由度}}$$

即
$$F=\frac{SS(TR)/(r-1)}{SSE/(n-r)}$$

当假设为真时，F 统计量遵从自由度为 $(r-1)$，$(n-r)$ 的 F 分布，记为$F\sim F_{r-1,n-r}$，零假设的拒绝域为：

$$F>F_{r-1,n-r}(\alpha)$$

将上述方法推广到多元，就是设有 r 个总体G_1，G_2，…，G_r，从这 r 个总体中抽取独立样本如下：

$$\boldsymbol{X}_1^{(1)},\boldsymbol{X}_2^{(1)},\cdots,\boldsymbol{X}_{n_1}^{(1)}\sim N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$$
$$\cdots\cdots\cdots\cdots$$
$$\boldsymbol{X}_1^{(r)},\boldsymbol{X}_2^{(r)},\cdots,\boldsymbol{X}_{n_r}^{(r)}\sim N_p(\boldsymbol{\mu}_r,\boldsymbol{\Sigma})$$

样本 $\{\boldsymbol{X}_j^{(k)}\}(k=1,2,\cdots,r;\ j=1,2,\cdots,n_k)$ 相互独立，要检验的假设就是

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_r,\quad H_1:\text{至少存在 } i\neq j,\text{使得 } \boldsymbol{\mu}_i\neq\boldsymbol{\mu}_j \tag{2.15}$$

用类似于一元方差分析的办法，前面所述的三个平方和就变成了矩阵，形式如下：

$$\boldsymbol{B}=SS(TR)=\sum_{k=1}^{r}n_k(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}})'$$

$$\boldsymbol{E}=SSE=\sum_{k=1}^{r}\sum_{j=1}^{n_k}(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}}_k)(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}}_k)' \tag{2.16}$$

$$\boldsymbol{W}=SST=\sum_{k=1}^{r}\sum_{j=1}^{n_k}(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}})'$$

很显然，$\boldsymbol{W}=\boldsymbol{B}+\boldsymbol{E}$。

关于 $\boldsymbol{B}$ 的检验可用 Wilks Λ 分布，再转化为 F 分布，具体可参考 1.5 节。

2.2 协方差阵的检验

上面讨论了多元正态分布均值的检验，但这仅研究了问题的一个方面，倘若要深究不同总体的平均水平（均值）的波动幅度，前面介绍的方法就无能为力了。本节所介绍的协方差阵的检验可以解决该类问题。

2.2.1 检验 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$

设 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_n$ 是来自正态总体 $N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ 的一个样本，$\boldsymbol{\Sigma}_0$ 是已知的正定矩阵，要检验

$$H_0:\ \boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0,\quad H_1:\ \boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}_0 \tag{2.17}$$

检验假设式（2.17）所用的统计量是

$$M=(n-1)[\ln|\boldsymbol{\Sigma}_0|-p-\ln|\hat{\boldsymbol{\Sigma}}|+\mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_0^{-1})] \tag{2.18}$$

式中，$\hat{\boldsymbol{\Sigma}}=\dfrac{\boldsymbol{L}}{n-1}$，是样本协方差阵。关于统计量 M 的推证过程见参考文献 [1]。柯林（Korin，1968）已导出 M 的极限分布和近似分布，并对小的 n 给出了表，即当 $p\leqslant 10$，$n\leqslant 75$，$\alpha=0.05$ 及 $\alpha=0.01$ 时 M 的 α 分位点表。当 $p>10$ 或 $n>75$ 时，M 近似于 $bF(f_1,\ f_2)$，记作

$$M\approx bF(f_1,f_2) \tag{2.19}$$

其中，$D_1=\dfrac{2p+1-\dfrac{2}{p+1}}{6(n-1)}$；$D_2=\dfrac{(p-1)(p+2)}{6(n-1)^2}$；$f_1=\dfrac{p(p+1)}{2}$；$f_2=\dfrac{f_1+2}{D_2-D_1^2}$；$b=\dfrac{f_1}{1-D_1-\dfrac{f_1}{f_2}}$。

2.2.2 检验 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_r$

上面讨论的检验 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$，是帮助我们分析当前的波动幅度与过去的波动情形有无显著差异。但在实际问题中我们往往面临多个总体，需要了解这多个总体之间的

波动幅度有无明显的差异。例如在研究职工工资构成时，若按工业行业分组，就有采掘业、制造业、文化教育、金融保险等，不同行业间工资总额的构成存在波动，研究波动是否存在显著的差异，就是做行业间协方差阵相等性的检验。用统计理论来描述就是：

设有 r 个总体，从各个总体中抽取样品如下：

$$\boldsymbol{X}_1^{(1)},\boldsymbol{X}_2^{(1)},\cdots,\boldsymbol{X}_{n_1}^{(1)} \sim N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$$
$$\cdots\cdots\cdots\cdots$$
$$\boldsymbol{X}_1^{(r)},\boldsymbol{X}_2^{(r)},\cdots,\boldsymbol{X}_{n_r}^{(r)} \sim N_p(\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r)$$
$$n = n_1 + n_2 + \cdots + n_r$$

此时要检验的假设是

$$H_0:\ \boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_r,\ H_1:\ \{\boldsymbol{\Sigma}_i\}\ \text{不全相等} \tag{2.20}$$

检验所用的统计量为：

$$M=(n-r)\ln\left|\frac{\boldsymbol{L}}{(n-r)}\right|-\sum_{i=1}^{r}(n_i-1)\ln\left|\frac{\boldsymbol{L}_i}{(n_i-1)}\right| \tag{2.21}$$

其中

$$\boldsymbol{L}_k=\sum_{i=1}^{n_k}(\boldsymbol{X}_i^{(k)}-\bar{\boldsymbol{X}}_k)(\boldsymbol{X}_i^{(k)}-\bar{\boldsymbol{X}}_k)'$$
$$\bar{\boldsymbol{X}}_k=\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_i^{(k)},\quad k=1,2,\cdots,r$$
$$\boldsymbol{L}=\sum_{k=1}^{r}\boldsymbol{L}_i$$

当 r，p，n 不大且 $n_1=n_2=\cdots=n_r=n_0$ 时，本书电子附表 4 中列出了 M 的上 α 分位点；当 r，p，n 较大且 $\{n_i\}$ 互不相等时，该表中未列出它们对应的临界值，此时可用 F 分布去近似，M 近似遵从 $bF(f_1,\ f_2)$，记作

$$M\approx bF(f_1,f_2) \tag{2.22}$$

其中

$$f_1=\frac{p(p+1)(r-1)}{2},\quad f_2=\frac{(f_1+2)}{(d_2-d_1^2)},\quad b=\frac{f_1}{(1-d_1-\frac{f_1}{f_2})}$$

$$d_1=\begin{cases}\dfrac{2p^2+3p-1}{6(p+1)(r-1)}\left(\displaystyle\sum_{i=1}^{r}\frac{1}{n_i-1}-\frac{1}{n-r}\right), & \text{至少有一对 } n_i\neq n_j\\ \dfrac{(2p^2+3p-1)(r-1)}{6(p+1)r(n-1)}, & n_1=n_2=\cdots=n_r\end{cases}$$

$$d_2=\begin{cases}\dfrac{(p-1)(p+2)}{6(r-1)}\left(\displaystyle\sum_{i=1}^{r}\frac{1}{(n_i-1)^2}-\frac{1}{(n-r)^2}\right), & \text{至少有一对 } n_i\neq n_j\\ \dfrac{(p-1)(p+2)(r^2+r+1)}{6r^2(n-1)^2}, & n_1=n_2=\cdots=n_r\end{cases}$$

关于协方差阵检验的具体例子可参见参考文献［4］。

2.3 有关检验的上机实现

例 2—1

1999 年国家财政部、经贸委、人事部和计委联合发布了《国有资本金效绩评价规则》。其中，竞争性工商企业的评价指标体系包括下面八大基本指标：净资产收益率、总资产报酬率、总资产周转率、流动资产周转率、资产负债率、已获利息倍数、销售增长率和资本积累率。下面我们借助于这一指标体系对我国上市公司的运营情况进行分析，以下数据（见表 2—1）为 35 家上市公司 2008 年年报数据，这 35 家上市公司分别来自于电力、煤气及水的生产和供应业，房地产业，信息技术业，在后面各章中也经常以该数据为例进行分析。

表 2—1

行业	公司简称	股票代码	净资产收益率	总资产报酬率	资产负债率	总资产周转率	流动资产周转率	已获利息倍数	销售增长率	资本积累率
电力、煤气及水的生产和供应业	深圳能源	000027	9.17	4.92	53.45	0.39	1.57	3.56	2.76	33.00
	深南电 A	000037	0.61	1.23	61.17	0.60	1.74	1.41	−12.81	−0.01
	富龙热电	000426	−11.30	−5.56	48.89	0.13	0.76	−0.34	−40.10	−9.93
	穗恒运 A	000531	−7.70	−1.53	70.25	0.57	2.70	0.61	−29.45	−7.15
	粤电力 A	000539	0.34	−1.15	54.84	0.48	2.42	0.52	11.78	−7.72
	韶能股份	000601	−2.95	−1.29	61.79	0.27	2.52	0.53	15.77	−4.67
	ST 惠天	000692	−1.86	−0.81	63.34	0.40	1.09	0.43	8.08	−1.82
	城投控股	600649	12.28	8.46	39.92	0.25	0.57	40.20	29.21	−2.19
	大连热电	600719	1.58	0.96	60.53	0.32	0.70	1.31	−3.44	0.75
	华电能源	600726	0.43	0.33	77.63	0.40	2.39	1.08	12.66	−6.04
	国电电力	600795	1.26	0.20	71.65	0.26	1.68	1.10	−5.88	5.68
房地产业	长春经开	600215	0.09	0.21	29.10	0.05	0.08	1.23	9.07	0.09
	大龙地产	600159	1.21	0.09	61.63	0.04	0.05	1.84	−57.90	−0.08
	金丰投资	600606	9.78	6.51	46.07	0.20	0.31	6.22	−51.99	−8.40
	新黄浦	600638	6.81	5.96	31.91	0.12	0.31	5.57	−18.48	4.99
	浦东金桥	600639	9.02	6.16	42.74	0.20	0.86	4.51	40.62	4.75
	外高桥	600648	6.90	2.09	78.11	0.70	2.47	7.04	19.88	5.21
	中华企业	600675	14.31	6.82	63.67	0.37	0.44	5.89	33.93	11.82
	渝开发 A	000514	6.53	5.14	31.61	0.14	0.40	4.42	−15.56	6.64
	莱茵置业	000558	21.22	7.95	73.67	0.44	0.52	1.04	−13.15	28.42
	粤宏远 A	000573	−8.47	−4.84	44.12	0.14	0.24	−3.90	−26.72	−7.81
	中国国贸	600007	8.40	6.21	48.06	0.12	3.04	1.10	1.20	5.06
	万科 A	000002	12.65	5.77	67.44	0.37	0.39	10.62	15.38	8.93
	三木集团	000632	1.96	1.05	80.12	0.88	0.95	1.74	−11.30	−9.55
	国兴地产	000838	2.97	2.21	44.34	0.17	0.17	30.65	−74.76	3.06
	中关村	000931	9.69	1.72	80.11	0.47	0.57	2.03	−7.90	1.59

续前表

行业	公司简称	股票代码	净资产收益率	总资产报酬率	资产负债率	总资产周转率	流动资产周转率	已获利息倍数	销售增长率	资本积累率
信息技术业	中兴通讯	000063	11.65	5.02	70.15	0.98	1.21	4.28	27.36	17.40
	长城电脑	000066	1.01	0.39	53.93	1.35	3.57	1.22	−6.99	−30.87
	南天信息	000948	9.48	6.61	45.43	1.06	1.41	4.62	15.13	110.72
	同方股份	600100	3.57	2.63	53.32	0.78	0.00	2.79	−4.77	26.72
	永鼎股份	600105	2.54	1.69	71.91	0.42	0.63	1.87	27.49	2.63
	宏图高科	600122	10.71	5.42	57.49	1.77	2.12	3.21	33.03	11.23
	新大陆	000997	4.54	3.74	31.88	0.86	1.09	7.49	18.42	−6.27
	方正科技	600601	4.42	3.16	43.95	1.40	4.67	3.06	−13.58	4.73
	复旦复华	600624	4.44	3.68	49.44	0.53	0.85	3.19	13.57	2.60

说明：1. 该表中数据均来自于合并会计报表。

2. 除净资产收益率、资产负债率指标直接取自会计年报，其他各指标均由各企业年报提供数据计算而得，各指标的计算公式如下：

a. 总资产报酬率$=\dfrac{\text{利润总额}+\text{财务费用}}{(\text{年初总资产}+\text{年末总资产})/2}\times 100\%$

b. 总资产周转率$=\dfrac{\text{主营业务收入}}{(\text{年初总资产}+\text{年末总资产})/2}$

c. 流动资产周转率$=\dfrac{\text{主营业务收入}}{(\text{年初流动资产}+\text{年末流动资产})/2}$

d. 已获利息倍数$=\dfrac{\text{利润总额}+\text{财务费用}}{\text{财务费用}}$

e. 销售增长率$=\dfrac{\text{本年主营业务收入}-\text{上年主营业务收入}}{\text{上年主营业务收入}}\times 100\%$

f. 资本积累率$=\dfrac{\text{年末股东权益}-\text{年初股东权益}}{\text{年初股东权益}}\times 100\%$

在 SPSS 软件的数据窗口依次定义变量，并输入以上数据。在表 2—1 的数据中，不同的行业可以看做不同的总体，因此，35 个数据分别来自 3 个总体。下面尝试对 3 个不同行业的上市公司的经营能力进行比较。

在进行比较分析之前，首先要对各数据是否遵从多元正态分布进行检验。然而，遗憾的是，多元正态性检验在常见的统计软件中并不容易实现。在实际工作中，往往借助于考察每一个变量的结果来对向量的分布做出判断，并且，当数据量较大且没有明显的证据表明所得数据不遵从多元正态分布时，通常认为数据来自多元正态总体。SPSS 软件提供了对单变量进行正态性检验的功能。

对上面的数据，依次点选 Analyze→Descriptive Statistics→Explore…进入 Explore 对话框，可以看到上市公司数据的所有变量名及变量标签均出现在左边的列表框中，将净资产收益率、总资产报酬率、资产负债率、总资产周转率、流动资产周转率、已获利息倍数、销售增长率及资本积累率 8 个变量选入 Dependent 框中，点击下方的 Plots…按钮进入 Plots 对话框，选中 Normality plots with tests 复选项以输出有关正态性检验的图表，点击 Continue 继续，点击 OK 运行，则可以得到如下结果（见输出结果 2—1，其他输出结果略）。

输出结果 2—1　　Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
净资产收益率	.113	35	.200*	.978	35	.677
总资产报酬率	.121	35	.200*	.964	35	.298
资产负债率	.086	35	.200*	.962	35	.265
总资产周转率	.180	35	.006	.864	35	.000
流动资产周转率	.164	35	.018	.885	35	.002
已获利息倍数	.281	35	.000	.551	35	.000
销售增长率	.103	35	.200*	.949	35	.104
资本积累率	.251	35	.000	.655	35	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction.

此表给出了对每一个变量进行正态性检验的结果，因为该例中样本数 $n=35<2\,000$，所以此处选用 Shapiro-Wilk 统计量。由 Sig. 值可以看到，总资产周转率、流动资产周转率、已获利息倍数及资本积累率均明显不遵从正态分布，因此，在下面的分析中，我们只对净资产收益率、总资产报酬率、资产负债率及销售增长率这四个指标进行比较，并认为这四个变量组成的向量遵从正态分布（尽管事实上也许并非如此）。这四个指标涉及公司的获利能力、资本结构及成长能力，我们认为这四个指标可以对公司运营能力作出近似的度量。

SPSS 的 GLM 模块可以完成多元正态分布有关均值与方差的检验。依次点选 Analyze→General Linear Model→Multivariate…进入 Multivariate 对话框，将净资产收益率、总资产报酬率、资产负债率及销售增长率这四个指标选入 Dependent 列表框，将行业指标选入 Fixed Factor(s) 框，点击 OK 运行，则可以得到如下结果（见输出结果 2—2）。

输出结果 2—2　　Between-Subjects Factors

		N
行业	电力、煤气及水的生产和供应业	11
	房地产业	15
	信息技术业	9

Multivariate Tests[c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.967	209.590[a]	4.000	29.000	.000
	Wilks' Lambda	.033	209.590[a]	4.000	29.000	.000
	Hotelling's Trace	28.909	209.590[a]	4.000	29.000	.000
	Roy's Largest Root	28.909	209.590[a]	4.000	29.000	.000
行业	Pillai's Trace	.481	2.374	8.000	60.000	.027
	Wilks' Lambda	.563	2.413[a]	8.000	58.000	.025
	Hotelling's Trace	.699	2.445	8.000	56.000	.024
	Roy's Largest Root	.560	4.197[b]	4.000	30.000	.008

a. Exact statistic.

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+行业.

上面第一张表是样本数据分别来自三个行业的个数。第二张表是多变量检验表，该表给出了几个统计量，由 Sig. 值可以看到，无论从哪个统计量来看，三个行业的运营能力（从净资产收益率、总资产报酬率、资产负债率及销售增长率这四个指标的整体来看）都是有显著差别的。实际上，GLM 模型是拟合了下面的模型：

$$\boldsymbol{Y}=\beta_0+\beta_1\boldsymbol{X}+\varepsilon$$

式中　$\boldsymbol{Y}=$(净资产收益率，总资产报酬率，资产负债率，销售增长率)′

$\boldsymbol{X}=$行业

上面的多变量检验表实际上是对该线性模型显著性的检验，此处有常数项 β_0 是因为不能肯定模型过原点。而模型通过了显著性检验，意味着行业的不同取值对 $\boldsymbol{Y}$ 的取值有显著影响，也就是说，不同行业的运营能力是不同的。

输出结果 2—3 给出了每个财务指标的分析结果，同时给出了每个财务指标的方差来源，包括校正模型、截距、主效应（行业）、误差及总的方差来源，还给出了自由度、均方、F 统计量及 Sig. 值。

输出结果 2—3　　Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	净资产收益率	306.300[a]	2	153.150	4.000	.028
	总资产报酬率	69.471[b]	2	34.736	3.319	.049
	资产负债率	302.366[c]	2	151.183	.680	.514
	销售增长率	2 904.588[d]	2	1 452.294	2.154	.133
Intercept	净资产收益率	615.338	1	615.338	16.073	0.000
	总资产报酬率	217.975	1	217.975	20.830	0.000
	资产负债率	105 315.459	1	105 315.459	473.833	0.000
	销售增长率	1.497	1	1.497	.002	.963
行业	净资产收益率	306.300	2	153.150	4.000	.028
	总资产报酬率	69.471	2	34.736	3.319	.049
	资产负债率	302.366	2	151.183	.680	.514
	销售增长率	2 904.588	2	1 452.294	2.154	.133
Error	净资产收益率	1 225.054	32	38.283		
	总资产报酬率	334.866	32	10.465		
	资产负债率	7 112.406	32	222.263		
	销售增长率	21 579.511	32	674.360		
Total	净资产收益率	2 238.216	35			
	总资产报酬率	641.682	35			
	资产负债率	117 585.075	35			
	销售增长率	24 585.045	35			
Corrected Total	净资产收益率	1 531.354	34			
	总资产报酬率	404.338	34			
	资产负债率	7 414.772	34			
	销售增长率	24 484.099	34			

a. R Squared = .200 (Adjusted R Squared = .150).

b. R Squared = .172 (Adjusted R Squared = .120).

c. R Squared = .041 (Adjusted R Squared = −.019).

d. R Squared = .119 (Adjusted R Squared = .064).

其中，第三列给出了用 Type Ⅲ方法计算的偏差平方和。SPSS 软件给出了四种计算偏差平方和的方法，可以根据方差分析中是否存在交互效应及设计是否平衡等不同情况选用不同的计算方法。此处只有一个因素即行业，使用默认方法即可。由该结果可以看到，四个指标的 Sig. 值分别为 0.028，0.049，0.514 及 0.133，说明三个行业在净资产收益率和总资产报酬率两个财务指标上存在显著差别，但在资产负债率和销售增长率两个财务指标上没有显著差别。

由 GLM 默认选项的输出结果可以得知，三个行业的运营能力存在明显的差别，并且分别考察四个指标在三个行业之间是否存在明显差别，发现净资产收益率、总资产报酬率这两个指标在三个行业也均有明显的差别。

在实际工作中，我们往往更希望知道差别主要来自哪些行业，或者进行不同行业运营能力的比较。对此，对 GLM 模块的选项作如下设置：在 GLM 主对话框中点击 Contrasts…按钮进入 Contrasts 对话框，在 Change Contrast 框架中，打开 Contrast 右侧的下拉框并选择 Simple，此时下侧的 Reference Category 被激活，默认是 Last 被选中，表明第一、二行业均与第三行业做比较，若选中 First，则将做第二、三行业与第一行业的比较。点击 Change 按钮，Continue 继续，OK 运行，则除上面的结果外，还可得到如下结果（见输出结果 2—4）。

输出结果 2—4　　Contrast Results (K Matrix)

行业 Simple Contrast[a]		Dependent Variable 净资产收益率	总资产报酬率	资产负债率	销售增长率
Level 1 vs. Level 3	Contrast Estimate	-5.649	-3.070	7.259	-13.223
	Hypothesized Value	0	0	0	0
	Difference (Estimate-Hypothesized)	-5.649	-3.070	7.259	-13.223
	Std. Error	2.781	1.454	6.701	11.672
	Sig.	.051	.043	.287	.266
	95% Confidence Interval for Difference Lower Bound	-11.313	-6.031	-6.390	-36.998
	Upper Bound	.016	-.108	20.908	10.552
Level 2 vs. Level 3	Contrast Estimate	1.054	-.056	1.791	-22.696
	Hypothesized Value	0	0	0	0
	Difference (Estimate-Hypothesized)	1.054	-.056	1.791	-22.696
	Std. Error	2.609	1.364	6.286	10.949
	Sig.	.689	.967	.778	.046
	95% Confidence Interval for Difference Lower Bound	-4.260	-2.834	-11.013	-44.999
	Upper Bound	6.368	2.722	14.595	-.394

a. Reference category = 3.

输出结果 2—4 表示，在 0.05 的显著性水平下，第一行业（电力、煤气及水的生产和供应业）与第三行业（信息技术业）的总资产报酬率指标存在显著差别，净资产收益率、资产负债率和销售增长率等财务指标无明显差别，但由第一栏可以看

到，电力、煤气及水的生产和供应业的净资产收益率、总资产报酬率和销售增长率均低于信息技术业，资产负债率高于信息技术业，似乎说明信息技术业作为新兴行业，其成长能力要更高一些。第二行业（房地产业）与第三行业的销售增长率指标有明显的差别，第三行业大于第二行业，说明信息技术业的获利能力高于房地产业。净资产收益率、总资产报酬率和资产负债率等财务指标没有显著差别。

输出结果 2—5 是上面多重比较可信性的度量，由 Sig. 值可以看到，比较检验是可信的。

输出结果 2—5　　**Multivariate Test Results**

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	.481	2.374	8.000	60.000	.027
Wilks' lambda	.563	2.413[a]	8.000	58.000	.025
Hotelling's trace	.699	2.445	8.000	56.000	.024
Roy's largest root	.560	4.197[b]	4.000	30.000	.008

a. Exact statistic.

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

输出结果 2—6 是对每一个指标在三个行业比较的结果，与输出结果 2—3 中的有关结果一致。

输出结果 2—6　　**Univariate Test Results**

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.
Contrast	净资产收益率	306.300	2	153.150	4.000	.028
	总资产报酬率	69.471	2	34.736	3.319	.049
	资产负债率	302.366	2	151.183	.680	.514
	销售增长率	2 904.588	2	1 452.294	2.154	.133
Error	净资产收益率	1 225.054	32	38.283		
	总资产报酬率	334.866	32	10.465		
	资产负债率	7 112.406	32	222.263		
	销售增长率	21 579.511	32	674.360		

在 Multivariate 主对话框中点击 Options…按钮，进入 Options 对话框，在上面 Estimated Marginal Means 框架中，把行业（chany）选入右边 Display Means for：列表框中以输出各行业各财务指标的均值，选中下方的 Compare main effects 复选框，则输出不同行业各财务指标比较的结果，在下方的 Display 框架中，提供了很多可选的统计量或中间结果，选中 Homogeneity tests 复选项进行各行业（总体）数据协方差阵相等的检验。Continue 继续，OK 运行，则还可以得到如下结果（见输出结果 2—7）。

输出结果 2—7　　**Box's Test of Equality of Covariance Matrices[a]**

Box's M	29.216
F	1.172
df1	20
df2	2 585.573
Sig.	.269

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design：Intercept＋行业.

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
净资产收益率	.500	2	32	.611
总资产报酬率	1.757	2	32	.189
资产负债率	4.537	2	32	.018
销售增长率	1.739	2	32	.192

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Design: Intercept + 行业.

上面第一张表是协方差阵相等的检验，检验统计量是 Box's M，由 Sig. 值可以认为三个行业（总体）的协方差阵是相等的。第二张表给出了各行业误差平方相等的检验，在 0.05 的显著性水平下，净资产收益率、总资产报酬率以及销售增长率的误差平方在三个行业间没有显著差别，而资产负债率的误差平方在三个行业间有显著差别。这似乎说明，除了行业因素，对资产负债率有显著影响的还有其他因素。这与此处均值比较没有太大的关系。

输出结果 2—8 给出了每一行业各财务指标描述统计量的估计，不再具体说明。

输出结果 2—8 **Estimates**

Dependent Variable	行业	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
净资产收益率	电力、煤气及水的生产和供应业	.169	1.866	-3.631	3.969
	房地产业	6.871	1.598	3.617	10.125
	信息技术业	5.818	2.062	1.617	10.019
总资产报酬率	电力、煤气及水的生产和供应业	.523	.975	-1.463	2.510
	房地产业	3.537	.835	1.835	5.238
	信息技术业	3.593	1.078	1.396	5.789
资产负债率	电力、煤气及水的生产和供应业	60.315	4.495	51.158	69.471
	房地产业	54.847	3.849	47.006	62.688
	信息技术业	53.056	4.969	42.933	63.178
销售增长率	电力、煤气及水的生产和供应业	-1.038	7.830	-16.987	14.911
	房地产业	-10.512	6.705	-24.170	3.146
	信息技术业	12.184	8.656	-5.448	29.816

输出结果 2—9 给出了不同行业各财务指标的比较与检验及检验的可信性统计量，其中，第一张表的结果与输出结果 2—4 相同，只不过比输出结果 2—4 更为具体，表中各项也很容易理解，不再说明；第二张表与输出结果 2—5 具有相同的作用，且结果完全相同。

输出结果 2—9

Pairwise Comparisons

Dependent Variable	(I)行业	(J)行业	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
净资产收益率	电力、煤气及水的生产和供应业	房地产业	-6.702*	2.456	.010	-11.705	-1.699
		信息技术业	-5.649	2.781	.051	-11.313	.016
	房地产业	电力、煤气及水的生产和供应业	6.702*	2.456	.010	1.699	11.705
		信息技术业	1.054	2.609	.689	-4.260	6.368
	信息技术业	电力、煤气及水的生产和供应业	-5.649	2.781	.051	-.016	11.313
		房地产业	-1.054	2.609	.689	-6.368	4.260
总资产报酬率	电力、煤气及水的生产和供应业	房地产业	-3.013*	1.284	.025	-5.629	-.398
		信息技术业	-3.070*	1.454	.043	-6.031	-.108
	房地产业	电力、煤气及水的生产和供应业	3.013*	1.284	.025	.398	5.629
		信息技术业	-.056	1.364	.967	-2.834	2.722
	信息技术业	电力、煤气及水的生产和供应业	3.070*	1.454	.043	.108	6.031
		房地产业	.056	1.364	.967	-2.722	2.834

(1)

续输出结果 2—9

资产负债率	电力、煤气及水的生产和供应业	房地产业	5.468	5.918	.362	−6.587	17.523
		信息技术业	7.259	6.701	.287	−6.390	20.908
	房地产业	电力、煤气及水的生产和供应业	−5.468	5.918	.362	−17.523	6.587
		信息技术业	1.791	6.286	.778	−11.013	14.595
	信息技术业	电力、煤气及水的生产和供应业	−7.259	6.701	.287	−20.908	6.390
		房地产业	1.791	6.286	.778	−14.595	11.013
销售增长率	电力、煤气及水的生产和供应业	房地产业	9.474	10.308	.365	−11.524	30.471
		信息技术业	−13.223	11.672	.266	−36.998	10.552
	房地产业	电力、煤气及水的生产和供应业	−9.474	10.308	.365	−30.471	11.524
		信息技术业	−22.696*	10.949	.046	−44.999	−.394
	信息技术业	电力、煤气及水的生产和供应业	13.223	11.672	.266	−10.552	36.998
		房地产业	22.696*	10.949	.046	.394	41.999

(1)

Based on estimated marginal means.

*. The mean difference is significant at the 0.05 level.

a. Adjustment for multiple comparisons: Least Significant Difference(equivalent to no adjustments).

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	.481	2.374	8.000	60.000	.027
Wilk's lambda	.563	2.413[a]	8.000	58.000	.025
Hotelling's trace	.699	2.445	8.000	56.000	.024
Roy's largest root	.560	4.197[b]	4.000	30.000	.008

(2)

Each F tests the multivariate effect of 行业. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic.

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

输出结果 2—10 与输出结果 2—3 中的有关检验部分及输出结果 2—6 是相同的，也是对三个行业中各财务指标相等的假设的检验，可以看到在 0.05 的显著性水平下，净资产收益率和总资产报酬率在三个行业中有明显的差别。

输出结果 2—10　　**Univariate Tests**

Dependent Variable		Sum of Squares	df	Mean Square	F	Sig.
净资产收益率	Contrast	306.300	2	153.150	4.000	.028
	Error	1 225.054	32	38.283		
总资产报酬率	Contrast	69.471	2	34.736	3.319	.049
	Error	334.866	32	10.465		
资产负债率	Contrast	302.366	2	151.183	.680	.514
	Error	7 112.406	32	222.263		
销售增长率	Contrast	2 904.588	2	1 452.294	2.154	.133
	Error	21 579.511	32	674.360		

The F tests the effect of 行业. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

在主对话框中，点击 Post Hoc…按钮，进入 Post Hoc Multiple Comparisons for Observed Means 对话框，还可以输出其他很多有用的结果，此处不再一一说明。

综上所述，我们对三个行业的运营能力进行了具体的比较分析，所得数据表明，从总体来看，信息技术业要稍好于电力、煤气及水的生产和供应业以及房地产业。原因可能是房地产业在前几年的快速发展后，由于进入企业过多，盲目上马项目过多，造成了不良局面，以致整个行业不景气，运营能力有所下降。而信息技术业作为新兴行业，发展较快，利润空间较大，从获利能力和成长能力来看发展良好，整体运营能力较强。对于每一财务指标的分析上面已有说明，此处不再赘述。

读者在分析具体问题时，一定要结合当时的实际背景。只有定性分析与定量分析相结合，才能得出更加令人信服的结果。

□ 参考文献

[1] 方开泰，张尧庭. 多元统计分析引论. 北京：科学出版社，1982

[2] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993

[3] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[4] 袁志发，宋世德. 多元统计分析. 北京：科学出版社，2009

[5] M. S. Srivastava，E. M. Carter. 应用之多变量统计学. 北京：世界图书出版公司，1989

[6] James H. Bray，Scott E. Maxwell. *Multivariate Analysis of Variance*. Beverly Hills：Sage Publications，1985

□ 思考与练习

1. 试举出两个可以运用多元均值检验的实际问题。

2. 试谈 Wilks 统计量在多元方差分析中的重要意义。

3. 现选取内蒙古、广西、贵州、云南、西藏、宁夏、新疆、甘肃和青海等 9 个内陆边远省区。选取人均 GDP、第三产业比重、人均消费支出、人口自然增长率及文盲半文盲人口占 15 岁以上人口的比例等 5 项能够较好地说明各地区社会经济发展水平的指标，验证边远及少数民族聚居区的社会经济发展水平与全国平均水平有无显著差异。

边远及少数民族聚居区社会经济发展水平的指标数据

地区	人均 GDP（元）	三产比重（%）	人均消费（元）	人口增长（%）	文盲半文盲（%）
内蒙古	5 068	31.1	2 141	8.23	15.83
广西	4 076	34.2	2 040	9.01	13.32
贵州	2 342	29.8	1 551	14.26	28.98
云南	4 355	31.1	2 059	12.1	25.48
西藏	3 716	43.5	1 551	15.9	57.97
宁夏	4 270	37.3	1 947	13.08	25.56
新疆	6 229	35.4	2 745	12.81	11.44
甘肃	3 456	32.8	1 612	10.04	28.65
青海	4 367	40.9	2 047	14.48	42.92

资料来源：中华人民共和国国家统计局：《中国统计年鉴（1998）》，北京，中国统计出版社，1998。

5 项指标的全国平均水平为：

$$\boldsymbol{\mu}_0 = (6\,212.01,\ 32.87,\ 2\,972,\ 9.5,\ 15.78)'$$

4. 试针对某一实际问题具体运用多元方差分析方法。

第3章 Chapter 3 聚类分析

学习目标

1. 了解适合用聚类分析解决的问题；
2. 理解对象之间的相似性是如何测量的；
3. 区别不同的距离；
4. 区分不同的聚类方法及其相应的应用；
5. 理解如何选择类的个数；
6. 简述聚类分析的局限。

人们往往会碰到通过划分同种属性的对象很好地解决问题的情形，而不论这些对象是个体、公司、产品甚至行为。如果没有一种客观的方法，基于在总体内区分群体的战略选择，比如市场细分将不可能。其他领域如从自然科学领域（比如为多种动物群体——昆虫、哺乳动物和爬行动物的区分建立生物分类学）到社会科学领域（比如分析不同精神病的特征），也会遇到类似的问题。所有情况下，研究者都在基于一个多维剖面的观测中寻找某种“自然”结构。

为此最常用的技巧是聚类分析。聚类分析将个体或对象分类，使得同一类中的对象之间的相似性比与其他类的对象的相似性更强。其目的在于使类内对象的同质性最大化和类与类间对象的异质性最大化。本章将介绍聚类分析的性质和目的，并且引导研究者使用各种聚类分析方法。

3.1 聚类分析的基本思想

3.1.1 导言

在古老的分类学中，人们主要靠经验和专业知识，很少利用统计学方法。随着

生产技术和科学的发展，分类越来越细，以致有时仅凭经验和专业知识不能进行确切分类，于是统计这个有用的工具逐渐引入分类学，形成了数值分类学。近年来，数理统计的多元分析方法有了迅速的发展，多元分析的技术自然被引入分类学中，于是从数值分类学中逐渐分离出聚类分析这个新的分支。

我们认为，所研究的样品或指标（变量）之间存在程度不同的相似性（亲疏关系）。于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间相似程度的统计量，以这些统计量作为划分类型的依据，把一些相似程度较大的样品（或指标）聚合为一类，把另外一些彼此之间相似程度较大的样品（或指标）聚合为另外一类……关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到把所有的样品（或指标）都聚合完毕，把不同的类型一一划分出来，形成一个由小到大的分类系统。最后再把整个分类系统画成一张分群图（又称谱系图），用它把所有的样品（或指标）间的亲疏关系表示出来。

在社会、经济、人口研究中，存在着大量分类研究、构造分类模式的问题。例如在经济研究中，为了研究不同地区城镇居民生活中的收入及消费状况，往往需要划分为不同的类型去研究；在人口研究中，需要构造人口生育分类模式、人口死亡分类函数，以此来研究人口的生育和死亡规律。过去，人们主要靠经验和专业知识做定性分类处理，导致许多分类带有主观性和任意性，不能很好地揭示客观事物内在的本质差别和联系，特别是对于多因素、多指标的分类问题，定性分类更难以实现准确分类。

聚类分析不仅可以用来对样品进行分类，也可以用来对变量进行分类。对样品的分类常称为 Q 型聚类分析，对变量的分类常称为 R 型聚类分析。与多元分析的其他方法相比，聚类分析的方法还是比较粗糙的，理论上也不算完善，但由于它能解决许多实际问题，所以很受实际研究者重视，同回归分析、判别分析一起称为多元分析的三大方法。

3.1.2 聚类的目的

在一些社会、经济问题中，我们面临的往往是比较复杂的研究对象，如果能把相似的样品（或指标）归成类，处理起来就大为方便，所以如前所述，聚类分析的目的就是把相似的研究对象归成类。首先来看一个简单的例子。

例 3—1

若我们需要将下列 11 户城镇居民按户主个人的收入进行分类，对每户做了如下的统计，结果如表 3—1 所示。在表中，“标准工资收入”、“职工奖金”、“职工津贴”、“性别”、“就业身份”等称为指标，每户称为样品。若对户主进行分类，还可以采用其他指标，如“子女个数”、“政治面貌”等。指标如何选择取决于聚类的目的。

表 3—1　　某市 2001 年城镇居民户主个人收入数据

X1　职工标准工资收入（元）				X5　单位得到的其他收入（元）			
X2　职工奖金收入（元）				X6　其他收入（元）			
X3　职工津贴收入（元）				X7　性别			
X4　其他工资性收入（元）				X8　就业身份			
X1	X2	X3	X4	X5	X6	X7	X8
540.00	0.0	0.0	0.0	0.0	6.00	男	国有
1 137.00	125.00	96.00	0.0	109.00	812.00	女	集体
1 236.00	300.00	270.00	0.0	102.00	318.00	女	国有
1 008.00	0.0	96.00	0.0	86.0	246.00	男	集体
1 723.00	419.00	400.00	0.0	122.00	312.00	男	国有
1 080.00	569.00	147.00	156.00	210.00	318.00	男	集体
1 326.00	0.0	300.00	0.0	148.00	312.00	女	国有
1 110.00	110.00	96.00	0.0	80.00	193.00	女	集体
1 012.00	88.00	298.00	0.0	79.00	278.00	女	国有
1 209.00	102.00	179.00	67.00	198.00	514.00	男	集体
1 101.00	215.00	201.00	39.00	146.00	477.00	男	集体

例 3—1 中的 8 个指标，前 6 个是定量的，后 2 个是定性的。如果分得更细一些，指标的类型有三种尺度。

（1）间隔尺度。变量用连续的量来表示，如各种奖金、各种津贴等。

（2）有序尺度。指标用有序的等级来表示，如文化程度分为文盲、小学、中学、中学以上等，有次序关系，但没有数量表示。

（3）名义尺度。指标用一些类来表示，这些类之间既没有等级关系，也没有数量关系，如例 3—1 中的性别和职业都是名义尺度指标。

不同类型的指标，在聚类分析中，处理的方式是大不一样的。总的来说，处理间隔尺度指标的方法较多，对另两种尺度的变量的处理方法不多。

聚类分析根据实际的需要可能有两个方向，一是对样品（如例 3—1 中的户主）聚类；一是对指标聚类。第一位重要的问题是“什么是类”。简单地讲，相似样品（或指标）的集合称为类。由于经济问题的复杂性，欲给类下一个严格的定义是困难的，在 3.3 节中，我们将给出一些待探讨的定义。

将例 3—1 抽象化，就得到表 3—2 所示的数据阵，其中 x_{ij} 表示第 i 个样品的第 j 个指标的值。我们的目的是从这些数据出发，将样品（或指标）进行分类。

表 3—2　　数据矩阵

No.	x_1	x_2	…	x_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	…	x_{np}

聚类分析给人们提供了丰富多彩的分类方法，这些方法大致可归纳为：

（1）系统聚类法。首先，将 n 个样品看成 n 类（一类包含一个样品），然后将性质最接近的两类合并成一个新类，得到 $n-1$ 类，再从中找出最接近的两类加以合

并，变成 $n-2$ 类，如此下去，最后所有的样品均在一类，将上述并类过程画成一张图（称为聚类图）便可决定分多少类，每类各有哪些样品。

（2）模糊聚类法。将模糊数学的思想观点用到聚类分析中产生的方法。该方法多用于定性变量的分类。

（3）K-均值法。K-均值法是一种非谱系聚类法，它是把样品聚集成 k 个类的集合。类的个数 k 可以预先给定或者在聚类过程中确定。该方法可应用于比系统聚类法适用的大得多的数据组。

（4）有序样品的聚类。n 个样品按某种原因（时间、地层深度等）排成次序，必须是次序相邻的样品才能聚成一类。

（5）分解法。它的程序正好和系统聚类法相反，首先所有的样品均在一类，然后用某种最优准则将它分为两类。再试图用同种准则将这两类各自分裂为两类，从中选一个使目标函数较好者，这样由两类变成三类。如此下去，一直分裂到每类只有一个样品为止（或用其他停止规则），将上述分裂过程画成图，由图便可求得各个类。

（6）加入法。将样品依次加入，每次加入后将它放到当前聚类图的应在位置上，全部加入后，即可得到聚类图。

本书将重点介绍系统聚类法和 K-均值法。对于其他方法有兴趣的读者可参阅参考文献［1］～［5］。

3.2 相似性度量

从一组复杂数据产生一个相当简单的类结构，必然要求进行相关性或相似性度量。在相似性度量的选择中，常常包含许多主观上的考虑，但最重要的考虑是指标性质（包括离散的、连续的）或观测的尺度（名义的、有序的、间隔的）以及有关的知识。

当对样品进行聚类时，“靠近”往往用某种距离来刻画。另一方面，当对指标聚类时，根据相关系数或某种关联性度量来聚类。

在表 3—2 中，每个样品有 p 个指标，故每个样品可以看成 p 维空间中的一个点，n 个样品就组成 p 维空间中的 n 个点，此时自然想用距离来度量样品之间的接近程度。

用 x_{ij} 表示第 i 个样品的第 j 个指标，数据矩阵如表 3—2 所示，第 j 个指标的均值和标准差记作 $\bar{x}_j$ 和 S_j。用 d_{ij} 表示第 i 个样品与第 j 个样品之间的距离，作为距离当然满足 1.2 节中的四条公理。

最常见、最直观的距离是：

$$d_{ij}(1)=\sum_{k=1}^{p}|x_{ik}-x_{jk}| \tag{3.1}$$

$$d_{ij}(2)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{1/2} \tag{3.2}$$

前者称为绝对值距离，后者称为欧氏距离，这两个距离统一成

$$d_{ij}(q)=\left[\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{q}\right]^{1/q} \tag{3.3}$$

它称为明考斯基（Minkowski）距离。当 $q=1$ 和 2 时就是上述的两个距离，当 q 趋于无穷时

$$d_{ij}(\infty)=\max_{1\leqslant k\leqslant p}|x_{ik}-x_{jk}| \tag{3.4}$$

称为切比雪夫距离。

可以验证，$d_{ij}(q)$ 满足距离的四条公理。

$d_{ij}(q)$ 在实际中应用广泛，但是有一些缺点，例如距离的大小与各指标的观测单位有关，具有一定的人为性；另一方面，它没有考虑指标之间的相关性。通常的改进办法有下面两种：

（1）当各指标的测量值相差较大时，先将数据标准化，然后用标准化后的数据计算距离。

令 $\overline{X}_j$，R_j 和 S_j 分别表示第 j 个指标的样本均值、样本极差和样本标准差，即

$$\overline{X}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$$

$$R_j=\max_{1\leqslant i\leqslant n}\{x_{ij}\}-\min_{1\leqslant i\leqslant n}\{x_{ij}\}$$

$$S_j=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{X}_j)^2\right]^{1/2}$$

则标准化后的数据为：

$$x'_{ij}=\frac{x_{ij}-\overline{X}_j}{S_j}$$

或　　$x^{*}_{ij}=\dfrac{x_{ij}-\overline{X}_j}{S_j}$，　　$i=1,2,\cdots,n;\ j=1,2,\cdots,p$

当 $x_{ij}>0(i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)$ 时，有人采用

$$d_{ij}(LW)=\frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik}-x_{jk}|}{x_{ik}+x_{jk}} \tag{3.5}$$

它最早是由兰斯（Lance）和威廉姆斯（Williams）提出的，称为兰氏距离。这个距离有助于克服 $d_{ij}(q)$ 的第一个缺点，但没有考虑指标间的相关性。

（2）一种改进的距离就是前面讨论过的马氏距离。

$$d^{2}_{ij}(M)=(\boldsymbol{x}_{(i)}-\boldsymbol{x}_{(j)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{(i)}-\boldsymbol{x}_{(j)}) \tag{3.6}$$

式中，$\boldsymbol{x}_{(i)}$ 表示矩阵行向量的转置；$\boldsymbol{\Sigma}$ 是数据矩阵的协方差阵。可以证明，它对一切线性变换是不变的，故不受指标量纲的影响。它对指标的相关性也做了考虑，我们仅用一个例子来说明。

例 3—2

已知一个二维正态总体 G 的分布为：

$$N_2\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)$$

求点 $\boldsymbol{A}=\begin{bmatrix}1\\1\end{bmatrix}$ 和点 $\boldsymbol{B}=\begin{bmatrix}1\\-1\end{bmatrix}$ 至均值 $\boldsymbol{\mu}=\begin{bmatrix}0\\0\end{bmatrix}$ 的距离。

由假设可算得

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{0.19}\begin{bmatrix}1 & -0.9\\-0.9 & 1\end{bmatrix}$$

从而

$$d^2_{A\mu}(M)=(1,1)\boldsymbol{\Sigma}^{-1}\begin{bmatrix}1\\1\end{bmatrix}=0.2/0.19$$

$$d^2_{B\mu}(M)=(1,-1)\boldsymbol{\Sigma}^{-1}\begin{bmatrix}1\\-1\end{bmatrix}=3.8/0.19$$

如果用欧氏距离，则有

$$d^2_{A\mu}(2)=2,\quad d^2_{B\mu}(2)=2$$

两者相等，而按马氏距离两者差 18 倍之多。由第 1 章讨论我们知道，本例的分布密度是

$$f(y_1,y_2)=\frac{1}{2\pi\sqrt{0.19}}\exp\left\{-\frac{1}{0.38}[y_1^2-1.8y_1y_2+y_2^2]\right\}$$

$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 两点的密度分别是

$$f(1,1)=0.215\ 7\text{ 和 }f(1,-1)=0.000\ 016\ 58$$

说明前者应当离均值近，后者离均值远，马氏距离正确地反映了这一情况，而欧氏距离则不然。这个例子告诉我们，正确地选择距离是非常重要的。

但是，在聚类分析之前，我们事先对研究对象有多少个不同类型的情况一无所知，马氏距离公式中的 $\boldsymbol{\Sigma}$ 如何计算呢？如果用全部数据计算的均值和协方差阵来计算马氏距离，效果也不是很理想。因此，通常人们还是喜欢应用欧氏距离聚类。

以上几种距离均适用于间隔尺度变量，如果指标是有序尺度或名义尺度的，也有一些定义距离的方法。下面通过一个实例来说明定义距离的较灵活的思想方法。

例3—3

欧洲各国的语言有许多相似之处。为了研究这些语言的历史关系，也许通过比较它们数字的表达比较恰当。表3—3列举了英语、挪威语、丹麦语、荷兰语、德语、法语、西班牙语、意大利语、波兰语、匈牙利语和芬兰语的1，2，…，10的拼法，希望计算这11种语言之间的距离。

表3—3　　11种欧洲语言的数词

英语 (English)	挪威语 (Norwegian)	丹麦语 (Danish)	荷兰语 (Dutch)	德语 (German)	法语 (French)
one	en	en	een	ein	un
two	to	to	twee	zwei	deux
three	tre	tre	drie	drei	trois
four	fire	fire	vier	vier	quatre
five	fem	fem	vijf	funf	einq
six	seks	seks	zes	sechs	six
seven	sju	syv	zeven	siebcn	sept
eight	ate	otte	acht	acht	huit
nine	ni	ni	negen	neun	neuf
ten	ti	ti	tien	zehn	dix

西班牙语 (Spanish)	意大利语 (Italian)	波兰语 (Polish)	匈牙利语 (Hungarian)	芬兰语 (Finnish)
uno	uno	jeden	egy	yksi
dos	due	dwa	ketto	kaksi
tres	tre	trzy	harom	kolme
cuatro	quattro	cztery	negy	neua
cinco	cinque	piec	ot	viisi
seix	sei	szesc	hat	kuusi
siete	sette	siedem	het	seitseman
ocho	otto	osiem	nyolc	kahdeksau
nueve	nove	dziewiec	kilenc	yhdeksan
diez	dieci	dziesiec	tiz	kymmenen

显然，此例无法直接用上述公式来计算距离。仔细观察表3—3，发现前三种语言（英、挪、丹）很相似，尤其每个单词的第一个字母，于是产生一种定义距离的办法：用两种语言的10个数词中的第一个字母不相同的个数来定义两种语言之间的距离，例如英语和挪威语中只有1和8的第一个字母不同，故它们之间的距离为

2。11 种语言之间两两的距离列于表 3—4 中。

表 3—4　　11 种欧洲语之间的距离

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0

当 p 个指标都是名义尺度时，例如 $p=5$，有两个样品的取值为：

$$\boldsymbol{X}_1 = (V,Q,S,T,K)'$$
$$\boldsymbol{X}_2 = (V,M,S,F,K)'$$

这两个样品的第一个指标都取 V，称为配合的；第二个指标一个取 Q，另一个取 M，称为不配合的。记配合的指标数为 m_1，不配合的指标数为 m_2，定义它们之间的距离为：

$$d_{12} = \frac{m_2}{m_1 + m_2} \tag{3.7}$$

在聚类分析中不仅需要将样品分类，也需要将指标分类。在指标之间也可以定义距离，更常用的是相似系数，用 C_{ij} 表示指标 i 和指标 j 之间的相似系数。C_{ij} 的绝对值越接近于 1，表示指标 i 和指标 j 的关系越密切；C_{ij} 的绝对值越接近于 0，表示指标 i 和指标 j 的关系越疏远。对于间隔尺度，常用的相似系数有夹角余弦和相关系数。

(1) 夹角余弦。这是受相似形的启发而来。图 3—1 中的曲线 AB 和 CD 尽管长度不一，但形状相似。当长度不是主要矛盾时，应定义一种相似系数使 AB 和 CD 呈现出比较密切的关系，而夹角余弦适合这一要求。它的定义是：

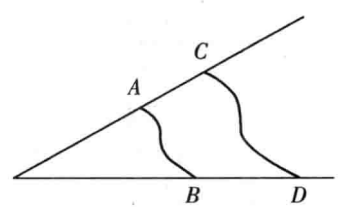


图 3—1

$$C_{ij}(1)=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[\left(\sum_{k=1}^{n}x_{ki}^2\right)\left(\sum_{k=1}^{n}x_{kj}^2\right)\right]^{1/2}} \tag{3.8}$$

它是指标向量（x_{1i}，x_{2i}，…，x_{ni}）和（x_{1j}，x_{2j}，…，x_{nj}）之间的夹角余弦。

（2）相关系数。这是大家最熟悉的统计量，它是将数据标准化后的夹角余弦。相关系数常用 r_{ij} 表示，为了和其他相似系数记号统一，这里记为$C_{ij}(2)$。它的定义是：

$$C_{ij}(2)=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)(x_{kj}-\overline{X}_j)}{\left[\sum_{k=1}^{n}(x_{ki}-\overline{X}_i)^2\sum_{k=1}^{n}(x_{kj}-\overline{X}_j)^2\right]^{1/2}} \tag{3.9}$$

名义尺度指标之间也可以定义相似系数，本书不做介绍，详见参考文献［6］、［7］。

有时指标之间也可用距离来描述它们的接近程度。实际上，距离和相似系数之间可以互相转化。若 d_{ij} 是一个距离，则 $C_{ij}=1/(1+d_{ij})$ 为相似系数。若 C_{ij} 为相似系数且非负，则 $d_{ij}=1-C_{ij}^2$ 可以看成距离（不一定符合距离的定义），或把 $d_{ij}=[2(1-C_{ij})]^{1/2}$ 看成距离。

如果指标均为取两值的名义尺度指标，也可定义相关系数，参见参考文献［8］。

3.3 类和类的特征

我们的目的是聚类，那么什么叫做类呢？由于客观事物千差万别，在不同的问题中类的含义是不尽相同的。因此企图给类下一个严格的定义，绝非一件容易的事情。下面给出类的几个定义，不同的定义适用于不同的场合。

用 G 表示类，设 G 中有 k 个元素，这些元素用 i，j 等表示。

定义 3.1 T 为一给定的阈值，如果对任意的 i，$j\in G$，有 $d_{ij}\leqslant T$（d_{ij} 为 i 和 j 的距离），则称 G 为一个类。

定义 3.2 对阈值 T，如果对每个 $i\in G$，有

$$\frac{1}{k-1}\sum_{j\in G}d_{ij}\leqslant T \tag{3.10}$$

则称 G 为一个类。

定义 3.3 对阈值 T，V，如果

$$\frac{1}{k(k-1)}\sum_{i\in G}\sum_{j\in G}d_{ij}\leqslant T \tag{3.11}$$

$d_{ij}\leqslant V$，对一切 i，$j\in G$，则称 G 为一个类。

定义 3.4 对阈值 T，若对任意一个 $i\in G$，一定存在 $j\in G$，使得 $d_{ij}\leqslant T$，则称 G 为一个类。

易见，定义 3.1 的要求是最高的，凡符合它的类，一定也是符合后三种定义的类。此外，凡符合定义 3.2 的类，也一定是符合定义 3.3 的类。

现在类 G 的元素用 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，…，$\boldsymbol{x}_m$ 表示，m 为 G 内的样品数（或指标数），可以从不同的角度来刻画 G 的特征。常用的特征有下面三种：

（1）均值 $\bar{\boldsymbol{x}}_G$（或称为 G 的重心）：

$$\bar{\boldsymbol{x}}_G=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$$

（2）样本离差阵及协方差阵：

$$\boldsymbol{L}_G=\sum_{i=1}^{m}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_G)(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_G)'$$

$$\boldsymbol{\Sigma}_G=\frac{1}{n-1}\boldsymbol{L}_G$$

（3）G 的直径。它有多种定义，例如

(a) $D_G=\sum_{i=1}^{m}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_G)'(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_G)=\mathrm{tr}(\boldsymbol{L}_G)$

(b) $D_G=\max\limits_{i,j\in G}d_{ij}$

在聚类分析中，不仅要考虑各个类的特征，而且要计算类与类之间的距离。由于类的形状是多种多样的，所以类与类之间的距离也有多种计算方法。令 G_p 和 G_q 中分别有 k 个和 m 个样品，它们的重心分别为 $\bar{\boldsymbol{x}}_p$ 和 $\bar{\boldsymbol{x}}_q$，它们之间的距离用 $D(p,\ q)$ 表示。下面是一些常用的定义。

（1）最短距离法（nearest neighbor 或 single linkage method）。

$$D_k(p,q)=\min\{d_{jl}\mid j\in G_p,l\in G_q\}\tag{3.12}$$

它等于类 G_p 与类 G_q 中最邻近的两个样品的距离。该准则下类的合并过程在图 3—2 中概要说明。

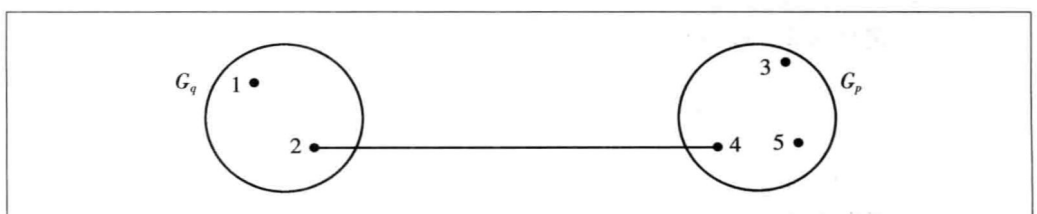


图 3—2 类间距离示意图——类群距离 $D_k(p,\ q)=d_{24}$

（2）最长距离法（farthest neighbor 或 complete linkage method）。

$$D_k(p,q)=\max\{d_{jl}\mid j\in G_p,l\in G_q\}\tag{3.13}$$

它等于类 G_p 与类 G_q 中最远的两个样品的距离。该准则下合并类的过程如图 3—3 所示。

图 3—3　类间距离示意图——类群距离 $D_k(p,q)=d_{13}$

（3）类平均法（group average method）。

$$D_G(p,q)=\frac{1}{lk}\sum_{i\in G_p}\sum_{j\in G_q}d_{ij} \tag{3.14}$$

它等于类 G_p 和类 G_q 中任两个样品距离的平均，式中的 l 和 k 分别为类 G_p 和类 G_q 中的样品数。该准则下合并类的过程如图 3—4 所示。

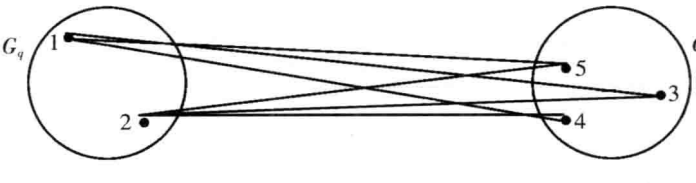


图 3—4　类间距离示意图——类群距离 $D_k(p,q)=\dfrac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$

（4）重心法（centroid method）。

$$D_c(p,q)=d_{\bar{X}_p\bar{X}_q} \tag{3.15}$$

它等于两个重心 $\bar{X}_p$ 和 $\bar{X}_q$ 间的距离。

（5）离差平方和法（sum of squares method）。若采用直径的第一种定义方法，用 D_p，D_q 分别表示类 G_p 和类 G_q 的直径，用 D_{p+q} 表示大类 D_{p+q} 的直径，则

$$D_p=\sum_{i\in G_p}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_p)'(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_p)$$

$$D_q=\sum_{j\in G_q}(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_q)'(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_q)$$

$$D_{p+q}=\sum_{j\in G_p\cup G_q}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'(\boldsymbol{x}_j-\bar{\boldsymbol{x}})$$

式中　$\bar{\boldsymbol{x}}=\dfrac{1}{k+m}\sum_{i\in G_p\cup G_q}\boldsymbol{x}_i$

用离差平方和法定义 G_p 和 G_q 之间的距离平方为：

$$D_w^2(p,q)=D_{p+q}-D_p-D_q \tag{3.16}$$

可以证明这种定义是有意义的。证明参见参考文献［7］。如果样品间的距离采用欧氏距离，同样可以证明下式成立：

$$D_w^2(p,q)=\frac{km}{k+m}D_c^2(p,q) \tag{3.17}$$

这表明，离差平方和法定义的类间距离 $D_w(p,\ q)$ 与重心法定义的距离 $D_c(p,\ q)$ 只差一个常数倍，这个倍数与两类的样品数有关。

3.4 系统聚类法

系统聚类法（hierarchical clustering method）是聚类分析诸方法中使用最多的。它包含下列步骤（见图 3—5）。

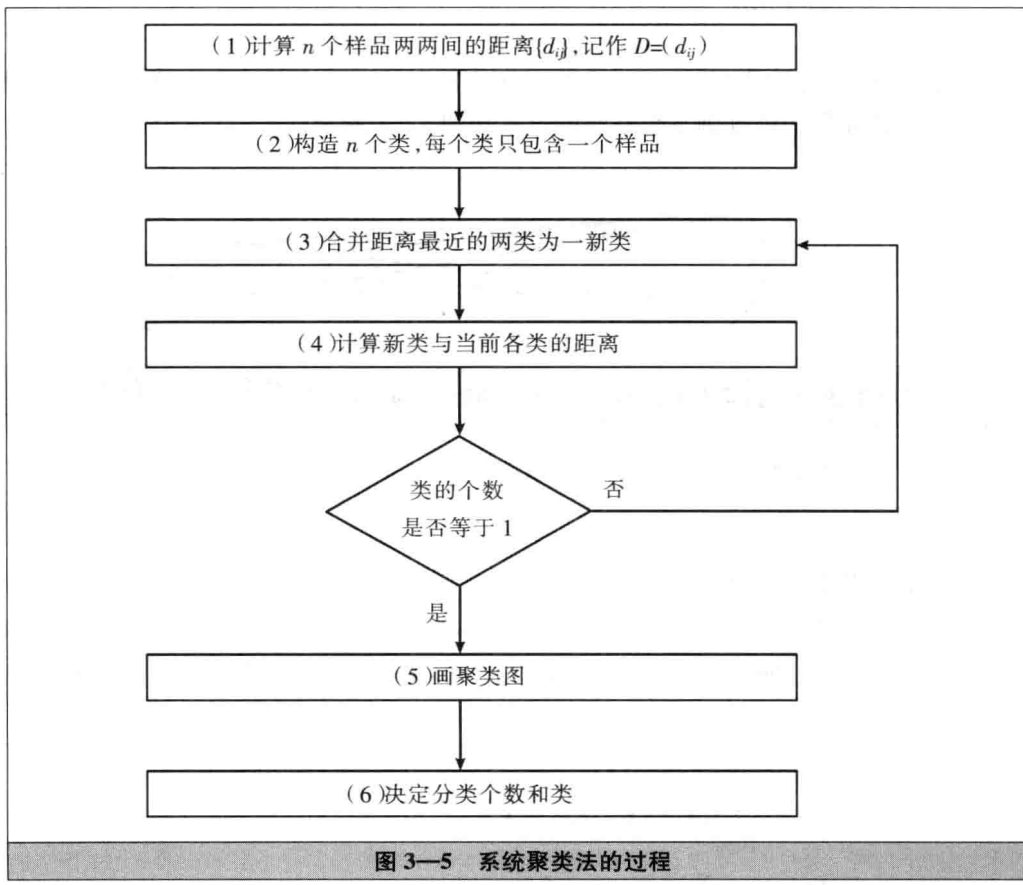


图 3—5 系统聚类法的过程

上节我们曾给出类与类之间的五种距离的定义，每一种定义用到上述系统聚类程序中，就得到一种系统聚类法。我们现在通过一个简单的例子来说明各种系统聚类法。

例 3—4

为了研究辽宁等 5 个省份 2000 年城镇居民消费支出的分布规律，根据调查资料做类型划分。指标名称及原始数据见表 3—5 和参考文献 [10]。

表 3—5　　2000 年 5 个省份城镇居民平均每人全年消费性支出数据

X1 食品支出（元/人）					X5 交通和通讯支出（元/人）			
X2 衣着支出（元/人）					X6 娱乐、教育和文化服务支出（元/人）			
X3 家庭设备、用品及服务支出（元/人）					X7 居住支出（元/人）			
X4 医疗保健支出（元/人）					X8 杂项商品和服务支出（元/人）			
	X1	X2	X3	X4	X5	X6	X7	X8
辽宁	1 772.14	568.25	298.66	352.20	307.21	490.83	364.28	202.50
浙江	2 752.25	569.95	662.31	541.06	623.05	917.23	599.98	354.39
河南	1 386.76	460.99	312.97	280.78	246.24	407.26	547.19	188.52
甘肃	1 552.77	517.16	402.03	272.44	265.29	563.10	302.27	251.41
青海	1 711.03	458.57	334.91	307.24	297.72	495.34	274.48	306.45

现在将表 3—5 中的每个省份看成一个样品，先计算 5 个省份之间的欧氏距离，用 $\boldsymbol{D}_0$ 表示相应的矩阵（由于矩阵对称，故只写出下三角部分）。

$$\boldsymbol{D}_0=\begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ \text{辽宁 } 1 & 0 & & & & \\ \text{浙江 } 2 & 1\,220.13 & 0 & & & \\ \text{河南 } 3 & 457.91 & 1\,580.69 & 0 & & \\ \text{甘肃 } 4 & 284.60 & 1\,390.71 & 356.80 & 0 & \\ \text{青海 } 5 & 195.14 & 1\,284.71 & 452.80 & 208.90 & 0 \end{bmatrix}$$

距离矩阵 $\boldsymbol{D}_0$ 中的各元素数值的大小就反映了 5 个省份间消费水平的接近程度。例如青海省和辽宁省之间的欧氏距离最小，为 195.14，反映了这两个省份城镇居民的消费水平最接近。

3.4.1　最短距离法和最长距离法

最短距离法是类与类之间的距离采用式（3.12）计算的系统聚类法。

开始例 3—4 有 5 类：$G_1=\{$辽宁 1$\}$，$G_2=\{$浙江 2$\}$，$G_3=\{$河南 3$\}$，$G_4=\{$甘肃 4$\}$，$G_5=\{$青海5$\}$，由最短距离法的定义，这时

$$D_k(i,j)=d_{ij},\qquad i,j=1,2,\cdots,5$$

即这五类之间的距离等于 5 个样品之间的距离。为了简化记号，下面用 $D(i,\ j)$ 代替 $D_k(i,\ j)$。我们发现 $\boldsymbol{D}_0$ 中最小的元素是 $D(1,\ 5)=195.14$，故将类 G_1 和类 G_5 合并成一个新类 $G_6=\{1,\ 5\}$，然后计算 G_6 与 G_2，G_3，G_4 之间的距离。利用

$$D(6,i)=\min\{D(1,i),D(5,i)\},\qquad i=2,3,4$$

其最近相邻的距离是：

$$d_{(1,5)2}=\min\{d_{12},d_{25}\}=\min\{1\ 220.13,1\ 284.71\}=1\ 220.13$$

$$d_{(1,5)3}=\min\{d_{13},d_{35}\}=\min\{457.91,452.80\}=452.80$$

$$d_{(1,5)4}=\min\{d_{14},d_{45}\}=\min\{284.60,208.90\}=208.90$$

在距离矩阵 $\boldsymbol{D}_0$ 中消去 1，5 所对应的行和列，并加入 {1，5} 这一新类对应的一行一列，得到新距离矩阵为：

$$\boldsymbol{D}_1=\begin{bmatrix} & G_6 & G_2 & G_3 & G_4 \\ G_6=\{1,5\} & 0 & & & \\ G_2 & 1\ 220.13 & 0 & & \\ G_3 & 452.80 & 1\ 580.69 & 0 & \\ G_4 & 208.90 & 1\ 390.71 & 356.80 & 0 \end{bmatrix}$$

然后，在 $\boldsymbol{D}_1$ 中发现类间最小距离是 $d_{64}=d_{(1,4,5)}=208.90$，合并类 {1，5} 和 G_4，得新类 $G_7=\{1，4，5\}$。再利用

$$D(7,i)=\min\{D(4,i),D(6,i)\},\qquad i=2,3$$

计算得

$$d_{(1,4,5)2}=\min\{d_{42},d_{(1,5)2}\}=\min\{1\ 390.71,1\ 220.13\}=1\ 220.13$$

$$d_{(1,4,5)3}=\min\{d_{43},d_{(1,5)3}\}=\min\{356.80,452.80\}=356.80$$

故得下一层次聚类的距离矩阵为：

$$\boldsymbol{D}_2=\begin{bmatrix} & G_7 & G_2 & G_3 \\ G_7=\{1,4,5\} & 0 & & \\ G_2 & 1\ 220.13 & 0 & \\ G_3 & 356.80 & 1\ 580.69 & 0 \end{bmatrix}$$

类间最小距离是 $d_{37}=356.80$，合并类 G_3 和类 G_7 得新类 $G_8=\{1，3，4，5\}$。此时，我们有两个不同的类 $G_8=\{1，3，4，5\}$ 和 G_2 合并，形成一个大类的聚类系统。

最后，决定类的个数与类。若用类的定义 3.1，从图上，分两类较为合适，这时阈值 $T=5$，这等价于在图 3—6 上距离为 5 处切一刀，得到两类为 {辽宁、青海、甘肃、河南} 与 {浙江}。

所谓最长距离法，是类与类之间的距离采用式（3.13）计算的系统聚类法。

上述两种方法中，主要的不同是计算新类与其他类的距离的递推公式不同。设某步将类 G_p 和 G_q 合并为 G_r，则 G_r 与其他类 G_l 的距离为：

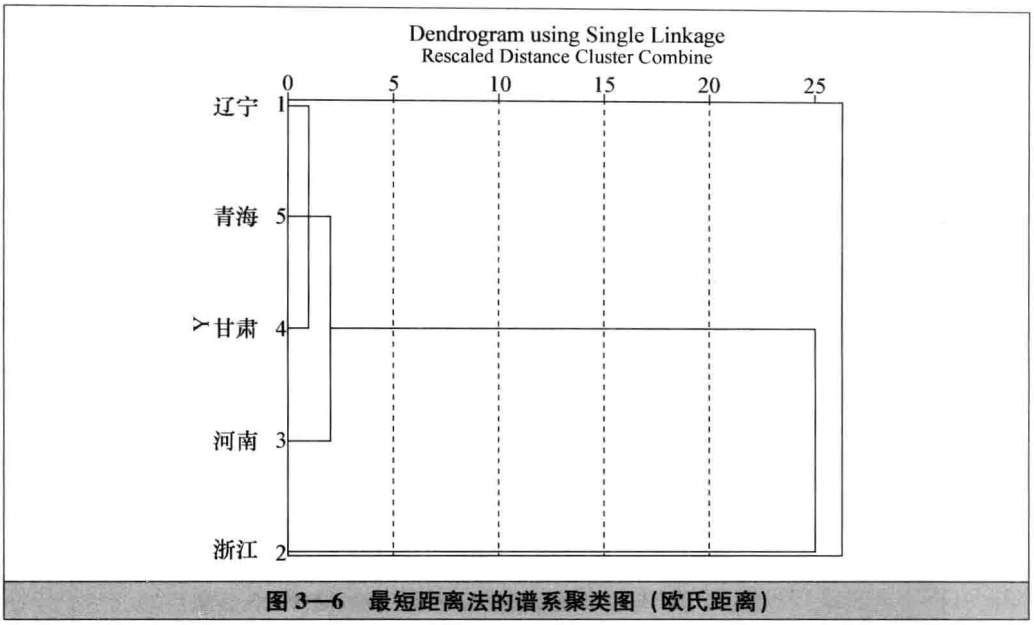


图 3—6　最短距离法的谱系聚类图（欧氏距离）

$$D_k(r,l) = \min\{D_k(p,l), D_k(q,l)\} \tag{3.18}$$

$$D_s(r,l) = \max\{D_s(p,l), D_s(q,l)\} \tag{3.19}$$

也就是说，在最长距离法中，选择最大的距离作为新类与其他类之间的距离，然后将类间距离最小的两类进行合并，一直合并到只有一类为止。

最短距离法也可用于对指标的分类，分类时可以用距离，也可以用相似系数。但用相似系数时应找最大的元素并类，计算新类与其他类的距离应使用式（3.19）。

最短距离法的主要缺点是它有链接聚合的趋势，因为类与类之间的距离为所有距离中的最短者，两类合并以后，它与其他类的距离缩小了，这样容易形成一个比较大的类，大部分样品都被聚在一类中，在树状聚类图中，会看到一个延伸的链状结构，所以最短距离法的聚类效果并不好，实践中不提倡使用。

最长距离法克服了最短距离法链接聚合的缺陷，两类合并以后与其他类的距离是原来两个类中的距离最大者，加大了合并后的类与其他类的距离（见图 3—7）。

我们看到，本例中最短距离法与最长距离法得到的结果是相同的。

3.4.2　重心法和类平均法

从物理的观点看，一个类用它的重心（该类样品的均值）做代表比较合理，类与类之间的距离就用重心之间的距离来代表。若样品之间采用欧氏距离，设某一步将类 G_p 与 G_q 合并成 G_r，它们各有 n_p，n_q，$n_r(n_r=n_p+n_q)$ 个样品，它们的重心用 $\bar{\boldsymbol{X}}_p$，$\bar{\boldsymbol{X}}_q$ 和 $\bar{\boldsymbol{X}}_r$ 表示，显然

$$\bar{\boldsymbol{X}}_r = \frac{1}{n_r}(n_p\bar{\boldsymbol{X}}_p + n_q\bar{\boldsymbol{X}}_q) \tag{3.20}$$

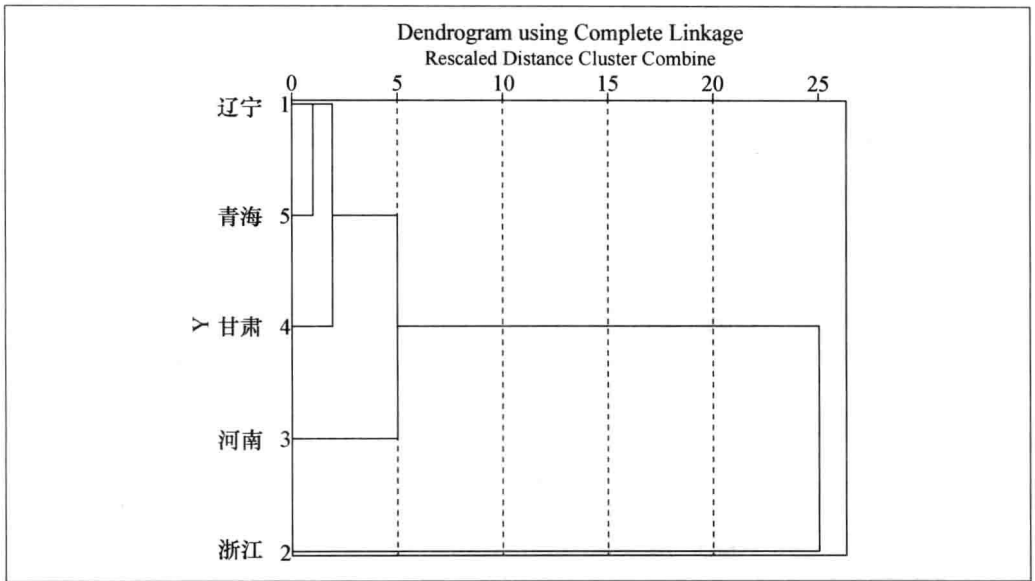


图 3—7 最长距离法的谱系聚类图（欧氏距离）

某一类 G_k 的重心为 $\bar{\boldsymbol{X}}_k$，它与新类 G_r 的距离是：

$$D_c^2(k,r)=(\bar{\boldsymbol{X}}_k-\bar{\boldsymbol{X}}_r)'(\bar{\boldsymbol{X}}_k-\bar{\boldsymbol{X}}_r) \tag{3.21}$$

可以证明（参见参考文献［6］），$D_c^2(k,\ r)$ 是如下的形式：

$$D_c^2(k,r)=\frac{n_p}{n_r}D_c^2(k,p)+\frac{n_q}{n_r}D_c^2(k,q)-\frac{n_p}{n_r}\frac{n_q}{n_r}D_c^2(p,q) \tag{3.22}$$

这就是重心法的距离递推公式。

重心法虽有很好的代表性，但并未充分利用各样本的信息。有学者将两类之间的距离平方定义为这两类元素两两之间的平均平方距离，即

$$\begin{aligned}D_G^2(k,r)&=\frac{1}{n_k n_r}\sum_{i\in G_k}\sum_{j\in G_r}d_{ij}^2\\&=\frac{1}{n_k n_r}\Big[\sum_{i\in G_k}\sum_{j\in G_p}d_{ij}^2+\sum_{i\in G_k}\sum_{j\in G_q}d_{ij}^2\Big]\end{aligned} \tag{3.23}$$

上式也可记为：

$$D_G^2(k,r)=\frac{n_p}{n_r}D_G^2(k,p)+\frac{n_q}{n_r}D_G^2(k,q) \tag{3.24}$$

这就是类平均法的递推公式。类平均法是聚类效果较好、应用比较广泛的一种聚类方法。它有两种形式，一种是组间联结法（between-groups linkage）；另一种是组内联结法（within-groups linkage）。组间联结法在计算距离时只考虑两类之间样品之间距离的平均；组内联结法在计算距离时把两组所有个案之间的距离都考虑在内。还有一种类平均法，它将类与类之间的距离定义为：

$$D_G^2(p,q) = \frac{1}{n_p n_q} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij}^2 \tag{3.25}$$

用类似的方法可导出这种定义下的距离递推公式如下：

$$D_G^2(k,r) = \frac{n_p}{n_r} D_G^2(k,p) + \frac{n_q}{n_r} D_G^2(k,q) \tag{3.26}$$

有人认为，类平均法是系统聚类法中比较好的方法之一。

在类平均法的递推公式中没有反映 D_{pq} 的影响，有学者将递推公式改为：

$$D_{kr}^2 = \frac{n_p}{n_r}(1-\beta) D_{kp}^2 + \frac{n_q}{n_r}(1-\beta) D_{kq}^2 + \beta D_{pq}^2 \tag{3.27}$$

式中，$\beta<1$。对应于式（3.27）的聚类法称为可变类平均法。

可变类平均法的分类效果与 β 的选择关系极大，有一定的人为性，因此在实践中使用尚不多。β 如果接近 1，一般分类效果不好，故 β 常取负值。重心法的谱系聚类图如图 3—8 所示。类平均法（组内联结法）的谱系聚类图（欧氏距离）如图 3—9 所示。

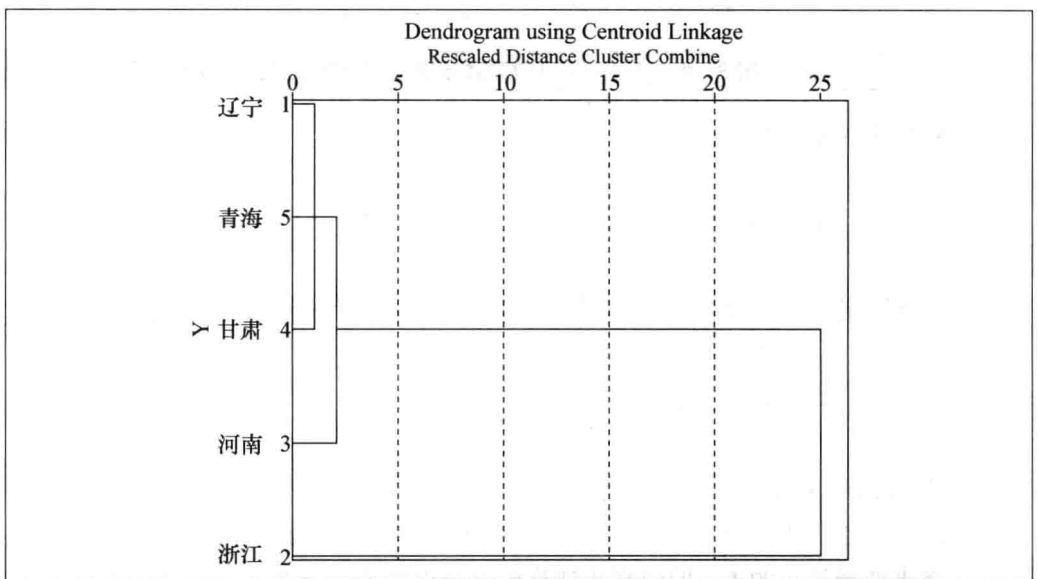


图 3—8　重心法的谱系聚类图（平方欧氏距离）

3.4.3　离差平方和法（或称 Ward 方法）

离差平方和法是由沃德（Ward）提出的，许多文献中称为 Ward 法。它的思想源于方差分析，即如果类分得正确，同类样品的离差平方和应当较小，类与类之间的离差平方和应当较大。

设将 n 个样品分成 k 类 G_1，G_2，…，G_k，用 $\boldsymbol{x}_{it}$ 表示类 G_t 中的第 i 个样品（注意 $\boldsymbol{x}_{it}$ 是 p 维向量），n_t 表示类 G_t 中的样品个数，$\bar{\boldsymbol{x}}_t$ 是类 G_t 的重心，则在类 G_t 中的样品的离差平方和为：

$$\boldsymbol{L}_t = \sum_{i=1}^{n_t} (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_t)'(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_t)$$

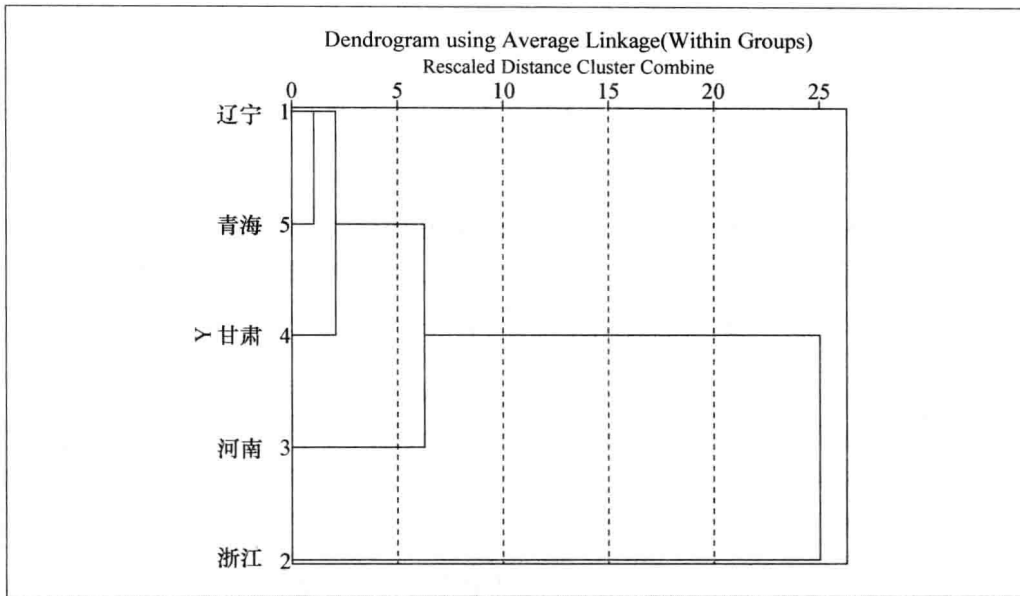


图 3—9 类平均法（组内联结法）的谱系聚类图（欧氏距离）

整个类内平方和为：

$$\boldsymbol{L}=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_t)'(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_t)=\sum_{t=1}^{k}\boldsymbol{L}_t$$

当 k 固定时，要选择使 $\boldsymbol{L}$ 达到极小的分类，n 个样品分成 k 类，一切可能的分法有：

$$R(n,k)=\frac{1}{k}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n \tag{3.28}$$

式（3.28）的证明见参考文献[6]。例如，当 $n=21$，$k=2$ 时，$R(21,2)=2^{21}-1=2\,097\,151$。当 n，k 更大时，$R(n,k)$ 就达到了天文数字。因此，要比较这么多分类来选择最小的 $\boldsymbol{L}$，即使高速计算机也难以完成。于是，只好放弃在一切分类中求 $\boldsymbol{L}$ 的极小值的要求，而是设计出某种规格：找到一个局部最优解，Ward 法就是寻找局部最优解的一种方法。其思想是先让 n 个样品各自成一类，然后每次缩小一类，每缩小一类，离差平方和就要增大，选择使 $\boldsymbol{L}$ 增加最小的两类合并，直到所有的样品归为一类为止。

若将某类 G_p 和 G_q 合并为 G_r，则类 G_k 与新类 G_r 的距离递推公式为：

$$D_w^2(k,r)=\frac{n_p+n_k}{n_r+n_k}D_w^2(k,p)+\frac{n_q+n_k}{n_r+n_k}D_w^2(k,q)-\frac{n_k}{n_r+n_k}D_w^2(p,q) \tag{3.29}$$

需要指出的是，离差平方和法只能得到局部最优解（见图 3—10）。至今还没有很好的办法以较少的计算求得精确最优解。

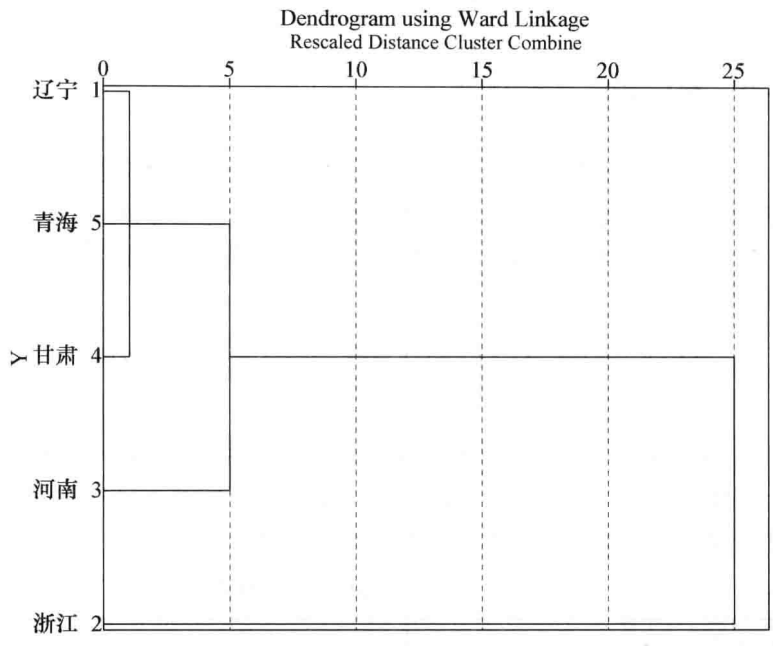


图 3—10　离差平方和法的谱系聚类图（平方欧氏距离）

3.4.4　分类数的确定

到目前为止，我们还没有讨论过如何确定分类数，聚类分析的目的是要对研究对象进行分类，因此，如何选择分类数成为各种聚类方法中的主要问题之一。在 K-均值聚类法中聚类之前需要指定分类数，谱系聚类法（系统聚类法）中我们最终得到的只是一个树状结构图，从图中可以看出存在很多类，但问题是如何确定类的最佳个数。

确定分类数是聚类分析中迄今为止尚未完全解决的问题之一，主要的障碍是对类的结构和内容很难给出一个统一的定义，这样就不能给出在理论上和实践中都可行的虚无假设。实际应用中人们主要根据研究的目的，从实用的角度出发，选择合适的分类数。德穆曼（Demirmen）曾提出根据树状结构图来分类的准则。

准则 1：任何类都必须在邻近各类中是突出的，即各类重心之间距离必须大。

准则 2：各类所包含的元素都不应过多。

准则 3：分类的数目应该符合使用的目的。

准则 4：若采用几种不同的聚类方法处理，则在各自的聚类图上应发现相同的类。

系统聚类中每次合并的类与类之间的距离也可以作为确定类数的一个辅助工具。在系统聚类过程中，首先把离得近的类合并，所以在并类过程中聚合系数（agglomeration coefficients）呈增加趋势，聚合系数小，表示合并的两类的相似程度较大，两个差异很大的类合到一起，会使该系数很大。如果以 y 轴为聚合系数，x 轴表示分类数，画出聚合系数随分类数的变化曲线，会得到类似于因子分析中的

碎石图，可以在曲线开始变得平缓的点选择合适的分类数。

3.4.5 系统聚类法的统一

上面介绍的五种系统聚类法，并类的原则和步骤是完全一样的，所不同的是类与类之间的距离有不同的定义，从而得到不同的递推公式。如果能将它们统一为一个公式，将大大有利于编制计算机程序。兰斯和威廉姆斯于 1967 年给出了一个统一的公式：

$$D^2(k,r)=\alpha_p D^2(k,p)+\alpha_q D^2(k,q)+\beta D^2(p,q)+\gamma\left|D^2(k,p)-D^2(k,q)\right| \tag{3.30}$$

式中，α_p，α_q，β，γ 对于不同的方法有不同的取值，表 3—6 列出了不同方法中这四个参数的取值。表中除了上述五种方法外，还列举了另三种系统聚类法，由于它们用得较少，这里不再详述，可参见参考文献［6］。

表 3—6　　系统聚类法参数表

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	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	$(n_k+n_p)/(n_k+n_r)$	$(n_k+n_p)/(n_k+n_r)$	$-n_k/(n_k+n_r)$	0

一般而言，不同聚类方法的结果不完全相同。最短距离法适用于条形的类。最长距离法、重心法、类平均法、离差平方和法适用于椭圆形的类。

现在许多统计软件都包含系统聚类法的程序，只要将数据输入，即可很方便地将上述八种方法定义的距离全部算出，并画出聚类图。本书将介绍采用 SPSS 软件实现聚类分析的过程。

由于上述聚类方法得到的结果不完全相同，于是产生一个问题：选择哪一个结果为好？为了解决这个问题，需要研究系统聚类法的性质，现简要介绍如下。

（1）单调性。令 D_r 为系统聚类法中第 r 次并类时的距离，如例 3—4，用最短距离时，有 $D_1=195.14$，$D_2=208.90$，$D_3=356.80$，$D_4=1\ 220.13$，此时 $D_1<D_2<D_3<\cdots$。一种系统聚类法若能保证 $\{D_r\}$ 是严格单调上升的，则称它具有单调性。由单调性画出的聚类图符合系统聚类的思想，先结合的类关系较近，后结合的类关系较疏远。显然，最短距离法和最长距离法具有并类距离的单调性。可以证明，类平均法、离差平方和法、可变法和可变类平均法都具有单调性，只有重心法

和中间距离法不具有单调性（证明见参考文献［6］）。

（2）空间的浓缩与扩张。对同一问题作聚类图时，横坐标（并类距离）的范围相差很远。最短距离法的范围较小，最长距离法的范围较大，类平均法则介于二者之间。范围小的方法区分类的灵敏度差，而范围太大的方法灵敏度又过高，会使支流淹没主流，这与收音机的灵敏度有相似之处。灵敏度太低的收音机接收的台少，灵敏度太高，台与台之间容易干扰，适中为好。按这一直观的想法引进如下的概念。

定义 3.5　设两个同阶矩阵 $\boldsymbol{A}=(a_{ij})$ 和 $\boldsymbol{B}=(b_{ij})$ 的元素非负，如果 $\boldsymbol{A}$ 的每一个元素不小于 $\boldsymbol{B}$ 相应的元素，若 $a_{ij}\geqslant b_{ij}(\forall i, j)$，则记作 $\boldsymbol{A}\geqslant\boldsymbol{B}$（请勿与非负定阵 $\boldsymbol{A}\geqslant\boldsymbol{B}$ 的意义相混淆，这个记号仅在本节中使用）。由定义推知，$\boldsymbol{A}\geqslant\boldsymbol{0}$，表示 $\boldsymbol{A}$ 的元素非负。

设有 A，B 两种系统聚类法，在第 k 步的距离阵记作 $\boldsymbol{A}_k$ 和 $\boldsymbol{B}_k(k=0, 1, \cdots, n-1)$，若 $\boldsymbol{A}_k\geqslant\boldsymbol{B}_k(k=1, 2, \cdots, n-1)$，则称 $\boldsymbol{A}$ 比 $\boldsymbol{B}$ 扩张或者 $\boldsymbol{B}$ 比 $\boldsymbol{A}$ 浓缩。对系统聚类法有如下的结论（参见参考文献［6］）：

$$(K)\leqslant(G)\leqslant(S)$$
$$(C)\leqslant(G)\leqslant(W)$$

式中，(K) 是最短距离法；(S) 是最长距离法；(C) 是重心法；(W) 是离差平方和法；(G) 是类平均法。归纳起来说，与类平均法相比，最短距离法、重心法使空间浓缩；最长距离法、离差平方和法使空间扩张。太浓缩的方法不够灵敏，太扩张的方法在样本大时容易失真。类平均法比较适中，相比其他方法，类平均法不太浓缩也不太扩张，故许多书推荐这种方法。

有关系统聚类法的性质，学者们还从其他角度提出了比较优劣的原则。欲将 n 个样品分为 k 类，有人定义一个分类函数（或叫损失函数），然后寻找这个函数的最优解，在某些条件下，最短距离法的解是最优的，而系统聚类法的其他方法都不具有这个性质（参见参考文献［6］、［7］），故最短距离法在实践中也很受推崇。系统聚类法的各种方法的比较仍是一个值得研究的课题，例如，有学者用随机模拟做了研究，发现最长距离法并不可取。

3.5　模糊聚类分析

模糊集的理论是 20 世纪 60 年代中期美国的自动控制专家查德（L. A. Zadeh）教授首先提出的。模糊集的理论已广泛应用于许多领域，将模糊集概念用到聚类分析中便产生了模糊聚类分析。

3.5.1　模糊聚类的几个基本概念

（1）特征函数。对于一个普通集合 A，空间中任一元素 x，要么 $x\in A$，要么

$x\notin A$，二者必居其一，这一特征用一个函数表示为：

$$A(x)=\begin{cases}1, & x\in A\\ 0, & x\notin A\end{cases}$$

则称 $A(x)$ 为集合 A 的特征函数。

如某工业企业完成年计划利润定义为 1，没有完成年计划利润则定义为 0，用特征函数来描述即为：

$$A(x)=\begin{cases}1, & x\in A \quad \text{完成}\\ 0, & x\notin A \quad \text{没完成}\end{cases}$$

（2）隶属函数。当我们要了解某企业完成年计划利润程度的大小时，仅用特征函数就不够了。模糊数学把它推广到 [0，1] 闭区间，即用 [0，1] 区间的一个数去度量它，这个数叫隶属度。当用函数表示隶属度的变化规律时，就叫做隶属函数。即

$$0\leqslant A(x)\leqslant 1$$

如果某企业完成年计划利润的 90%，可以说，这个企业完成年计划利润的隶属度是 0.9。显然，隶属度概念是特征函数概念的拓广。特征函数描述空间的元素之间是否有关联，而隶属度描述了元素之间的关联是多少。

用集合语言来描述隶属函数的概念就为：设 x 为全域，若 A 为 x 上取值 [0，1] 的一个函数，则称 A 为模糊集。

若一个矩阵元素取值于 [0，1] 范围内，则称该矩阵为模糊矩阵。

（3）模糊矩阵的运算法则。如果 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是 $n\times p$ 和 $p\times m$ 的模糊矩阵，则乘积 $\boldsymbol{C}=\boldsymbol{A}\cdot\boldsymbol{B}$ 为 $n\times m$ 阵，其元素为：

$$C_{ij}=\overset{p}{\underset{k=1}{\mathrm{V}}}(a_{ik}\Lambda b_{kj}),\quad i=1,2,\cdots,n;j=1,2,\cdots,m$$

符号“V”和“Λ”的含义为：

$$a\mathrm{V}b=\max(a,b)$$
$$a\Lambda b=\min(a,b)$$

3.5.2 模糊分类关系

（1）乘积空间。n 个样品的所有全体所组成的集合 x 作为全域，令 $X\times Y=\{(x,\ y)\mid x\in X,\ y\in Y\}$，则称 $X\times Y$ 为 X 的全域乘积空间。

（2）分类关系。设 R 为 $X\times Y$ 上的一个集合，并且满足：

1）反身性：$(x,\ x)\in R$，即集合中每个元素和它自己同属一类。

2）对称性：若 $(x,\ y)\in R$，则 $(y,\ x)\in R$。

3）传递性：若 $(x,\ y)\in R$，$(y,\ z)\in R$，则有 $(x,\ z)\in R$。

这三条性质称为等价关系，满足这三条性质的集合 R 为一分类关系。

模糊聚类分析的实质就是根据研究对象本身的属性构造模糊矩阵，在此基础上根据一定的隶属度来确定其分类关系。

如果水平 λ_1 和 λ_2 满足 $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant 1$，则按水平 λ_2 分出的每一类必是按水平 λ_1 分出的某一类的子类。这就是模糊分类的基本原理。下面举一个简单的数值例子来说明其应用。

设 $X=\{x_1, x_2, x_3\}$ 上的模糊矩阵

$$\boldsymbol{R}=\begin{bmatrix} 1 & 0.4 & 0.6 \\ 0.4 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{bmatrix}$$

是一个模糊分类关系，现在从 $\boldsymbol{R}$ 出发对 X 进行分类。

当 $0.6<\lambda \leqslant 1$ 时，有

$$\boldsymbol{R}_\lambda=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

可知 $X=\{x_1\} \cup \{x_2\} \cup \{x_3\}$，即 x_1，x_2，x_3 各为一类。

当 $0.4<\lambda \leqslant 0.6$ 时，有

$$\boldsymbol{R}_\lambda=\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

可知 $X=\{x_1, x_3\} \cup \{x_2\}$，即 x_1，x_3 为一类，x_2 为另一类。

当 $0 \leqslant \lambda \leqslant 0.4$ 时，有

$$\boldsymbol{R}_\lambda=\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

可知 $X=\{x_1, x_2, x_3\}$，即 x_1，x_2，x_3 为一类。

3.5.3 模糊聚类分析计算步骤

(1) 对原始数据进行变换。变换方法通常有标准化变换、极差变换、对数变换等。

(2) 计算模糊相似矩阵。选取在 $[-1, 1]$ 区间中的普通相似系数 $r_{ij}^* = \cos(\theta)$ 构成相似系数矩阵，在此基础上做变换

$$r_{ij}=\frac{1+r_{ij}^*}{2}$$

使得 r_{ij}^* 被压缩到 $[0, 1]$ 区间内，$\boldsymbol{R}=(r_{ij})$ 构成了一个模糊矩阵。

（3）建立模糊等价矩阵。对模糊矩阵进行褶积运算：$\boldsymbol{R}\rightarrow\boldsymbol{R}^2\rightarrow\boldsymbol{R}^3\rightarrow\cdots\rightarrow\boldsymbol{R}^n$，经过有限次的褶积后使得 $\boldsymbol{R}^n\cdot\boldsymbol{R}=\boldsymbol{R}^n$，由此得到模糊分类关系 $\boldsymbol{R}^n$。

（4）进行聚类。给定不同的置信水平 λ，求 $\boldsymbol{R}_\lambda$ 截阵，找出 $\boldsymbol{R}$ 的 λ 显示，得到普通的分类关系 $\boldsymbol{R}_\lambda$。当 $\lambda=1$ 时，每个样品自成一类，随着 λ 值的减小，由细到粗逐渐并类。聚类结果也可像前面系统聚类一样画出树形聚类图。

3.6 *K*-均值聚类和有序样品的聚类

3.6.1 *K*-均值法（快速聚类法）

非谱系聚类法是把样品（而不是变量）聚集成 K 个类的集合。类的个数 K 可以预先给定，或者在聚类过程中确定。因为在计算机计算过程中无须确定距离（或相似系数矩阵），也无须存储数据，所以，非谱系方法可应用于比系统聚类法大得多的数据组。

非谱系聚类法或者一开始就对元素分组，或者从一个构成各类核心的“种子”集合开始。选择好的初始构形，将能消除系统的偏差。一种方法是从所有项目中随机地选择“种子”点或者随机地把元素分成若干个初始类。

这里讨论一种更特殊的非谱系过程，即 K-均值法，又叫快速聚类法。

麦克奎因（Macqueen）于 1967 年提出了 K-均值法（参见参考文献 [11]）。这种聚类方法的思想是把每个样品聚集到其最近形心（均值）类中。在它的最简单说明中，这个过程由下列三步所组成：

（1）把样品粗略分成 K 个初始类。

（2）进行修改，逐个分派样品到其最近均值的类中（通常用标准化数据或非标准化数据计算欧氏距离）。重新计算接受新样品的类和失去样品的类的形心（均值）。

（3）重复第（2）步，直到各类无元素进出。

若不在一开始就粗略地把样品分到 K 个预先指定的类（第（1）步），也可以指定 K 个最初形心（“种子”点），然后进行第（2）步。

样品的最终聚类在某种程度上依赖于最初的划分，或种子点的选择。

为了检验聚类的稳定性，可用一个新的初始分类重新检验整个聚类算法。如果最终分类与原来一样，则不必再行计算；否则，须另行考虑聚类算法。参见参考文献 [11]。

关于 K-均值法，对其预先不固定类数 K 这一点有很大的争论，其中包括下面几点：

（1）如果有两个或多个“种子”点无意中跑到一个类内，则其聚类结果将很难区分。

(2) 局外干扰的存在将至少产生一个样品非常分散的类。

(3) 即使已知总体由 K 个类组成，抽样方法也可造成属于最稀疏类的数据不出现在样本中。强行把这些数据分成 K 个类会导致无意义的聚类。

许多聚类算法都要求给定 K，而选择几种算法进行反复检验，对于结果的分析也许是有好处的。其他非谱系聚类过程的讨论可参见参考文献 [11]。

3.6.2 有序样品的聚类

在前几节的讨论中，分类的样品是相互独立的，分类时彼此是平等的。但在有些实际问题中，要研究的现象与时间的顺序密切相关。例如我们想要研究，1949—2011 年，国民收入可以划分为几个阶段。阶段的划分必须以年份顺序为依据，总的想法是要将国民收入接近的年份划分到一个段内。要完成类似这样问题的研究，用前几节分类的方法显然是不行的。对于这类有序样品的分类，实质上是需要找出一些分点，将它们划分成几个分段，每个分段看作一类，称这种分类为分割。显然，分点在不同位置可以得到不同的分割。这样就存在一个如何决定分点，使其达到所谓最优分割的问题。即要求一个分割能使各段内部样品间的差异最小，而各段之间样品的差异最大。这就是决定分割点的依据。

假设用 x_1，x_2，…，x_n 表示 n 个有顺序的样品，有序样品的分类结果要求每一类必须呈：$\{x_i, x_{i+1}, \cdots, x_{i+j}\}$ ($i \geqslant 1$，$j \geqslant 0$)。增加了有序这个约束条件，会对分类带来哪些影响呢?

1. 可能的分类数目

n 个样品分成 k 类，如果样品是彼此平等的，则一切可能的分法有：

$$R(n,k) = \sum_{\substack{i_1+\cdots+i_k=n \\ i_j \geqslant 1, j=1,\cdots,k}} \frac{n!}{i_1!\cdots i_k!} \tag{3.31}$$

而对于有序样品，n 个样品分成 k 类的一切可能的分法有：

$$R'(n,k) = \binom{n-1}{k-1} \tag{3.32}$$

这是容易证明的。n 个有序样品有 $(n-1)$ 个间隔，分成两类就是在这 $(n-1)$ 个间隔中插上一根“棍子”，故有 $(n-1)=\binom{n-1}{1}$ 种可能；若要分成三类，就是在这 $(n-1)$ 个间隔中任意插上两根“棍子”，故有 $\binom{n-1}{2}$ 种可能；要分成 k 类，就是插上 $k-1$ 根“棍子”，故有 $\binom{n-1}{k-1}$ 种可能。容易证明，$R(n,k)=o(k^n)$，$R'(n,k)=o(nk)$，当 n 较大时，$R'(n,k) \ll R(n,k)$，故有序样品的聚类问题要简单一些。

2. 最优分割法

这种方法用来分类的依据是离差平方和，但由于 $R'(n, k)$ 比 $R(n, k)$ 小得多，因此与系统聚类法中的离差平方和法又有所不同，前者可以求得精确最优解，而后者只能求得局部最优解。这种方法首先是由费歇（Fisher）提出的，又称为 Fisher 算法。

设样品依次是 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，…，$\boldsymbol{x}_n$（每个是 m 维向量），最优分割法的步骤大致如下：

（1）定义类的直径。设某一类 G_{ij} 是 $\{\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_j\}(j>i)$，它们的均值记成 $\bar{\boldsymbol{x}}_{ij}$：

$$\bar{\boldsymbol{x}}_{ij} = \frac{1}{j-i+1}\sum_{l=i}^{j}\boldsymbol{x}_l$$

G_{ij} 的直径用 $D(i, j)$ 表示，常用的直径是：

$$D(i,j) = \sum_{l=i}^{j}(\boldsymbol{x}_l - \bar{\boldsymbol{x}}_{ij})'(\boldsymbol{x}_l - \bar{\boldsymbol{x}}_{ij}) \tag{3.33}$$

当 $m=1$ 时，有时用直径

$$D(i,j) = \sum_{l=i}^{j}|(x_l - \tilde{x}_{ij})| \tag{3.34}$$

式中，$\tilde{x}_{ij}$ 是 $(x_i, x_{i+1}, \cdots, x_j)$ 的中位数。

（2）定义目标函数。将 n 个样品分成 k 类，设某一种分法是：$P(n, k)$：$\{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_1+1}, \cdots, \boldsymbol{x}_{i_2-1}\}$，$\{\boldsymbol{x}_{i_2}, \boldsymbol{x}_{i_2+1}, \cdots, \boldsymbol{x}_{i_3-1}\}$，…，$\{\boldsymbol{x}_{i_k}, \boldsymbol{x}_{i_k+1}, \cdots, \boldsymbol{x}_n\}$，或简记成

$$P(n,k):\{i_1, i_1+1, \cdots, i_2-1\}, \{i_2, i_2+1, \cdots, i_3-1\}, \cdots, \{i_k, i_k+1, \cdots, n\} \tag{3.35}$$

其中分点 $1=i_1<i_2<\cdots<i_k\leqslant i_{k+1}=n+1$。定义这种分类的目标函数为：

$$e[P(n,k)] = \sum_{j=1}^{k}D(i_j, i_{j+1}-1) \tag{3.36}$$

当 n，k 固定时，$e[P(n, k)]$ 越小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 $P(n, k)$ 使目标函数达到极小，以下 $P(i, j)$ 一般表示使 $e[P(n, k)]$ 达到极小的分类。

（3）精确最优解的求法。容易验证有如下递推公式：

$$e[P(n,2)] = \min_{2\leqslant j\leqslant n}\{D(1, j-1) + D(j, n)\} \tag{3.37}$$

$$e[P(n,k)] = \min_{k\leqslant j\leqslant n}\{e[P(j-1, k-1)] + D(j, n)\} \tag{3.38}$$

当我们要分 k 类时，首先找 j_k 使式（3.38）达到最小，即

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

于是 $G_k=\{j_k, j_k+1, \cdots, n\}$，然后找 j_k-1 使它满足

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

得到类 $G_{k-1}=\{j_{k-1}, \cdots, j_k-1\}$。采用类似的方法得到所有类 $G_1, \cdots, G_k$，这就是我们要求的最优解。

3.7　计算步骤与上机实现

本书以 SPSS 22.0 和 S-Plus 8 两种软件来说明前面讲述的几种聚类法的实现过程。具体步骤如下：

（1）分析所需要研究的问题，确定聚类分析所需要的多元变量；

（2）选择对样品聚类还是对指标聚类；

（3）选择合适的聚类方法；

（4）选择所需的输出结果。

我们将实现过程用逻辑框图表示为图 3—11。

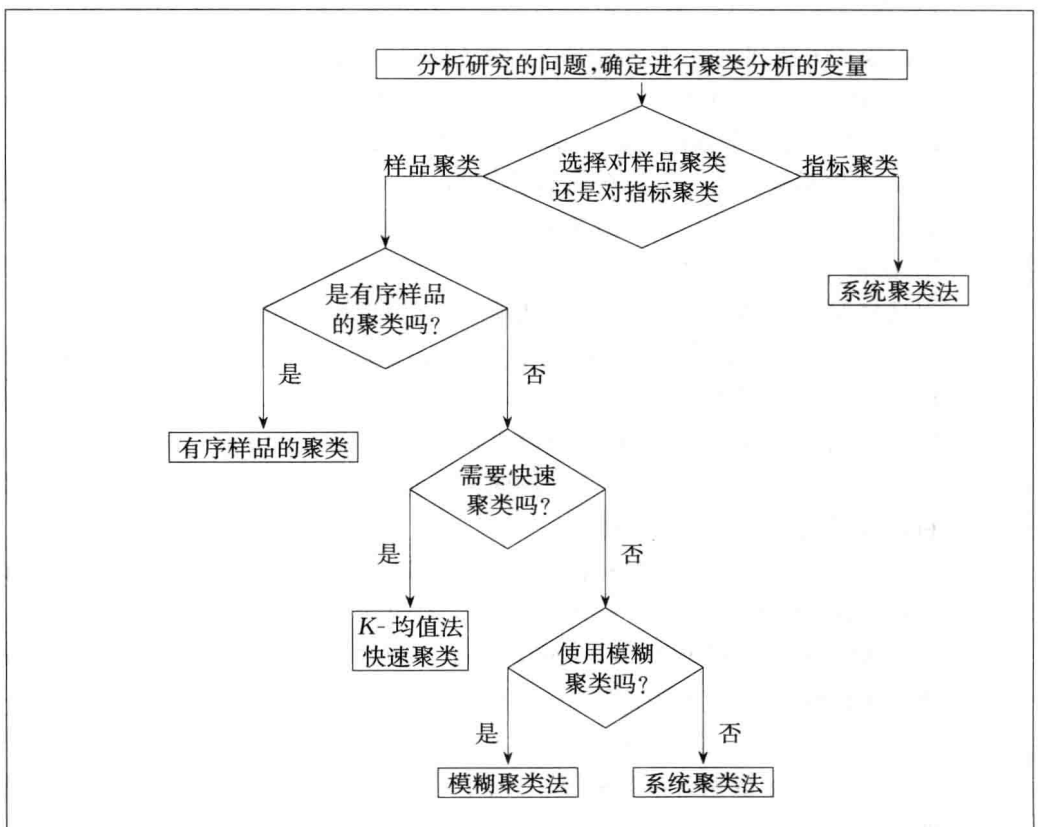


图 3—11　聚类分析流程图

3.7.1　系统聚类法

下面用 World95.sav 来做一个实例分析。为了研究亚洲国家或地区的经济发

展水平和文化教育水平，以便对亚洲国家和地区进行分类研究，我们进行聚类分析（在 World95. sav 数据中筛选出亚洲国家和地区，使用 Data→Select Cases→If condition is satisfied，选入 region=3）。详细步骤如下：

（1）打开数据。使用菜单中 File→Open 命令，然后选中要分析的数据 World95. sav。

（2）在菜单的选项中选择 Analyze→Classify 命令。Classify 命令下有两个聚类分析命令，一是 K-Means Cluster（K-均值聚类）；二是 Hierarchical Cluster（系统聚类法）。这里，我们选择系统聚类法。

（3）在系统聚类法中，我们看到 Cluster 下有两个选项，Cases（样品聚类或 Q 型聚类）和 Variables（变量聚类或 R 型聚类）。这里，我们选择对样品进行聚类。

（4）Display 下面有两个选项，分别是 Statistics（统计量）和 Plots（输出图形），我们可以选择所需要输出的统计量和图形。

（5）在系统聚类法中有四个按钮，分别是 Statistics，Plots，Method，Save。1）在 Statistics 中，有 Agglomeration schedule（每一阶段聚类的结果），Proximity matrix（样品间的相似性矩阵）。由 Cluster membership 可以指定聚类的个数，None 选项不指定聚类个数，Single solution 指定一个确定类的个数，Range of solution 指定类的个数的范围（如从分 3 类到分 5 类）。2）在 Plots 中，有 Dendrogram（谱系聚类图，也称树状聚类图），Icicle（冰柱图），Orientation（指冰柱图的方向，Horizontal 水平方向，Vertical 垂直方向）。3）在 Method 中，Cluster Method 可以选择聚类方法，Measure 中可以选择计算的距离。4）在 Save 中，可以选择保存聚类结果。选好每个选项后，点 OK 就可以执行了。

在这个数据文件中，选择的变量（Variables(s)）有 Urban（城市人口比例），Lifeexpf（女性平均寿命），Lifeexpm（男性平均寿命），Literacy（有读写能力的人所占比例），Gdp _ cap（人均国内生产总值），以 Country（国家或地区）来标识（Label Cases）本例中的 17 个亚洲国家或地区，并以其他 5 个变量进行 Q 型聚类分析，即对国家或地区进行聚类。

这里我们将原始变量标准化（在 Method 选项下 Transform Values 的 Standardize 空白框内，选择 Z scores），在 Statistics 选项中选择 Agglomeration schedule，聚类方法选择 Within-group linkage（组内联结法），计算距离选择平方欧氏距离，输出样本间的接近度矩阵、冰柱图和树状聚类图。得到的结果如表 3—7、表 3—8 和图 3—12、图 3—14 所示。

将表 3—8 的聚合系数利用 Excel 作出聚合系数随分类数变化曲线，如图 3—13 所示。

输出结果中，表 3—7 表示接近度矩阵，是反映样品之间相似性或者相异性的矩阵。本例中由于计算距离使用的是平方欧氏距离，所以样品间距离越大，样品越相异。如果我们计算距离选择 Pearson 相关系数，则接近度矩阵是相似性矩阵。由表中矩阵可以看出，Bangladesh（孟加拉国）与 Cambodia（柬埔寨）的距离是最小的，因此它们最先聚为一类。

表 3—7

接近度矩阵

Proximity Matrix

Case	Squared Euclidean Distance																
	1: Afghanistan	2: Bangladesh	3: Cambodia	4: China Mainland	5: Hong Kong	6: India Mainland	7: Indonesia	8: Japan	9: Malaysia	10: N. Korea	11: Pakistan	12: Philippines	13: S. Korea	14: Singapore	15: Taiwan	16: Thailand	17: Vietnam
1:Afghanistan	0.000	1.586	0.969	15.500	38.743	5.032	10.875	46.572	17.161	23.368	3.843	15.768	26.247	39.898	30.028	18.419	14.616
2:Bangladesh	1.586	0.000	0.146	7.777	28.022	1.211	5.054	34.838	9.276	14.806	0.736	8.893	17.440	29.571	19.862	10.357	7.666
3:Cambodia	0.969	0.146	0.000	9.394	31.080	1.951	6.159	37.827	11.024	16.823	1.438	10.343	19.653	32.513	22.447	11.824	8.893
4:China Mainland	15.500	7.777	9.394	0.000	12.675	2.975	0.623	15.967	0.617	2.259	5.298	0.806	4.317	13.585	4.889	0.554	0.421
5:Hong Kong	38.743	28.022	31.080	12.675	0.000	19.389	14.578	1.854	8.222	8.137	20.958	11.357	3.734	0.299	2.564	12.781	15.260
6:India	5.032	1.211	1.951	2.975	19.389	0.000	1.424	25.088	3.875	7.636	0.522	3.655	9.883	20.563	11.666	4.802	3.102
7:Indonesia	10.875	5.054	6.159	0.623	14.578	1.424	0.000	18.277	1.119	2.958	3.421	0.637	5.058	15.091	6.514	1.119	0.424
8:Japan	46.572	34.838	37.827	15.967	1.854	25.088	18.277	0.000	11.458	11.089	28.114	14.534	5.850	1.633	4.633	14.441	17.785
9:Malaysia	17.161	9.276	11.024	0.617	8.222	3.875	1.119	11.458	0.000	1.173	5.954	0.595	1.962	8.835	2.421	0.945	1.253
10:N. Korea	23.368	14.806	16.823	2.259	8.137	7.636	2.958	11.089	1.173	0.000	10.649	0.904	1.012	7.875	1.715	1.905	2.589
11:Pakistan	3.843	0.736	1.438	5.298	20.958	0.522	3.421	28.114	5.954	10.649	0.000	6.246	12.564	22.563	14.263	7.977	5.929
12:Philippines	15.768	8.893	10.343	0.806	11.357	3.655	0.637	14.534	0.595	0.904	6.246	0.000	2.559	11.365	3.838	0.765	0.665
13:S. Korea	26.247	17.440	19.653	4.317	3.734	9.883	5.058	5.850	1.962	1.012	12.564	2.559	0.000	3.369	0.390	3.821	5.083
14:Singapore	39.898	29.571	32.513	13.585	0.299	20.563	15.091	1.633	8.835	7.875	22.563	11.365	3.369	0.000	2.658	13.150	15.688
15:Taiwan	30.028	19.862	22.447	4.889	2.564	11.666	6.514	4.633	2.421	1.715	14.263	3.838	0.390	2.658	0.000	4.596	6.289
16:Thailand	18.419	10.357	11.824	0.554	12.781	4.802	1.119	14.441	0.945	1.905	7.977	0.765	3.821	13.150	4.596	0.000	0.294
17:Vietnam	14.616	7.666	8.893	0.421	15.260	3.102	0.424	17.785	1.253	2.589	5.929	0.665	5.083	15.688	6:289	0.294	0.000

This is a dissimilarity matrix.

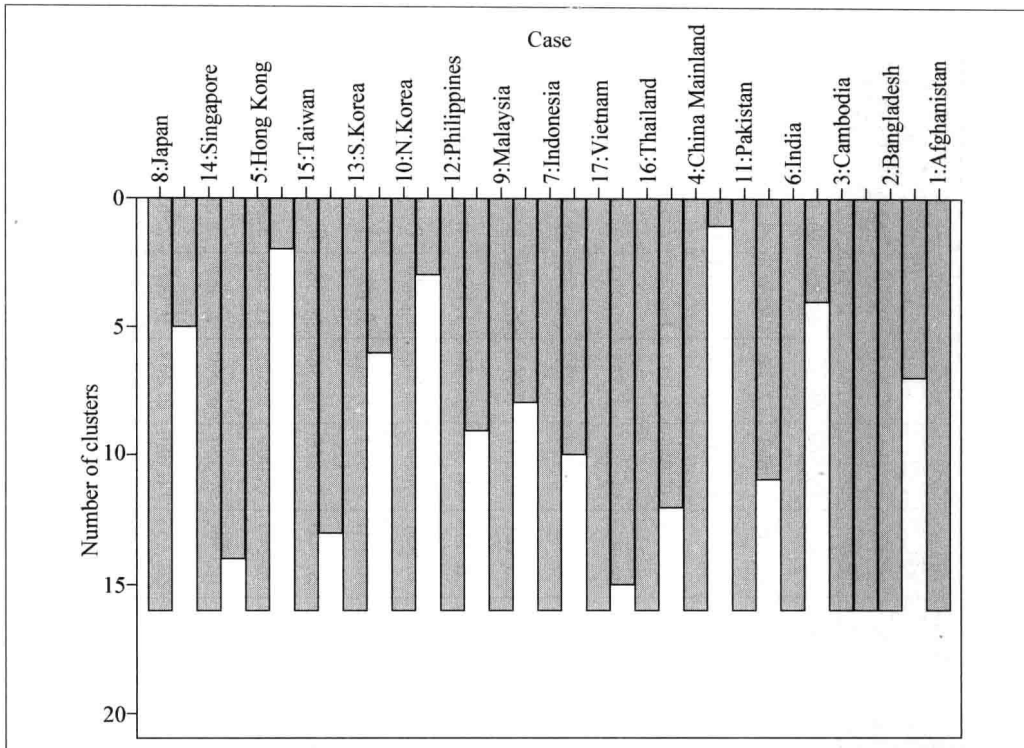


图 3—12 冰柱图

表 3—8 聚类过程的结果

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	0.146	0	0	10
2	16	17	0.294	0	0	5
3	5	14	0.299	0	0	12
4	13	15	0.390	0	0	11
5	4	16	0.423	0	2	7
6	6	11	0.522	0	0	13
7	4	7	0.573	5	0	9
8	9	12	0.595	0	0	9
9	4	9	0.723	7	8	14
10	1	2	0.901	0	1	13
11	10	13	1.039	0	4	14
12	5	8	1.262	3	0	15
13	1	6	1.744	10	6	16
14	4	10	2.141	9	11	15
15	4	5	5.694	14	12	16
16	1	4	10.000	13	15	0

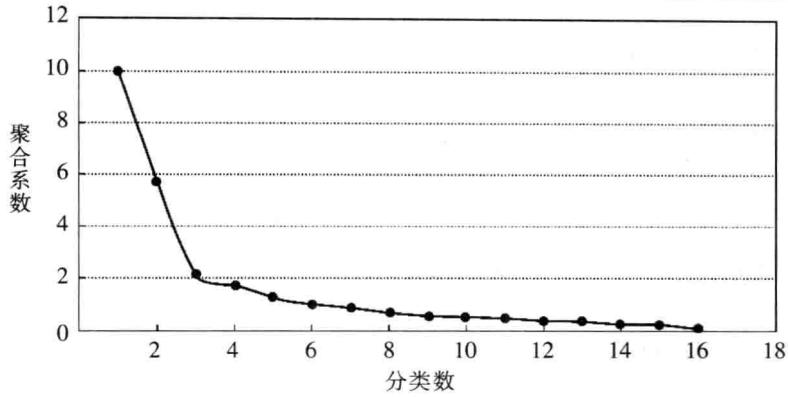


图 3—13　聚合系数随分类数变化曲线

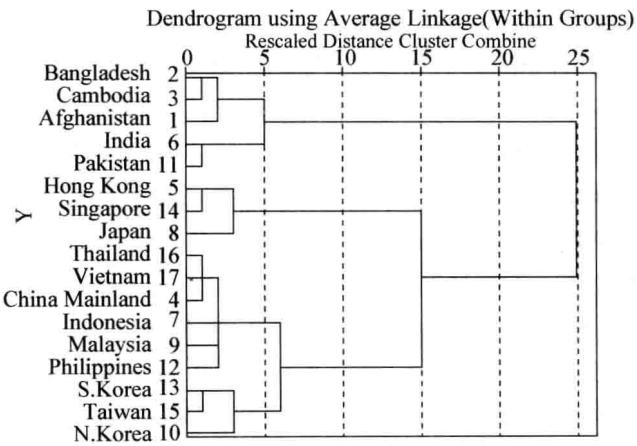


图 3—14　树状聚类图

图 3—12 是冰柱图，也是反映样品聚类情况的图。如果按照设定的类数，在该类数的行上从左到右就可以找到各类所包含的样品。比如我们希望分为三类，最左边的类数应选 3，每个样品右边都有一列冰柱，如果某个样品右边的列冰柱长度小于 3，那么它和前面冰柱长度大于 3 的样品聚为一类，如此下去，直到找到全部三类为止。例如，Hong Kong 右边的列冰柱长度为 2，那么它就与 Japan 和 Singapore 聚为一类了，而 China Mainland 右边的列冰柱长度为 1，那么从 Taiwan 到 China Mainland 又被聚为一类，后面样品聚为另一类。

表 3—8 是对每一阶段聚类结果的反映，Coefficients 表示聚合系数，第 2 列和第 3 列表示聚合的类，比如第一阶段（Stage＝1）时第 2 个样品——Bangladesh（孟加拉国）与第 3 个样品——Cambodia（柬埔寨）聚为一类，注意这时有 16 类（17－1＝16）。因此，某阶段的分类数等于总的样品数减去这个阶段的序号。

图 3—13 是聚合系数随分类数变化的曲线。由图可以看出，当分类数为 3 或 4 时，曲线变得比较平缓，这个分类数也符合我们分类的目的。

图 3—14 是树状聚类图，从图中可以由分类个数得到分类情况。如果我们选择分类数为 3，就从距离大概为 13 的地方往下切，得到分类结果如下。{1：孟加拉国，柬埔寨，阿富汗，印度，巴基斯坦}；{2：中国香港，新加坡，日本}；{3：泰国，越南，中国大陆，印度尼西亚，马来西亚，菲律宾，韩国，中国台湾和朝鲜}。我们可以从经济发展水平和文化教育水平来理解所做的分类。第 2 类应该是亚洲经济发达程度最高的国家或地区，第 1 类的经济水平和文教水平都比较低，第 3 类国家和地区的经济水平和文教水平居中。

3.7.2 快速聚类法

同样，我们使用上面的数据文件 World95. sav，从中筛选出亚洲国家或地区，试图将亚洲国家或地区按经济和文教水平分为三类。可以使用快速聚类法（*K*-Means cluster）对样品进行聚类。

我们使用的变量有 Country（国家或地区），Urban（城市人口比例），Lifeexpf（女性平均寿命），Lifeexpm（男性平均寿命），Literacy（有读写能力的人所占比例），Gdp-cap（人均国内生产总值），以 Country 来标识本例中的 17 个亚洲国家或地区，并以其他 5 个变量进行 Q 型聚类分析，即对国家或地区进行聚类。

在 SPSS 软件中选择 Analyze→Classify→K-Means Cluster。进入 *K*-均值聚类对话框以后，将上面 5 个变量选入 Variable，将 Country 用于标识（Label cases by）。将分类数（Fixed Number of clusters）指定为 3。我们可以在 Option 选项中选择 Initial cluster centers（最初分类重心），ANOVA（方差分析表），Cluster information for each case（每个样品的分类信息）。得到如下分类结果，如表3—9 至表 3—12 所示。

表 3—9　　最初各类的重心

Initial Cluster Centers

	Cluster		
	1	2	3
People living in cities (%)	18	77	71
Average female life expectancy	44	82	78
Average male life expectancy	45	76	72
People who read (%)	29	99	91
Gross domestic product/capita	205	19 860	7 055

输出结果中，表 3—9 表示最初各类的重心，也就是种子点。表 3—10 是样品的分类情况。这里我们看到快速聚类法将亚洲国家或地区分为这样三类：1：{阿富汗，孟加拉国，柬埔寨，中国大陆，印度，印度尼西亚，马来西亚，朝鲜，巴基斯坦，菲律宾，泰国，越南}；2：{中国香港，日本，新加坡}；3：{韩国，中国台湾}。我们也可以对分类结果做分析。第 1 类国家或地区经济和文教卫生水平较低。第 2 类国家或地区是亚洲国家或地区中的佼佼者，其经济发达程度和文教卫生水平都很高。第 3 类国家或地区处于两者之间。这个结果可以结合表 3—11（最后各类

的重心）来分析，可以看到，第 2 类的人均 GDP 比另外两组要高。

表 3—10　　样品的分类情况

Cluster Membership

Case Number	country	Cluster	Distance
1	Afghanistan	1	571.615
8	Bangladesh	1	573.924
19	Cambodia	1	516.229
24	China Mainland	1	398.151
47	Hong Kong	2	1 856.036
50	India	1	500.047
51	Indonesia	1	94.543
57	Japan	2	3 363.045
66	Malaysia	1	2 220.274
69	N. Korea	1	230.069
76	Pakistan	1	370.165
80	Philippines	1	96.542
86	S. Korea	3	214.034
89	Singapore	2	1 507.033
96	Taiwan	3	214.034
98	Thailand	1	1 025.608
108	Vietnam	1	545.396

表 3—11　　最后各类的重心

Final Cluster Centers

	Cluster		
	1	2	3
People living in cities (%)	29	90	72
Average female life expectancy	63	80	76
Average male life expectancy	60	75	70
People who read (%)	66	88	94
Gross domestic product/capita	775	16 497	6 841

表 3—12　　方差分析表

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
People living in cities (%)	5 336.488	2	169.577	14	31.469	.000
Average female life expectancy	454.600	2	70.494	14	6.449	.010
Average male life expectancy	321.326	2	41.113	14	7.816	.005
People who read (%)	1 073.096	2	570.625	14	1.881	.189
Gross domestic product/capita	304 156 058.2	2	1 780 295.690	14	170.846	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

表 3—12 是方差分析表，但是应当注意 F 值只能作为描述使用，不能根据该值判断各类均值是否有显著差异。从方差分析表可以看出，有 4 个变量对分类贡献显著。

3.7.3 模糊聚类法

我们继续使用上面的例子，希望将亚洲国家或地区分成三类进行分析研究。这里使用 S-Plus 8 软件。

进入 S-Plus 8 软件以后，首先打开上述数据文件，可以用 File→Import Data→From File，然后选择数据形式为 *.sav（SPSS 数据）。打开数据后，使用 Statistics→Cluster Analysis→ Fuzzy Partitioning 实现模糊聚类分析。

在 Variables 中选择 Urban（城市人口比例），Lifeexpf（女性平均寿命），Lifeexpm（男性平均寿命），Literacy（有读写能力的人所占比例），Gdp-cap（人均国内生产总值）进行 Q 型聚类分析，即对国家或地区进行聚类。在 Option 选项中指定类的个数为 3。在 Subset Rows with 选项中输出 region==“Pacific/Asia”，选择好变量以后，点击“OK”就可以得到结果。我们还选择了 Plot 选项中的 Cluster Plot（分类图）和 Silhouette Plot（侧影图）两个图形输出。得到以下结果，如表 3—13、表 3—14、图 3—15、图 3—16 所示。

表 3—13　　类别系数

Membership coefficients:

	[,1]	[,2]	[,3]
1	0.968 906 30	0.008 877 134	0.022 216 57
8	0.968 851 97	0.008 894 486	0.022 253 55
19	0.975 595 75	0.006 939 114	0.017 465 14
24	0.978 011 33	0.006 197 143	0.015 791 53
47	0.025 768 24	0.926 998 319	0.047 233 44
50	0.976 991 73	0.006 534 693	0.016 473 57
51	0.956 754 01	0.011 896 147	0.031 349 84
57	0.112 426 19	0.718 083 367	0.169 490 44
66	0.536 486 58	0.096 610 105	0.366 903 31
69	0.912 292 06	0.023 466 117	0.064 241 83
76	0.977 292 52	0.006 385 690	0.016 321 79
80	0.934 002 79	0.017 868 250	0.048 128 96
86	0.020 328 92	0.014 122 369	0.965 548 71
89	0.020 134 16	0.943 660 437	0.036 205 40
96	0.023 940 73	0.018 822 767	0.957 236 50
98	0.760 195 42	0.059 092 610	0.180 711 97
108	0.971 278 11	0.008 184 565	0.020 537 32

表 3—14　　分类情况

Closest hard clustering:

1	8	19	24	47	50	51	57	66	69	76	80	86	89	96	98	108
1	1	1	1	2	1	1	2	1	1	1	1	3	2	3	1	1

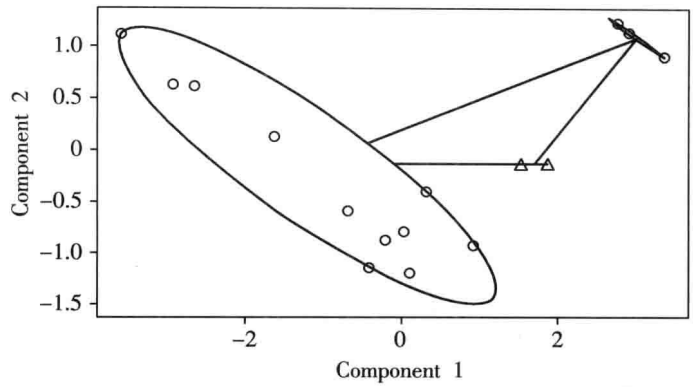


图 3—15　样品的分类图

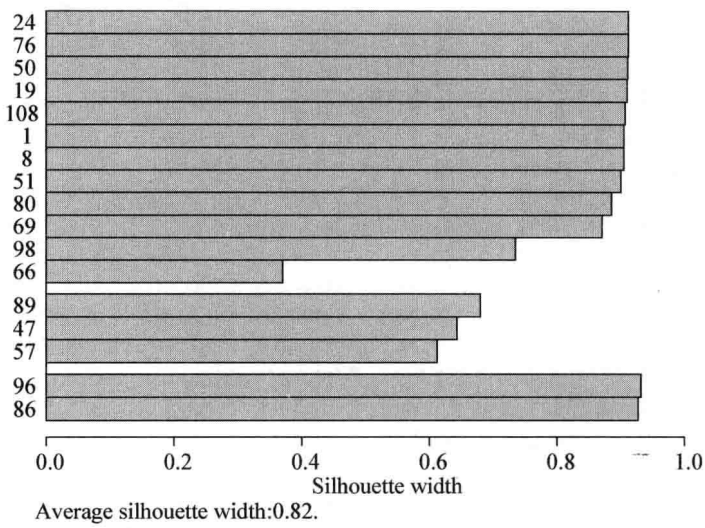


图 3—16　样品的侧影图

表 3—13 是各类的分类系数。由于我们指定分为三类，某个样品在这三类中的某类上系数最大，则将该样品聚为该类。比如第 1 个样品在第 1 类的系数最大，说明第 1 个样品在第 1 类中。

表 3—14 是聚类结果。由结果可以看出，与 K-均值快速聚类得到的结果是完全一致的。

图 3—15 是样品的分类图。由图可以看出，各类很明显地被分开。

图 3—16 是样品的侧影图。类似于水平的冰柱图或者树形图，可以看出三类中，各类包含哪些样品。

我们看到，此例中由模糊聚类得到的结果与采用 K-均值聚类得到的结果是一样的。同时应该看到，这种分类带有较强的主观性，而且分类结果也比较粗糙，一般仅适合对大量数据的快速聚类。

3.8 社会经济案例研究

例 3—5

城镇居民消费水平通常用表 3—15 中的八项指标来描述，八项指标间存在一定的线性相关。为研究城镇居民的消费结构，需将相关性强的指标归并到一起，这实际上就是对指标聚类。原始数据列于表 3—15。将原始数据录入 SPSS，并依次点击 Analyze→Correlate→Bivariate，打开 Bivariate Correlations 对话框，把八个变量选入 Variables 栏中，单击“OK”，得到这八个指标对应的相关系数，列于表 3—16。

表 3—15　　2012 年分地区城镇居民家庭平均每人全年现金消费支出

地区	食品	衣着	居住	家庭设备及用品	交通通信	文教娱乐	医疗保健	其他
	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$
北京	7 535.29	2 638.90	1 970.94	1 610.70	3 781.51	3 695.98	1 658.37	1 154.18
天津	7 343.64	1 881.43	1 854.22	1 151.16	3 083.37	2 254.22	1 556.35	899.87
河北	4 211.16	1 541.99	1 502.41	876.10	1 723.75	1 203.80	1 047.28	424.63
山西	3 855.56	1 529.47	1 438.88	832.52	1 672.29	1 506.20	905.88	470.72
内蒙古	5 463.18	2 730.23	1 583.56	1 242.64	2 572.93	1 971.78	1 354.09	798.68
辽宁	5 809.39	2 042.40	1 433.28	1 069.65	2 323.29	1 843.89	1 309.62	762.07
吉林	4 635.27	2 044.80	1 594.14	871.46	1 780.67	1 642.70	1 447.50	597.00
黑龙江	4 687.23	1 806.92	1 336.85	742.22	1 462.61	1 216.56	1 180.67	550.51
上海	9 655.60	2 111.17	1 790.48	1 906.49	4 563.80	3 723.74	1 016.65	1 485.53
江苏	6 658.37	1 915.97	1 437.08	1 288.42	2 689.51	3 077.76	1 058.11	700.06
浙江	7 552.02	2 109.58	1 551.69	1 161.39	4 133.50	2 996.59	1 228.02	812.39
安徽	5 814.92	1 540.66	1 396.97	811.23	1 809.72	1 932.74	1 142.96	562.44
福建	7 317.42	1 634.21	1 753.86	1 254.71	2 961.78	2 104.83	773.22	793.17
江西	5 071.61	1 476.63	1 173.91	966.23	1 501.34	1 487.30	670.71	427.93
山东	5 201.32	2 196.98	1 572.35	1 125.99	2 370.23	1 655.91	1 005.25	650.21
河南	4 607.47	1 885.99	1 190.81	1 145.42	1 730.35	1 525.33	1 085.47	562.13
湖北	5 837.93	1 783.41	1 371.15	978.26	1 476.98	1 651.92	1 029.55	366.78
湖南	5 441.63	1 624.57	1 301.60	1 034.30	2 084.15	1 737.64	918.41	466.65
广东	8 258.44	1 520.59	2 099.75	1 467.20	4 176.66	2 954.13	1 048.28	871.30
广西	5 552.56	1 146.46	1 377.26	1 125.39	2 088.64	1 626.05	883.56	444.06
海南	6 556.10	864.96	1 521.04	777.20	2 004.34	1 319.54	993.24	420.13
重庆	6 870.23	2 228.76	1 177.02	1 196.03	1 903.24	1 470.64	1 101.56	625.66
四川	6 073.86	1 651.14	1 284.09	1 097.93	1 946.72	1 587.43	772.75	635.62
贵州	4 992.85	1 399.00	1 013.53	849.94	1 891.03	1 396.00	654.53	388.82
云南	5 468.17	1 759.89	973.76	634.09	2 264.23	1 434.30	939.13	410.35
西藏	5 517.69	1 361.57	845.18	474.69	1 387.45	550.48	467.23	580.05
陕西	5 550.71	1 789.06	1 322.22	986.82	1 788.38	2 078.52	1 212.44	604.69
甘肃	4 602.33	1 631.40	1 287.93	833.15	1 575.67	1 388.21	1 049.65	478.72
青海	4 667.34	1 512.24	1 232.39	923.70	1 549.76	1 097.21	906.14	457.51
宁夏	4 768.91	1 875.70	1 193.37	929.01	2 110.41	1 515.91	1 063.09	610.74
新疆	5 238.89	2 031.14	1 166.59	950.17	1 660.27	1 280.81	1 027.60	536.24

表 3—16 相似性系数（相关系数）矩阵

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1	.227	.612**	.748**	.859**	.787**	.213	.797**
x2	.227	1	.305	.508**	.385*	.470**	.646**	.568**
x3	.612**	.305	1	.708**	.742**	.736**	.584**	.676**
x4	.748**	.508**	.708**	1	.802**	.857**	.367*	.830**
x5	.859**	.385*	.742**	.802**	1	.890**	.362*	.849**
x6	.787**	.470**	.736**	.857**	.890**	1	.488**	.824**
x7	.213	.646**	.584**	.367*	.362*	.488**	1	.443*
x8	.797**	.568**	.676**	.830**	.849**	.824**	.443*	1

表 3—16 中最大的相关系数为 $r_{5,6}=0.890$，将 $G5$ 和 $G6$ 并成一新类 $G9$，然后计算 $G9$ 与各类的相关系数，再找最大的相关系数，每次缩小一类，得到图 3—17。我们可以看出全国城镇居民的消费结构大致可以分为四个方面，一类是食品、家庭设备及用品、交通通信和文教娱乐支出，这是在消费结构中起主导作用的方面；一类是居住支出；一类是衣着支出；还有一类是医疗保健支出。

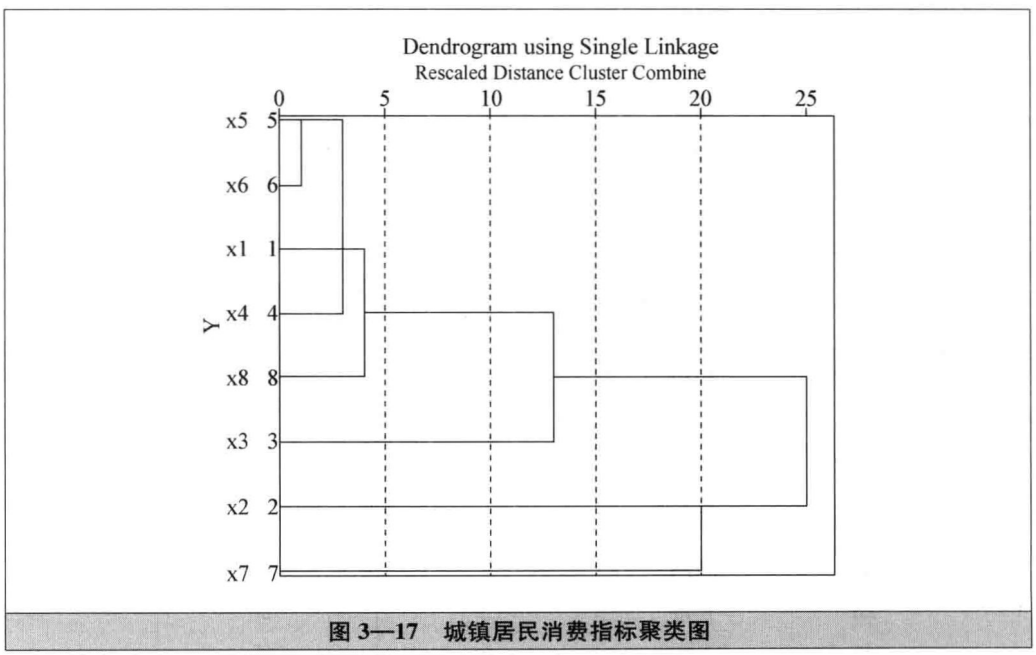


图 3—17 城镇居民消费指标聚类图

上面介绍的几种系统聚类方法，并类的原则和步骤基本一致，所不同的是类与类的距离有不同的定义。其实可以把这几种方法统一起来，有利于在计算机上灵活地选择更有意义的谱系图。

表 3—17 是不同聚类方法聚类结果的对照表。

表 3—17　　不同聚类方法聚类结果对照表

序号	地区	类平均法距离法类标记	最短距离法类标记	最长距离法类标记
1	北京	1	1	1
2	天津	1	1	1
3	河北	2	2	2
4	山西	2	2	2
5	内蒙古	2	2	2
6	辽宁	2	2	2
7	吉林	2	2	2
8	黑龙江	2	2	2
9	上海	3	3	3
10	江苏	1	1	1
11	浙江	1	1	1
12	安徽	2	2	2
13	福建	1	1	1
14	江西	2	2	2
15	山东	2	2	2
16	河南	2	2	2
17	湖北	2	2	2
18	湖南	2	2	2
19	广东	1	1	1
20	广西	2	2	2
21	海南	2	2	2
22	重庆	2	2	2
23	四川	2	2	2
24	贵州	2	2	2
25	云南	2	2	2
26	西藏	2	2	2
27	陕西	2	2	2
28	甘肃	2	2	2
29	青海	2	2	2
30	宁夏	2	2	2
31	新疆	2	2	2

对例 3—5，我们采用欧氏距离，分别运用类平均法、最短距离法、最长距离法，对 31 个省、直辖市、自治区分类。类平均法聚类在 SPSS 中的操作为：点选

Analyze→Classify→Hierarchical Cluster，打开 Hierarchical Cluster Analysis 对话框，将八个聚类指标选入 Variables 栏中，将表示地区的变量选入 Label Cases by 栏中，点击“Plots”按钮，在弹出的窗口中选中 Dendrogram（谱系图）选项，点击“Continue”返回主对话框，再点击“Method”按钮，在 Cluster Method 下拉菜单中选择 Between-groups linkage（组间连接法，即类平均法）选项，返回主对话框后点击“OK”即可得到聚类结果。最短距离法和最长距离法的操作步骤与类平均法一样，只不过要在 Cluster Method 下拉菜单中分别选择 Nearest neighbor 和 Furthest neighbor 选项。图 3—18、图 3—19、图 3—20 分别显示了三种方法的分类结果。为便于对照，将三种方法的分类结果综合列于表 3—17。

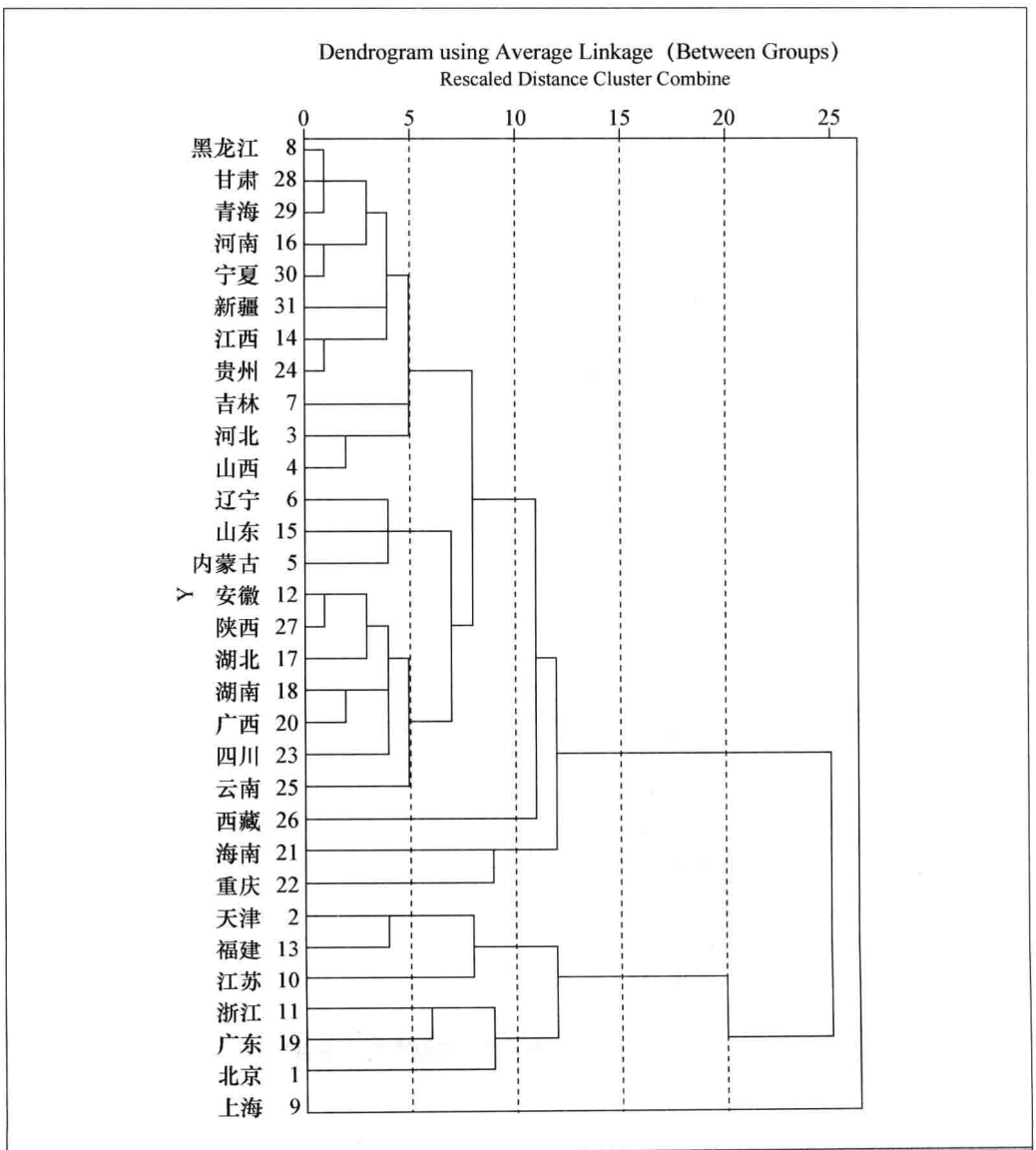


图 3—18 类平均法谱系图

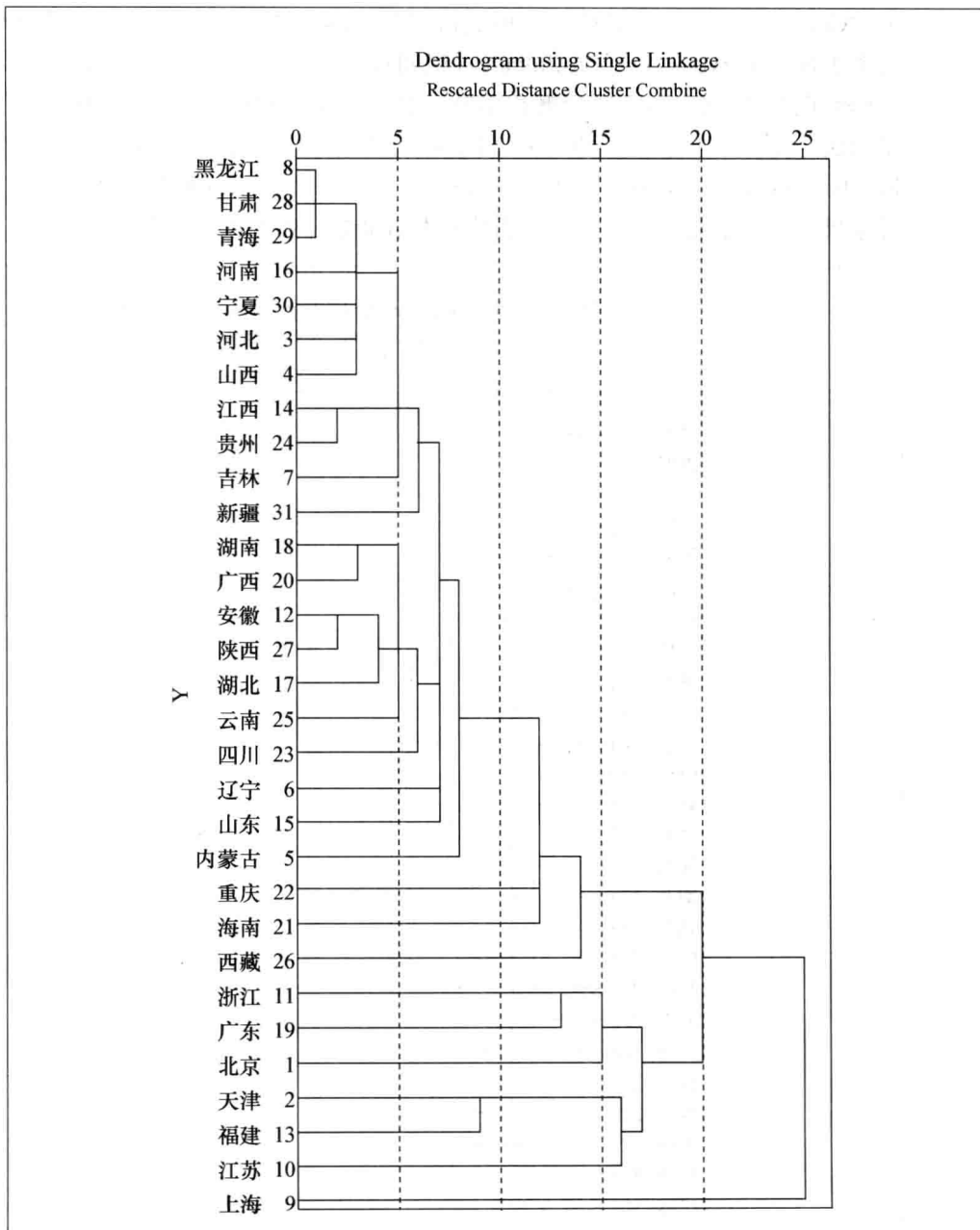


图 3—19 最短距离法谱系图

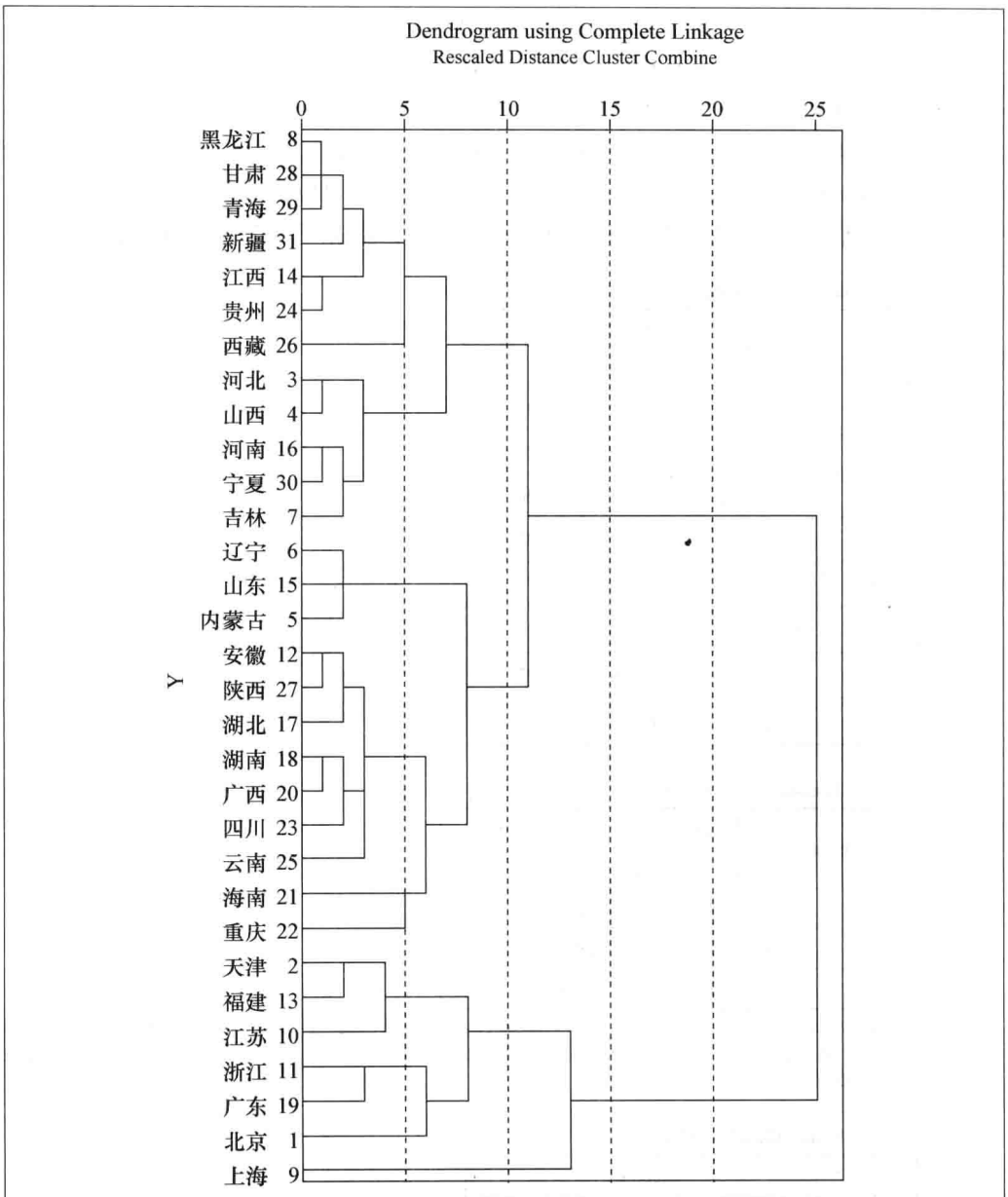


图 3—20　最长距离法谱系图

从表 3—17 和图 3—18、图 3—19 以及图 3—20 可以看出，三种方法得到的结果是一致的，即 {9} 为一类，{1，2，10，11，13，19} 为一类，其余的为一类。

更多的聚类分析方法请参见参考文献 [1]。

例 3—6

仍以 2005 年 31 个省、直辖市、自治区的城镇居民月平均消费支出数据为例，在 SPSS 中利用 K-均值法对 31 个省、直辖市、自治区的城镇居民消费水平进行聚类分析。

在 SPSS 中依次点击 Analyze→Classify→K-Means Cluster，打开 K-Means Cluster Analysis 对话框，将八个变量选入 Variable 框中，将表示地区的变量选入 Label Cases by 栏中，将分类数（Fixed Number of clusters）定为 3。在 Save 按钮中可以选择保存样本的聚类结果（Cluster membership）和各样本与各自中心点的距离（Distance from cluster center）；在 Options 按钮中可以选择输出初始类中心点、方差分析表等结果，读者可以根据实际情况来选择。点击“OK”得到聚类结果如下：

Initial Cluster Centers

	Cluster		
	1	2	3
x1	7 317. 42	9 655. 60	3 855. 56
x2	1 634. 2	2 111. 2	1 529. 5
x3	1 753. 86	1 790. 48	1 438. 88
x4	1 254. 7	1 906. 5	832. 5
x5	2 961. 78	4 563. 80	1 672. 29
x6	2 104. 83	3 723. 74	1 506. 20
x7	773. 22	1 016. 65	905. 88
x8	793. 17	1 485. 53	470. 72

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	723. 933	. 000	1 213. 456
2	545. 656	963. 058	191. 805
3	. 000	. 000	. 000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is . 000. The current iteration is 3. The minimum distance between initial centers is 3 441. 903.

Final Cluster Centers

	Cluster		
	1	2	3
x1	7 212. 83	8 957. 02	5 201. 13
x2	2 068. 1	1 815. 9	1 705. 5
x3	1 624. 14	1 945. 12	1 309. 27
x4	1 277. 1	1 686. 8	925. 1
x5	3 092. 15	4 370. 23	1 859. 79
x6	2 600. 00	3 338. 94	1 506. 53
x7	1 229. 27	1 032. 47	1 002. 90
x8	830. 89	1 178. 42	530. 73

Number of Cases in each Cluster

Cluster	1	6.000
	2	2.000
	3	23.000
Valid		31.000
Missing		.000

其中，第一个表显示了3个类的初始类中心情况，可以看出，第二类的各指标值总体上是最优的，其后依次为第一类和第三类。第二个表展示了3个类中心点每次迭代的偏移情况，可知第一次迭代3个类的中心点分别偏移了723.933，0.000，1 213.456，直到第三次迭代3个类的中心点偏移才达到指定判定标准(0)。第三个表展示了3个类的最终类中心情况，总体来看，第二类各指标值仍是最优的。最后一个表给出了各类中的样品数目，第一类包括6个地区，第二类包括2个地区，第三类包括23个地区。如果在操作过程中选择了保存样本的聚类结果，可以返回数据表，看到名为QCL_1的变量，其各值表示对应地区所属的类别：北京、天津、江苏、浙江、福建和重庆第一类；上海和广东为第二类；其他的为第三类。

例3—7

采用例2—1中35家上市公司的2008年年报数据，研究电力、煤气及水的生产和供应业、房地产业和信息技术业三类行业股份公司的资产周转及盈利情况。这里使用聚类分析对这35家公司进行分类。我们首先将原始变量标准化，采用组间的类平均法，距离计算选择平方欧氏距离，Transform下面的Standardize空白框选择Z scores，对样品（Cases）进行聚类。

详细步骤如下：

(1) 打开数据。使用菜单中File→Open命令，然后选中要分析的数据文件。

(2) 在菜单中的选项中选择Analyze→Classify→Hierarchical Cluster命令。

(3) 在Cluster选择Cases（样品聚类或Q型聚类）。

(4) Display下面有两个选项，选择Statistics（统计量）和Plots（输出图形）。

(5) 点击Method，Cluster Method下选择Between-groups linkage；Measure下选择Interval中的Squared Euclidean distance；在Transform Values下选择Standardize中的Z scores。点击Continue。

(6) 在Plots中，选中Dendrogram（谱系聚类图，也称树状聚类图），点击Continue。

选好每个选项后，点“OK”就可以执行了。

本案例的部分操作窗口如图3—21和图3—22所示。

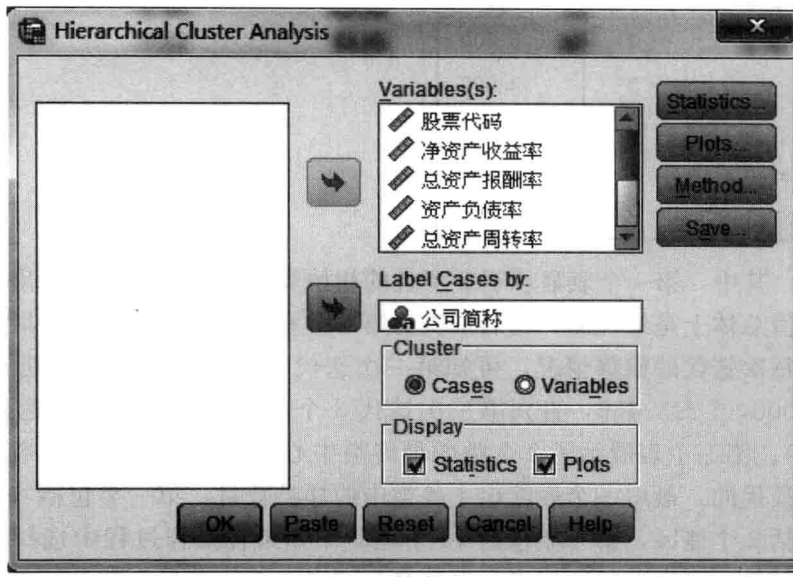


图 3—21 SPSS 中的聚类分析窗口

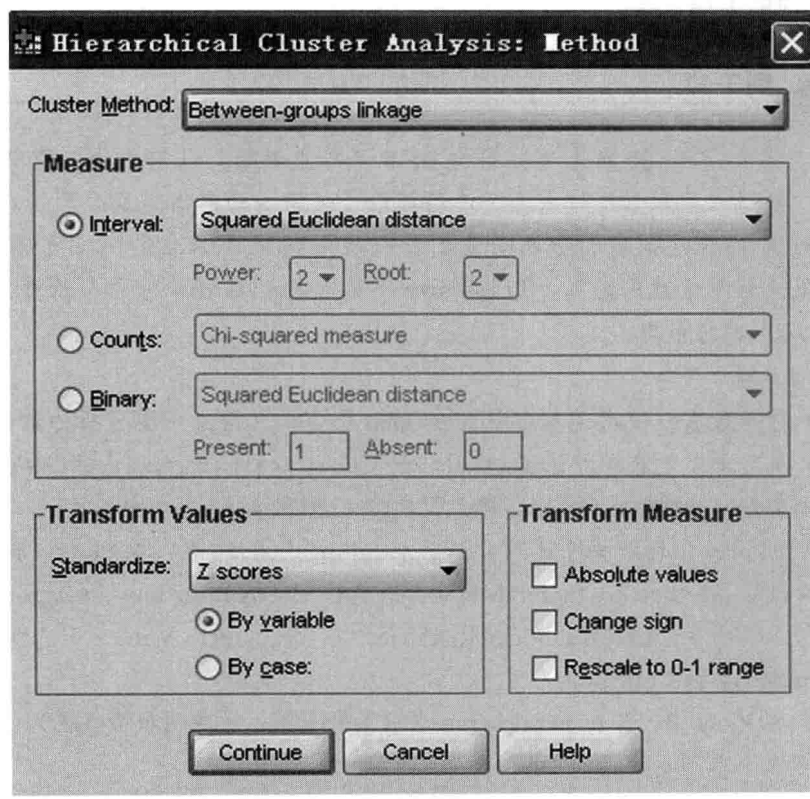


图 3—22 SPSS 中的聚类分析 Method 窗口

输出的聚类分析结果如表 3—18 所示。

表 3—18 **Agglomeration Schedule**

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	6	.773	0	0	5
2	3	21	.987	0	0	30
3	9	11	1.463	0	0	6
4	2	7	1.706	0	0	5
5	2	5	2.274	4	1	14
6	9	31	2.444	3	0	13
7	10	17	2.591	0	0	13
8	24	26	2.749	0	0	19
9	30	35	2.815	0	0	17
10	14	15	3.121	0	0	21
11	16	18	3.173	0	0	20
12	23	27	3.776	0	0	15
13	9	10	4.049	6	7	17
14	2	4	4.578	5	0	19
15	1	23	4.723	0	12	18
16	19	33	5.781	0	0	24
17	9	30	6.177	13	9	23
18	1	20	6.796	15	0	25
19	2	24	6.975	14	8	23
20	16	22	7.941	11	0	25
21	13	14	8.522	0	10	26
22	28	34	9.010	0	0	28
23	2	9	9.055	19	17	27
24	12	19	9.480	0	16	26
25	1	16	11.017	18	20	27
26	12	13	11.490	24	21	29
27	1	2	13.272	25	23	29
28	28	32	14.062	22	0	31
29	1	12	14.423	27	26	30
30	1	3	21.127	29	2	31
31	1	28	22.535	30	28	33
32	8	25	25.769	0	0	33
33	1	8	33.343	31	32	34
34	1	29	38.695	33	0	0

由于接近度矩阵占较多篇幅，这里不给出。根据每阶段聚类结果（见图 3—23）和聚合系数随分类数变化的曲线图（见图 3—24），分为六类以后，聚合系数的变化

趋于逐渐平坦，因此可以结合研究目的，将35家公司分为6类。由图3—23，在距离为10的地方往下切，得到如下的分类结果：1：{南天信息}；2：{城投控股}；3：{国兴地产}；4：{宏图高科，方正科技，长城电脑}；5：{粤宏远A，富龙热力}；其他的公司分到第6类。而从各个公司的实际指标值看，南天信息、城投控股、国兴地产分属3个不同的行业，而且它们在8项经济效益指标上都与其他公司存在显著差异，各自归为一类比较合适。宏图高科、方正科技和长城电脑都属信息技术业，除销售增长率一项外，其他指标值差异不大，划为一类也比较适宜。

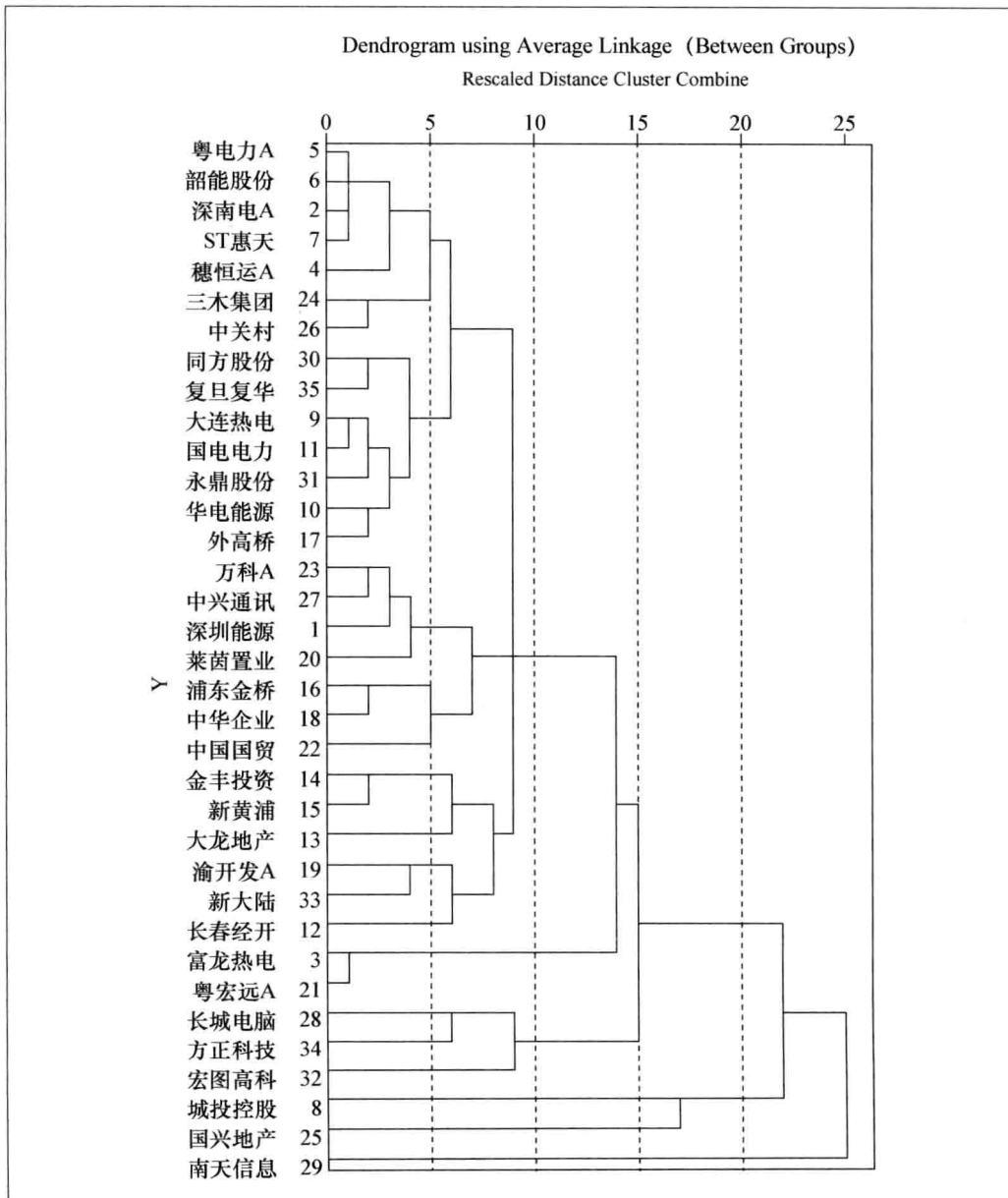


图3—23 树状聚类图

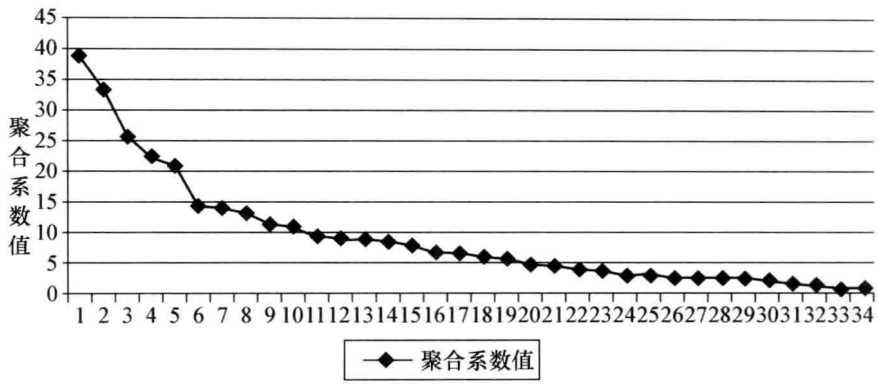


图3—24 聚合系数随分类数变化的曲线图

□ 参考文献

[1] 王国梁，何晓群．多变量经济数据统计分析．西安：陕西科学出版社，1993

[2] 方开泰，潘恩沛．聚类分析．北京：地质出版社，1982

[3] Gordon, A. D. *Classification*. Chapman and Hall, London, 1975

[4] Hartigan, J. A. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, 1975

[5] Ryzin, J. Van [ed]. *Classification and Clustering*. Academic Press, New York, 1977

[6] Arthanavi, T. S. and Yadolah Dodge. *Mathematical Programming in Statistics*. John Wiley & Sons, Inc., New York, 1981

[7] 张尧庭，方开泰．多元统计分析引论．北京：科学出版社，1982

[8] 方开泰．实用多元统计分析．上海：华东师范大学出版社，1989

[9] 王学仁，王松桂．实用多元统计分析．上海：上海科学技术出版社，1990

[10] 中华人民共和国国家统计局．中国统计年鉴（2013）．北京：中国统计出版社，2014

[11] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations, the 5th Berkley Symposium on Mathematics. *Statistics and Probability*, 1967, 1 (1)

□ 思考与练习

1. 聚类分析的基本思想和功能是什么？

2. 试述系统聚类法的原理和具体步骤。
3. 试述 K-均值聚类的方法原理。
4. 试述模糊聚类的思想方法。
5. 试运用 SPSS 软件进行一个实际问题的分类研究。

第4章

Chapter 4 判别分析

学习目标

1. 掌握应该使用线性判别函数而不使用多元回归的情形；
2. 理解判别分析用于实际问题时的基本假定；
3. 掌握判别分析应用时的要点；
4. 描述判别分析的计算方法及其应用场合；
5. 掌握如何解释线性判别函数的性质，即用显著的判别力去判定被解释变量；
6. 掌握如何通过 SPSS 软件实现判别分析。

回归模型普及性的基础在于用它去预测和解释度量（metric）变量。但是，对于非度量（nonmetric）变量，一般的多元回归不适合解决此类问题。本章介绍的判别分析适用于被解释变量是非度量变量的情形。在这种情况下，人们对于预测和解释影响一个对象所属类别的关系感兴趣，比如为什么某人是或者不是消费者，一家公司成功还是破产等。本章的目的主要有两个：（1）介绍判别分析的内在性质、基本原理和应用条件；（2）举例说明这些方法的应用和结果的解释。

判别分析在主要目的是识别一个个体所属类别的情况下有着广泛的应用。潜在的应用包括预测新产品的成功或失败，决定一个学生是否被录取，按职业兴趣对学生分组，确定某人信用风险的种类，或者预测一个公司能否成功。

4.1 判别分析的基本思想

有时会遇到包含属性被解释变量和几个度量解释变量的问题，这时需要选择一种合适的分析方法。比如，我们希望区分好和差的信用风险。如果有信用风险的度

量指标，就可以使用多元回归。但若需要判断某人是在好的或者差的一类，这就不是多元回归分析所要求的度量类型。

当被解释变量是属性变量而解释变量是度量变量时，判别分析是合适的统计分析方法。在很多情况下，被解释变量包含两组或者两类，比如，雄性与雌性、高与低。另外，有多于两组的情况，比如低、中、高的分类。判别分析能够解决两组或者更多组的情况。当包含两组时，称作两组判别分析。当包含三组或者三组以上时，称作多组判别分析（multiple discriminant analysis）。

判别分析最基本的要求是：分组类型在两组以上；每组案例的规模必须至少在一个以上；解释变量必须是可测量的，这样才能够计算其平均值和方差，使其能合理地应用于统计函数。

与其他多元线性统计模型类似，判别分析的假设之一是，每一个判别变量（解释变量）不能是其他判别变量的线性组合。这时，为其他变量线性组合的判别变量不能提供新的信息，更重要的是在这种情况下无法估计判别函数。不仅如此，有时一个判别变量与另外的判别变量高度相关，或与另外的判别变量的线性组合高度相关，虽然能求解，但参数估计的标准误将很大，以至于参数估计统计上不显著。这就是通常所说的多重共线性问题。

判别分析的假设之二是，各组变量的协方差矩阵相等。判别分析最简单和最常用的形式是采用线性判别函数，它们是判别变量的简单线性组合。在各组协方差矩阵相等的假设条件下，可以使用很简单的公式来计算判别函数和进行显著性检验。

判别分析的假设之三是，各判别变量遵从多元正态分布，即每个变量对于所有其他变量的固定值有正态分布。在这种条件下可以精确计算显著性检验值和分组归属的概率。当违背该假设时，计算的概率将非常不准确。

4.2 距离判别

4.2.1 两总体情况

设有两个总体 G_1 和 G_2，$\boldsymbol{x}$ 是一个 p 维样品，若能定义样品到总体 G_1 和 G_2 的距离 $d(\boldsymbol{x}, G_1)$ 和 $d(\boldsymbol{x}, G_2)$，则可用如下的规则进行判别：若样品 $\boldsymbol{x}$ 到总体 G_1 的距离小于到总体 G_2 的距离，则认为样品 $\boldsymbol{x}$ 属于总体 G_1，反之，则认为样品 $\boldsymbol{x}$ 属于总体 G_2；若样品 $\boldsymbol{x}$ 到总体 G_1 和 G_2 的距离相等，则让它待判。这个准则的数学模型可作如下描述：

$$\begin{cases} \boldsymbol{x}\in G_1, & d(\boldsymbol{x},G_1)<d(\boldsymbol{x},G_2) \\ \boldsymbol{x}\in G_2, & d(\boldsymbol{x},G_1)>d(\boldsymbol{x},G_2) \\ \text{待判}, & d(\boldsymbol{x},G_1)=d(\boldsymbol{x},G_2) \end{cases} \tag{4.1}$$

当总体 G_1 和 G_2 为正态总体且协方差相等时，选用马氏距离，即

$$d^2(\boldsymbol{x},G_1)=(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) \tag{4.2}$$

$$d^2(\boldsymbol{x},G_2)=(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \tag{4.3}$$

这里，$\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ 分别为总体 G_1 和 G_2 的均值和协方差阵。当总体不是正态总体时，有时也可以用马氏距离来描述 $\boldsymbol{x}$ 到总体的远近。

若 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$，这时

$$d^2(\boldsymbol{x},G_2)-d^2(\boldsymbol{x},G_1)=2\left(\boldsymbol{x}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

令

$$\bar{\boldsymbol{\mu}}=\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$$

$$\boldsymbol{\alpha}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

$$W(\boldsymbol{x})=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{\mu}})=\boldsymbol{\alpha}'(\boldsymbol{x}-\bar{\boldsymbol{\mu}}) \tag{4.4}$$

于是判别规则可表示为：

$$\begin{cases}\boldsymbol{x}\in G_1, & W(\boldsymbol{x})>0\\ \boldsymbol{x}\in G_2, & W(\boldsymbol{x})<0\\ \text{待判}, & W(\boldsymbol{x})=0\end{cases} \tag{4.5}$$

这个规则取决于 $W(\boldsymbol{x})$ 的值，通常称 $W(\boldsymbol{x})$ 为判别函数，由于它是线性函数，又称为线性判别函数，$\boldsymbol{\alpha}$ 称为判别系数（类似于回归系数）。线性判别函数使用最方便，在实际应用中也最广泛。

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\boldsymbol{\Sigma}$ 未知时，可通过样本来估计。设 $\boldsymbol{x}_1^{(1)}$，…，$\boldsymbol{x}_{n_1^{(1)}}$ 是来自 G_1 的样本，$\boldsymbol{x}_1^{(2)}$，…，$\boldsymbol{x}_{n_2^{(2)}}$ 是来自 G_2 的样本，可以得到以下估计：

$$\hat{\boldsymbol{\mu}}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i^{(1)}=\bar{\boldsymbol{x}}^{(1)}$$

$$\hat{\boldsymbol{\mu}}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{x}_i^{(2)}=\bar{\boldsymbol{x}}^{(2)}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n_1+n_2-2}(\boldsymbol{A}_1+\boldsymbol{A}_2)$$

其中，$\boldsymbol{A}_a=\sum_{j=1}^{n_a}(\boldsymbol{x}_j^{(a)}-\bar{\boldsymbol{x}}^{(a)})(\boldsymbol{x}_j^{(a)}-\bar{\boldsymbol{x}}^{(a)})'$，$a=1,2$。

当两个总体协差阵 $\boldsymbol{\Sigma}_1$ 与 $\boldsymbol{\Sigma}_2$ 不等时，可用

$$W(\boldsymbol{x})=d^2(\boldsymbol{x},G_2)-d^2(\boldsymbol{x},G_1)=(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$$

作为判别函数，这时它是 $\boldsymbol{x}$ 的二次函数。

4.2.2　多总体情况

1. 协差阵相同

设有 k 个总体 G_1，G_2，…，G_k，它们的均值分别是 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，…，$\boldsymbol{\mu}_k$，协差阵均为 $\boldsymbol{\Sigma}$。类似于两总体的讨论，判别函数为：

$$W_{ij}(\boldsymbol{x})=\left(\boldsymbol{x}-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2}\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j),\quad i,j=1,2,\cdots,k$$

相应的判别规则是

$$\begin{cases}\boldsymbol{x}\in G_i, & W_{ij}(\boldsymbol{x})>0,\forall\, j\neq i\\ \text{待判}, & \text{某个 } W_{ij}(\boldsymbol{x})=0\end{cases}$$

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\cdots$，$\boldsymbol{\mu}_k$，$\boldsymbol{\Sigma}$ 未知时，设从 G_a 中抽取的样本为 $\boldsymbol{x}_1^{(a)}$，$\cdots$，$\boldsymbol{x}_{n_a}^{(a)}$（$a=1$，$2$，$\cdots$，$k$），则它们的估计为：

$$\hat{\boldsymbol{\mu}}_a=\bar{\boldsymbol{x}}^{(a)}=\frac{1}{n_a}\sum_{j=1}^{n_a}\boldsymbol{x}_j^{(a)}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n-k}\sum_{1}^{k}\boldsymbol{A}_a$$

式中　$n=n_1+n_2+\cdots+n_k$

$$\boldsymbol{A}_a=\sum_{j=1}^{n_a}(\boldsymbol{x}_j^{(a)}-\bar{\boldsymbol{x}}^{(a)})(\boldsymbol{x}_j^{(a)}-\bar{\boldsymbol{x}}^{(a)})'$$

2. 协差阵不相同

这时判别函数为：

$$V_{ij}(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-(\boldsymbol{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)$$

判别规则为：

$$\begin{cases}\boldsymbol{x}\in G_i, & V_{ij}(\boldsymbol{x})<0,\forall\, j\neq i\\ \text{待判}, & \text{某个 } V_{ij}=0\end{cases}$$

当 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$\cdots$，$\boldsymbol{\mu}_k$，$\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$，$\cdots$，$\boldsymbol{\Sigma}_k$ 未知时，$\hat{\boldsymbol{\mu}}_a$的估计与协差阵相同时的估计一致，而

$$\hat{\boldsymbol{\Sigma}}_a=\frac{1}{n_a-1}\boldsymbol{A}_a,\quad a=1,2,\cdots,k$$

式中，$\boldsymbol{A}_a$与协差阵相同时的估计一致。

线性判别函数容易计算，二次判别函数计算比较复杂，为此需要一些计算方法。因 $\boldsymbol{\Sigma}_i>\boldsymbol{0}$，存在唯一的下三角阵 $\boldsymbol{V}_i$，其对角线元素均为正，使得

$$\boldsymbol{\Sigma}_i=\boldsymbol{V}_i\boldsymbol{V}_i'$$

从而

$$\boldsymbol{\Sigma}_i^{-1}=(\boldsymbol{V}_i')^{-1}\boldsymbol{V}_i^{-1}=\boldsymbol{L}_i'\boldsymbol{L}_i$$

$\boldsymbol{L}_i$ 仍为下三角阵。我们可事先将 $\boldsymbol{L}_1$，$\boldsymbol{L}_2$，$\cdots$，$\boldsymbol{L}_k$ 算出。令 $\boldsymbol{Z}_i=\boldsymbol{L}_i(\boldsymbol{x}-\boldsymbol{\mu}_i)$，则

$$d^2(\boldsymbol{x},G_i)=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{L}_i'\boldsymbol{L}_i(\boldsymbol{x}-\boldsymbol{\mu}_i)=\boldsymbol{Z}_i'\boldsymbol{Z}_i$$

用这样的方法计算就比较方便。

4.3　贝叶斯判别

贝叶斯（Bayes）统计的思想是：假定对研究的对象已有一定的认识，常用先验概率分布来描述这种认识，然后我们取得一个样本，用样本来修正已有的认识（先验概率分布），得到后验概率分布，各种统计推断都通过后验概率分布来进行。将贝叶斯思想用于判别分析，就得到贝叶斯判别。

设有 k 个总体 G_1，G_2，…，G_k，分别具有 p 维密度函数 $p_1(\boldsymbol{x})$，$p_2(\boldsymbol{x})$，…，$p_k(\boldsymbol{x})$，已知出现这 k 个总体的先验分布为 q_1，q_2，…，q_k，我们希望建立判别函数和判别规则。

用 D_1，D_2，…，D_k 表示 R^p 的一个划分，即 D_1，D_2，…，D_k 互不相交，且 $D_1 \cup \cdots \cup D_k = R^p$。如果这个划分取得适当，正好对应于 k 个总体，这时判别规则可以表示为：

$$\boldsymbol{x} \in G_i, \quad \boldsymbol{x} \text{ 落入 } D_i, \quad i=1,2,\cdots,k$$

问题是如何获得这个划分。用 $c(j \mid i)$ 表示样品来自 G_i 而误判为 G_j 的损失，这一误判的概率为：

$$p(j \mid i) = \int_{D_j} p_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

于是有以上判别规则，所带来的平均损失 ECM（expected cost of misclassification）为：

$$ECM(D_1, D_2, \cdots, D_k) = \sum_{i=1}^{k} q_i \sum_{j=1}^{k} c(j \mid i) p(j \mid i)$$

我们总是定义 $c(i \mid i)=0$，目的是求 D_1，D_2，…，D_k，使 ECM 达到最小。

关于贝叶斯判别具体的性质、详细的数学证明及推导见参考文献 [2]。

4.4　费歇判别

费歇判别的思想是投影，将 k 组 p 维数据投影到某一个方向，使得组与组之间的投影尽可能地分开。如何衡量组与组之间尽可能地分开呢？他借用了一元方差分析的思想。

设从 k 个总体分别取得 k 组 p 维观察值如下：

$$\begin{aligned} G_1 &: \boldsymbol{x}_1^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)} \\ &\vdots \qquad\qquad n = n_1 + n_2 + \cdots + n_k \\ G_k &: \boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_{n_k}^{(k)} \end{aligned}$$

令 $\boldsymbol{a}$ 为 R^p 中的任一向量，$u(\boldsymbol{x})=\boldsymbol{a}'\boldsymbol{x}$ 为 $\boldsymbol{x}$ 向以 $\boldsymbol{a}$ 为法线方向的投影，这时，上述数据的投影为：

$$\begin{aligned} G_1 &: \boldsymbol{a}'\boldsymbol{x}_1^{(1)},\cdots,\boldsymbol{a}'\boldsymbol{x}_{n_1}^{(1)} \\ &\vdots \\ G_k &: \boldsymbol{a}'\boldsymbol{x}_1^{(k)},\cdots,\boldsymbol{a}'\boldsymbol{x}_{n_k}^{(k)} \end{aligned}$$

它正好组成一元方差分析的数据，其组间平方和为：

$$SSG=\sum_{i=1}^{k}n_i(\boldsymbol{a}'\bar{\boldsymbol{x}}^{(i)}-\boldsymbol{a}'\bar{\boldsymbol{x}})^2=\boldsymbol{a}'\Big[\sum_{i=1}^{k}n_i(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}})'\Big]\boldsymbol{a}=\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}$$

式中，$\boldsymbol{B}=\sum_{i=1}^{k}n_i(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}})'$，$\bar{\boldsymbol{x}}^{(i)}$ 和 $\bar{\boldsymbol{x}}$ 分别为第 i 组均值和总均值向量。

组内平方和为：

$$\begin{aligned} SSE&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{a}'\boldsymbol{x}_j^{(i)}-\boldsymbol{a}'\bar{\boldsymbol{x}}^{(i)})^2 \\ &=\boldsymbol{a}'\Big[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})'\Big]\boldsymbol{a}=\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a} \end{aligned}$$

式中，$\boldsymbol{E}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})'$。如果 k 组均值有显著差异，则

$$F=\frac{SSG/(k-1)}{SSE/(n-k)}=\frac{n-k}{k-1}\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

应充分地大，或者

$$\Delta(\boldsymbol{a})=\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

应充分地大。所以我们可以求 $\boldsymbol{a}$，使得 $\Delta(\boldsymbol{a})$ 达到最大。显然，这个 $\boldsymbol{a}$ 并不唯一，因为如果 $\boldsymbol{a}$ 使 $\Delta(\cdot)$ 达到极大，则 $c\boldsymbol{a}$ 也使 $\Delta(\cdot)$ 达到极大，c 为任意不等于零的实数。由矩阵知识，我们知道 $\Delta(\cdot)$ 的极大值为 λ_1，它是 $|\boldsymbol{B}-\lambda\boldsymbol{E}|=0$ 的最大特征根，$\boldsymbol{l}_1$，$\boldsymbol{l}_2$，…，$\boldsymbol{l}_r$ 为相应的特征向量，当 $\boldsymbol{a}=\boldsymbol{l}_1$ 时，可使 $\Delta(\cdot)$ 达到最大。由于 $\Delta(\boldsymbol{a})$ 的大小可衡量判别函数 $u(\boldsymbol{x})=\boldsymbol{a}'\boldsymbol{x}$ 的效果，故称 $\Delta(\boldsymbol{a})$ 为判别效率。综上所述，得到如下的定理。

定理 4.1 费歇准则下的线性判别函数 $u(\boldsymbol{x})=\boldsymbol{a}'\boldsymbol{x}$ 的解 $\boldsymbol{a}$ 为方程 $|\boldsymbol{B}-\lambda\boldsymbol{E}|=0$ 的最大特征根 λ_1 所对应的特征向量 $\boldsymbol{l}_1$，且相应的判别效率为 $\Delta(\boldsymbol{l}_1)=\lambda_1$。

在有些问题中，仅用一个线性判别函数不能很好地区分各个总体，可取 λ_2 对应的特征向量 $\boldsymbol{l}_2$，建立第二个判别函数 $\boldsymbol{l}_2'\boldsymbol{x}$。如还不够，可建立第三个线性判别函数 $\boldsymbol{l}_3'\boldsymbol{x}$，依此类推。

迄今为止，我们仅仅给出了费歇准则下的判别函数，没有给出判别规则。前面曾讲过，在费歇准则下的判别函数并不唯一，若 $u(\boldsymbol{x})=\boldsymbol{l}'\boldsymbol{x}$ 为判别函数，则

$au(\boldsymbol{x})+\beta$ 为与 $u(\boldsymbol{x})$ 具有相同判别效率的判别函数。不唯一性对于制定判别规则并没有妨碍，我们可从中任取一个。一旦取定了判别函数，根据它就可以确定判别规则。

关于费歇判别具体的性质、详细的数学证明及推导见参考文献［2］。

4.5　逐步判别

在多元回归中，变量选择的好坏直接影响回归的效果，而在判别分析中也有类似的问题。如果在某个判别问题中，将其中最主要的指标忽略了，由此建立的判别函数效果一定不好。但是，在许多问题中，事先并不十分清楚哪些指标是主要的。这时，是否将有关的指标尽量收集加入计算才好呢？理论和实践证明，指标太多，不仅带来大量的计算，同时许多对判别无作用的指标反而会干扰我们的视线。因此，适当筛选变量就成为一件很重要的事情。凡具有筛选变量能力的判别方法统称为逐步判别法。和通常的判别分析一样，逐步判别也有许多不同的原则，从而产生各种方法。有关逐步判别法的理论基础详见参考文献［1］、［2］所讨论指标的附加信息检验。

逐步判别的原则为：

（1）在 x_1，x_2，…，x_m（即 m 个自变量）中先选出一个自变量，它使维尔克斯统计量 $\Lambda_i(i=1, 2, \cdots, m)$ 达到最小。为了叙述的方便，又不失一般性，假定挑选的变量次序是按自然的次序，即第 r 步正好选中 x_r，第一步选中 x_1，则有 $\Lambda_1=\min\limits_{1\leqslant i\leqslant m}\{\Lambda_i\}$，并考察 Λ_1 是否落入接受域，如果不显著，则表明一个变量也选不中，不能用判别分析；如果显著，则进入下一步。

（2）在未选中的变量中，计算它们与已选中的变量 x_1 配合的 Λ 值。选择使 $\Lambda_{1i}(2\leqslant i\leqslant m)$ 达到最小的作为第二个变量。仿此，如已选入了 r 个变量，不妨设为 x_1，x_2，…，x_r，则在未选中的变量中逐次选一个与它们配合，计算 $\Lambda_{1,2,\cdots,r,l}(r<l\leqslant m)$，选择使其达到极小的变量作为第 $r+1$ 个变量，并检验新选的第 $r+1$ 个变量能否提供附加信息，如果不能则转入（4），否则转入（3）。

（3）在已选入的 r 个变量中，要考虑较早选的变量中其重要性有没有较大的变化，应及时把不能提供附加信息的变量剔除出去。剔除的原则等同于引进的原则。例如在已进入的 r 个变量中要考察 $x_l(1\leqslant l\leqslant r)$ 是否应剔除，就是计算 $\Lambda_{l\cdot 1,\cdots,l-1,l+1,\cdots,r}$，选择达到极小（大）的 l，看是否显著，如不显著将该变量剔除，仍回到（3），继续考察余下的变量是否需要剔除，如显著则回到（2）。

（4）这时既不能选进新变量，又不能剔除已选进的变量，利用已选中的变量建立判别函数。

有关逐步判别的计算方法和案例参见参考文献［1］、［2］。

4.6 判别分析应用的几个例子

判别分析的逻辑步骤框图如图 4—1 所示。

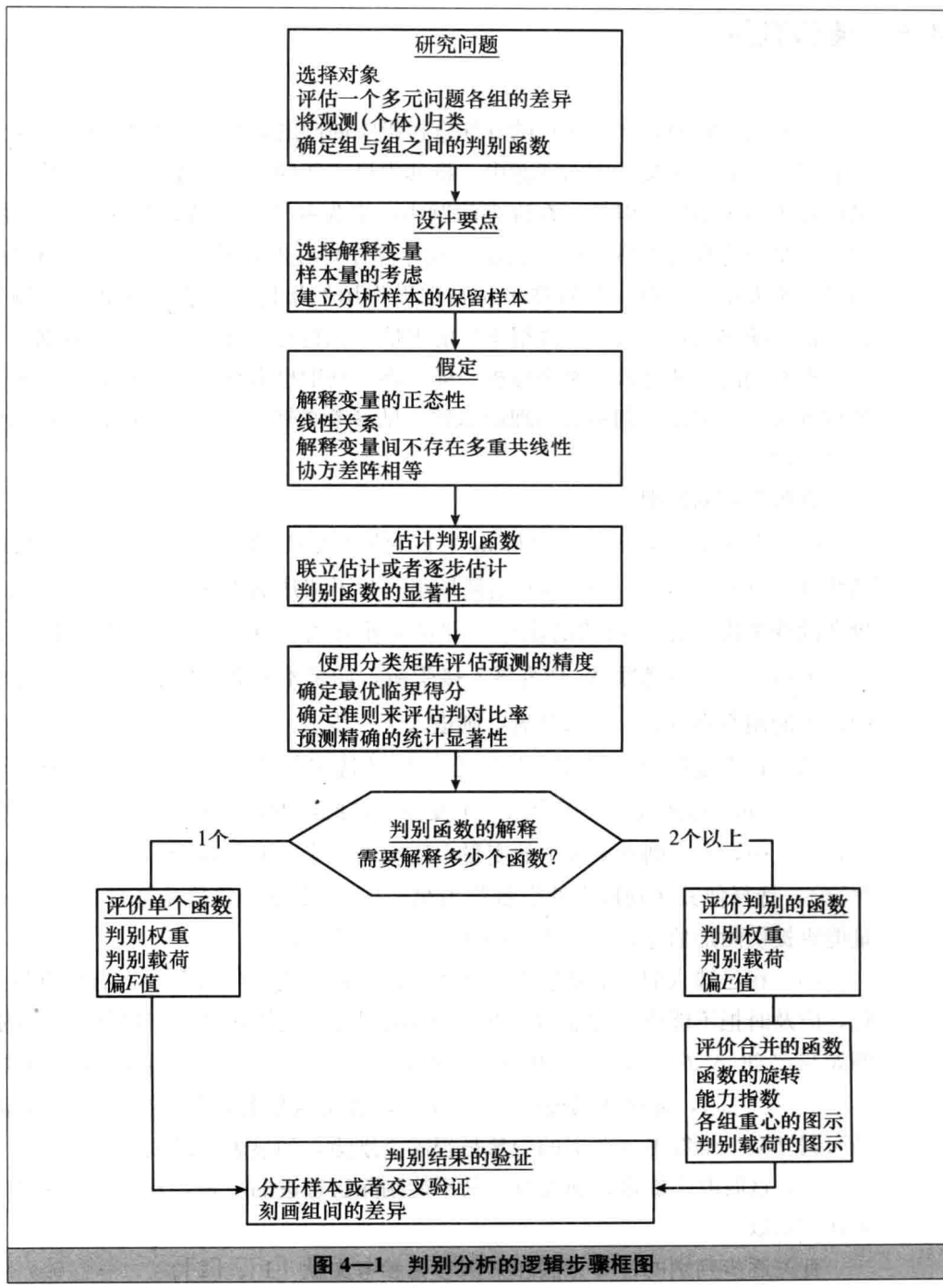


图 4—1 判别分析的逻辑步骤框图

有关判别分析逻辑步骤框图的具体解释参见参考文献［12］。

下面用 SPSS 软件中的 Discriminant 模块来实现判别分析。

例 4—1

判别分析的一个重要应用是动植物的分类，最著名的一个例子是 1936 年费歇的鸢尾花数据（Iris Data）（参见参考文献［13］）。鸢尾花为法国的国花，Setosa，Versicolor，Virginica 是三种有名的鸢尾花，其萼片是绚丽多彩的，和向上的花瓣不同，花萼是下垂的。这三种鸢尾花很像，人们试图建立模型，根据萼片和花瓣的四个度量来对鸢尾花分类。该数据给出 150 朵鸢尾花的萼片长（sepal length）、萼片宽（sepal width）、花瓣长（petal length）、花瓣宽（petal width）以及这些花分别属于的种类（Species）共五个变量。萼片和花瓣的长宽为四个定量变量，而种类为分类变量（取三个值：Setosa，Versicolor，Virginica）。这里三种鸢尾花各有 50 个观测值。

定义新的变量 y 为被解释变量，用“1”代表 Setosa 鸢尾花，用“2”代表 Versicolor 鸢尾花，用“3”代表 Virginica 鸢尾花，将萼片长（sepal length）、萼片宽（sepal width）、花瓣长（petal length）和花瓣宽（petal width）四个变量作为解释变量。

使用 SPSS 软件中的 Analyze→Classify→Discriminant，就进入了判别分析的对话框。分组变量（Grouping Variable）选择 y，然后定义其区域，最小值是 1，最大值是 3。解释变量（Independents）选择 sepal. length，sepal. width，petal. length 和 petal. width。

统计量（Statistics）选项中选择描述统计量 Means，Univariate ANOVAs 和 Box's M，函数选择 Fisher 和非标准化函数，矩阵选择 Within-groups correlation。分类（Classify）选项中选择先验概率（所有组相等或根据组的大小计算概率），因为三个品种都是 50 朵，因此两种选择的效果一样，子选项显示（Display）中选择每个个体的结果（Casewise results），综合表（Summary table）和“留一个在外”（Leave-one-out classification）的验证原则，协方差矩阵选择 Within-groups，作图选择 Combined-groups。

保存（Save）选项中可以选择预测的分类、判别得分以及所属类别的概率。如果采用逐步判别法，我们还可以选择判别的方法（Method）。得到分析结果如下（见输出结果 4—1）。

输出结果 4—1　　**Discriminant**

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		150	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		150	100.0

Group Statistics

品种		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Setosa	萼片长	5.006	.352 5	50	50.000
	萼片宽	3.428	.379 1	50	50.000
	花瓣长	1.462	.173 7	50	50.000
	花瓣宽	.246	.105 4	50	50.000
Versicolor	萼片长	5.936	.516 2	50	50.000
	萼片宽	2.770	.313 8	50	50.000
	花瓣长	4.260	.469 9	50	50.000
	花瓣宽	1.326	.197 8	50	50.000
Virginica	萼片长	6.588	.635 9	50	50.000
	萼片宽	2.974	.322 5	50	50.000
	花瓣长	5.552	.551 9	50	50.000
	花瓣宽	2.026	.274 7	50	50.000
Total	萼片长	5.843	.828 1	150	150.000
	萼片宽	3.057	.435 9	150	150.000
	花瓣长	3.758	1.765 3	150	150.000
	花瓣宽	1.199	.762 2	150	150.000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
萼片长	.381	119.265	2	147	.000
萼片宽	.599	49.160	2	147	.000
花瓣长	.059	1 180.161	2	147	.000
花瓣宽	.071	960.007	2	147	.000

输出结果 4—1 分析的是各组的描述统计量和对各组均值是否相等的检验。第 1 张表反映的是有效样本量及变量缺失的情况。第 2 张表是各组变量的描述统计分析。第 3 张表是对各组均值是否相等的检验。由第 3 张表可以看出，在 0.01 的显著性水平上我们拒绝变量萼片长（sepal length）、萼片宽（sepal width）、花瓣长（petal length）和花瓣宽（petal width）在三组的均值相等的假设，即认为变量萼片长（sepal length）、萼片宽（sepal width）、花瓣长（petal length）和花瓣宽（petal width）在三组的均值是有显著差异的。

输出结果 4—2 是对各组协方差矩阵是否相等的 Box's M 检验。第 1 张表反映协方差矩阵的秩和行列式的对数值。由行列式值可以看出，协方差矩阵不是病态矩阵。第 2 张表是对各总体协方差阵是否相等的统计检验。由 F 值及其显著水平，我们在 0.05 的显著性水平下拒绝原假设（原假设假定各总体协方差阵相等）。因此，在分类（Classify）选项中的协方差矩阵选择可以考虑采用 Separate-groups，以检验采用 Within-groups 和 Separate-groups 两种协方差所得出的结果是否存在显著差异。如果存在显著差异，就应该采用 Separate-groups 协方差矩阵；反之，就用 Within-groups 协方差矩阵。

输出结果 4—2　Box's Test of Equality of Covariance Matrices

Log Determinants

品种	Rank	Log Determinant
Setosa	4	-13.067
Versicolor	4	-10.874
Virginica	4	-8.927
Pooled within-groups	4	-9.959

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		146.663
F	Approx.	7.045
	df1	20
	df2	77 566.751
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

输出结果 4—3 分析的是典型判别函数。第 1 张表反映判别函数的特征根、解释方差的比例和典型相关系数。第一判别函数解释了 99.1%的方差，第二判别函数解释了 0.9%的方差，两个判别函数解释了全部方差。第 2 张表是对两个判别函数的显著性检验。由 Wilks' Lambda 检验，认为两个判别函数在 0.05 的显著性水平上是显著的。

输出结果 4—3　Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	32.192[a]	99.1	99.1	.985
2	.285[a]	.9	100.0	.471

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function (s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.023	546.115	8	.000
2	.778	36.530	3	.000

输出结果 4—4 显示的是判别函数、判别载荷和各组的重心。第 1 张表是标准化的判别函数，表示为：

$$y_1 = -0.427 Sepal.Length^* - 0.521 Sepal.Width^* + 0.947 Petal.Length^* + 0.575 Petal.Width^*$$

$$y_2 = 0.012Sepal.Length^* + 0.735Sepal.Width^* - 0.401Petal.Length^* + 0.581Petal.Width^*$$

这里 * 表示标准化变量，标准化变量的系数也就是前面讲的判别权重。第 2 张表是结构矩阵，即判别载荷。由判别权重和判别载荷可以看出，哪些解释变量对判别函数的贡献较大。第 3 张表是非标准化的判别函数，表示为：

$$y_1 = -2.105 - 0.829Sepal.Length - 1.534Sepal.Width + 2.201Petal.Length + 2.810Petal.Width$$

$$y_2 = -6.661 + 0.024Sepal.Length + 2.165Sepal.Width - 0.932Petal.Length + 2.839Petal.Width$$

我们可以根据这个判别函数计算每个观测的判别 Z 得分。第 4 张表是反映判别函数在各组的重心。根据结果，判别函数在 $y=1$ 这一组的重心为（−7.608，0.215），在 $y=2$ 这一组的重心为（1.825，−0.728），在 $y=3$ 这一组的重心为（5.783，0.513）。这样，我们就可以根据每个观测的判别 Z 得分对观测进行分类。

输出结果 4—4

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
萼片长	−.427	.012
萼片宽	−.521	.735
花瓣长	.947	−.401
花瓣宽	.575	.581

Structure Matrix

	Function	
	1	2
花瓣长	.706*	.168
萼片宽	−.119	.864*
花瓣宽	.633	.737*
萼片长	.223	.311*

Pooled within groups correlations between discriminating variables and standardized canonical discriminant functions.

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function.

Canonical Discriminant Function Coefficients

	Function	
	1	2
萼片长	−.829	.024
萼片宽	−1.534	2.165
花瓣长	2.201	−.932
花瓣宽	2.810	2.839
(Constant)	−2.105	−6.661

Unstandardized coefficients.

Functions at Group Centroids

品种	Function	
	1	2
Setosa	-7.608	.215
Versicolor	1.825	-.728
Virginica	5.783	.513

Unstandardized canonical discriminant functions evaluated at group means.

输出结果 4—5 是分类的统计结果。第 1 张表概括了分类过程，说明 150 个观测都参与分类。第 2 张表说明各组的先验概率，我们在 Classify 选项中选择的是所有组的先验概率相等。第 3 张表是每组的分类函数（区别于判别函数），也称费歇线性判别函数，由表中的结果可以说明：$y=1$ 这一组的分类函数是

$$f_1=-86.308+23.544Sepal.Length+23.588Sepal.Width-16.431Petal.Length-17.398Petal.Width$$

$y=2$ 这组的分类函数是

$$f_2=-72.853+15.698Sepal.Length+7.073Sepal.Width+5.211Petal.Length+6.434Petal.Width$$

$y=3$ 这组的分类函数是

$$f_3=-104.368+12.446Sepal.Length+3.685Sepal.Width+12.767Petal.Length+21.079Petal.Width$$

我们可以计算出每个观测在各组的分类函数值，然后将观测分类到较大的分类函数值中。第 4 张表是分类矩阵表。Predicted Group Membership 表示预测的所属组关系，Original 表示原始数据的所属组关系，Cross-validated 表示交叉验证的所属组关系，这里交叉验证是采用“留一个在外”的原则，即每个观测是通过除了这个观测以外的其他观测推导出来的判别函数来分类的。由第 4 张表可以看出，通过判别函数预测，有 147 个观测是分类正确的，其中，$y=1$ 组 50 个观测全部被判对，$y=2$ 组 50 个观测中有 48 个观测被判对，$y=3$ 组 50 个观测中有 49 个观测被判对，从而有 147/150=98%的原始观测被判对。在交叉验证中，$y=1$ 组 50 个观测全部被判对，$y=2$ 组 50 个观测中有 48 个观测被判对，$y=3$ 组 50 个观测中有 49 个观测被判对，从而交叉验证有 147/150=98%的原始观测被判对。还可以通过分类结果分析判对和判错的百分比。最后为分类结果图，从图中可以看到，Setosa 鸢尾花与 Versicolor 鸢尾花和 Virginica 鸢尾花可以很清晰地区分开，而 Versicolor 鸢尾花和 Virginica 鸢尾花这两种之间存在重合区域，即存在误判。

输出结果 4—5　　Classification Statistics

Classification Processing Summary

Processed		150
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		150

(1)

Prior Probabilities for Groups

品种	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Setosa	.333	50	50.000
Versicolor	.333	50	50.000
Virginica	.333	50	50.000
Total	1.000	150	150.000

(2)

Classification Function Coefficients

	品种		
	Setosa	Versicolor	Virginica
萼片长	23.544	15.698	12.446
萼片宽	23.588	7.073	3.685
花瓣长	−16.431	5.211	12.767
花瓣宽	−17.398	6.434	21.079
(Constant)	−86.308	−72.853	−104.368

Fisher's linear discriminant functions

(3)

Classification Results[b,c]

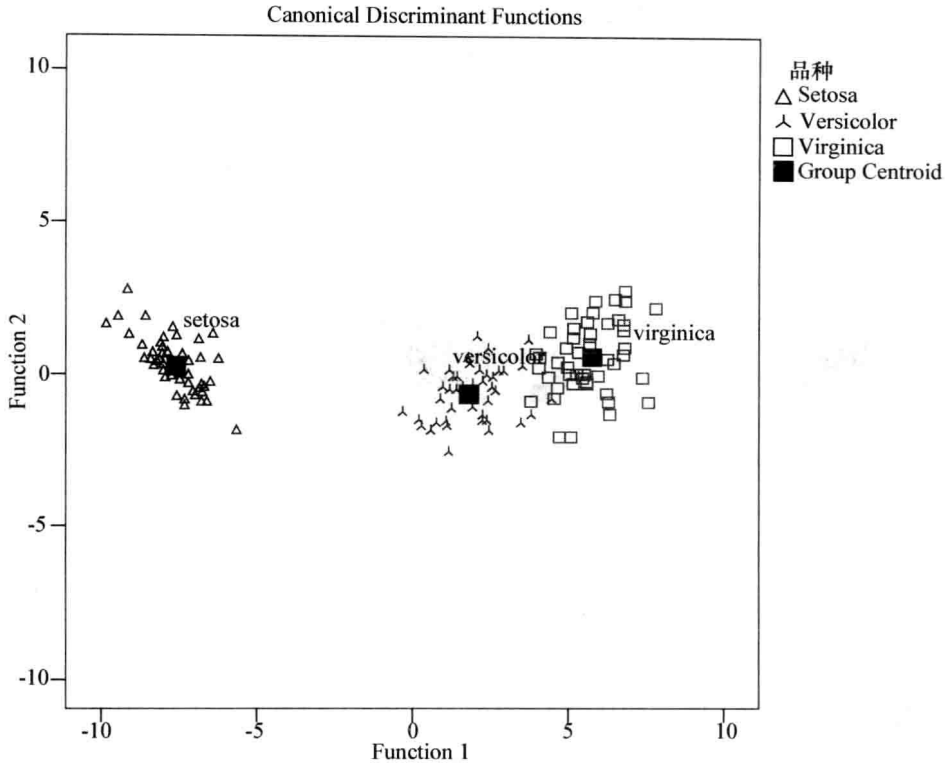
		品种	Predicted Group Membership			Total
			Setosa	Versicolor	Virginica	
Original	Count	Setosa	50	0	0	50
		Versicolor	0	48	2	50
		Virginica	0	1	49	50
	%	Setosa	100.0	.0	.0	100.0
		Versicolor	.0	96.0	4.0	100.0
		Virginica	.0	2.0	98.0	100.0
Cross-validated[a]	Count	Setosa	50	0	0	50
		Versicolor	0	48	2	50
		Virginica	0	1	49	50
	%	Setosa	100.0	.0	.0	100.0
		Versicolor	.0	96.0	4.0	100.0
		Virginica	.0	2.0	98.0	100.0

(4)

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 98.0% of original grouped cases correctly classified.

c. 98.0% of cross-validated grouped cases correctly classified.



(5)

我们还可以通过保存（Save）选项选择预测的类别关系和判别得分等，对观测进行诊断。

由前面分析发现，协方差矩阵不等，可以考虑采用 Separate-groups 协方差矩阵。选择 Separate-groups 协方差矩阵，其他选择同上，得到分类结果如下（见输出结果 4—6）。

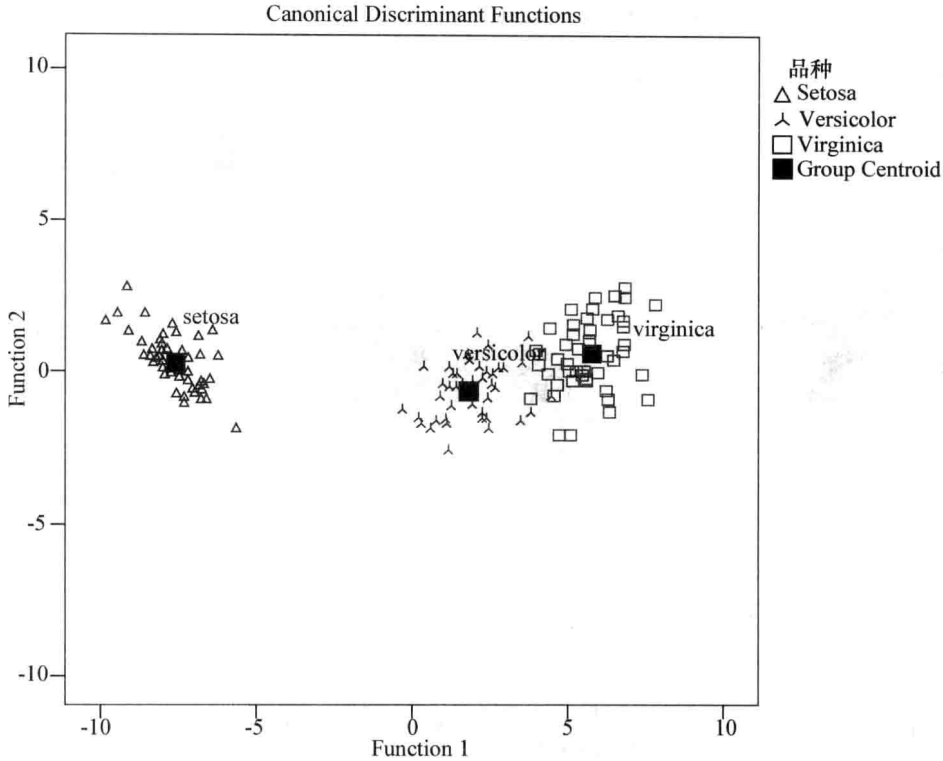
输出结果 4—6

Classification Results[a]

		品种	Predicted Group Membership			Total
			Setosa	Versicolor	Virginica	
Original	Count	Setosa	50	0	0	50
		Versicolor	0	47	3	50
		Virginica	0	1	49	50
	%	Setosa	100.0	.0	.0	100.0
		Versicolor	.0	94.0	6.0	100.0
		Virginica	.0	2.0	98.0	100.0

a. 97.3% of original grouped cases correctly classified.

(1)



(2)

由输出结果 4—6 的表（1）可以看出，通过判别函数预测，有 146 个观测是分类正确的，其中，$y=1$ 组 50 个观测全部被判对，$y=2$ 组 50 个观测中有 47 个观测被判对，$y=3$ 组 50 个观测中有 49 个观测被判对，从而有 146/150＝97.3%的原始观测被判对。输出结果 4—6（2）为分类结果图，从图中可以看到，Setosa 鸢尾花与 Versicolor 鸢尾花和 Virginica 鸢尾花可以很清晰地区分开，而 Versicolor 鸢尾花和 Virginica 鸢尾花这两种之间存在重合区域，即存在误判。

由输出结果 4—6 可以看出，采用 Separate-groups 协方差矩阵与采用 Within-groups 协方差矩阵的预测效果没有明显的差别，因此，可以采用 Within-groups 协方差矩阵来进行判别。

例 4—2

距离判别案例。为了研究 2012 年全国各地区农村居民家庭人均消费支出情况，按人均收入、人均 GDP 以及消费支出将 29 个省、直辖市、自治区（除福建和陕西以外）分为三种类型，设置 Group 变量取值分别为 1，2，3。试建立判别函数，判定福建、陕西分别属于哪个消费水平类型。判别指标及原始数据如表 4—1 所示。

表 4—1　　2012 年 31 个省、直辖市、自治区农村居民家庭平均每人生活消费支出　　元/人

x_1　人均食品支出
x_2　人均衣着支出
x_3　人均住房支出
x_4　人均家庭设备及服务支出
x_5　人均交通和通信支出
x_6　人均文教娱乐用品及服务支出
x_7　人均医疗保健支出
x_8　其他商品及服务支出

序号	地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Group
1	上海	4 629.16	704.43	1 827.63	646.13	1 704.83	952.10	1 028.96	253.39	1
2	北京	3 898.95	947.97	2 196.37	773.54	1 398.80	1 152.67	1 125.25	334.49	1
3	广东	3 155.91	319.41	1 107.61	378.45	760.07	466.63	446.46	232.66	1
4	浙江	3 796.89	750.69	1 936.64	602.54	1 499.95	902.23	746.05	251.04	1
5	江苏	2 723.01	610.70	1 477.81	532.76	1 311.05	1 184.18	724.23	232.74	1
6	安徽	1 836.79	331.77	1 081.74	346.90	516.60	385.92	510.06	144.00	2
7	天津	2 988.83	780.72	1 263.51	451.30	1 066.27	766.08	760.41	228.40	2
8	江西	1 580.03	264.88	1 009.66	278.31	494.46	342.70	380.45	105.59	2
9	山东	2 059.81	454.40	1 399.90	405.63	937.55	500.98	635.34	120.20	2
10	湖北	1 532.06	315.21	1 173.49	397.63	496.10	394.63	591.87	169.68	2
11	湖南	1 748.29	317.87	1 069.13	372.96	481.58	400.22	497.24	136.56	2
12	广西	1 373.88	156.47	1 144.20	274.63	453.01	270.24	383.95	108.82	2
13	海南	2 122.36	178.86	775.63	207.47	435.58	253.97	306.54	155.20	2
14	重庆	1 564.04	380.13	550.18	413.52	489.31	394.23	482.24	85.85	2
15	四川	1 653.61	338.52	769.18	333.20	463.94	329.29	498.29	101.90	2
16	贵州	1 045.87	226.79	708.78	211.36	371.35	226.44	282.51	84.30	2
17	云南	1 331.01	241.07	728.32	247.00	470.19	289.22	362.63	65.82	2
18	西藏	930.39	372.54	249.35	173.30	363.95	40.86	82.67	90.52	2
19	河北	1 647.68	396.48	1 115.19	349.84	604.33	358.49	543.75	156.73	3
20	山西	1 657.57	501.74	1 137.44	298.29	625.99	498.02	490.25	149.75	3
21	内蒙古	1 808.01	481.71	1 000.88	268.10	912.25	513.97	588.87	157.42	3
22	辽宁	2 019.41	517.86	928.39	250.52	668.71	556.56	548.77	176.23	3
23	吉林	1 887.33	478.16	745.07	251.93	699.03	606.26	840.52	204.15	3
24	黑龙江	1 990.33	544.64	664.23	228.51	611.34	518.04	727.02	167.67	3
25	河南	1 474.20	424.10	1 035.81	361.58	525.11	343.83	468.81	146.21	3
26	甘肃	1 210.04	303.10	663.71	250.43	436.03	327.30	398.01	100.41	3
27	青海	1 300.42	403.52	1 204.04	256.63	683.73	283.28	520.06	121.62	3
28	宁夏	1 453.62	463.35	1 033.17	304.95	620.79	373.36	492.14	172.21	3
29	新疆	1 391.16	426.60	1 289.31	215.09	646.42	261.74	444.18	110.21	3
1	福建	3 030.19	471.44	1 136.31	426.37	794.98	565.83	380.60	193.07	
2	陕西	1 293.38	332.72	1 253.98	298.69	503.34	445.47	619.94	136.37	

资料来源：中华人民共和国国家统计局：《中国统计年鉴（2013）》，北京，中国统计出版社，2014。

距离判别法无法在SPSS中直接实现，用R软件可以运行，参见参考文献［11］。

解：本例中组数 $k=3$，判别指标 $p=8$，各组中样本为：$n_1=5$，$n_2=13$，$n_3=11$，待判样品个数为2。

总体协方差阵的逆矩阵：

$$\hat{\boldsymbol{\Sigma}}^{-1}=\begin{bmatrix} 1.77 & 0.89 & 0.58 & -0.47 & -3.04 & 2.13 & -1.75 & -11.97 \\ 0.89 & 22.51 & 3.27 & -16.04 & -10.40 & -1.77 & 3.40 & -33.50 \\ 0.58 & 3.27 & 3.17 & -7.28 & -3.97 & 1.59 & 0.35 & -10.70 \\ -0.47 & -16.04 & -7.28 & 52.02 & 9.14 & -2.43 & -12.55 & 28.10 \\ -3.04 & -10.40 & -3.97 & 9.14 & 15.47 & -6.17 & -1.48 & 37.57 \\ 2.13 & -1.77 & 1.59 & -2.43 & -6.17 & 15.28 & -7.83 & -20.53 \\ -1.75 & 3.40 & 0.35 & -12.55 & -1.48 & -7.83 & 18.45 & -10.76 \\ -11.97 & -33.50 & -10.70 & 28.10 & 37.57 & -20.53 & -10.76 & 252.03 \end{bmatrix}\times 10^{-5}$$

将原29个样品的回判结果列于表4—2，两个待判样品的判别结果列于表4—3。福建省应判归第一类消费水平，陕西省归入第三类消费水平为宜。本例的回判准确率高，回判正确率为89.66%，说明各地区农村居民的消费水平划分为三种类型是合适的。由于SPSS中的判别分析没有距离判别这一方法，因此距离判别法无法在SPSS中直接实现，本书是用R软件程序实现的，距离判别程序略。

注：为编写程序方便，在录入数据时，将“地区”指标用其英文“area”代替，用“s”代替“Group”。

表4—2　　已知样品回判结果

地区	原属类号	最小距离及归类		正误判标志（正=0；误=1）
上海	1	12.419 685 03	1	0
北京	1	12.163 187 09	1	0
广东	1	11.387 574 96	1	0
浙江	1	3.503 847 439	1	0
江苏	1	16.463 007 39	1	0
安徽	2	1.145 772 097	2	0
天津	2	13.216 134 32	2	0
江西	2	2.037 410 535	2	0
山东	2	5.629 110 774	2	0
湖北	2	7.797 146 765	2	0
湖南	2	1.858 969 281	2	0
广西	2	4.392 380 106	2	0
海南	2	10.045 045 2	2	0
重庆	2	11.096 830 75	2	0
四川	2	3.032 019 602	2	0
贵州	2	2.691 399 426	2	0
云南	2	3.525 650 047	2	0

续前表

地区	原属类号	最小距离及归类		正误判标志（正=0；误=1）
西藏	2	17.277 214 68	3	1
河北	3	3.680 968 33	3	0
山西	3	4.642 705 205	3	0
内蒙古	3	5.868 870 16	3	0
辽宁	3	5.117 014 206	3	0
吉林	3	12.042 977 85	3	0
黑龙江	3	9.573 854 488	3	0
河南	3	3.374 577 21	2	1
甘肃	3	2.630 002 219	2	1
青海	3	4.984 155 379	3	0
宁夏	3	3.619 441 075	3	0
新疆	3	8.734 159 148	3	0

表 4—3　　未知样品回判结果

地区	最小距离及归类	
福建	14.766 884 91	1
陕西	6.458 671 807	3

这里顺便指出，回判的误判率并不是“误判概率”，而且前者通常要小些，回判情况仅供使用时参考。

例 4—3

2005 年全国城镇居民月平均消费状况可划分为两类，分类后的数据如表 4—4 所示。试建立费歇线性判别函数，并将广东、西藏两个待判省区归类。

表 4—4　　2005 年 31 个省、直辖市、自治区城镇居民月平均消费类型划分数据

x_1 人均粮食支出（元/人）　　x_5 人均衣着支出（元/人）

x_2 人均副食支出（元/人）　　x_6 人均日用杂品支出（元/人）

x_3 人均烟、酒、饮料支出（元/人）　　x_7 人均水电燃料支出（元/人）

x_4 人均其他副食支出（元/人）　　x_8 人均其他非商品支出（元/人）

序号	地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Group
1	北京	21.3	124.89	35.43	73.98	93.01	20.58	43.97	433.73	1
2	上海	21.13	168.69	40.81	70.12	74.32	15.46	50.9	422.74	1
3	浙江	19.96	142.24	43.33	50.74	101.77	12.92	53.44	394.55	1
4	天津	21.5	122.39	29.08	51.64	55.04	11.3	54.88	288.13	2
5	河北	18.25	90.21	24.45	32.44	62.48	7.45	47.5	178.84	2
6	山西	21.84	66.38	18.05	31.32	74.48	8.19	34.97	177.45	2
7	内蒙古	21.37	67.08	20.28	35.27	81.07	10.94	39.46	182.2	2
8	辽宁	22.74	115.88	28.21	42.44	58.07	9.63	48.65	194.85	2

续前表

序号	地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Group
9	吉林	20.22	88.94	18.54	35.63	65.72	8.81	50.29	186.52	2
10	黑龙江	21.33	75.5	14	29.56	69.29	8.24	42.08	165.9	2
11	江苏	18.61	122.51	27.07	42.5	63.47	15.38	36.14	240.92	2
12	安徽	19.61	107.13	32.85	35.77	61.34	7.53	34.6	142.23	2
13	福建	25.56	171.65	22.3	40.53	57.13	12.6	54.03	225.08	2
14	江西	18.75	104.68	15.55	35.61	51.8	11.18	36.27	142.72	2
15	山东	18.27	88.34	19.07	43.19	72.97	12.59	42.16	200.18	2
16	河南	19.07	73.18	18.01	29.38	64.51	8.91	38.14	155.45	2
17	湖北	18.76	102.67	21.87	30.47	64.33	11.99	42.14	168.17	2
18	湖南	20.25	104.45	20.72	38.15	62.98	12.67	39.16	213.56	2
19	广西	18.7	131.35	11.69	32.06	41.54	10.84	42.77	178.51	2
20	海南	16.16	139.92	12.98	23.58	24.87	10.76	32.35	144.21	2
21	重庆	18.18	120.39	26.18	37.94	68.16	11.64	38.48	246.37	2
22	四川	18.53	109.95	21.49	33.04	50.98	10.88	33.96	183.85	2
23	贵州	18.33	92.43	25.38	32.19	56.32	14	38.57	144.82	2
24	云南	22.3	99.08	33.36	32.01	52.06	7.04	32.85	190.04	2
25	陕西	20.03	70.75	19.75	34.95	53.29	10.55	38.2	189.41	2
26	甘肃	18.68	72.74	23.72	38.69	62.41	9.65	35.26	170.12	2
27	青海	20.33	75.64	20.88	33.85	53.81	10.06	32.82	171.32	2
28	宁夏	19.75	70.24	18.67	36.71	61.75	10.08	40.26	165.22	2
29	新疆	21.03	78.55	14.35	34.33	64.98	9.83	33.87	161.67	2
1	广东	23.68	173.30	17.43	43.59	53.66	16.86	65.02	385.94	
2	西藏	29.67	146.90	64.51	54.36	86.10	14.77	32.19	193.10	

解：（1）计算总体 G_1 和 G_2 各判别变量的均值。

$$\bar{\boldsymbol{x}}_1=(20.797,145.273,39.857,64.947,89.700,16.320,49.437,417.007)'$$

$$\bar{\boldsymbol{x}}_2=(19.929,98.540,21.481,35.510,59.802,10.490,39.995,184.913)'$$

$$\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2=(0.868,46.733,18.376,29.437,29.898,5.830,9.442,232.094)'$$

$$\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2=(40.726,243.813,61.337,100.457,149.502,26.810,89.431,601.920)'$$

（2）计算协方差阵 $\boldsymbol{\Sigma}$ 的估计值 $\boldsymbol{S}_p$ 的逆阵。

$$\boldsymbol{S}_p^{-1}=\begin{bmatrix}43.44 & -0.95 & 1.56 & -5.78 & -2.43 & 15.31 & -1.22 & 0.001\\ -0.95 & 0.48 & -0.57 & 0.75 & 0.69 & -2.11 & -0.73 & -0.10\\ 1.56 & -0.57 & 4.91 & -1.94 & -0.94 & 6.99 & 1.37 & -0.15\\ -5.78 & 0.75 & -1.94 & 7.67 & 1.09 & -9.89 & -1.45 & -0.70\\ -2.43 & 0.69 & -0.94 & 1.09 & 1.86 & -2.86 & -1.10 & -0.16\\ 15.31 & -2.11 & 6.99 & -9.89 & -2.86 & 46.08 & 5.74 & -0.31\\ -1.22 & -0.73 & 1.37 & -1.45 & -1.10 & 5.74 & 4.84 & -0.18\\ 0.001 & -0.10 & -0.15 & -0.70 & -0.16 & -0.31 & -0.18 & 0.25\end{bmatrix}\times 10^{-2}$$

（3）计算费歇样本判别函数。

$$y=(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}\boldsymbol{X}$$
$$=-1.431x_1+0.117x_2-0.009x_3+0.196x_4+0.373x_5-0.833x_6$$
$$-0.474x_7+0.215x_8$$

(4) 计算两个一元总体均值的中点 m 的估计值 $\hat{m}$。

$$\hat{m}=\frac{1}{2}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2)=54.946$$

(5) 计算统计检验量 F 值。

马氏距离为：

$$D^2=(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)=61.588$$

F 检验统计量为：

$$F=\left[\frac{n_1+n_2-p-1}{(n_1+n_2-2)\ p}\right]\left[\frac{n_1n_2}{n_1+n_2}\right]D^2=15.335$$

其第一自由度 $p=8$，第二自由度 $n_1+n_2-p-1=20$，查 F 分布表有

$$F>F_{0.05}(8,20)=2.45$$

故在 0.05 水平上两个总体的均值有一定的差异，即判别函数有效。

(6) 回判及待判样品的归类。

1) 计算两个一元总体均值的中点 m 的估计值 $\hat{m}$。

$$\hat{m}=\frac{1}{2}\ (\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}\ (\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2)\ =54.946$$

2) 计算原 29 个样品的线性判别函数值 y_0。

$$y_0=(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p^{-1}\boldsymbol{x}_0$$

判别函数值 y_0 也列于表 4—5，于是费歇判别法则为：

若 $y_0=\hat{\boldsymbol{l}}'\boldsymbol{x}_0\geqslant 54.946$，判 $\boldsymbol{x}_0$ 来自总体 G_1。

若 $y_0=\hat{\boldsymbol{l}}'\boldsymbol{x}_0<54.946$，判 $\boldsymbol{x}_0$ 来自总体 G_2。

对于两个待判省区，判别函数值 y_0 小于 54.946，故都判归低消费总体。将原 29 个省、直辖市、自治区的回判结果也列于表 4—5，此例没有误判，回判准确率很高。

表 4—5　　**样品回判结果**

序号	地区	原属类号	判别函数值及归类		正误判标志（正=0；误=1）
1	北京	1	88.331 210	1	0
2	上海	1	84.521 977	1	0
3	浙江	1	84.366 627	1	0
4	天津	2	40.484 475	2	0
5	河北	2	23.616 179	2	0

续前表

序号	地区	原属类号	判别函数值及归类		正误判标志（正=0；误=1）
6	山西	2	25.037 676	2	0
7	内蒙古	2	25.608 075	2	0
8	辽宁	2	21.555 527	2	0
9	吉林	2	21.732 365	2	0
10	黑龙江	2	18.687 482	2	0
11	江苏	2	41.336 157	2	0
12	安徽	2	21.972 639	2	0
13	福建	2	24.837 676	2	0
14	江西	2	15.755 588	2	0
15	山东	2	32.287 809	2	0
16	河南	2	18.859 534	2	0
17	湖北	2	21.133 793	2	0
18	湖南	2	30.837 323	2	0
19	广西	2	19.352 961	2	0
20	海南	2	13.723 759	2	0
21	重庆	2	45.743 151	2	0
22	四川	2	26.019 176	2	0
23	贵州	2	12.867 934	2	0
24	云南	2	24.506 296	2	0
25	陕西	2	20.008 588	2	0
26	甘肃	2	24.265 621	2	0
27	青海	2	19.187 159	2	0
28	宁夏	2	18.068 700	2	0
29	新疆	2	20.458 099	2	0
1	广东	待判	52.925 987	2	
2	西藏	待判	30.882 179	2	

在SPSS中进行费歇判别分析是十分快捷的。首先按照表4—4把数据输入SPSS数据表中，然后依次点击Analyze→Classify→Discriminant，打开Discriminant Analysis对话框，将对话框左侧变量列表中的group选入Grouping Variable框，并点击“Define Range”，在弹出的Discriminant Analysis：Define Range对话框中，定义判别原始数据的类别区间，本例为两类，故在Minimum处输入1，在Maximum处输入2，点击“Continue”返回Discriminant Analysis对话框。再从对话框左侧的变量列表中将八个变量选入Independents框，作为判别分析的基础数据变量。点击“Statistics”，弹出Discriminant Analysis：Statistics对话框，在Descriptive栏中选Means项，要求对各组的各变量作均值与标准差的描述；在Function Coefficients栏中选Unstandardized项（注意，不是Fisher's项!），要求显示费

歇判别法建立的非标准化系数。之后，点击“Continue”返回 Discriminant Analysis 对话框。点击“Save”，弹出 Discriminant Analysis：Save 对话框，选 Predicted group membership 项要求将回判的结果存入原始数据库中。点击“Continue”返回 Discriminant Analysis 对话框，其他项目不变，点击“OK”即完成分析。在输出结果中，可以看到各组均值、标准差、协方差阵等描述统计结果以及判别函数。返回数据表中，可以看到判别结果已经作为一个新的变量被保存，广东和西藏均被划分到第二大类。受篇幅所限，各输出结果在此不再列示。

在 SPSS 中进行贝叶斯判别分析时，操作步骤与例 4—3 中的费歇判别相同，但是在 Discriminant Analysis：Statistics 对话框的 Function Coefficients 栏中要选 Fisher's 项而不是 Unstandardized 项（因为 Bayes 判别思想是由 Fisher 提出来的，故 SPSS 以此命名）。Save 项还增加 Probabilities of group member 项，点击“OK”后得出分析结果。

逐步判别法也可以在 SPSS 中实现。操作步骤仍与例 4—3 类似，不同之处在于点击 Analyze→Classify→Discriminant，打开 Discriminant Analysis 对话框后，将 Independents 栏下的“Enter independents together”项改选为“Use stepwise method”，此时窗口右侧的“Method”按钮被激活，点击后进入 Discriminant Analysis：Stepwise Method 对话框，在 Method 栏中选中 Mahalanobis distance 项，即采用马氏距离，其他选项保持不变，返回主对话框后，其他操作仍与前面的例子类似。

□ 参考文献

[1] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993

[2] 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982

[3] R. E. Frank，W. E. Massey，and D. G. Morrison. Bias in Multiple Discriminant Analysis. *Journal of Marketing Research*，1965，2 (3)

[4] P. E. Green and J. D. Carroll. *Mathematical Tools for Applied Multivariate Analysis*. New York：Academic Press，1978

[5] W. D. Perreault，D. N. Behrman，and G. M. Armstrong. Alternative Approaches for Interpreting of Multiple Discriminant Analysis in Marketing Research. *Journal of Business Research*，1979 (7)

[6] Guy Gessner，N. K. Maholtra，W. A. Kamakura，and M. Zmijewski. Estimating Models with Binary Dependent Variables：Some Theoretical and Empirical Observations. *Journal of Business Research*，1988 (16)

[7] C. J. Huberty. Issues in the Use and Interpretation of Discriminant Analysis. *Psychological Bulletin*，1984

[8] N. Johnson and D. Wichren. *Applied Multivariate Statistical Analysis*. Upper Saddle River, N. J.: Prentice-Hall, 1982

[9] 王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990

[10] 何晓群，黄旭安，陈少杰. 中国上市公司财务危机的个案诊断. 中国经济评论，2003 (4)

[11] 何晓群. 应用多元统计分析. 北京：中国统计出版社，2010

[12] 何晓群. 多元统计分析（第二版）. 北京：中国人民大学出版社，2008

[13] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*. 1936, 7 (2)

□ 思考与练习

1. 应用判别分析应该具备什么样的条件？
2. 试述贝叶斯判别方法的思路。
3. 试述费歇判别方法的思想。
4. 什么是逐步判别分析？
5. 简要叙述判别分析的步骤及流程。
6. 为研究某地区人口死亡状况，已按某种方法将 15 个已知样品分为 3 类，指标及原始数据如下表所示，试建立判别函数，并判定另外 4 个待判样品属于哪类。

x_1：0 岁组死亡概率　　x_4：55 岁组死亡概率
x_2：1 岁组死亡概率　　x_5：80 岁组死亡概率
x_3：10 岁组死亡概率　　x_6：平均预期寿命

组别	序号	x_1	x_2	x_3	x_4	x_5	x_6
第一组	1	34.16	7.44	1.12	7.87	95.19	69.30
	2	33.06	6.34	1.08	6.77	94.08	69.70
	3	36.26	9.24	1.04	8.97	97.30	68.80
	4	40.17	13.45	1.43	13.88	101.20	66.20
	5	50.06	23.03	2.83	23.74	112.52	63.30
第二组	1	33.24	6.24	1.18	22.90	160.01	65.40
	2	32.22	4.22	1.06	20.70	124.70	68.70
	3	41.15	10.08	2.32	32.84	172.06	65.85
	4	53.04	25.74	4.06	34.87	152.03	63.50
	5	38.03	11.20	6.07	27.84	146.32	66.80
第三组	1	34.03	5.41	0.07	5.20	90.10	69.50
	2	32.11	3.02	0.09	3.14	85.15	70.80
	3	44.12	15.12	1.08	15.15	103.12	64.80
	4	54.17	25.03	2.11	25.15	110.14	63.70
	5	28.07	2.01	0.07	3.02	81.22	68.30
待判样品	1	50.22	6.66	1.08	22.54	170.60	65.20
	2	34.64	7.33	1.11	7.78	95.16	69.30
	3	33.42	6.22	1.12	22.95	160.31	68.30
	4	44.02	15.36	1.07	16.45	105.30	64.20

第5章

Chapter 5 主成分分析

学习目标

1. 理解主成分分析的基本理论与方法；
2. 了解主成分的性质；
3. 理解主成分的求解方法；
4. 掌握用 SPSS 软件求解主成分的方法；
5. 正确理解软件输出结果并对结果进行分析。

主成分分析（principal components analysis）也称主分量分析，是由霍特林于1933年首先提出的。主成分分析是利用降维的思想，在损失很少信息的前提下，把多个指标转化为几个综合指标的多元统计方法。通常把转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能。这样在研究复杂问题时就可以只考虑少数几个主成分而不至于损失太多信息，从而更容易抓住主要矛盾，揭示事物内部变量之间的规律性，同时使问题得到简化，提高分析效率。本章主要介绍主成分分析的基本理论和方法、主成分分析的计算步骤及主成分分析的上机实现。

5.1 主成分分析的基本原理

5.1.1 主成分分析的基本思想

在对某一事物进行实证研究时，为了更全面、准确地反映事物的特征及其发展

规律，人们往往要考虑与其有关系的多个指标，这些指标在多元统计中也称为变量。这样就产生了如下问题：一方面人们为了避免遗漏重要的信息而考虑尽可能多的指标，另一方面考虑指标的增多增加了问题的复杂性，同时由于各指标均是对同一事物的反映，不可避免地造成信息的大量重叠，这种信息的重叠有时甚至会抹杀事物的真正特征与内在规律。基于上述问题，人们就希望在定量研究中涉及的变量较少，而得到的信息量又较多。主成分分析正是研究如何通过原来变量的少数几个线性组合来解释原来变量绝大多数信息的一种多元统计方法。

既然研究某一问题涉及的众多变量之间有一定的相关性，就必然存在着起支配作用的共同因素。根据这一点，通过对原始变量相关矩阵或协方差矩阵内部结构关系的研究，利用原始变量的线性组合形成几个综合指标（主成分），在保留原始变量主要信息的前提下起到降维与简化问题的作用，使得在研究复杂问题时更容易抓住主要矛盾。一般来说，利用主成分分析得到的主成分与原始变量之间有如下基本关系：

（1）每一个主成分都是各原始变量的线性组合。

（2）主成分的数目大大少于原始变量的数目。

（3）主成分保留了原始变量的绝大多数信息。

（4）各主成分之间互不相关。

通过主成分分析，可以从事物之间错综复杂的关系中找出一些主要成分，从而能有效利用大量统计数据进行定量分析，揭示变量之间的内在关系，得到对事物特征及其发展规律的一些深层次的启发，把研究工作引向深入。

5.1.2 主成分分析的基本理论

设对某一事物的研究涉及 p 个指标，分别用 X_1，X_2，…，X_p 表示，这 p 个指标构成的 p 维随机向量为 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$。设随机向量 $\boldsymbol{X}$ 的均值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$。

对 $\boldsymbol{X}$ 进行线性变换，可以形成新的综合变量，用 $\boldsymbol{Y}$ 表示，也就是说，新的综合变量可以由原来的变量线性表示，即满足下式：

$$\begin{cases} Y_1=u_{11}X_1+u_{21}X_2+\cdots+u_{p1}X_p \\ Y_2=u_{12}X_1+u_{22}X_2+\cdots+u_{p2}X_p \\ \cdots\cdots\cdots\cdots \\ Y_p=u_{1p}X_1+u_{2p}X_2+\cdots+u_{pp}X_p \end{cases} \tag{5.1}$$

由于可以任意地对原始变量进行上述线性变换，由不同的线性变换得到的综合变量 $\boldsymbol{Y}$ 的统计特性也不尽相同。因此为了取得较好的效果，我们总是希望 $Y_i=\boldsymbol{u}_i'\boldsymbol{X}$ 的方差尽可能大且各 Y_i 之间互相独立，由于

$$\mathrm{var}(Y_i)=\mathrm{var}(\boldsymbol{u}_i'\boldsymbol{X})=\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

而对任意的常数 c，有

$$\mathrm{var}(c\boldsymbol{u}_i'\boldsymbol{X})=c^2\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

因此对 $\boldsymbol{u}_i$ 不加限制时，可使 $\mathrm{var}(Y_i)$ 任意增大，问题将变得没有意义。我们将线性变换约束在下面的原则之下：

（1）$\boldsymbol{u}_i'\boldsymbol{u}_i=1$（$i=1, 2, \cdots, p$）。

（2）Y_i 与 Y_j 相互无关（$i\neq j$；$i, j=1, 2, \cdots, p$）。

（3）Y_1 是 $X_1, X_2, \cdots, X_p$ 的一切满足原则（1）的线性组合中方差最大者；Y_2 是与 Y_1 不相关的 $X_1, X_2, \cdots, X_p$ 所有线性组合中方差最大者；……Y_p 是与 $Y_1, Y_2, \cdots, Y_{p-1}$ 都不相关的 $X_1, X_2, \cdots, X_p$ 的所有线性组合中方差最大者。

基于以上三条原则确定的综合变量 $Y_1, Y_2, \cdots, Y_p$ 分别称为原始变量的第一、第二……第 p 个主成分。其中，各综合变量在总方差中所占的比重依次递减。在实际研究工作中，通常只挑选前几个方差最大的主成分，从而达到简化系统结构、抓住问题实质的目的。

5.1.3　主成分分析的几何意义

由 5.1.1 节的介绍我们知道，在处理涉及多个指标问题的时候，为了提高分析的效率，可以不直接对 p 个指标构成的 p 维随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 进行分析，而是先对向量 $\boldsymbol{X}$ 进行线性变换，形成少数几个新的综合变量 $Y_1, Y_2, \cdots, Y_p$，使得各综合变量之间相互独立且能解释原始变量尽可能多的信息，这样，在以损失很少部分信息为代价的前提下，达到简化数据结构、提高分析效率的目的。这一节，我们着重讨论主成分分析的几何意义。为了方便，我们仅在二维空间中讨论主成分的几何意义，所得结论可以很容易地扩展到多维的情况。

设有 N 个样品，每个样品有两个观测变量 X_1，X_2，这样，在由变量 X_1，X_2 组成的坐标空间中，N 个样品散布的情况如带状（见图 5—1）。

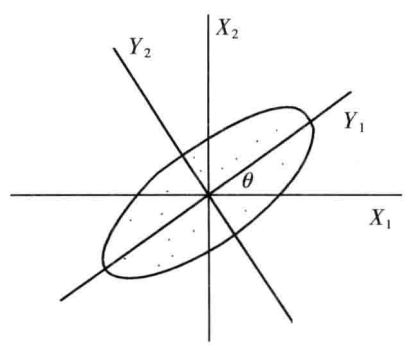


图 5—1

由图可以看出，这 N 个样品无论沿 X_1 轴方向还是沿 X_2 轴方向，均有较大的离散性，其离散程度可以分别用观测变量 X_1 的方差和 X_2 的方差定量地表示。显然，若只考虑 X_1 和 X_2 中的任何一个，原始数据中的信息均会有较大的损失。我们的目的是考虑 X_1 和 X_2 的线性组合，使原始样品数据可以由新的变量 Y_1 和 Y_2 来刻画。在几何上表示就是将坐标轴按逆时针方向旋转 θ 角度，得到新坐标轴 Y_1 和 Y_2，坐标旋转公式如下：

$$\begin{cases} Y_1 = X_1\cos\theta + X_2\sin\theta \\ Y_2 = -X_1\sin\theta + X_2\cos\theta \end{cases}$$

其矩阵形式为：

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \boldsymbol{UX}$$

式中，$\boldsymbol{U}$ 为旋转变换矩阵，由上式可知它是正交阵，即满足

$$\boldsymbol{U}' = \boldsymbol{U}^{-1}, \qquad \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$$

经过这样的旋转之后，N 个样品点在 Y_1 轴上的离散程度最大，变量 Y_1 代表了原始数据的绝大部分信息，这样，有时在研究实际问题时，即使不考虑变量 Y_2 也无损大局。因此，经过上述旋转变换就可以把原始数据的信息集中到 Y_1 轴上，对数据中包含的信息起到了浓缩的作用。主成分分析的目的就是找出变换矩阵 $\boldsymbol{U}$，而主成分分析的作用与几何意义也就很明了了。下面我们用服从正态分布的变量进行分析，以使主成分分析的几何意义更为明显。为方便起见，我们以二元正态分布为例。对于多元正态总体的情况，有类似的结论。

设变量 X_1，X_2 服从二元正态分布，分布密度为：

$$f(X_1, X_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2\sigma_1^2\sigma_2^2\sqrt{1-\rho^2}} [(X_1-\mu_1)^2\sigma_2^2 - 2\sigma_1\sigma_2\rho(X_1-\mu_1)(X_2-\mu_2) + \sigma_1^2(X_2-\mu_2)^2] \right\}$$

令 $\boldsymbol{\Sigma}$ 为变量 X_1，X_2 的协方差矩阵，其形式如下：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

令 $$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

则上述二元正态分布的密度函数有如下矩阵形式：

$$f(X_1, X_2) = \frac{1}{2\pi \left|\boldsymbol{\Sigma}\right|^{1/2}} \mathrm{e}^{-1/2(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}$$

考虑 $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = d^2$（$d$ 为常数），为方便，不妨设 $\boldsymbol{\mu}=\boldsymbol{0}$，上式有如下展开

形式：

$$\frac{1}{1-\rho^2}\left[\left(\frac{X_1}{\sigma_1}\right)^2-2\rho\left(\frac{X_1}{\sigma_1}\right)\left(\frac{X_2}{\sigma_2}\right)+\left(\frac{X_2}{\sigma_2}\right)^2\right]=d^2$$

令 $Z_1=X_1/\sigma_1$，$Z_2=X_2/\sigma_2$，则上面的方程变为：

$$Z_1^2-2\rho Z_1Z_2+Z_2^2=d^2(1-\rho^2)$$

这是一个椭圆的方程，长短轴分别为 $2d\sqrt{1\pm\rho}$。

又令 $\lambda_1\geqslant\lambda_2>0$ 为 $\boldsymbol{\Sigma}$ 的特征根，$\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$ 为相应的标准正交特征向量。

$\boldsymbol{P}=(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2)$，则 $\boldsymbol{P}$ 为正交阵，$\boldsymbol{\Lambda}=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}$，有

$$\boldsymbol{\Sigma}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}',\ \boldsymbol{\Sigma}^{-1}=\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}'$$

因此有　$d^2=(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})=\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$　　$(\boldsymbol{\mu}=\boldsymbol{0})$

$$=\boldsymbol{X}'(\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}')\boldsymbol{X}=\boldsymbol{X}'\left(\frac{1}{\lambda_1}\boldsymbol{\gamma}_1\boldsymbol{\gamma}_1'+\frac{1}{\lambda_2}\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2'\right)\boldsymbol{X}$$

$$=\frac{1}{\lambda_1}(\boldsymbol{\gamma}_1'\boldsymbol{X})^2+\frac{1}{\lambda_2}(\boldsymbol{\gamma}_2'\boldsymbol{X})^2$$

$$=\frac{Y_1^2}{\lambda_1}+\frac{Y_2^2}{\lambda_2}$$

与上面一样，这也是一个椭圆方程，且在 Y_1，Y_2 构成的坐标系中，其主轴的方向恰恰是 Y_1，Y_2 坐标轴的方向。因为 $Y_1=\boldsymbol{\gamma}_1'\boldsymbol{X}$，$Y_2=\boldsymbol{\gamma}_2'\boldsymbol{X}$，所以，$Y_1$，$Y_2$ 就是原始变量 X_1，X_2 的两个主成分，它们的方差分别为 λ_1，λ_2，在 Y_1 方向上集中了原始变量 X_1 的变差，在 Y_2 方向上集中了原始变量 X_2 的变差，经常有 λ_1 远大于 λ_2，这样，我们就可以只研究原始数据在 Y_1 方向上的变化而不至于损失过多信息，而 $\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$ 就是椭圆在原始坐标系中的主轴方向，也是坐标轴转换的系数向量。对于多维的情况，上面的结论依然成立。

这样，我们就对主成分分析的几何意义有了一个充分的了解。主成分分析的过程无非就是坐标系旋转的过程，各主成分表达式就是新坐标系与原坐标系的转换关系，在新坐标系中，各坐标轴的方向就是原始数据变差最大的方向。

5.2　总体主成分及其性质

由上面的讨论可知，求解主成分的过程就是求满足三个原则的原始变量 X_1，X_2，…，X_p 的线性组合的过程。本节从总体出发，介绍求解主成分的一般方法及主成分的性质，下节介绍样本主成分的导出。

主成分分析的基本思想就是在保留原始变量尽可能多的信息的前提下达到降维的目的，从而简化问题的复杂性并抓住问题的主要矛盾。而这里对于随机变量 X_1，

X_2，…，X_p 而言，其协方差矩阵或相关矩阵正是对各变量离散程度与变量之间的相关程度的信息的反映，而相关矩阵不过是将原始变量标准化后的协方差矩阵。我们所说的保留原始变量尽可能多的信息，也就是指生成的较少的综合变量（主成分）的方差和尽可能接近原始变量方差的总和。因此在实际求解主成分的时候，总是从原始变量的协方差矩阵或相关矩阵的结构分析入手。一般来说，从原始变量的协方差矩阵出发求得的主成分与从原始变量的相关矩阵出发求得的主成分是不同的。下面我们分别就协方差矩阵与相关矩阵进行讨论。

5.2.1 从协方差矩阵出发求解主成分

引论：设矩阵 $\boldsymbol{A}'=\boldsymbol{A}$，将 $\boldsymbol{A}$ 的特征根 λ_1，λ_2，…，λ_n 依大小顺序排列，不妨设 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$，$\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$，…，$\boldsymbol{\gamma}_p$ 为矩阵 $\boldsymbol{A}$ 各特征根对应的标准正交特征向量，则对任意向量 $\boldsymbol{x}$，有

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}=\lambda_1,\ \min_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}=\lambda_n \tag{5.2}$$

结论：设随机向量 $\boldsymbol{X}=(X_1,\ X_2,\ \cdots,\ X_p)'$ 的协方差矩阵为 $\boldsymbol{\Sigma}$，$\lambda_1,\lambda_2,\cdots,\lambda_p(\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p)$ 为 $\boldsymbol{\Sigma}$ 的特征根，$\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_p$ 为矩阵 $\boldsymbol{A}$ 各特征根对应的标准正交特征向量，则第 i 个主成分为：

$$Y_i=\gamma_{1i}X_1+\gamma_{2i}X_2+\cdots+\gamma_{pi}X_p,\quad i=1,2,\cdots,p$$

此时

$$\begin{aligned}&\operatorname{var}(Y_i)=\boldsymbol{\gamma}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_i=\lambda_i\\&\operatorname{cov}(Y_i,Y_j)=\boldsymbol{\gamma}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_j=0,\quad i\neq j\end{aligned} \tag{5.3}$$

令 $\boldsymbol{P}=(\boldsymbol{\gamma}_1,\ \boldsymbol{\gamma}_2,\ \cdots,\ \boldsymbol{\gamma}_p)$，$\boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_p)$。

由以上结论，我们把 X_1，X_2，…，X_p 的协方差矩阵 $\boldsymbol{\Sigma}$ 的非零特征根 λ_1，λ_2，…，λ_p（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$）对应的标准化特征向量 $\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$，…，$\boldsymbol{\gamma}_p$ 分别作为系数向量，$Y_1=\boldsymbol{\gamma}_1'\boldsymbol{X}$，$Y_2=\boldsymbol{\gamma}_2'\boldsymbol{X}$，…，$Y_p=\boldsymbol{\gamma}_p'\boldsymbol{X}$ 分别称为随机向量 $\boldsymbol{X}$ 的第一主成分、第二主成分……第 p 主成分。$\boldsymbol{Y}$ 的分量 Y_1，Y_2，…，Y_p 依次是 $\boldsymbol{X}$ 的第一主成分、第二主成分……第 p 主成分的充分必要条件是：

（1）$\boldsymbol{Y}=\boldsymbol{P}'\boldsymbol{X}$，即 $\boldsymbol{P}$ 为 p 阶正交阵；

（2）$\boldsymbol{Y}$ 的分量之间互不相关，即 $D(\boldsymbol{Y})=\operatorname{diag}(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_p)$；

（3）$\boldsymbol{Y}$ 的 p 个分量按方差由大到小排列，即 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$。

注：无论 $\boldsymbol{\Sigma}$ 的各特征根是否存在相等的情况，对应的标准化特征向量 $\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$，…，$\boldsymbol{\gamma}_p$ 总是存在的，我们总可以找到对应各特征根的彼此正交的特征向量。这样，求主成分的问题应变成求特征根与特征向量的问题（参见参考文献［2］和［3］）。

5.2.2　主成分的性质

性质 1　$\boldsymbol{Y}$ 的协方差阵为对角阵 $\boldsymbol{\Lambda}$。

这一性质可由上述结论容易得到，证明略。

性质 2　记 $\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$，有 $\sum\limits_{i=1}^{p}\lambda_i=\sum\limits_{i=1}^{p}\sigma_{ii}$。

证明：由 $\boldsymbol{P}=(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_p)$，则有

$$\boldsymbol{\Sigma}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$$

于是

$$\sum_{i=1}^{p}\sigma_{ii}=\operatorname{tr}(\boldsymbol{\Sigma})=\operatorname{tr}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}')=\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{P})=\operatorname{tr}(\boldsymbol{\Lambda})=\sum_{i=1}^{p}\lambda_i$$

定义 5.1　称 $\alpha_k=\dfrac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$（$k=1,2,\cdots,p$）为第 k 个主成分 Y_k 的方差贡献率，称 $\dfrac{\sum\limits_{i=1}^{m}\lambda_i}{\sum\limits_{i=1}^{p}\lambda_i}$ 为主成分 $Y_1,Y_2,\cdots,Y_m$ 的累积贡献率。

由此进一步可知，主成分分析是把 p 个随机变量的总方差 $\sum\limits_{i=1}^{p}\sigma_{ii}$ 分解为 p 个不相关的随机变量的方差之和，使第一主成分的方差达到最大。第一主成分是以变化最大的方向向量各分量为系数的原始变量的线性函数，最大方差为 λ_1。$\alpha_1=\dfrac{\lambda_1}{\sum\lambda_i}$ 表明了 λ_1 的方差在全部方差中的比值，称 α_1 为第一主成分的贡献率。这个值越大，表明 Y_1 这个新变量综合 $X_1,X_2,\cdots,X_p$ 信息的能力越强，也即由 Y_1 的差异来解释随机向量 $\boldsymbol{X}$ 的差异的能力越强。正因如此，才把 Y_1 称为 $\boldsymbol{X}$ 的主成分，进而我们就更清楚为什么主成分的位次是按特征根 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 取值的大小排序的。

进行主成分分析的目的之一是减少变量的个数，所以一般不会取 p 个主成分，而是取 m（$m<p$）个主成分。m 取多少比较合适，是一个很实际的问题，通常以所取 m 使得累积贡献率达到 85%以上为宜，即

$$\frac{\sum\limits_{i=1}^{m}\lambda_i}{\sum\limits_{i=1}^{p}\lambda_i}\geqslant 85\% \tag{5.4}$$

这样，既能使信息损失不太多，又能达到减少变量、简化问题的目的。另外，选取主成分还可根据特征根的变化来确定。图 5—2 为 SPSS 统计软件生成的碎石图。

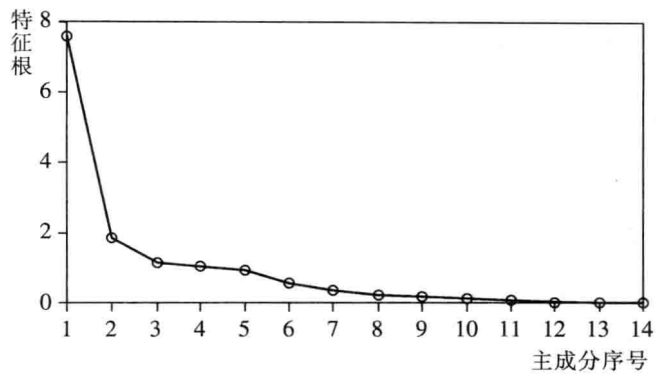


图 5—2 碎石图

由图 5—2 可知，第二个及第三个特征根变化的趋势已经开始趋于平稳，所以，取前两个或前三个主成分是比较合适的。这种方法确定的主成分个数与按累积贡献率确定的主成分个数往往是一致的。在实际应用中，有些研究工作者习惯于保留特征根大于 1 的那些主成分，但这种方法缺乏完善的理论支持。在大多数情况下，当 $m=3$ 时即可使所选主成分保持信息总量的比重达到 85%以上。

定义 5.2 第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k, X_i)$ 称为因子负荷量。

因子负荷量是主成分解释中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。在下一章中还将对因子负荷量的统计意义给出更详细的解释。由下面的性质我们可以看到，因子负荷量与系数向量成正比。

性质 3 $\rho(Y_k, X_i)=\gamma_{ik}\sqrt{\lambda_k}/\sqrt{\sigma_{ii}}$，　$k,i=1,2,\cdots,p$　　(5.5)

证明：$\sqrt{\mathrm{var}(Y_k)}=\sqrt{\lambda_k}$　　$\sqrt{\mathrm{var}(X_i)}=\sqrt{\sigma_{ii}}$

令 $\boldsymbol{e}_i=(0,\cdots,0,1,0,\cdots,0)'$ 为单位向量，则

$$X_i=\boldsymbol{e}_i'\boldsymbol{X}$$

又　　$Y_k=\boldsymbol{\gamma}_k'\boldsymbol{X}$

于是　　$\mathrm{cov}(Y_k, X_i)=\mathrm{cov}(\boldsymbol{\gamma}_k'\boldsymbol{X}, \boldsymbol{e}_i'\boldsymbol{X})=\boldsymbol{e}_i'D(\boldsymbol{X})\boldsymbol{\gamma}_k=\boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{\gamma}_k=\lambda_k\boldsymbol{e}_i'\boldsymbol{\gamma}_k=\lambda_k\gamma_{ik}$

$$\rho(Y_k, X_i)=\frac{\mathrm{cov}(Y_k, X_i)}{\sqrt{\mathrm{var}(Y_k)}\sqrt{\mathrm{var}(X_i)}}=\frac{\gamma_{ik}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$$

由性质 3 知，因子负荷量 $\rho(Y_k, X_i)$ 与系数 γ_{ik} 成正比，与 X_i 的标准差成反比关系，因此，绝不能将因子负荷量与系数向量混为一谈。在解释主成分的成因或第 i 个变量对第 k 个主成分的重要性时，应当根据因子负荷量而不能仅仅根据 Y_k 与 X_i 的变换系数 γ_{ik}。

性质 4 $\sum_{i=1}^{p}\rho^2(Y_k, X_i)\sigma_{ii}=\lambda_k$　　(5.6)

证明：由性质 3 有

$$\sum_{i=1}^{p}\rho^2(Y_k,X_i)\sigma_{ii}=\sum_{i=1}^{p}\lambda_k\gamma_{ik}^2=\lambda_k\sum_{i=1}^{p}\gamma_{ik}^2=\lambda_k \tag{5.7}$$

性质 5　$\sum_{k=1}^{p}\rho^2(Y_k,X_i)=\frac{1}{\sigma_{ii}}\sum_{k=1}^{p}\lambda_k\gamma_{ik}^2=1$

证明：因为向量 $\boldsymbol{Y}$ 是随机向量 $\boldsymbol{X}$ 的线性组合，因此 X_i 也可以精确表示成 Y_1，Y_2，…，Y_p 的线性组合。由回归分析知识知，X_i 与 Y_1，Y_2，…，Y_p 的全相关系数的平方和等于 1，而因为 Y_1，Y_2，…，Y_p 之间互不相关，所以 X_i 与 Y_1，Y_2，…，Y_p 的全相关系数的平方和也就是 $\sum_{k=1}^{p}\rho^2(Y_k,X_i)$，因此，性质 5 成立。

定义 5.3　X_i 与前 m 个主成分 Y_1，Y_2，…，Y_m 的全相关系数平方和称为 Y_1，Y_2，…，Y_m 对原始变量 X_i 的方差贡献率 v_i，即

$$v_i=\frac{1}{\sigma_{ii}}\sum_{k=1}^{m}\lambda_k\gamma_{ik}^2,\quad i=1,2,\cdots,p \tag{5.8}$$

这一定义说明了前 m 个主成分提取了原始变量 X_i 中 v_i 的信息，由此可以判断我们提取的主成分说明原始变量的能力。

5.2.3　从相关矩阵出发求解主成分

考虑如下的数学变换：

令　$Z_i=\dfrac{X_i-\mu_i}{\sqrt{\sigma_{ii}}}$，　$i=1,2,\cdots,p$

式中，μ_i 与 σ_{ii} 分别表示变量 X_i 的期望与方差。于是有

$$E(Z_i)=0,\qquad \mathrm{var}(Z_i)=1$$

令　$$\boldsymbol{\Sigma}^{1/2}=\begin{bmatrix}\sqrt{\sigma_{11}} & 0 & \cdots & 0\\ 0 & \sqrt{\sigma_{22}} & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}}\end{bmatrix}$$

于是，对原始变量 $\boldsymbol{X}$ 进行如下标准化：

$$\boldsymbol{Z}=(\boldsymbol{\Sigma}^{1/2})^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

经过上述标准化后，显然有

$$E(\boldsymbol{Z})=\boldsymbol{0}$$

$$\mathrm{cov}(\boldsymbol{Z}) = (\boldsymbol{\Sigma}^{1/2})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{1/2})^{-1} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} = \boldsymbol{R}$$

由于上面的变换过程，原始变量 X_1，X_2，…，X_p 的相关阵实际上就是对原始变量标准化后的协方差矩阵，因此，由相关矩阵求主成分的过程与主成分个数的确定准则实际上是与由协方差矩阵出发求主成分的过程与主成分个数的确定准则相一致的，在此不再赘述。仍用 λ_i，$\boldsymbol{\gamma}_i$ 分别表示相关阵 $\boldsymbol{R}$ 的特征根与对应的标准正交特征向量，此时，求得的主成分与原始变量的关系式为：

$$Y_i = \boldsymbol{\gamma}_i'\boldsymbol{Z} = \boldsymbol{\gamma}_i'(\boldsymbol{\Sigma}^{1/2})^{-1}(\boldsymbol{X}-\boldsymbol{\mu}), \quad i = 1,2,\cdots,p \tag{5.9}$$

5.2.4 由相关阵求主成分时主成分性质的简单形式

由相关阵出发所求得的主成分依然具有上面所述的各种性质，不同的是在形式上要简单，这是由相关阵 $\boldsymbol{R}$ 的特性决定的。我们将由相关阵得到的主成分的性质总结如下：

（1）$\boldsymbol{Y}$ 的协方差矩阵为对角阵 $\boldsymbol{\Lambda}$；

（2）$\sum_{i=1}^{p}\mathrm{var}(Y_i) = \mathrm{tr}(\boldsymbol{\Lambda}) = \mathrm{tr}(\boldsymbol{R}) = p = \sum_{i=1}^{p}\mathrm{var}(Z_i)$；

（3）第 k 个主成分的方差占总方差的比例，即第 k 个主成分的方差贡献率为 $\alpha_k = \lambda_k/p$，前 m 个主成分的累积方差贡献率为 $\sum_{i=1}^{m}\lambda_i/p$；

（4）$\rho(Y_k, Z_i) = \gamma_{ik}\sqrt{\lambda_k}$。

注意到 $\mathrm{var}(Z_i)=1$，且 $\mathrm{tr}(\boldsymbol{R})=p$，结合前面从协方差矩阵出发求主成分部分对主成分性质的说明，可以很容易地得出上述性质。虽然主成分的性质在这里有更简单的形式，但应注意其实质与前面的结论并没有区别。需要注意的一点是，判断主成分的成因或原始变量（这里，原始变量指的是标准化以后的随机向量 $\boldsymbol{Z}$）对主成分的重要性有更简单的方法，因为由上面第（4）条性质知，这里因子负荷量仅依赖于由 Z_i 到 Y_k 的转换向量系数 γ_{ik}（因为对不同的 Z_i，因子负荷量表达式的后半部分 $\sqrt{\lambda_k}$ 是固定的）。

5.3 样本主成分的导出

在实际研究工作中，总体协方差阵 $\boldsymbol{\Sigma}$ 与相关阵 $\boldsymbol{R}$ 通常是未知的，因此需要通过样本数据来估计。设有 n 个样品，每个样品有 p 个指标，这样共得到 np 个数据，原始资料矩阵如下：

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

记　　$$\boldsymbol{S}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{ki}-\bar{x}_i)'$$

$$\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ki},\qquad i=1,2,\cdots,p$$

$$\boldsymbol{R}=(r_{ij})_{p\times p},\quad r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

$\boldsymbol{S}$ 为样本协方差矩阵，作为总体协方差阵 $\boldsymbol{\Sigma}$ 的无偏估计；$\boldsymbol{R}$ 是样本相关矩阵，为总体相关矩阵的估计。由前面的讨论知，若原始资料阵 $\boldsymbol{X}$ 是经过标准化处理的，则由矩阵 $\boldsymbol{X}$ 求得的协方差阵就是相关矩阵，即 $\boldsymbol{S}$ 与 $\boldsymbol{R}$ 完全相同。因为由协方差矩阵求解主成分的过程与由相关矩阵出发求解主成分的过程是一致的，所以下面我们仅介绍由相关阵 $\boldsymbol{R}$ 出发求解主成分。

根据总体主成分的定义，主成分 $\boldsymbol{Y}$ 的协方差是：

$$\mathrm{cov}(\boldsymbol{Y})=\boldsymbol{\Lambda}$$

式中，$\boldsymbol{\Lambda}$ 为对角阵。

$$\boldsymbol{\Lambda}=\begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

假定资料矩阵 $\boldsymbol{X}$ 为已标准化后的数据矩阵，则可由相关矩阵代替协方差矩阵，于是上式可表示为：

$$\boldsymbol{P}'\boldsymbol{R}\boldsymbol{P}=\boldsymbol{\Lambda}$$

于是，所求的新的综合变量（主成分）的方差 $\lambda_i(i=1,2,\cdots,p)$ 是

$$|\boldsymbol{R}-\lambda\boldsymbol{I}|=0$$

的 p 个根，λ 为相关矩阵的特征根，相应的各个 γ_{ij} 是其特征向量的分量。

因为 $\boldsymbol{R}$ 为正定矩阵，所以其特征根都是非负实数，将它们依大小顺序排列 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$，其相应的特征向量记为 $\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$，$\cdots$，$\boldsymbol{\gamma}_p$，则相对于 Y_1 的方差为：

$$\mathrm{var}(Y_1)=\mathrm{var}(\boldsymbol{\gamma}_1'\boldsymbol{X})=\lambda_1$$

同理有

$$\mathrm{var}(Y_i)=\mathrm{var}(\boldsymbol{\gamma}_i'\boldsymbol{X})=\lambda_i$$

即对于 Y_1 有最大方差，Y_2 有次大方差……并且协方差为：

$$\begin{aligned}\mathrm{cov}(Y_i,Y_j)&=\mathrm{cov}(\boldsymbol{\gamma}_i'\boldsymbol{X},\boldsymbol{\gamma}_j'\boldsymbol{X})=\boldsymbol{\gamma}_i'\boldsymbol{R}\boldsymbol{\gamma}_j\\&=\boldsymbol{\gamma}_i'\Big(\sum_{\alpha=1}^{p}\lambda_\alpha\boldsymbol{\gamma}_\alpha\boldsymbol{\gamma}_\alpha'\Big)\boldsymbol{\gamma}_j\\&=\sum_{\alpha=1}^{p}\lambda_\alpha(\boldsymbol{\gamma}_i'\boldsymbol{\gamma}_\alpha)(\boldsymbol{\gamma}_\alpha'\boldsymbol{\gamma}_j)=0,\ i\neq j\end{aligned}$$

由此可有新的综合变量（主成分）Y_1，Y_2，…，Y_p 彼此不相关，并且 Y_i 的方差为 λ_i，则 $Y_1=\boldsymbol{\gamma}_1'\boldsymbol{X}$，$Y_2=\boldsymbol{\gamma}_2'\boldsymbol{X}$，…，$Y_p=\boldsymbol{\gamma}_p'\boldsymbol{X}$ 分别称为第一、第二…… 第 p 个主成分。由上述求主成分的过程可知，主成分在几何图形中的方向实际上就是 $\boldsymbol{R}$ 的特征向量的方向；主成分的方差贡献就等于 $\boldsymbol{R}$ 的相应特征根。这样，我们利用样本数据求解主成分的过程实际上就转化为求相关阵或协方差阵的特征根和特征向量的过程。

5.4 有关问题的讨论

5.4.1 关于由协方差矩阵或相关矩阵出发求解主成分

由前面的讨论可知，求解主成分的过程实际就是对矩阵结构进行分析的过程，也就是求解特征根的过程。在实际分析过程中，我们可以从原始数据的协方差矩阵出发，也可以从原始数据的相关矩阵出发，其求主成分的过程是一致的。但是，从协方差阵出发和从相关阵出发所求得的主成分一般来说是有差别的，而且这种差别有时候还很大。这方面的例子见参考文献［7］。

一般而言，对于度量单位不同的指标或取值范围彼此差异非常大的指标，不直接由其协方差矩阵出发进行主成分分析，而应该考虑将数据标准化。比如，在对上市公司的财务状况进行分析时，常常会涉及利润总额、市盈率、每股净利率等指标，其中利润总额取值常常从几十万元到上百万元，市盈率取值一般从 5 到六七十之间，而每股净利率在 1 以下，不同指标取值范围相差很大，这时若是直接从协方差矩阵入手进行主成分分析，利润总额将明显起到重要支配作用，而其他两个指标的作用很难在主成分中体现出来，此时应该考虑对数据进行标准化处理。

但是，对原始数据进行标准化处理后倾向于各个指标的作用在主成分的构成中相等。对于取值范围相差不大或度量相同的指标进行标准化处理后，其主成分分析的结果仍与由协方差阵出发求得的结果有较大区别。其原因是由于对数据进行标准化的过程实际上也就是抹杀原始变量离散程度差异的过程，标准化后的各变量方差相等，均为 1，而实际上方差也是对数据信息的重要概括，也就是说，对原始数据进行标准化后抹杀了一部分重要信息，因此才使得标准化后各变量在对主成分构成

中的作用趋于相等。由此看来，对同度量或取值范围在同量级的数据，还是直接从协方差矩阵求解主成分为宜。

对于从什么出发求解主成分，现在还没有一个定论，但是我们应该看到，不考虑实际情况就对数据进行标准化处理或者直接从原始变量的相关矩阵出发求解主成分是有其不足之处的，这一点需要注意。建议在实际工作中分别从不同角度出发求解主成分并研究其结果的差别，看看是否发生明显差异且这种差异产生的原因在何处，以确定用哪种结果更为可信。

5.4.2　主成分分析不要求数据来自于正态总体

由上面的讨论可知，无论是从原始变量协方差矩阵出发求解主成分，还是从相关矩阵出发求解主成分，均没有涉及总体分布的问题。也就是说，与很多多元统计方法不同，主成分分析不要求数据来自于正态总体。实际上，主成分分析就是对矩阵结构的分析，其中用到的主要是矩阵运算的技术及矩阵对角化和矩阵的谱分解技术。我们知道，对多元随机变量而言，其协方差矩阵或相关矩阵均是非负定的，这样，就可以按照求解主成分的步骤求出其特征根、标准正交特征向量，进而求出主成分，达到缩减数据维数的目的。同时，由主成分分析的几何意义可以看到，对来自多元正态总体的数据，我们得到了合理的几何解释，即主成分就是按数据离散程度最大的方向进行坐标轴旋转。

主成分分析的这一特性大大扩展了其应用范围，对多维数据，只要是涉及降维的处理，我们都可以尝试用主成分分析，而不用花太多精力考虑其分布情况。

5.4.3　主成分分析与重叠信息

首先应当认识到，主成分分析方法适用于变量之间存在较强相关性的数据，如果原始数据相关性较弱，运用主成分分析不能起到很好的降维作用，即所得的各个主成分浓缩原始变量信息的能力差别不大。一般认为，当原始数据大部分变量的相关系数都小于 0.3 时，运用主成分分析不会取得很好的效果。

很多研究者在运用主成分分析方法时，都或多或少地存在对主成分分析消除原始变量重叠信息的期望，这样，在实际工作之初就可以把与某一研究问题相关而可能得到的变量（指标）都纳入分析过程，再用少数几个主成分浓缩这些有用信息（假定已剔除了重叠信息），然后对主成分进行深入分析。在对待重叠信息方面，生成的新的综合变量（主成分）是有效剔除了原始变量中的重叠信息，还是仅按原来的模式将原始信息中的绝大部分用几个不相关的新变量表示出来，这一点还有待讨论。

为说明这个问题，有必要再回顾一下主成分的求解过程。我们仅就从协方差矩阵出发求主成分的过程予以说明，对从相关阵出发有类似的情况。

对于 p 维指标的情况，得到其协方差矩阵如下：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}$$

现在考虑一种极端情况，即有两个指标完全相关，不妨设第一个指标在进行主成分分析时考虑了两次。则协方差矩阵变为：

$$\boldsymbol{\Sigma}_1=\begin{bmatrix}\sigma_{11} & \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{11} & \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ \sigma_{p1} & \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}$$

此时主成分分析实际上是由$(p+1)\times(p+1)$维矩阵$\boldsymbol{\Sigma}_1$进行。$\boldsymbol{\Sigma}_1$的行列式的值为零但仍满足非负定，只不过其最小的特征根为零，由$\boldsymbol{\Sigma}_1$出发求解主成分，其方差总和不再是$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp}$，而是变为$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp}+\sigma_{11}$。也就是说，第一个指标在分析过程中起到了加倍的作用，其重叠信息完全像其他指标提供的信息一样在起作用。这样求得的主成分已经与没有第一个指标重叠信息时不一样了，因为主成分方差的总和已经变为$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp}+\sigma_{11}$而不是$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp}$，每个主成分解释方差的比例也相应发生变化，而整个分析过程没有对重叠信息做任何特殊处理。也就是说，由于对第一个指标罗列了两次，其在生成的主成分构成中也起到了加倍的作用。这一点尤其应该引起注意，这意味着主成分分析对重叠信息的剔除是无能为力的，同时主成分分析还损失了一部分信息。对此，参考文献［4］举例进行了说明。

这就告诉我们，在实际工作中，在选取初始变量进入分析时应该小心，对原始变量存在多重共线性的问题，在应用主成分分析方法时一定要慎重。应该考虑所选取的初始变量是否合适，是否真实地反映了事物的本来面目，如果是出于避免遗漏某些信息的原因而特意选取了过多的存在重叠信息的变量，就要特别注意应用主成分分析所得到的结果。

如果所得到的样本协方差矩阵（或相关阵）最小的特征根接近于零，那么就有

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_p=(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\gamma}_p=\lambda_p\boldsymbol{\gamma}_p\approx\boldsymbol{0} \tag{5.10}$$

进而推出

$$(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\gamma}_p\approx 0 \tag{5.11}$$

这就意味着，中心化以后的原始变量之间存在着多重共线性，即原始变量存在着不可忽视的重叠信息。因此，在进行主成分分析得出协方差阵或是相关阵，发现最小特征根接近于零时，应该注意对主成分的解释，或者考虑对最初纳入分析的指标进行筛选。由此可以看出，虽然主成分分析不能有效地剔除重叠信息，但它至少可以发现原始变量是否存在重叠信息，这对减少分析中的失误是有帮助的。

5.5　主成分分析步骤及框图

5.5.1　主成分分析步骤

由前面的讨论大体上可以明了进行主成分分析的步骤，对此进行归纳如下：

（1）根据研究问题选取初始分析变量；

（2）根据初始变量特性判断由协方差阵求主成分还是由相关阵求主成分；

（3）求协方差阵或相关阵的特征根与相应标准特征向量；

（4）判断是否存在明显的多重共线性，若存在，则回到第（1）步；

（5）得到主成分的表达式并确定主成分个数，选取主成分；

（6）结合主成分对研究问题进行分析并深入研究。

5.5.2　主成分分析的逻辑框图

主成分分析的逻辑框图如图 5—3 所示。

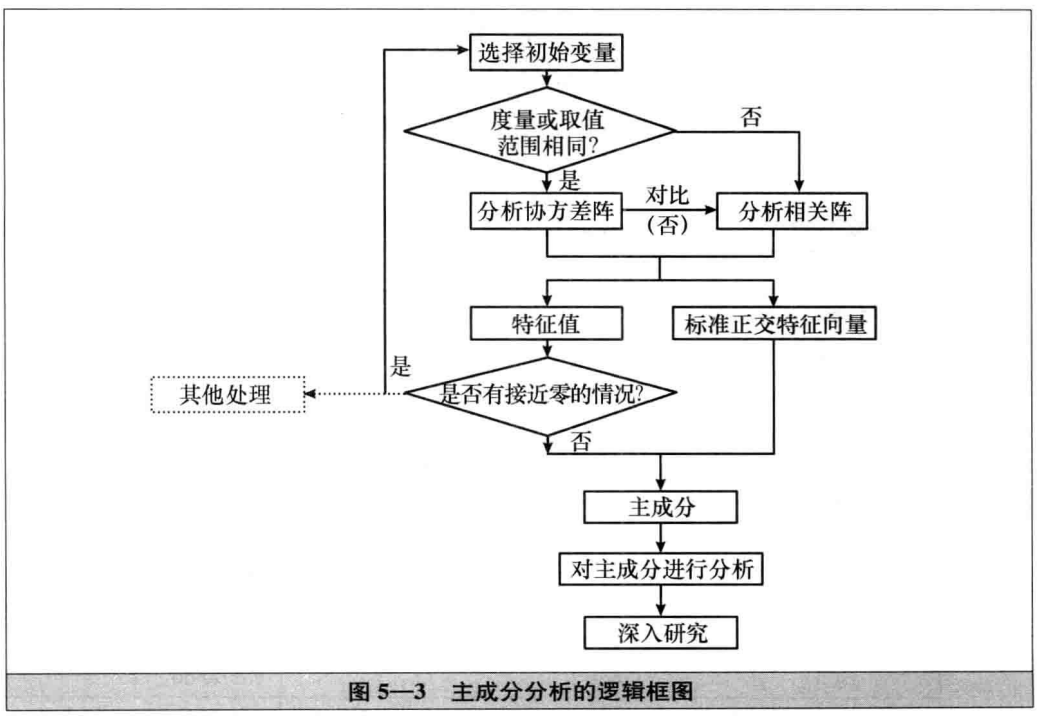


图 5—3　主成分分析的逻辑框图

5.6　主成分分析的上机实现

SPSS 软件 FACTOR 模块提供了主成分分析的功能。下面先以之前版本 SPSS

10.0 自带的数据 Employee data.sav 为例介绍主成分分析的上机实现方法，在 SPSS 软件的安装目录下可以找到该数据集；然后，举一个实际的例子介绍主成分分析的具体应用。

例 5—1

数据集 Employee data 为 Midwestern 银行在 1969—1971 年之间雇员情况的数据，共包括 474 条观测及如下 10 个变量：Id（观测号），Gender（性别），Bdate（出生日期），Educ（受教育程度（年数）），Jobcat（工作种类），Salary（目前年薪），Salbegin（开始受聘时的年薪），Jobtime（受雇时间（月）），Prevexp（受雇以前的工作时间（月）），Minority（是否少数族裔）。下面我们用主成分分析方法处理该数据，以期用少数变量来描述该银行的雇用情况。

进入 SPSS 软件，打开数据集 Employee data.sav。依次点选 Analyze→Dimension Reduction→Factor…进入 Factor Analysis（因子分析）对话框。（在 SPSS 软件中，主成分分析与因子分析均在 Factor Analysis 模块中完成。）此时，数据集 Employee data.sav 中的变量名均已显示在左边的窗口中，依次选中变量 Educ，Salary，Salbegin，Jobtime，Prevexp 并点击向右的箭头按钮，这五个变量便进入 Variables 窗口（此时若选中 Variables 窗口中的变量，则窗口左侧的箭头按钮即转向左侧，点此按钮即可剔除所选中变量）。点击下方的 OK 按钮，即可得到输出结果 5—1。

输出结果 5—1

Communalities

	Initial	Extraction
Educational Level (years)	1.000	0.754
Current Salary	1.000	0.896
Beginning Salary	1.000	0.916
Months since Hire	1.000	0.999
Previous Experience (Months)	1.000	0.968

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.477	49.541	49.541	2.477	49.541	49.541
2	1.052	21.046	70.587	1.052	21.046	70.587
3	1.003	20.070	90.656	1.003	20.070	90.656
4	0.365	7.299	97.955			
5	0.102	2.045	100.00			

Extraction method: Principal Component Analysis.

Component Matrix*

	Component		
	1	2	3
Educational Level (years)	0.846	−0.194	−1.4E−02
Current Salary	0.940	0.104	2.857E−02
Beginning Salary	0.917	0.264	−7.7E−02
Months since Hire	6.806E−02	−5.2E−02	0.996
Previous Experience (Months)	−0.178	0.965	6.901E−02

Extraction Method: Principal Component Analysis.

* 3 components extracted.

其中，Communalities 表给出了该次分析从每个原始变量中提取的信息，表格下面的表注表明，该次分析是用 Factor Analysis 模块默认的信息提取方法即主成分分析完成的。可以看到除受教育程度信息损失较大外，主成分几乎包含了各个原始变量至少 90%的信息。Total Variance Explained 表则显示了各主成分解释原始变量总方差的情况，SPSS 默认保留特征根大于 1 的主成分，在本例中看到保留 3 个主成分为宜，这 3 个主成分集中了 5 个原始变量信息的 90.656%，可见效果比较好。实际上，主成分解释总方差的百分比也可以由 Communalities 表计算得出，即 (0.896+0.916+0.999+0.968+0.754)/5=90.66%。Component Matrix 表中给出了标准化原始变量用求得的主成分线性表示的近似表达式，我们以表中 Current Salary 一行为例，不妨用 *prin*1，*prin*2，*prin*3 来表示各个主成分，则由 Component Matrix 表可以得到

$$\text{标准化的 } salary \approx 0.940 \times prin1 + 0.104 \times prin2 + (2.857\text{E}-02) \times prin3$$

在上面的主成分分析中，SPSS 默认是从相关阵出发求解主成分，且默认保留特征根大于 1 的主成分。实际上，对主成分的个数我们可以自己确定，方法为：进入 Factor Analysis 对话框并选择好变量之后，点击 Extraction 选项，在弹出的对话框中有一个 Extract 选择框，默认是选择 Eigenvalues greater than 1，也就是保留特征根大于 1 的主成分，可以输入其他数值来改变 SPSS 软件保留的特征根的大小；另外，还可以选择 Fixed Number of Factors 选项直接确定主成分的个数。在实际进行主成分分析时，可以先按照默认设置做一次主成分分析，然后根据输出结果确定应保留主成分的个数，用该方法进行设定后重新分析。

因为上面的结果是默认从相关阵出发得到的，而对于由相关阵出发求得的主成分，其性质有简单的表达形式，我们可以方便地加以验证。

由 Component Matrix 表中的结果可以得到

$$0.940^2+0.917^2+(6.806\text{E}-02)^2+(-0.178)^2+0.846^2$$
$$=2.477\,031=\text{第一主成分的方差}$$

这就验证了性质 4。又有

$$0.940^2+0.104^2+(2.857\text{E}-02)^2=0.896$$

这恰好与 Communalities 表中三个主成分提取 Salary 变量的信息相等。重做一次主

成分分析，此次将5个主成分全部保留，得到新的Component Matrix表，如输出结果5—2所示。

输出结果 5—2 **Component Matrix***

	Component				
	1	2	3	4	5
Educational Level (years)	0.846	−0.194	−1.4E−02	0.496	7.733E−03
Current Salary	0.940	0.104	2.857E−02	−0.234	0.222
Beginning Salary	0.917	0.264	−7.7E−02	−0.183	−0.225
Months since Hire	6.806E−02	−5.2E−02	0.996	−1.3E−02	−2.6E−02
Previous Experience (Months)	−0.178	0.965	6.901E−02	0.174	3.769E−02

Extraction Method: Principal Component Analysis.

* 5 components extracted.

可以看到，前三个主成分的相应结果与输出结果5—1中的对应部分结果是一致的。对输出结果5—2有如下关系式：

$$0.940^2+0.104^2+(2.857\text{E}-02)^2+(-0.234)^2+0.222^2=1$$

这就验证了性质5。由此表还可以得到标准化原始变量用各主成分线性表示的精确的表达式。仍以Current Salary为例，有

$$\text{标准化的 } salary=0.940\times prin1+0.104\times prin2+(2.857\text{E}-02)\times prin3-0.234\times prin4+0.222\times prin5$$

由SPSS软件默认选项输出的结果，我们还不能得到用原始变量表示出主成分的表达式，这是因为Component Matrix表中表示的是因子载荷矩阵而不是主成分的系数矩阵，因此要对SPSS的因子分析模块运行结果进行调整。将Component Matrix表中的第i列的每个元素分别除以第i个特征根的平方根$\sqrt{\lambda_i}$，就可以得到主成分分析的第i个主成分的系数。主成分的系数矩阵如输出结果5—3所示。

输出结果 5—3

	*prin*1	*prin*2	*prin*3
Educational Level (years)	0.537 65	−0.188 98	−0.013 96
Current Salary	0.597 457	0.101 834	0.028 523
Beginning Salary	0.582 45	0.256 952	−0.076 77
Months since Hire	0.043 243	−0.050 93	0.994 159
Previous Experience (months)	−0.113 4	0.940 903	0.068 887

由此表可以写出各个主成分用标准化后的原始变量表示的表达式。

$$\begin{aligned}prin1 =&0.537\,65\times\text{标准化的 }educ+0.597\,457\times\text{标准化的 }salary\\&+0.582\,45\times\text{标准化的 }salbegin+0.043\,243\times\text{标准化的 }jobtime\\&-0.113\,4\times\text{标准化的 }prevexp\\prin2 =&-0.188\,98\times\text{标准化的 }educ+0.101\,834\times\text{标准化的 }salary\\&+0.256\,952\times\text{标准化的 }salbegin-0.050\,93\times\text{标准化的 }jobtime\\&+0.940\,903\times\text{标准化的 }prevexp\end{aligned}$$

$$prin3 = -0.013\,96 \times \text{标准化的 } educ + 0.028\,523 \times \text{标准化的 } salary - 0.076\,77 \times \text{标准化的 } salbegin - 0.994\,159 \times \text{标准化的 } jobtime + 0.068\,887 \times \text{标准化的 } prevexp$$

例5—2

在企业经济效益的评价中，设计的指标往往很多。为了简化系统结构，抓住经济效益评价中的主要问题，我们可由原始数据矩阵出发求主成分。在对我国部分省、直辖市、自治区独立核算的工业企业的经济效益评价中，涉及9项指标，原始数据如表5—1所示，即样品数 $n=28$，变量数 $p=9$。

表5—1

地区	百元固定资产原值实现值(%)	百元固定资产原值实现利税(%)	百元资金实现利税(%)	百元工业总产值实现利税(%)	百元销售收入实现利税(%)	每吨标准煤实现工业产值(元)	每千瓦时电力实现工业产值(元)	全员劳动生产率(元/人·年)	百元流动资金实现产值(元)
北京(1)	119.29	30.98	29.92	25.97	15.48	2 178	3.41	21 006	296.7
天津(2)	143.98	31.59	30.21	21.94	12.29	2 852	4.29	20 254	363.1
河北(3)	94.8	17.2	17.95	18.14	9.37	1 167	2.03	12 607	322.2
山西(4)	65.8	11.08	11.06	12.15	16.84	8.82	1.65	10 166	284.7
内蒙古(5)	54.79	9.24	9.54	16.86	6.27	894	1.8	7 564	225.4
辽宁(6)	94.51	21.12	22.83	22.35	11.28	1 416	2.36	13.386	311.7
吉林(7)	80.49	13.36	13.76	16.6	7.14	1 306	2.07	9 400	274.1
黑龙江(8)	75.86	15.82	16.67	20.86	10.37	1 267	2.26	9 830	267
上海(9)	187.79	45.9	39.77	24.44	15.09	4 346	4.11	31 246	418.6
江苏(10)	205.96	27.65	22.58	13.42	7.81	3 202	4.69	23 377	407.2
浙江(11)	207.46	33.06	25.78	15.94	9.28	3 811	4.19	22 054	385.5
安徽(12)	110.78	20.7	20.12	18.69	6.6	1 468	2.23	12 578	341.1
福建(13)	122.76	22.52	19.93	18.34	8.35	2 200	2.63	12 164	301.2
江西(14)	94.94	14.7	14.18	15.49	6.69	1 669	2.24	10 463	274.4
山东(15)	117.58	21.93	20.89	18.65	9.1	1 820	2.8	17 829	331.1
河南(16)	85.98	17.3	17.18	20.12	7.67	1 306	1.89	11 247	276.5
湖北(17)	103.96	19.5	18.48	18.77	9.16	1 829	2.75	15 745	308.9
湖南(18)	104.03	21.47	21.28	20.63	8.72	1 272	1.98	13 161	309
广东(19)	136.44	23.64	20.83	17.33	7.85	2 959	3.71	16 259	334
广西(20)	100.72	22.04	20.9	21.88	9.67	1 732	2.13	12 441	296.4
四川(21)	84.73	14.35	14.17	16.93	7.96	1 310	2.34	11 703	242.5
贵州(22)	59.05	14.48	14.35	24.53	8.09	1 068	1.32	9 710	206.7
云南(23)	73.72	21.91	22.7	29.72	9.38	1 447	1.94	12 517	295.8
陕西(24)	78.02	13.13	12.57	16.83	9.19	1 731	2.08	11 369	220.3
甘肃(25)	59.62	14.07	16.24	23.59	11.34	926	1.13	13 084	246.8
青海(26)	51.66	8.32	8.26	16.11	7.05	1 055	1.31	9 246	176.49
宁夏(27)	52.95	8.25	8.82	15.57	6.58	834	1.12	10 406	245.4
新疆(28)	60.29	11.26	13.14	18.68	8.39	1 041	2.9	10 983	266

首先对原始数据进行标准化，标准化后的数据如表 5—2 所示。

表 5—2

0.423 523	1.338 405	1.590 282	1.687 556	2.239 634	0.481 971	0.954 746	1.260 371	0.048 805
0.995 199	1.409 649	1.631 453	0.667 228	1.065 873	1.188 758	1.855 394	1.133 844	1.200 166
−0.143 52	−0.271	−0.109 06	−0.294 87	−0.008 54	−0.578 21	−0.457 63	−0.152 79	0.490 97
−0.814 99	−0.985 77	−1.087 21	−1.811 43	2.740 046	−1.792 73	−0.846 55	−0.563 49	−0.159 27
−1.069 92	−1.200 67	−1.303	−0.618 94	−1.149 19	−0.864 49	−0.693 03	−1.001 29	−1.187 52
−0.150 24	0.186 827	0.583 737	0.771 033	0.694 243	−0.317 1	−0.119 89	−2.271 7	0.308 902
−0.474 86	−0.719 49	−0.703 9	−0.684 77	−0.829 07	−0.432 45	−0.416 7	−0.692 38	−0.343 07
−0.582 06	−0.432 18	−0.290 78	0.393 79	0.359 408	−0.473 34	−0.222 24	−0.620 03	−0.466 19
2.009 583	3.080 956	2.988 656	1.300 186	2.096 133	2.755 433	1.671 171	2.983 284	2.162 524
2.430 294	0.949 485	0.548 246	−1.489 89	−0.582 54	1.555 783	2.264 78	1.659 299	1.964 851
2.465 025	1.581 335	1.002 539	−0.851 87	−0.041 66	2.194 408	1.753 048	1.436 7	1.588 578
0.226 481	0.137 774	0.199 007	−0.155 62	−1.027 76	−0.262 57	−0.252 94	−0.157 67	0.818 691
0.503 868	0.350 337	0.172 033	−0.244 23	−0.383 85	0.505 041	0.156 444	−0.227 32	0.126 834
−0.140 28	−0.562 98	−0.644 28	−0.965 8	−0.994 65	−0.051 79	−0.242 71	−0.513 52	−0.337 87
0.383 929	0.281 429	0.308 322	−0.165 74	−0.107 89	0.106 557	0.330 433	0.725 83	0.645 294
−0.347 74	−0.259 32	−0.218 38	0.206 435	−0.634 06	−0.432 45	−0.600 92	−0.381 61	−0.301 46
0.068 569	−0.002 38	−0.033 82	−0.135 36	−0.085 81	0.115 994	0.279 26	0.375 19	0.260 351
0.070 19	0.227 705	0.363 689	0.335 558	−0.247 71	−0.468 1	−0.508 81	−0.059 58	0.262 085
0.820 617	0.481 145	0.299 804	−0.499 95	−0.567 83	1.300 963	1.261 785	0.461 673	0.695 579
−0.006 45	0.294 277	0.309 741	0.652 037	0.101 843	0.014 276	−0.355 29	−0.180 72	0.043 603
−0.376 69	−0.603 86	−0.645 7	−0.601 22	−0.527 35	−0.428 25	−0.140 36	−0.304 89	−0.891 01
−0.971 28	−0.588 68	−0.620 14	1.322 972	−0.479 52	−0.682 02	−1.184 29	−0.640 22	−1.511 77
−0.631 61	0.279 093	0.565 282	2.636 993	−0.004 86	−0.284 59	−0.549 75	−0.167 93	0.033 199
−0.532 05	−0.746 35	−0.872 84	−0.626 54	−0.074 77	0.013 227	−0.406 46	−0.361 09	−1.275 95
−0.958 09	−0.636 56	−0.351 82	1.084 98	0.716 32	−0.830 93	−1.378 75	−0.072 53	−0.816 45
−1.142 39	−1.308 12	−1.484 72	−0.808 83	−0.862 19	−0.695 66	−1.194 53	−0.718 29	−2.035 61
−1.112 52	−1.316 3	−1.405 22	−0.945 55	−1.035 12	−0.927 41	−1.388 99	−0.523 11	−0.840 73
−0.942 57	−0.964 75	−0.791 92	−0.158 15	−0.369 13	−0.710 34	0.432 779	−0.426 03	−0.483 53

将表 5—2 中的数据导入 SPSS 软件，依次点选 Analyze→Dimension Reduction→Factor…进入 Factor Analysis 对话框。（在 SPSS 中，主成分分析与因子分析均在 Factor Analysis 模块中完成。）

点击 Descriptives 按钮，在弹出的对话框中，在 Correlation Matrix 中选择 Coefficients。回到原对话框点击下方的 OK，即可得到输出结果 5—4 和输出结果 5—5。

输出结果 5—4　　**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.150	68.332	68.332	6.150	68.332	68.332
2	1.473	16.365	84.698	1.473	16.365	84.698
3	.697	7.749	92.447			
4	.318	3.531	95.978			
5	.190	2.112	98.090			
6	.116	1.289	99.379			
7	.029	.324	99.703			
8	.024	.270	99.973			
9	.002	.027	100.000			

Extraction Method: Principal Component Analysis.

输出结果 5—5　　**Correlation Matrix**

		x1	x2	x3	x4	x5	x6	x7	x8	x9
Correlation	x1	1.000	.869	.770	−.053	.211	.920	.899	.795	.896
	x2	.869	1.000	.978	.387	.472	.886	.804	.814	.849
	x3	.770	.978	1.000	.523	.531	.797	.736	.740	.811
	x4	−.053	.387	.523	1.000	.323	.115	−.023	.125	.051
	x5	.211	.472	.531	.323	1.000	.175	.260	.371	.317
	x6	.920	.886	.797	.115	.175	1.000	.877	.815	.768
	x7	.899	.804	.736	−.023	.260	.877	1.000	.757	.818
	x8	.795	.814	.740	.125	.371	.815	.757	1.000	.715
	x9	.896	.849	.811	.051	.317	.768	.818	.715	1.000

由输出结果 5—4 看到，前两个主成分 y_1，y_2 的方差和占全部方差的比例为 84.7%。我们就选取 y_1 为第一主成分，y_2 为第二主成分，且这两个主成分的方差和占全部方差的 84.7%，即基本上保留了原来指标的信息，这样由原来的 9 个指标转化为 2 个新指标，起到了降维的作用。

SPSS 软件得到因子载荷矩阵如输出结果 5—6 所示。

输出结果 5—6　Component Matrix[a]

	Component	
	1	2
x1	.931	−.315
x2	.976	.163
x3	.931	.322
x4	.232	.863
x5	.433	.596
x6	.923	−.200
x7	.897	−.274
x8	.871	−.064
x9	.899	−.154

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

对 SPSS 的因子分析模块运行结果输出的 Component Matrix 的第 i 列的每个元素分别除以第 i 个特征根的平方根$\sqrt{\lambda_i}$，就得到主成分分析的第 i 个主成分的系数，结果如表 5—3 所示。

表 5—3

	主成分 1	主成分 2
x1	0.375 558 6	−0.259 51
x2	0.393 395 8	0.134 374
x3	0.375 255 6	0.265 294
x4	0.093 534 3	0.711 329
x5	0.174 559	0.491 327
x6	0.372 149 4	−0.164 96
x7	0.361 616 5	−0.225 4
x8	0.351 316 1	−0.052 36
x9	0.362 594 2	−0.126 75

由上表得到前两个主成分 y_1，y_2 的线性组合为：

$$\begin{aligned} y_1 =& 0.375\,558\,6x_1^* + 0.393\,395\,8x_2^* + 0.375\,255\,6x_3^* + 0.093\,534\,3x_4^* \\ &+ 0.174\,559x_5^* + 0.372\,149\,4x_6^* + 0.361\,616\,5x_7^* + 0.351\,316\,1x_8^* \\ &+ 0.362\,594\,2x_9^* \\ y_2 =& -0.259\,51x_1^* + 0.134\,374x_2^* + 0.265\,294x_3^* + 0.711\,329x_4^* \\ &+ 0.491\,327x_5^* - 0.164\,96x_6^* - 0.225\,4x_7^* - 0.052\,36x_8^* \\ &- 0.126\,75x_9^* \end{aligned} \tag{5.12}$$

式中，x_1^*，x_2^*，x_3^*，x_4^*，x_5^*，x_6^*，x_7^*，x_8^*，x_9^* 表示对原始变量标准化后的变量。

对所选主成分做经济解释。主成分分析的关键在于能否给主成分赋予新的意义，给出合理的解释，这个解释应根据主成分的计算结果结合定性分析来进行。主成分是原来变量的线性组合，在这个线性组合中，各变量的系数有大有小，有正有负，有的大小相当，因而不能简单地认为这个主成分是某个原变量的属性的作用。线性组合中各变量的系数的绝对值大者表明该主成分主要综合了绝对值大的变量，有几个变量系数大小相当时，应认为这一主成分是这几个变量的总和，这几个变量综合在一起应赋予怎样的经济意义，要结合经济专业知识，给出恰如其分的解释，才能达到深刻分析经济成因的目的。

我们所举的例子中有 9 个指标，这 9 个指标有很强的依赖性，通过主成分计算后，我们选择了 2 个主成分，这 2 个主成分具有明显的经济意义。第一主成分的线性组合中除了百元工业总产值实现利税和百元销售收入实现利税外，其余变量的系数相当，所以第一主成分可看成 x_1，x_2，x_3，x_6，x_7，x_8，x_9 的综合变量。可以解释为第一主成分反映了工业生产中投入的资金、劳动力所产生的效果，它是“投入”与“产出”之比。第一主成分所占信息总量为 68.3%，在我国目前的工业企业

中，经济效益首先反映在投入与产出之比上，其中固定资产所产生的经济效益更大一些。第二主成分是把工业生产中所得总量（即工业总产值和销售收入）与局部量（即利税）进行比较，反映了“产出”对国家所作的贡献。这样，在抓企业经济效益活动中，就应注重投入与产出之比和产出对国家所作的贡献。抓住了这两个方面，经济效益就一定会提高。

通常为了分析各样品在主成分所反映的经济意义方面的情况，还将标准化后的原始数据代入主成分表达式计算出各样品的主成分得分，由各样品的主成分得分（当主成分个数为 2 时）就可在二维空间中描出各样品的分布情况。

将表 5—2 中的数据代入式（5.12）中，得到 28 个省、直辖市、自治区的主成分得分，如表 5—4 所示。将这 28 个样品在平面直角坐标系上描出来，进而可进行样品分类。主成分得分图如图 5—4 所示。

表 5—4

样品号	主成分得分	
	第一主成分得分	第二主成分得分
1	2.816 239 8	2.425 742 1
2	3.735 83	0.536 482 9
3	−0.486 829	−0.097 749
4	−2.021 99	0.384 575 1
5	−2.976 295	−0.732 496
6	−0.418 013	1.267 645 6
7	−1.613 557	−0.809 631
8	−1.041 544	0.692 268 3
9	7.037 719 6	1.378 588 3
10	3.944 274 6	−2.806 731
11	4.368 435 3	−1.821 382
12	0.072 251 4	−0.598 317
13	0.512 410 2	−0.523 112
14	−1.189 847	−1.252 904
15	0.984 384 2	−0.362 783
16	−1.027 545	0.097 747 2
17	0.354 847 2	−0.300 258
18	−0.043 54	0.387 657 5
19	1.818 933 6	−1.314 742
20	0.137 497 9	0.718 906 3
21	−1.409 933	−0.610 274
22	−2.244 18	1.318 488 9
23	−0.021 156	2.400 203 4
24	−1.624 188	−0.406 119
25	−1.645 068	1.748 612
26	−3.408 248	−0.592 533
27	−3.068 128	−0.842 123
28	−1.542 767	−0.285 761

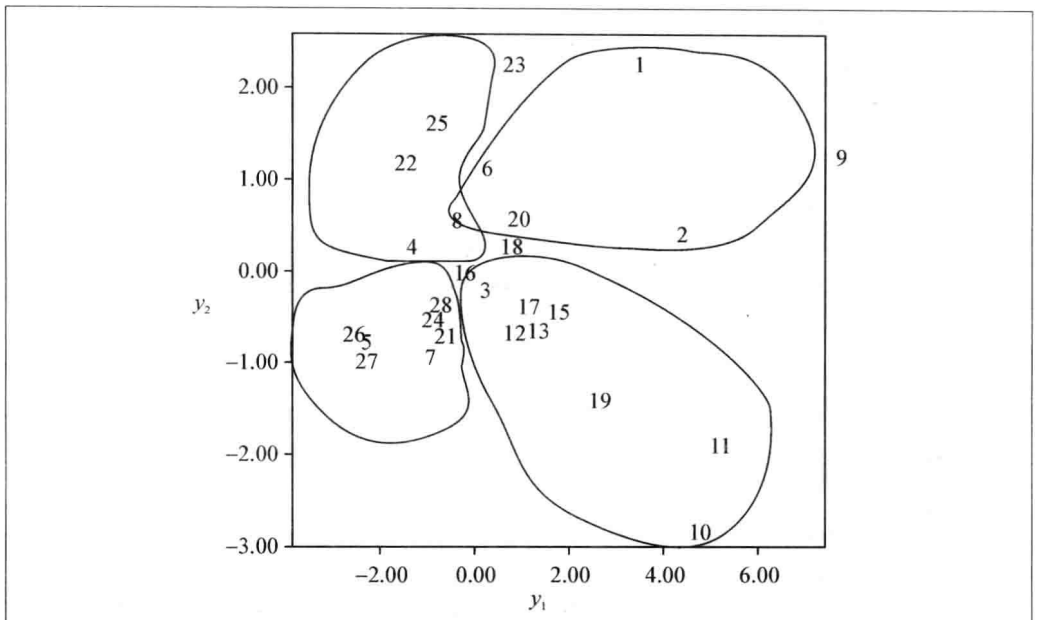


图 5—4

由图 5—4 可看出，分布在第一象限的是上海、北京、天津、广西 4 个省区，这 4 个省区的经济效益在全国来说属于比较好的，其中上海的经济效益最好。分布在第四象限的是江苏、浙江、安徽、福建、山东、湖北、广东 7 个省区。因为第四象限的主要特征是第一主成分，第一主成分占信息总量的比重最大，所以这 7 个省区的经济效益也算比较好。分布在第二象限和第三象限的地区可属同一类，经济效益较差。

例 5—3

全国重点水泥企业经济效益综合评价。

利用主成分综合评价全国重点水泥企业的经济效益。原始数据（数据来自 1984 年《中国统计年鉴》）见表 5—5。

表 5—5

厂家编号及指标		固定资产利税率	资金利税率	销售收入利税率	资金利润率	固定资产产值率	流动资金周转天数	万元产值能耗	全员劳动生产率
1	琉璃河	16.68	26.75	31.84	18.4	53.25	55	28.83	1.75
2	邯郸	19.7	27.56	32.94	19.2	59.82	55	32.92	2.87
3	大同	15.2	23.4	32.98	16.24	46.78	65	41.69	1.53
4	哈尔滨	7.29	8.97	21.3	4.76	34.39	62	39.28	1.63
5	华新	29.45	56.49	40.74	43.68	75.32	69	26.68	2.14
6	湘乡	32.93	42.78	47.98	33.87	66.46	50	32.87	2.6
7	柳州	25.39	37.82	36.76	27.56	68.18	63	35.79	2.43

续前表

厂家编号及指标		固定资产利税率	资金利税率	销售收入利税率	资金利润率	固定资产产值率	流动资金周转天数	万元产值能耗	全员劳动生产率
8	峨嵋	15.05	19.49	27.21	14.21	6.13	76	35.76	1.75
9	耀县	19.82	28.78	33.41	20.17	59.25	71	39.13	1.83
10	永登	21.13	35.2	39.16	26.52	52.47	62	35.08	1.73
11	工源	16.75	28.72	29.62	19.23	55.76	58	30.08	1.52
12	抚顺	15.83	28.03	26.4	17.43	61.19	61	32.75	1.6
13	大连	16.53	29.73	32.49	20.63	50.41	69	37.57	1.31
14	江南	22.24	54.59	31.05	37	67.95	63	32.33	1.57
15	江油	12.92	20.82	25.12	12.54	51.07	66	39.18	1.83

将指标“流动资金周转天数”和“万元产值能耗”取倒数，经标准化后的数据取名为“重点水泥厂”，如表 5—6 所示。

表 5—6

x1	x2	x3	x4	x5	x6	x7	x8
−0.376 75	−0.357 95	−0.113 56	−0.366 69	−0.038 79	1.193 347	1.426 821	−0.277 12
0.088 158	−0.293 88	0.050 803	−0.287 32	0.356 157	1.193 347	0.289 035	2.253 119
−0.604 58	−0.622 92	0.056 779	−0.581	−0.427 73	−0.373 28	−1.398 1	−0.774 13
−1.822 27	−1.764 31	−1.688 44	−1.719 97	−1.172 55	0.043 644	−1.009 54	−0.548 22
1.589 096	1.994 436	1.216 277	2.141 428	1.287 927	−0.872 79	2.164 81	0.603 944
2.124 815	0.909 999	2.298 075	1.168 142	0.755 315	2.211 656	0.301 235	1.643 15
0.964 09	0.517 672	0.621 586	0.542 104	0.858 711	−0.099 74	−0.354 1	1.259 096
−0.627 67	−0.932 2	−0.805 37	−0.782 4	−2.871 37	−1.620 4	−0.347 91	−0.277 12
0.106 631	−0.197 38	0.121 03	−0.191 09	0.321 892	−1.101 43	−0.983 78	−0.096 39
0.308 295	0.310 434	0.980 193	0.438 921	−0.085 68	0.043 644	−0.204 79	−0.322 3
−0.365 97	−0.202 12	−0.445 27	−0.284 35	0.112 093	0.666 636	1.046 255	−0.796 72
−0.507 6	−0.256 7	−0.926 4	−0.462 93	0.438 513	0.191 732	0.330 666	−0.615 99
−0.399 84	−0.122 23	−0.016 44	−0.145 45	−0.209 52	−0.872 79	−0.703 61	−1.271 14
0.479 171	1.844 149	−0.231 6	1.478 681	0.844 885	−0.099 74	0.435 397	−0.683 77
−0.955 57	−0.827	−1.117 66	−0.948 09	−0.169 84	−0.503 83	−0.992 39	−0.096 39

导入 SPSS 中计算出其相关矩阵，见输出结果 5—7。

输出结果 5—7　　**Correlation Matrix**

		x1	x2	x3	x4	x5	x6	x7	x8
Correlation	x1	1.000	.849	.923	.902	.651	.312	.489	.598
	x2	.849	1.000	.690	.988	.723	.107	.595	.265
	x3	.923	.690	1.000	.774	.544	.366	.342	.531
	x4	.902	.988	.774	1.000	.688	.121	.596	.329
	x5	.651	.723	.544	.688	1.000	.399	.442	.359
	x6	.312	.107	.366	.121	.399	1.000	.343	.480
	x7	.489	.595	.342	.596	.442	.343	1.000	.226
	x8	.598	.265	.531	.329	.359	.480	.226	1.000

在确定主成分个数之前，采用与例5—2相同的SPSS操作，得到软件输出结果5—8。

输出结果 5—8 **Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.861	60.758	60.758	4.861	60.758	60.758
2	1.269	15.865	76.623	1.269	15.865	76.623
3	.837	10.463	87.085			
4	.517	6.464	93.549			
5	.378	4.727	98.276			
6	.115	1.443	99.719			
7	.021	.264	99.984			
8	.001	.016	100.000			

Extraction Method: Principal Component Analysis.

从上表可看出，前3个主成分解释了全部方差的87.085%，即包含原始数据的信息总量达到了87.085%，这说明前3个主成分代表原来的8个指标评价企业的经济效益已经有足够的把握。设这3个主成分分别用 y_1，y_2，y_3 来表示，按照例5—2的操作，只不过在点击Extraction按钮时，在Fixed Number of Factors中填写3，即可得到相关矩阵的前3个特征根的特征向量，如输出结果5—9所示。

输出结果 5—9 **Component Matrix[a]**

	Component		
	1	2	3
x1	.957	−.019	−.239
x2	.899	−.396	.037
x3	.862	.081	−.338
x4	.928	−.350	−.038
x5	.787	.000	.182
x6	.422	.773	.345
x7	.640	−.078	.642
x8	.571	.615	−.313

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

对输出结果5—9的第 i 列的每个元素分别除以第 i 个特征根的平方根 $\sqrt{\lambda_i}$，就得到主成分分析的第 i 个主成分的系数。结果如表5—7所示。

表 5—7

	主成分 1	主成分 2	主成分 3
x1	0.434 067	−0.016 487	−0.261 157
x2	0.407 766	−0.351 123	0.040 662
x3	0.390 911	0.072 240 8	−0.369 59
x4	0.420 727	−0.311 078	−0.041 12

续前表

	主成分 1	主成分 2	主成分 3
x5	0.356 854	0.000 221 4	0.198 903
x6	0.191 631	0.686 453	0.377 485
x7	0.290 517	−0.069 141	0.701 934
x8	0.258 889	0.546 309 6	−0.341 87

由上表可得 3 个主成分的线性组合如下：

$$\begin{aligned} y_1 =& 0.434\,067x_1^* + 0.407\,766x_2^* + 0.390\,911x_3^* + 0.420\,727x_4^* \\ &+ 0.356\,854x_5^* + 0.191\,631x_6^* + 0.290\,517x_7^* + 0.258\,889x_8^* \\ y_2 =& -0.016\,487x_1^* - 0.351\,123x_2^* + 0.072\,240\,8x_3^* - 0.311\,078x_4^* \\ &+ 0.000\,221\,4x_5^* + 0.686\,453x_6^* - 0.069\,141x_7^* + 0.546\,309\,6x_8^* \\ y_3 =& -0.261\,157x_1^* + 0.040\,662x_2^* - 0.369\,59x_3^* - 0.041\,12x_4^* \\ &+ 0.198\,903x_5^* + 0.377\,485x_6^* + 0.701\,934x_7^* - 0.341\,87x_8^* \end{aligned} \tag{5.13}$$

其中，x_1^*，x_2^*，x_3^*，x_4^*，x_5^*，x_6^*，x_7^*，x_8^* 表示对原始变量标准化后的变量。

主成分的经济意义由各线性组合中权数较大的几个指标的综合意义来确定。综合因子 y_1 中，x_1^*，x_2^*，x_3^*，x_4^* 的系数远大于其他变量的系数，所以，y_1 主要是固定资产利税率、资金利税率、销售收入利税率、资金利润率这 4 个指标的综合反映，它代表经济效益的盈利方面，刻画了企业的盈利能力。因为由 y_1 来评价企业的经济效益已有 60.76%的把握，所以这 4 项指标是反映企业经济效益的主要指标。同时，从 y_1 的线性组合中可以看到，前 4 个单项指标在综合因子 y_1 中所占的比重相当，说明这 4 项指标用于考核评价企业经济效益，每一项都是必不可少的。y_2 主要是流动资金周转天数和全员劳动生产率的综合反映，它标志着企业的资金和人力的利用水平，以资金和个人的利用率作用于企业的经济效益。资金和人力利用得好，劳动生产率就提高，资金周转就加快，从而提高企业经济效益。y_3 主要反映万元产值能耗，从改进生产工艺、勤俭节约方面作用于企业经济效益。这 3 个综合因子从三个影响企业经济效益的主要方面刻画企业经济效益，用它们来考核企业经济效益具有 87.085%的可靠性。

关于用样本主成分得分进行排序的问题，目前常用的方法是利用主成分 y_1，y_2，…，y_m 做线性组合，并以每个主成分 y_k 的方差贡献率 α_k 作为权数构造一个综合评价函数：$F=\alpha_1 y_1+\alpha_2 y_2+\cdots+\alpha_m y_m$，依据计算出的 F 值大小进行排序或分类划级。这一方法目前在一些专业文献中都有介绍，但在实践中有时应用效果并不理想，一直以来存在较大争议，主要原因是产生主成分的特征向量的各级分量符号不一致，很难进行排序评价。因此，有下面的改进：只用第一主成分作评价指数，理由是第一主成分与原变量综合相关度最强，并且第一主成分 y_1 对应于数据变异最大的方向，也就是使数据信息损失最小、精度最高的一维综合变量。但值得指出的是，使用这种方法是有前提条件的，即当主成分系数全为正的时候，也就是要求所有评价指标变量都是正相关的时候，第一主成分才可以用来进行排序。如果

系数中有正有负或近似为零，则说明第一主成分是无序指数，不能用来作为排序评价指数。而如果第一主成分系数全为正，则第二、第三……主成分由于与第一主成分正交，系数肯定有正有负，因而一般来说均为无序指数，不能用来作为排序评价指数。

依据第一主成分得分对各个水泥企业经济效益做综合评价，将标准化后的原始数据代入式（5.13）的第一个表达式中，计算出各样品的第一主成分得分并排名，如表 5—8 所示。

表 5—8

	$\hat{y}_1$	名次
琉璃河	0.049 451	7
邯郸	0.840 47	5
大同	−1.569 43	12
哈尔滨	−3.739 37	15
华新	3.957 057	1
湘乡	3.889 461	2
柳州	1.611 051	4
峨嵋	−2.804 58	14
耀县	−0.474 25	9
永登	0.663 081	6
工源	−0.269 52	8
抚顺	−0.752 11	10
大连	−1.066 54	11
江南	1.723 424	3
江油	−2.058 21	13

在表 5—8 的经济效益得分中，有许多企业的得分是负数，但并不表明企业的经济效益就为负，这里的正负仅表示该企业与平均水平的位置关系，企业经济效益的平均水平算作零点，这是我们在整个过程中将数据标准化的结果。

从表 5—8 可看到，华新水泥厂的综合经济效益最好，是第一名；湘乡水泥厂的综合经济效益为第二名；哈尔滨水泥厂的综合经济效益最差。

虽然此处可以根据各公司的主成分得分对各公司运营情况进行一些比较分析或分类研究，但因此处主成分的意义不十分明朗，我们把更深入的分析放到下一章，以期得到更合理、更容易解释的结果。

主成分分析一个非常重要的应用是解决回归建模的多重共线性问题，即所谓的主成分回归，可参见参考文献［6］。

□ 参考文献

［1］张尧庭，方开泰．多元分析引论．北京：科学出版社，1982

[2] 方开泰. 实用多元分析. 上海：华东师范大学出版社，1989

[3] 王静龙. 多元统计分析. 北京：科学出版社，2008

[4] 王惠文. 偏最小二乘回归方法及应用. 北京：国防工业出版社，1999

[5] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York, Inc.，1986

[6] 何晓群，刘文卿. 应用回归分析（第三版）. 北京：中国人民大学出版社，2011

[7] 何晓群. 多元统计分析（第二版）. 北京：中国人民大学出版社，2008

□ 思考与练习

1. 主成分的基本思想是什么？
2. 主成分在应用中的主要作用是什么？
3. 由协方差阵出发和由相关阵出发求主成分有什么不同？
4. 读者自己找一个实际问题的数据，应用 SPSS 软件试做主成分分析。

第 6 章

Chapter 6 因子分析

学习目标

1. 理解因子分析方法的思想；
2. 了解因子分析的基本理论；
3. 掌握求解因子的方法步骤；
4. 分辨因子分析与主成分分析的异同；
5. 能够用 SPSS 软件进行因子分析，并正确理解系统输出结果。

因子分析（factor analysis）模型是主成分分析的推广。它也是利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。相比主成分分析，因子分析更倾向于描述原始变量之间的相关关系，因此，因子分析的出发点是原始变量的相关矩阵。因子分析的思想始于 1904 年查尔斯·斯皮尔曼（Charles Spearman）对学生考试成绩的研究。近年来，随着电子计算机的高速发展，人们将因子分析的理论成功地应用于心理学、医学、气象、地质、经济学等各个领域，也使得因子分析的理论和方法更加丰富。本章主要介绍因子分析的基本理论及方法、运用因子分析方法分析实际问题的主要步骤及因子分析的上机实现等内容。

6.1 因子分析的基本理论

6.1.1 因子分析的基本思想

因子分析的基本思想是根据相关性大小把原始变量分组，使得同组内的变量之

间相关性较高，而不同组的变量间的相关性则较低。每组变量代表一个基本结构，并用一个不可观测的综合变量表示，这个基本结构就称为公共因子。对于所研究的某一具体问题，原始变量可以分解成两部分之和的形式，一部分是少数几个不可测的所谓公共因子的线性函数，另一部分是与公共因子无关的特殊因子。在经济统计中，描述一种经济现象的指标可以有很多，比如要反映物价的变动情况，对各种商品的价格做全面调查固然可以达到目的，但这样做显然耗时耗力，为实际工作者所不取。实际上，某一类商品中很多商品的价格之间存在明显的相关性或相互依赖性，只要选择几种主要商品的价格，进而对这几种主要商品的价格进行综合，得到某一种假想的“综合商品”的价格，就足以反映某一类物价的变动情况，这里，“综合商品”的价格就是提取出来的因子。这样，对各类商品物价或仅对主要类别商品的物价进行类似分析然后加以综合，就可以反映出物价的整体变动情况。这一过程也就是从一些有错综复杂关系的经济现象中找出少数几个主要因子，每一个主要因子代表经济变量间相互依赖的一种经济作用。抓住这些主要因子就可以帮助我们对复杂的经济问题进行分析和解释。

因子分析还可用于对变量或样品的分类处理，我们在得出因子的表达式之后，可以把原始变量的数据代入表达式得出因子得分值，根据因子得分在因子所构成的空间中把变量或样品点画出来，形象直观地达到分类的目的。

因子分析不仅可以用来研究变量之间的相关关系，还可以用来研究样品之间的相关关系，通常将前者称为 R 型因子分析，后者称为 Q 型因子分析。下面着重介绍 R 型因子分析。

6.1.2　因子分析的基本理论及模型

1. 查尔斯·斯皮尔曼提出因子分析时用到的例子

为了对因子分析的基本理论有一个完整的认识，我们先给出查尔斯·斯皮尔曼 1904 年用到的例子。斯皮尔曼在该例中研究了 33 名学生古典语（C）、法语（F）、英语（E）、数学（M）、判别（D）和音乐（Mu）6 门考试成绩之间的相关性，并得到如下相关矩阵：

$$
\begin{array}{c}
\\
\mathrm{C} \\
\mathrm{F} \\
\mathrm{E} \\
\mathrm{M} \\
\mathrm{D} \\
\mathrm{Mu}
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
\mathrm{C} & \mathrm{F} & \mathrm{E} & \mathrm{M} & \mathrm{D} & \mathrm{Mu}
\end{array} \\
\left[\begin{array}{cccccc}
1.00 & 0.83 & 0.78 & 0.70 & 0.66 & 0.63 \\
0.83 & 1.00 & 0.67 & 0.67 & 0.65 & 0.57 \\
0.78 & 0.67 & 1.00 & 0.64 & 0.54 & 0.51 \\
0.70 & 0.67 & 0.64 & 1.00 & 0.45 & 0.51 \\
0.66 & 0.65 & 0.54 & 0.45 & 1.00 & 0.40 \\
0.63 & 0.57 & 0.51 & 0.51 & 0.40 & 1.00
\end{array}\right]
\end{array}
$$

斯皮尔曼注意到上面相关矩阵中一个有趣的规律，即如果不考虑对角元素的话，任意两列的元素大致成比例，对 C 列和 E 列有

$$\frac{0.83}{0.67}\approx\frac{0.70}{0.64}\approx\frac{0.66}{0.54}\approx\frac{0.63}{0.51}\approx 1.2$$

于是斯皮尔曼指出每一科目的考试成绩都遵从以下形式：

$$X_i=a_iF+e_i \tag{6.1}$$

式中，X_i 为第 i 门科目标准化后的考试成绩，均值为 0，方差为 1；F 为公共因子，对各科考试成绩均有影响，也是均值为 0，方差为 1；e_i 为仅对第 i 门科目考试成绩有影响的特殊因子，F 与 e_i 相互独立。也就是说，每一门科目的考试成绩都可以看作一个公共因子（可以认为是一般智力）与一个特殊因子的和。在满足以上假定的条件下，就有

$$\operatorname{cov}(X_i,X_j)=E[(a_iF+e_i)(a_jF+e_j)]=a_ia_j\operatorname{var}(F)=a_ia_j$$

于是，有

$$\frac{\operatorname{cov}(X_i,X_j)}{\operatorname{cov}(X_i,X_k)}=\frac{a_j}{a_k} \tag{6.2}$$

式（6.2）与 i 无关，与在相关矩阵中所观察到的比例关系相一致。

此外，还可以得到如下有关 X_i 方差的关系式：

$$\begin{aligned}\operatorname{var}(X_i)&=\operatorname{var}(a_iF+e_i)=\operatorname{var}(a_iF)+\operatorname{var}(e_i)\\&=a_i^2\operatorname{var}(F)+\operatorname{var}(e_i)\\&=a_i^2+\operatorname{var}(e_i)\end{aligned}$$

因为 a_i 是一个常数，F 与 e_i 相互独立，且 F 与 X_i 的方差均被假定为 1，于是有

$$1=a_i^2+\operatorname{var}(e_i) \tag{6.3}$$

因此，常数 a_i 的意义就在于其平方表示了公共因子 F 解释 X_i 方差的比例，因此称为因子载荷，而 a_i^2 称为共同度。

对斯皮尔曼的例子进行推广，假定每一门科目的考试成绩都受到 m 个公共因子的影响及一个特殊因子的影响，于是式（6.1）就变成了如下因子分析模型的一般形式：

$$X_i=a_{i1}F_1+a_{i2}F_2+\cdots+a_{im}F_m+e_i \tag{6.4}$$

式中，X_i 为标准化后的第 i 门科目的考试成绩，均值为 0，方差为 1；F_1，F_2，…，F_m 是彼此独立的公共因子，都满足均值为 0，方差为 1；e_i 为特殊因子，与每一个公共因子均不相关且均值为 0；a_{i1}，a_{i2}，…，a_{im} 为对第 i 门科目考试成绩的因子载荷。对该模型，有

$$\operatorname{var}(X_i)=a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2+\operatorname{var}(e_i)=1 \tag{6.5}$$

式中，$a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2$ 表示公共因子解释 X_i 方差的比例，称为 X_i 的共同度；相

对的，$\mathrm{var}(e_i)$可称为 X_i 的特殊度或剩余方差，表示 X_i 的方差中与公共因子无关的部分。因为共同度不会大于 1，因此，$-1\leqslant a_{ij}\leqslant 1$。由模型（6.4）还可以很容易地得到如下 X_i 与 X_j 相关系数的关系式：

$$r_{ij}=a_{i1}a_{j1}+a_{i2}a_{j2}+\cdots+a_{im}a_{jm} \tag{6.6}$$

所以当 X_i 与 X_j 在某一公共因子上的载荷均较大时，也就表明了 X_i 与 X_j 的相关性较强。

2. 一般因子分析模型

下面我们给出更为一般的因子分析模型：设有 n 个样品，每个样品观测 p 个指标，这 p 个指标之间有较强的相关性（要求 p 个指标相关性较强的理由是很明确的，只有相关性较强，才能从原始变量中提取出“公共”因子）。为了便于研究，并消除由于观测量纲的差异及数量级不同所造成的影响，对样本观测数据进行标准化处理，使标准化后的变量均值为 0，方差为 1。为方便，把原始变量及标准化后的变量向量均用 $\boldsymbol{X}$ 表示，用 F_1，F_2，…，F_m（$m<p$）表示标准化的公共因子。如果：

（1）$\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$是可观测随机向量，且均值向量 $E(\boldsymbol{X})=\boldsymbol{0}$，协方差矩阵 $\mathrm{cov}(\boldsymbol{X})=\boldsymbol{\Sigma}$，且协方差矩阵 $\boldsymbol{\Sigma}$ 与相关阵 $\boldsymbol{R}$ 相等；

（2）$\boldsymbol{F}=(F_1, F_2, \cdots, F_m)'$（$m<p$）是不可观测的变量，其均值向量 $E(\boldsymbol{F})=\boldsymbol{0}$，协方差矩阵 $\mathrm{cov}(\boldsymbol{F})=\boldsymbol{I}$，即向量 $\boldsymbol{F}$ 的各分量是相互独立的；

（3）$\boldsymbol{\varepsilon}=(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$与 $\boldsymbol{F}$ 相互独立，且 $E(\boldsymbol{\varepsilon})=\boldsymbol{0}$，$\boldsymbol{\varepsilon}$ 的协方差矩阵 $\boldsymbol{\Sigma}_\varepsilon$ 是对角方阵

$$\mathrm{cov}(\boldsymbol{\varepsilon})=\boldsymbol{\Sigma}_\varepsilon=\begin{bmatrix}\sigma_{11}^2 & & & 0\\ & \sigma_{22}^2 & & \\ & & \ddots & \\ 0 & & & \sigma_{pp}^2\end{bmatrix}$$

即 $\boldsymbol{\varepsilon}$ 的各分量之间也是相互独立的，则模型

$$\begin{cases}X_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1\\ X_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\varepsilon_2\\ \cdots\cdots\cdots\\ X_p=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\varepsilon_p\end{cases} \tag{6.7}$$

称为因子模型。模型（6.7）的矩阵形式为：

$$\boldsymbol{X}=\boldsymbol{AF}+\boldsymbol{\varepsilon} \tag{6.8}$$

其中　$\boldsymbol{A}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1m}\\ a_{21} & a_{22} & \cdots & a_{2m}\\ \vdots & \vdots & & \vdots\\ a_{p1} & a_{p2} & \cdots & a_{pm}\end{bmatrix}$

由模型（6.7）及其假设前提知，公共因子 F_1，F_2，…，F_m 相互独立且不可测，是在原始变量的表达式中都出现的因子。公共因子的含义必须结合实际问题的具体意义确定。ε_1，ε_2，…，ε_p 叫做特殊因子，是向量 $\boldsymbol{X}$ 的分量 $X_i(i=1,2,\cdots,p)$ 所特有的因子。各特殊因子之间以及特殊因子与所有公共因子之间也都是相互独立的。矩阵 $\boldsymbol{A}$ 中的元素 a_{ij} 称为因子载荷，a_{ij} 的绝对值越大（$|a_{ij}|\leqslant 1$），表明 X_i 与 F_j 的相依程度越大，或称公共因子 F_j 对于 X_i 的载荷量越大，进行因子分析的目的之一就是要求出各个因子载荷的值。经过后面的分析会看到，因子载荷的概念与上一章主成分分析中的因子负荷量相对等，实际上，由于因子分析与主成分分析非常类似，在模型（6.7）中，若把 ε_i 看做 $a_{i,m+1}F_{m+1}+a_{i,m+2}F_{m+2}+\cdots+a_{i,p}F_p$ 的综合作用，则除了此处的因子为不可测变量这一区别，因子载荷与主成分分析中的因子负荷量是一致的。很多人对这两个概念并不加以区分而都称作因子载荷。矩阵 $\boldsymbol{A}$ 称为因子载荷矩阵。

为了更好地理解因子分析方法，有必要讨论一下载荷矩阵 $\boldsymbol{A}$ 的统计意义以及公共因子与原始变量之间的关系。

（1）因子载荷 a_{ij} 的统计意义。由模型（6.7）

$$\begin{aligned}\operatorname{cov}(X_i,F_j)&=\operatorname{cov}\Big(\sum_{j=1}^{m}a_{ij}F_j+\varepsilon_i,F_j\Big)\\&=\operatorname{cov}\Big(\sum_{j=1}^{m}a_{ij}F_j,F_j\Big)+\operatorname{cov}(\varepsilon_i,F_j)\\&=a_{ij}\end{aligned}$$

即 a_{ij} 是 X_i 与 F_j 的协方差，而注意到，X_i 与 F_j（$i=1,2,\cdots,p$；$j=1,2,\cdots,m$）都是均值为 0，方差为 1 的变量，因此，a_{ij} 同时也是 X_i 与 F_j 的相关系数。请读者对比主成分分析一章有关因子负荷量的论述并对两者进行比较。

（2）变量共同度与剩余方差。在上面斯皮尔曼的例子中我们提到了共同度与剩余方差的概念，对一般因子模型（6.7）的情况，我们重新总结这两个概念如下：

称 $a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2$ 为变量 X_i 的共同度，记为 $h_i^2(i=1,2,\cdots,p)$。由因子分析模型的假设前提，易得

$$\operatorname{var}(X_i)=1=h_i^2+\operatorname{var}(\varepsilon_i) \tag{6.9}$$

记 $\operatorname{var}(\varepsilon_i)=\sigma_i^2$，则

$$\operatorname{var}(X_i)=1=h_i^2+\sigma_i^2 \tag{6.10}$$

上式表明共同度 h_i^2 与剩余方差 σ_i^2 有互补的关系，h_i^2 越大，表明 X_i 对公共因子的依赖程度越大，公共因子能解释 X_i 方差的比例越大，因子分析的效果也就越好。

（3）公共因子 F_j 的方差贡献。共同度考虑的是所有公共因子 F_1，F_2，…，F_m 与某一个原始变量的关系，与此类似，考虑某一个公共因子 F_j 与所有原始变量 X_1，X_2，…，X_p 的关系。

记 $g_j^2=a_{1j}^2+a_{2j}^2+\cdots+a_{pj}^2(j=1,2,\cdots,m)$，则 g_j^2 表示的是公共因子 F_j 对于 $\boldsymbol{X}$ 的每一分量 $X_i(i=1,2,\cdots,p)$ 所提供的方差的总和，称为公共因子 F_j 对原始变量向量 $\boldsymbol{X}$ 的方差贡献，它是衡量公共因子相对重要性的指标。g_j^2 越大，表明公共因子 F_j 对 $\boldsymbol{X}$ 的贡献越大，或者说对 $\boldsymbol{X}$ 的影响和作用就越大。如果将因子载荷矩阵 $\boldsymbol{A}$ 的所有 $g_j^2(j=1,2,\cdots,m)$ 都计算出来，并按其大小排序，就可以依此提炼出最有影响的公共因子。

6.2　因子载荷的求解

因子分析可以分为确定因子载荷、因子旋转及计算因子得分三个步骤。首要的步骤即为确定因子载荷或者根据样本数据确定因子载荷矩阵 $\boldsymbol{A}$。有很多方法可以完成这项工作，如主成分法、主轴因子法、最小二乘法、极大似然法、α 因子提取法等。这些方法求解因子载荷的出发点不同，所得的结果也不完全相同。下面着重介绍比较常用的主成分法、主轴因子法与极大似然法。

6.2.1　主成分法

用主成分法确定因子载荷是在进行因子分析之前先对数据进行一次主成分分析，然后把前几个主成分作为未旋转的公共因子。相对于其他确定因子载荷的方法而言，主成分法比较简单。但是，由于用这种方法所得的特殊因子 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$ 之间并不相互独立，因此，用主成分法确定因子载荷不完全符合因子模型的假设前提，也就是说所得的因子载荷并不完全正确。当共同度较大时，特殊因子所起的作用较小，特殊因子之间的相关性所带来的影响几乎可以忽略。事实上，很多有经验的分析人员在进行因子分析时，总是先用主成分法进行分析，然后再尝试其他的方法。

用主成分法寻找公共因子的方法如下：假定从相关阵出发求解主成分，设有 p 个变量，则可以找出 p 个主成分。将所得的 p 个主成分按由大到小的顺序排列，记为 $Y_1,Y_2,\cdots,Y_p$，则主成分与原始变量之间存在如下关系式：

$$\begin{cases} Y_1=\gamma_{11}X_1+\gamma_{12}X_2+\cdots+\gamma_{1p}X_p \\ Y_2=\gamma_{21}X_1+\gamma_{22}X_2+\cdots+\gamma_{2p}X_p \\ \cdots\cdots\cdots\cdots \\ Y_p=\gamma_{p1}X_1+\gamma_{p2}X_2+\cdots+\gamma_{pp}X_p \end{cases} \tag{6.11}$$

式中，γ_{ij} 为随机向量 $\boldsymbol{X}$ 的相关矩阵的特征根所对应的特征向量的分量，因为特征向量之间彼此正交，从 $\boldsymbol{X}$ 到 $\boldsymbol{Y}$ 的转换关系是可逆的，很容易得出由 $\boldsymbol{Y}$ 到 $\boldsymbol{X}$ 的转换关系为：

$$\begin{cases} X_1=\gamma_{11}Y_1+\gamma_{21}Y_2+\cdots+\gamma_{p1}Y_p \\ X_2=\gamma_{12}Y_1+\gamma_{22}Y_2+\cdots+\gamma_{p2}Y_p \\ \cdots\cdots\cdots\cdots \\ X_p=\gamma_{1p}Y_1+\gamma_{2p}Y_2+\cdots+\gamma_{pp}Y_p \end{cases} \tag{6.12}$$

对上面每一等式只保留前 m 个主成分而把后面的部分用 ε_i 代替，则式（6.12）转化为：

$$\begin{cases} X_1=\gamma_{11}Y_1+\gamma_{21}Y_2+\cdots+\gamma_{m1}Y_m+\varepsilon_1 \\ X_2=\gamma_{12}Y_1+\gamma_{22}Y_2+\cdots+\gamma_{m2}Y_m+\varepsilon_2 \\ \cdots\cdots\cdots\cdots \\ X_p=\gamma_{1p}Y_1+\gamma_{2p}Y_2+\cdots+\gamma_{mp}Y_m+\varepsilon_p \end{cases} \tag{6.13}$$

式（6.13）在形式上已经与因子模型（6.7）相一致，并且 $Y_i(i=1,2,\cdots,m)$ 之间相互独立，Y_i 与 ε_i 之间相互独立。为了把 Y_i 转化成合适的公共因子，现在要做的工作只是把主成分 Y_i 变成方差为 1 的变量。为完成此变换，必须将 Y_i 除以其标准差，由上一章主成分分析的知识知其标准差即为特征根的平方根 $\sqrt{\lambda_i}$。于是，令 $F_i=Y_i/\sqrt{\lambda_i}$，$a_{ij}=\sqrt{\lambda_j}\gamma_{ji}$，则式（6.13）变为：

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

这与因子模型（6.7）完全一致，这样，就得到了载荷矩阵 $\boldsymbol{A}$ 和一组初始公共因子（未旋转）。

一般设 $\lambda_1,\lambda_2,\cdots,\lambda_p$（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$）为样本相关阵 $\boldsymbol{R}$ 的特征根，$\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_p$ 为对应的标准正交化特征向量。设 $m<p$，则因子载荷矩阵 $\boldsymbol{A}$ 的一个解为：

$$\hat{\boldsymbol{A}}=(\sqrt{\lambda_1}\boldsymbol{\gamma}_1,\sqrt{\lambda_2}\boldsymbol{\gamma}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{\gamma}_m) \tag{6.14}$$

共同度的估计为：

$$\hat{h}_i^2=\hat{a}_{i1}^2+\hat{a}_{i2}^2+\cdots+\hat{a}_{im}^2 \tag{6.15}$$

那么如何确定公共因子的数目 m 呢？一般而言，这取决于问题的研究者本人。对于同一问题进行因子分析时，不同的研究者可能会给出不同的公共因子数。当然，有时候由数据本身的特征可以很明确地确定因子数目。当用主成分法进行因子分析时，也可以借鉴确定主成分个数的准则，如所选取的公共因子的信息量的和达到总体信息量的一个合适比例为止。但对这些准则不应生搬硬套，应具体问题具体分析，总之要使所选取的公共因子能够合理地描述原始变量相关阵的结构，同时要有利于因子模型的解释。

6.2.2　主轴因子法

主轴因子法也比较简单，而且在实际应用中比较普遍。用主轴因子法求解因子载荷矩阵的方法，其思路与主成分法有类似的地方，两者均是从分析矩阵的结构入手，不同的地方在于，主成分法是在所有的 p 个主成分都能解释标准化原始变量所有方差的基础之上进行分析的，而主轴因子法中，假定 m 个公共因子只能解释原始变量的部分方差，利用公共因子方差（或共同度）来代替相关矩阵主对角线上的元素 1，并以这个新得到的矩阵（称为调整相关矩阵）为出发点，对其分别求解特征根与特征向量，从而得到因子解。

在因子模型（6.7）中，不难得到如下关于 $\boldsymbol{X}$ 的相关矩阵 $\boldsymbol{R}$ 的关系式：

$$\boldsymbol{R}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{\Sigma}_{\varepsilon}$$

式中，$\boldsymbol{A}$ 为因子载荷矩阵；$\boldsymbol{\Sigma}_{\varepsilon}$ 为对角阵，其对角元素为相应特殊因子的方差。则称 $\boldsymbol{R}^{*}=\boldsymbol{R}-\boldsymbol{\Sigma}_{\varepsilon}=\boldsymbol{A}\boldsymbol{A}'$ 为调整相关矩阵，显然 $\boldsymbol{R}^{*}$ 的主对角元素不再是 1，而是共同度 h_i^2。分别求解 $\boldsymbol{R}^{*}$ 的特征根与标准正交特征向量，进而求出因子载荷矩阵 $\boldsymbol{A}$。此时，$\boldsymbol{R}^{*}$ 有 m 个正的特征根。设 λ_1^{*}，λ_2^{*}，…，λ_m^{*}（$\lambda_1^{*}\geqslant\lambda_2^{*}\geqslant\cdots\geqslant\lambda_m^{*}$）为 $\boldsymbol{R}^{*}$ 的特征根，$\boldsymbol{\gamma}_1^{*}$，$\boldsymbol{\gamma}_2^{*}$，…，$\boldsymbol{\gamma}_m^{*}$ 为对应的标准正交化特征向量。$m<p$，则因子载荷矩阵 $\boldsymbol{A}$ 的一个主轴因子解为：

$$\hat{\boldsymbol{A}}=(\sqrt{\lambda_1^{*}}\,\boldsymbol{\gamma}_1^{*},\sqrt{\lambda_2^{*}}\,\boldsymbol{\gamma}_2^{*},\cdots,\sqrt{\lambda_m^{*}}\,\boldsymbol{\gamma}_m^{*}) \tag{6.16}$$

注意到，上面的分析是以首先得到调整相关矩阵 $\boldsymbol{R}^{*}$ 为基础的，而实际上，$\boldsymbol{R}^{*}$ 与共同度（或相对的剩余方差）都是未知的，需要先进行估计。一般先给出一个初始估计，然后估计出载荷矩阵 $\boldsymbol{A}$，再给出较好的共同度或剩余方差的估计。得到初始估计的方法有很多，可尝试对原始变量先进行一次主成分分析，给出初始估计值。

6.2.3　极大似然法

如果假定公共因子 $\boldsymbol{F}$ 和特殊因子 $\boldsymbol{\varepsilon}$ 服从正态分布，则能够得到因子载荷和特殊因子方差的极大似然估计。设 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 为来自正态总体 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的随机样本，其中 $\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{\Sigma}_{\varepsilon}$。从似然函数的理论知

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{np/2}\,|\,\boldsymbol{\Sigma}\,|^{n/2}}\mathrm{e}^{-1/2\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}[\sum_{j=1}^{n}(\boldsymbol{X}_j-\bar{\boldsymbol{X}})(\boldsymbol{X}_j-\bar{\boldsymbol{X}})'+n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})(\bar{\boldsymbol{X}}-\boldsymbol{\mu})']\}} \tag{6.17}$$

它通过 $\boldsymbol{\Sigma}$ 依赖于 $\boldsymbol{A}$ 和 $\boldsymbol{\Sigma}_{\varepsilon}$。但式（6.17）并不能唯一确定 $\boldsymbol{A}$，为此，添加如下条件：

$$\boldsymbol{A}'\boldsymbol{\Sigma}_{\varepsilon}^{-1}\boldsymbol{A}=\boldsymbol{\Lambda} \tag{6.18}$$

这里，$\boldsymbol{\Lambda}$ 是一个对角阵，用数值极大化的方法可以得到极大似然估计 $\hat{\boldsymbol{A}}$ 和

$\hat{\boldsymbol{\Sigma}}_\varepsilon$。极大似然估计 $\hat{\boldsymbol{A}}$，$\hat{\boldsymbol{\Sigma}}_\varepsilon$ 和 $\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{X}}$，将使 $\hat{\boldsymbol{A}}'\hat{\boldsymbol{\Sigma}}_\varepsilon^{-1}\hat{\boldsymbol{A}}$ 为对角阵，且使式（6.17）达到最大。

6.2.4 因子旋转

不管用何种方法确定初始因子载荷矩阵 $\boldsymbol{A}$，它们都不是唯一的。设 F_1，F_2，…，F_m 是初始公共因子，则可以建立它们的如下线性组合得到新的一组公共因子 F_1'，F_2'，…，F_m'，使得 F_1'，F_2'，…，F_m' 彼此相互独立，同时也能很好地解释原始变量之间的相关关系。

$$
\begin{aligned}
F_1' &= d_{11}F_1 + d_{12}F_2 + \cdots + d_{1m}F_m \\
F_2' &= d_{21}F_1 + d_{22}F_2 + \cdots + d_{2m}F_m \\
&\cdots\cdots\cdots \\
F_m' &= d_{m1}F_1 + d_{m2}F_2 + \cdots + d_{mm}F_m
\end{aligned}
$$

这样的线性组合可以找到无数组，由此便引出了因子分析的第二个步骤——因子旋转。建立因子分析模型的目的不仅在于找到公共因子，更重要的是知道每一个公共因子的意义，以便对实际问题进行分析。然而，我们得到的初始因子解各主因子的典型代表变量不是很突出，容易使因子的意义含糊不清，不便于对实际问题进行分析。出于这种考虑，可以对初始公共因子进行线性组合，即进行因子旋转，以期找到意义更为明确、实际意义更明显的公共因子。经过旋转后，公共因子对 X_i 的贡献 h_i^2 并不改变，但由于载荷矩阵发生变化，公共因子本身就可能发生很大的变化，每一个公共因子对原始变量的贡献 g_i^2 不再与原来相同，经过适当的旋转，我们就可以得到比较令人满意的公共因子。

因子旋转分为正交旋转与斜交旋转。正交旋转由初始载荷矩阵 $\boldsymbol{A}$ 右乘一正交阵得到。经过正交旋转而得到的新的公共因子仍然保持彼此独立的性质。而斜交旋转则放弃了因子之间彼此独立这个限制，因而可能达到更为简洁的形式，其实际意义也更容易解释。但不论是正交旋转还是斜交旋转，都应当使新的因子载荷系数要么尽可能地接近于零，要么尽可能地远离零。因为一个接近于零的载荷 a_{ij} 表明 X_i 与 F_j 的相关性很弱；而一个绝对值比较大的载荷 a_{ij} 则表明公共因子 F_j 在很大程度上解释了 X_i 的变化。这样，如果任一原始变量都与某些公共因子存在较强的相关关系，而与另外的公共因子几乎不相关的话，公共因子的实际意义就会比较容易确定。

对于一个具体问题做因子旋转，有时需要进行多次才能得到满意效果。每一次旋转后，矩阵各列平方的相对方差之和总会比上一次有所增加。如此继续下去，当总方差的改变不大时，就可以停止旋转，这样就得到了新的一组公共因子及相应的因子载荷矩阵，使得其各列元素平方的相对方差之和最大。

6.2.5　因子得分

当因子模型建立之后，往往需要反过来考察每一个样品的性质及样品之间的相互关系。比如当关于企业经济效益的因子模型建立之后，我们希望知道每一个企业经济效益的优劣，或者把诸企业划分归类，如哪些企业经济效益较好，哪些企业经济效益一般，哪些企业经济效益较差等。这就需要进行因子分析的第三个步骤，即计算因子得分。顾名思义，因子得分就是公共因子 F_1，F_2，…，F_m 在每一个样品点上的得分。这需要我们给出公共因子用原始变量表示的线性表达式，这样的表达式一旦能够得到，就可以很方便地把原始变量的取值代入表达式中，求出各因子的得分值。

在上一章的分析中曾给出了主成分得分的概念，其意义和作用与因子得分相似。但是在此处，公共因子用原始变量线性表示的关系式并不易得到。在主成分分析中，主成分是原始变量的线性组合，当取 p 个主成分时，主成分与原始变量之间的变换关系是可逆的，只要知道了原始变量用主成分线性表示的表达式，就可以方便地得到用原始变量表示主成分的表达式；而在因子模型中，公共因子的个数少于原始变量的个数，且公共因子是不可观测的隐变量，载荷矩阵 $\boldsymbol{A}$ 不可逆，因而不能直接求得公共因子用原始变量表示的精确线性组合。解决该问题的一种方法是用回归的思想求出线性组合系数的估计值，即建立如下以公共因子为因变量、原始变量为自变量的回归方程：

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p, \qquad j = 1,2,\cdots,m \tag{6.19}$$

此处因为原始变量与公共因子变量均为标准化变量，所以回归模型中不存在常数项。在最小二乘意义下，可以得到 $\boldsymbol{F}$ 的估计值：

$$\hat{\boldsymbol{F}} = \boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{X} \tag{6.20}$$

式中，$\boldsymbol{A}$ 为因子载荷矩阵；$\boldsymbol{R}$ 为原始变量的相关阵；$\boldsymbol{X}$ 为原始变量向量。这样，在得到一组样本值后，就可以代入上面的关系式求出公共因子的估计得分，从而用少数公共因子去描述原始变量的数据结构，用公共因子得分去描述原始变量的取值。在估计出公共因子得分后，可以利用因子得分进行进一步的分析，如样本点之间的比较分析，对样本点的聚类分析等。当因子数 m 较少时，还可以方便地把各样本点在图上标示出来，直观地描述样本的分布情况，从而便于把研究工作引向深入。

6.2.6　主成分分析与因子分析的区别

(1) 因子分析把展示在我们面前的诸多变量看成由对每一个变量都有作用的一些公共因子和一些仅对某一个变量有作用的特殊因子线性组合而成。因此，我们的目的就是要从数据中探查能对变量起解释作用的公共因子和特殊因子，以及公共因

子和特殊因子组合系数。主成分分析则简单一些，它只是从空间生成的角度寻找能解释诸多变量绝大部分变异的几组彼此不相关的新变量（主成分）。

（2）因子分析中，把变量表示成各因子的线性组合，而主成分分析中，把主成分表示成各变量的线性组合。

（3）主成分分析中不需要有一些专门假设，因子分析则需要一些假设。因子分析的假设包括：各个公共因子之间不相关，特殊因子之间不相关，公共因子和特殊因子之间不相关。

（4）提取主因子的方法不仅有主成分法，还有极大似然法等，基于这些不同算法得到的结果一般也不同。而主成分只能用主成分法提取。

（5）主成分分析中，当给定的协方差矩阵或者相关矩阵的特征根唯一时，主成分一般是固定的；而因子分析中，因子不是固定的，可以旋转得到不同的因子。

（6）在因子分析中，因子个数需要分析者指定（SPSS 根据一定的条件自动设定，只要是特征根大于 1 的因子都进入分析），随指定的因子数量不同而结果不同。在主成分分析中，主成分的数量是一定的，一般有几个变量就有几个主成分。

（7）和主成分分析相比，由于因子分析可以使用旋转技术帮助解释因子，在解释方面更加有优势。而如果想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进行后续的分析，则可以使用主成分分析。当然，这种情况也可以通过计算因子得分处理。所以，这种区分不是绝对的。

6.3 因子分析的步骤与逻辑框图

上面介绍了因子分析的基本思想及基本的理论方法，下面我们把因子分析的步骤及逻辑框图总结如下，以使读者能更加清楚因子分析各步骤之间的脉络关系，更好地运用因子分析方法解决实际问题。

6.3.1 因子分析的步骤

进行因子分析应包括如下几步：

（1）根据研究问题选取原始变量。

（2）对原始变量进行标准化并求其相关阵，分析变量之间的相关性。

（3）求解初始公共因子及因子载荷矩阵。

（4）因子旋转。

（5）计算因子得分。

（6）根据因子得分值进行进一步分析。

6.3.2　因子分析的逻辑框图

因子分析的逻辑框图如图 6—1 所示。

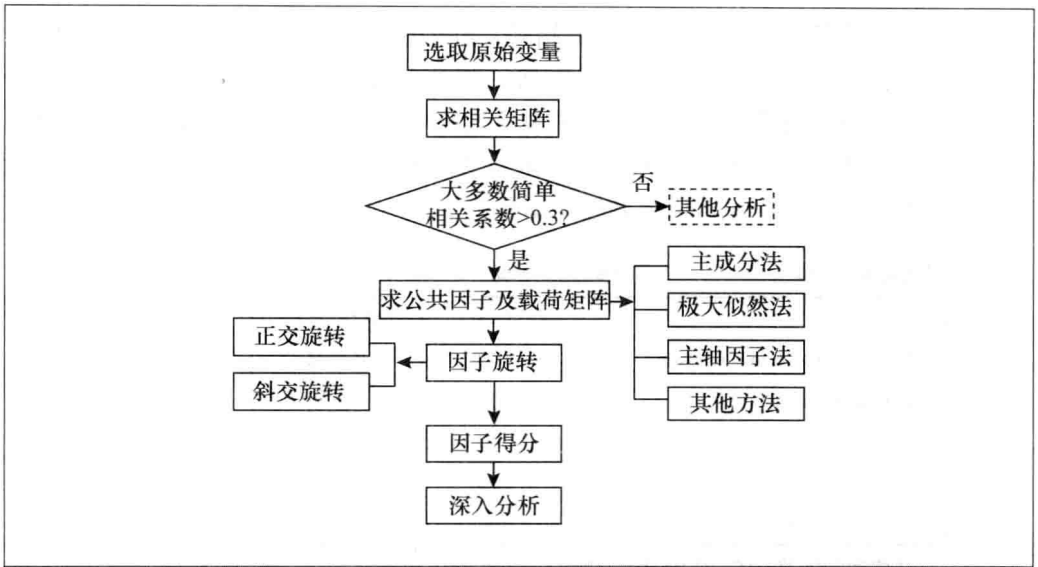


图 6—1

6.4　因子分析的上机实现

在上一章中，我们用 SPSS 的 Factor Analysis 模块实现了主成分分析，实际上，Factor Analysis 主要是 SPSS 软件进行因子分析的模块。由于主成分分析与因子分析（特别是因子分析中的主成分法）之间有密切的关系，SPSS 软件将这两种分析方法放在同一模块中。

下面先用之前版本 SPSS 10.0 自带的数据说明利用 Factor Analysis 模块进行因子分析的方法，然后给出一个具体案例。为了与主成分分析进行比较，此处仍沿用 Employee data. sav 数据集。

例 6—1

数据集 Employee data. sav 中各变量的解释说明见上一章主成分分析，用 Factor Analysis 模块进行因子分析。

打开 Employee data. sav 数据集并依次点选 Analyze→Dimension Reduction→

Factor…进入 Factor Analysis 对话框，选取 Educ，Salary，Salbegin，Jobtime，Prevexp 变量进入 Variables 窗口。

点击对话框的 Extraction 进入 Extraction 对话框，在 Method 选项框我们看到 SPSS 默认用主成分法提取因子，在 Analyze 框架中看到是从分析相关阵的结构出发求解公共因子。点击 Continue 按钮继续。如果这样交由程序运行的话，将得到与上一章输出结果 5—1 相同的结果，其中包括公共因子解释方差的比例、因子载荷矩阵（即 Component Matrix）等。选中 Display factor score coefficient matrix 复选框，我们在主成分分析中也选中了该选项，它要求 SPSS 输出因子得分矩阵，即标准化主成分（因子）用原始变量线性表示的系数矩阵。点击 Continue 继续，点击 OK 按钮运行，可以得到输出结果 6—1。

输出结果 6—1

Communalities

	Initial	Extraction
Educational Level (years)	1.000	0.754
Current Salary	1.000	0.896
Beginning Salary	1.000	0.916
Months since Hire	1.000	0.999
Previous Experience (months)	1.000	0.968

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.477	49.541	49.541	2.477	49.541	49.541
2	1.052	21.046	70.587	1.052	21.046	70.587
3	1.003	20.070	90.656	1.003	20.707	90.656
4	0.365	7.299	97.955			
5	0.102	2.045	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix*

	Component		
	1	2	3
Educational Level (years)	0.846	−0.194	−1.4E−02
Current Salary	0.940	0.104	2.857E−02
Beginning Salary	0.917	0.264	−7.7E−02
Months since Hire	6.806E−02	−5.2E−02	0.996
Previous Experience (months)	−0.178	0.965	6.901E−02

Extraction Method: Principal Component Analysis.

* 3 components extracted.

Component Score Coefficient Matrix

	Component		
	1	2	3
Educational Level (years)	0.342	−0.184	−0.014
Current Salary	0.380	0.099	0.028
Beginning Salary	0.370	0.250	−0.077
Months since Hire	0.027	−0.050	0.992
Previous Experience (months)	−0.072	0.917	0.069

Extraction Method: Principal Component Analysis.
Component Scores.

上面这几张表在主成分分析中也得到过，实际上，用主成分法求解公共因子与载荷矩阵，是求主成分的逆运算，这在前面有所表述。其中 Component Matrix 是因子载荷矩阵，是用标准化后的主成分（公共因子）近似表示标准化原始变量的系数矩阵，用 $fac1$，$fac2$，$fac3$ 表示各公共因子，以 Current Salary 为例，即有

$$\text{标准化的 } salary \approx 0.940 \times fac1 + 0.104 \times fac2 + (2.857\text{E}-02) \times fac3$$

由上一章知，当保留 5 个主成分时，标准化原始变量与公共因子之间有如下精确的关系式：

$$\text{标准化的 } salary = 0.940 \times prin1 + 0.104 \times prin2 + (2.857\text{E}-02) \times prin3 - 0.234 \times prin4 + 0.222 \times prin5$$

可见，主成分法求解公共因子就是把后面不重要的部分 $-0.234 \times prin4 + 0.222 \times prin5$ 作为特殊因子反映在因子模型中，由 Communalities 表可知，特殊因子的方差（特殊度）为 $1-0.896=0.104$。

因子得分系数矩阵（component score coefficient matrix）是用原始变量表示标准化主成分（公共因子）的系数矩阵，其关系式已在上一章给出，此处不再赘述。这里想说明的是用主成分求解公共因子时因子得分系数与因子载荷之间的关系。如上面表中因子得分系数中第一个元素为 0.342，它与第一主成分的方差 2.477，因子载荷矩阵中第一个元素 0.846 之间有如下关系式：

$$0.846 = 0.342 \times 2.477$$

此处之所以是乘以 2.477 而不是它的平方根，是因为此处主成分已经经过标准化了。同理，有 $-0.184 \times 1.052 = -0.194$。可见用主成分法进行因子分析与主成分分析是完全可逆的，因此，有些研究者也用主成分求解因子分析的结果来进行主成分分析。

实际上，在进行因子分析之前，我们往往先要了解变量之间的相关性，以判断进行因子分析是否合适。对此，进入 Factor Analysis 对话框后，点击 Descriptives 按钮，进入 Descriptives 对话框，在 Statistics 框架中选择 Univariate descriptives 会给出每个变量的均值、方差等统计量的值，在下部 Correlation Matrix 框架中，选中 Coefficients 选项以输出原始变量的相关矩阵，选中 Significance levels 以输出原

始变量各相关系数的显著性水平。Correlation Matrix 框架中还有其他一些选项可以帮助我们进行判断，此处不再详细说明。点击 Continue 按钮继续，点击 OK 运行，可以得到输出结果 6—2。

输出结果 6—2 **Correlation Matrix**

		Educational Level (years)	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Correlation	Educational Level (years)	1.000	0.661	0.633	0.047	−0.252
	Current Salary	0.661	1.000	0.880	0.084	−0.097
	Beginning Salary	0.633	0.880	1.000	−0.020	0.045
	Months since Hire	0.047	0.084	−0.020	1.000	0.003
	Previous Experience (months)	−0.252	−0.097	0.045	0.003	1.000
Sig. (1-tailed)	Educational Level (years)		0.000	0.000	0.152	0.000
	Current Salary	0.000		0.000	0.034	0.017
	Beginning Salary	0.000	0.000		0.334	0.163
	Months since Hire	0.152	0.034	0.334		0.474
	Previous Experience (months)	0.000	0.017	0.163	0.474	

由上面的结果可知，原始变量之间有较强的相关性，进行因子分析是合适的。

得到初始载荷矩阵与公共因子后，为了解释方便，往往需要对因子进行旋转，设置好其他选项后点击 Factor Analysis 对话框 Rotation…按钮，进入 Rotation 对话框，在 Method 框架中可以看到 SPSS 给出了多种进行旋转的方法，系统默认为不旋转。可以选择的旋转方法有 Varimax（方差最大正交旋转），Direct Oblimin（直接斜交旋转），Quartimax（四次方最大正交旋转），Equamax（平均正交旋转）及 Promax（斜交旋转），选中 Varimax 选项，此时，Display 框架中 Rotated solution 选项处于活动状态，选中该选项以输出旋转结果。点击 Continue 按钮继续，点击 OK 运行，除上面的结果外，还可得到输出结果 6—3。

输出结果 6—3 **Rotated Component Matrix***

	Component		
	1	2	3
Educational Level(years)	0.812	−0.306	3.616E−02
Current Salary	0.944	−2.1E−02	6.552E−02
Beginning Salary	0.946	0.133	−5.0E−02
Months since Hire	2.285E−02	2.928E−03	0.999
Previous Experience(months)	−4.7E−02	0.983	4.355E−03

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

* Rotation converged in 4 iterations.

Component Transformation Matrix

Component	1	2	3
1	0.990	−0.134	0.046
2	0.137	0.989	−0.058
3	−0.038	0.064	0.997

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Score Coefficient Matrix

	Component		
	1	2	3
Educational Level (years)	0.314	−0.229	0.013
Current Salary	0.388	0.049	0.040
Beginning Salary	0.403	0.193	−0.074
Months since Hire	−0.017	0.011	0.994
Previous Experience (months)	0.051	0.921	0.012

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

由结果可以看到，旋转后公共因子解释原始数据的能力没有提高，但因子载荷矩阵及因子得分系数矩阵都发生了变化，因子载荷矩阵中的元素更倾向于 0 或者±1。

有时为了使公共因子的实际意义更容易解释，往往需要放弃公共因子之间互不相关的约束而进行斜交旋转，最常用的斜交旋转方法为 Promax 方法。对此例进行斜交旋转，可得到输出结果 6—4。

输出结果 6—4

Pattern Matrix*

	Component		
	1	2	3
Educational Level (years)	0.797	−0.266	1.913E−02
Current Salary	0.946	2.770E−02	4.936E−02
Beginning Salary	0.960	0.181	−6.5E−02
Months since Hire	1.565E−03	1.667E−02	1.000
Previous Experience (months)	9.555E−03	0.985	1.577E−02

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
* Rotation converged in 4 iterations.

Structure Matrix

	Component		
	1	2	3
Educational Level (years)	0.827	−0.353	5.839E−02
Current Salary	0.945	−7.7E−02	8.681E−02
Beginning Salary	0.937	7.818E−02	−3.1E−02
Months since Hire	4.011E−02	−1.0E−02	0.999
Previous Experience (months)	−9.7E−02	0.984	−1.0E−02

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

Component Correlation Matrix

Component	1	2	3
1	1.000	−0.109	4.037E−02
2	−0.109	1.000	−2.7E−02
3	4.037E−02	−2.7E−02	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

可以看到，与正交旋转不同，斜交旋转的输出结果中没有 Rotated Component Matrix，而代之以 Pattern Matrix 和 Structure Matrix。这里，Pattern Matrix 即因子载荷矩阵，而 Structure Matrix 为公共因子与原始变量的相关阵。也就是说，在斜交旋转中，因子载荷系数不再等于公共因子与原始变量的相关系数。上面三个表格存在如下关系：

Structure Matrix=Pattern Matrix×Correlation Matrix

为了得到因子得分值，进行如下操作：在 Factor Analysis 对话框，点击下方的 Scores 按钮，进入 Factor Scores（因子得分）对话框，选中 Save as variables 复选框，即把原始数据各样本点的因子得分值存为变量，可以看到系统默认用回归方法求因子得分系数（Method 框架中 Regression 选项被自动选中），保留此设置。在此例中，我们还选中了 Save as variables 复选框，这一选项要求输出估计的因子得分值，该结果出现在数据窗口。在数据窗口可以看到，在原始变量后面出现了三个新的变量，变量名分别为 fac1 _ 1，fac2 _ 1，fac3 _ 1。这三个变量即为各个样品的第一公共因子、第二公共因子、第三公共因子的得分。在前面的分析中曾提过，这些得分是经过标准化的，这一点可以用下面的方法简单地验证。

依次点选 Analyze→Descriptive Statistics→Descriptives…进入 Descriptives 对话框，选中 fac1 _ 1，fac2 _ 1，fac3 _ 1 三个变量，点击 OK 按钮运行，可得到输出结果 6—5。

输出结果 6—5 **Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
REGR factor score 1 for analysis 1	474	−1.585 96	5.818 49	1.11E−16	1.000 000 0
REGR factor score 2 for analysis 1	474	−1.229 37	3.598 99	1.16E−16	1.000 000 0
REGR factor score 3 for analysis 1	474	−1.893 37	1.888 00	8.04E−16	1.000 000 0
Valid N (listwise)	474				

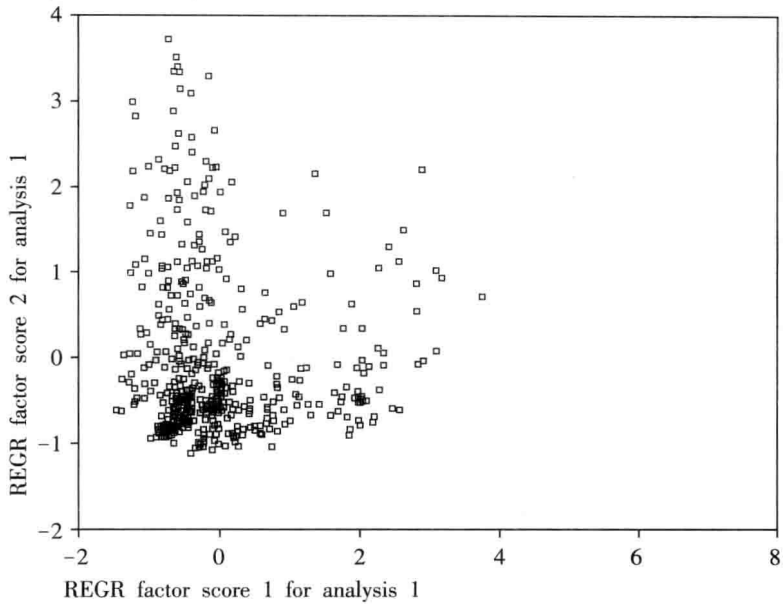
可以看到，三个变量的标准差均为 1（此处由于舍入原因，变量的均值不绝对等于零，而是有细微差别）。

得到各个样品的因子得分后，就可以对样本点进行分析，如用因子得分值代替原始数据进行归类分析或者回归分析等。同时，还可以在一张二维图上画出各数据点，描述各样本点之间的相关关系。

依次点选 Graphs→Legacy Dialogs→Scatter/Dot…进入 Scatter/Dot 对话框，选择 Simple/Scatter，点击 Define 按钮，在弹出的 Simple Scatterplot 对话框中，分

别选择 fac1 _ 1，fac2 _ 1 作为 X 轴与 Y 轴，点击 OK 交由程序运行，可得如下散点图（见输出结果 6—6）。

输出结果 6—6



由此可以直观地描述原始数据的散布情况。为了研究需要，还可以很方便地输出第一因子与第三因子、第二因子与第三因子的散点图或同时生成三个因子的散点图，这只需选择不同的变量或图形类型即可，在此不再详述。

例 6—2

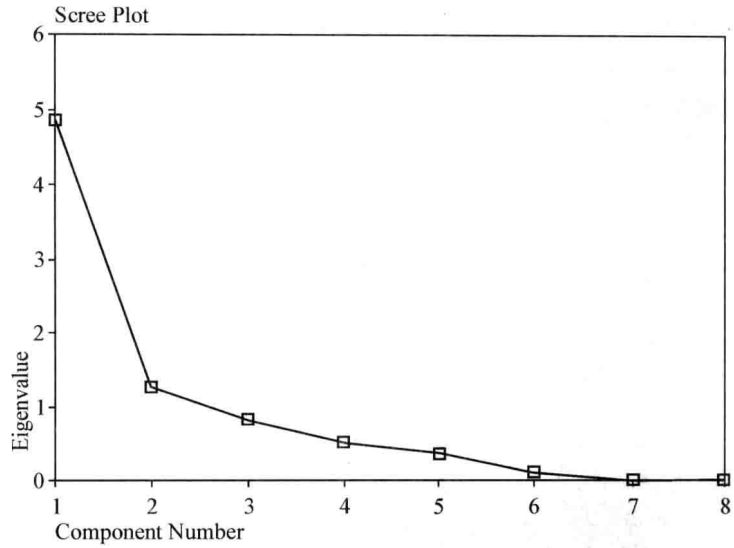
（数据见表 5—5）对企业经济效益指标体系的八项指标建立因子分析模型（详细因子分析上机实现见例 6—3）。

由 SPSS 输出方差解释表及碎石图（见输出结果 6—7）可看出，前三个特征根较大，其余五个特征根均较小。前三个公共因子对样本方差的贡献和为 87.085%，于是我们选取前三个公共因子建立因子载荷阵（即输出结果 5—9）。这里采用主成分法提取因子。

输出结果 6—7　　**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.861	60.758	60.758	4.861	60.758	60.758
2	1.269	15.865	76.623	1.269	15.865	76.623
3	.837	10.463	87.085	.837	10.463	87.085
4	.517	6.464	93.549			
5	.378	4.727	98.276			
6	.115	1.443	99.719			
7	.021	.264	99.984			
8	.001	.016	100.000			

Extraction Method: Principal Component Analysis.



对因子载荷阵进行方差最大化（Varimax）正交旋转，得到输出结果 6—8。

输出结果 6—8　Rotated Component Matrix[a]

	Component		
	1	2	3
x1	.914	.332	.167
x2	.881	−.056	.431
x3	.839	.398	.029
x4	.921	.000	.370
x5	.592	.266	.481
x6	−.048	.851	.411
x7	.307	.115	.849
x8	.399	.794	−.115

Extraction Method: Principal Component Analysis.

a. Rotation converged in 7 iterations.

由上表可得出企业经济效益指标体系的因子分析模型（特殊因子忽略不计）：

$$
\begin{aligned}
x_1 &= 0.914F_1 + 0.332F_2 + 0.167F_3 \\
x_2 &= 0.881F_1 - 0.056F_2 + 0.431F_3 \\
x_3 &= 0.839F_1 + 0.398F_2 + 0.029F_3 \\
x_4 &= 0.921F_1 + 0.370F_3 \\
x_5 &= 0.592F_1 + 0.266F_2 + 0.481F_3 \\
x_6 &= -0.048F_1 + 0.851F_2 + 0.411F_3 \\
x_7 &= 0.307F_1 + 0.115F_2 + 0.849F_3 \\
x_8 &= 0.399F_1 + 0.794F_2 - 0.115F_3
\end{aligned}
\tag{6.21}
$$

由因子分析模型可知，第一个主因子 F_1 主要由固定资产利税率、资金利税率、销售收入利税率、资金利润率等四个指标决定，这四个指标在主因子 F_1 上的载荷均在 0.85 以上，它代表着企业经济活动中的盈利能力，而且主因子 F_1 对 x_1 的方

差贡献已达60%之多，所以更说明是企业经济效益指标体系中的主要方面。此外，固定资产产值率对 F_1 的贡献相对也较大，这也是反映企业经济活动的盈利能力的主要指标。企业要提高经济效益，就要在这个主因子方面狠下工夫。

第二个主因子 F_2 主要由流动资金周转天数和全员劳动生产率决定，是代表企业经营效率的指标。经营效率主要反映企业的运营能力，企业改进管理方法，提高科学管理水平，也是提高经济效益的重要途径。

第三个主因子 F_3 主要反映了企业的产值和能耗，产值和能耗反映的是投入与产出的关系。企业要提高经济效益，就不能忽视降低生产成本。

例6—3

中心城市的综合发展是带动周边地区经济发展的重要动力。在我国经济发展进程中，各个中心城市一直是该地区经济和社会发展的“引路者”。因而，分析评价全国35个中心城市的综合发展水平，无论是对城市自身的发展，还是对周边地区的进步，都具有十分重要的意义。下面应用因子分析模型，选取反映城市综合发展水平的12个指标作为原始变量，运用SPSS软件，对全国35个中心城市的综合发展水平作分析评价。

（1）原始数据及指标解释。我们选取了反映城市综合发展水平的12个指标，其中包括8个社会经济指标，分别为：x_1——非农业人口数（万人）；x_2——工业总产值（万元）；x_3——货运总量（万吨）；x_4——批发零售住宿餐饮业从业人数（万人）；x_5——地方政府预算内收入（万元）；x_6——城乡居民年底储蓄余额（万元）；x_7——在岗职工人数（万人）；x_8——在岗职工工资总额（万元）。

4个城市公共设施水平的指标，分别为：x_9——人均居住面积（平方米）；x_{10}——每万人拥有公共汽车数（辆）；x_{11}——人均拥有铺装道路面积（平方米）；x_{12}——人均公共绿地面积（平方米）。

指标的选取参考了《中国城市统计年鉴》中指标的设置。数据来源于《中国城市统计年鉴（2004）》。数据如表6—1所示。

表6—1

城市	x_1	x_2	x_3	x_4	x_5	x_6
北京	830.8	38 103 630	30 671.14	127.4	5 925 388	64 413 910
天津	549.74	40 496 103	34 679	15.38	2 045 295	18 253 200
石家庄	331.33	11 981 505	10 008.48	8.07	493 429	10 444 919
太原	222.63	5 183 200	15 248.11	2.43	333 473	6 601 300
呼和浩特	97.81	2 407 794	4 155.1	2	205 779	2 554 496
沈阳	440.6	10 643 612	14 635.74	7.3	810 889	14 229 575
长春	313.05	15 115 270	10 891.98	6.94	459 709	8 313 564
哈尔滨	454.52	7 215 089	9 517.8	24.99	763 600	11 536 951

续前表

城市	x_1	x_2	x_3	x_4	x_5	x_6
上海	1 041.39	1.03E+08	63 861	35.22	8 992 850	60 546 000
南京	391.67	25 093 816	14 804.68	7.62	1 364 788	11 336 202
杭州	263.67	32 025 226	16 815.2	8.36	1 503 888	14 664 200
合肥	160.18	5 348 605	4 640.84	3.39	358 694	3 592 488
福州	205.43	12 889 573	8 250.39	4.69	674 522	8 762 245
南昌	195.46	4 149 169	4 454.45	3.62	314 094	4 828 029
济南	297.21	13 185 425	14 354.4	6.6	761 054	7 583 525
郑州	249.72	9 270 494	7 846.91	8.77	658 737	10 484 859
武汉	474.98	13 344 938	16 610.34	13.58	804 368	12 855 341
长沙	205.83	5 339 304	10 630.5	6.31	598 930	7 048 500
广州	493.32	40 178 324	28 859.45	21.47	2 747 707	37 273 276
南宁	167.99	2 083 763	5 893.09	4.95	362 435	4 514 961
海口	76.05	2 025 643	3 304.4	2.72	122 541	2 843 664
成都	386.23	9 700 976	28 798.2	8.06	895 752	14 944 197
贵阳	165.27	3 569 419	5 317.55	5.75	403 855	3 449 487
昆明	205.34	5 809 573	12 337.86	7.07	601 101	7 085 278
西安	312.88	6 386 627	9 392	12.21	648 037	12 105 607
兰州	175.54	5 215 490	5 580.8	3.7	205 660	4 683 830
西宁	105.13	1 148 959	2 037.15	1.24	84 397	1 749 293
银川	79.2	1 464 867	2 127.17	1.65	122 605	1 930 771
乌鲁木齐	142.94	3 110 943	12 754.02	3.94	409 119	4 203 000
大连	297.48	15 468 641	21 081.47	6.6	1 105 405	13 101 986
宁波	168.81	26 302 862	13 797.38	4.8	1 394 162	10 596 339
厦门	83.74	13 201 500	3 054.82	2.83	701 456	3 971 559
青岛	329.96	25 588 695	30 552.6	6.72	1 201 398	9 084 693
深圳	122.39	52 451 037	6 792.66	10.84	2 908 370	21 994 500
重庆	753.92	15 889 928	32 450.2	12.83	1 615 618	18 965 569
城市	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
北京	434.15	10 989 365	15	17.3	8.56	44.94
天津	174.5	3 254 148	18	7.99	7.23	17.45
石家庄	86.74	1 067 432	18	7.23	8.28	21.56
太原	74.55	945 212	16	5.06	7.88	20.58
呼和浩特	28.9	407 963	18	3.81	8.92	26.58
沈阳	101.7	1 521 548	15	9.32	6.7	28.36
长春	89.7	1 244 167	15	11.87	7.03	18.75

续前表

城市	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
哈尔滨	168.83	2 102 165	14	12.75	6.34	18.51
上海	281.51	7 686 511	19	14.57	12.92	19.11
南京	87.91	1 950 742	16	9.06	12.13	136.72
杭州	75.72	1 867 776	17	8.93	6.5	23.19
合肥	37.88	526 577	17	14.11	15.72	28.74
福州	71.3	1 073 262	18	9.65	7.9	31.6
南昌	49.79	692 717	17	7.37	7.67	23.98
济南	78.38	1 256 160	19	7.77	10.62	19.54
郑州	83.99	1 137 056	19	10.11	7.63	17.77
武汉	136.08	1 868 350	17	6.87	4.16	8.34
长沙	60.04	1 019 924	18	10.09	9.1	29.1
广州	182.16	5 247 087	17	11.16	12.76	178.76
南宁	50.79	668 976	18	9.91	9.32	35.12
海口	22.97	340 392	20	5.09	7.07	15.79
成都	124.03	1 894 496	17	8.95	10.17	25.59
贵阳	54.53	664 234	16	9.37	3.11	105.35
昆明	73.34	1 045 469	15	15.33	4.49	23.33
西安	113.73	1 535 896	15	7.32	4.48	8.82
兰州	54.91	740 661	15	10.33	6.3	11.22
西宁	20.6	301 364	17	11.47	4.92	14.2
银川	29.12	393 035	15	9.26	10.43	40.21
乌鲁木齐	47.42	782 873	19	22.89	6.49	20.53
大连	82.13	1 442 215	14	13.79	6.24	40.21
宁波	59.88	1 418 635	17	9.88	6.81	17.65
厦门	54.78	1 042 111	20	15.5	8.15	26.44
青岛	104.55	1 603 305	15	14.78	11.41	35.78
深圳	104.98	3 259 900	21	114.91	47.29	177.62
重庆	203.79	2 535 070	21	4.94	4.24	10.8

（2）计算运行结果。将标准化后的数据导入 SPSS 软件，依次点选 Analyze→Dimension Reduction→Factor…，进入 Factor Analysis 对话框。把 12 个指标变量选入 Variables 中，点击 Extraction 按钮，在 Method 选项中选择 Principal components（这时，因子分析等同于主成分分析，如果是主成分分析，则只能选择此项），在 Display 选项中选中 Scree plot，点击 Continue 按钮，回到主对话框点击 OK。

按照特征根大于 1 的原则，选入 3 个公共因子，其累计方差贡献率为 87.1%，

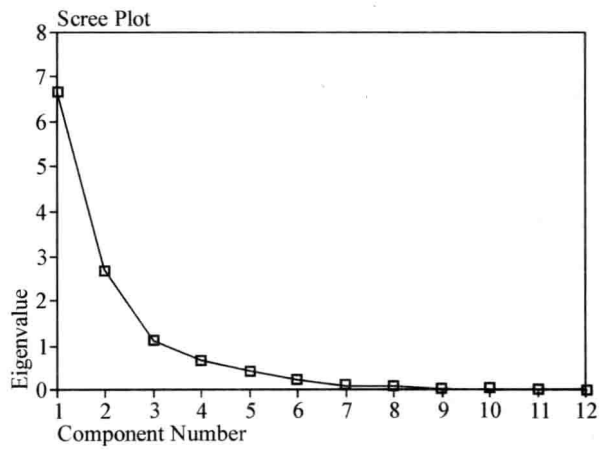
特征根及累计贡献率、碎石图、因子载荷矩阵见输出结果 6—9。

输出结果 6—9

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.671	55.589	55.589	6.671	55.589	55.589
2	2.675	22.293	77.882	2.675	22.293	77.882
3	1.107	9.224	87.105	1.107	9.224	87.105
4	.670	5.581	92.686			
5	.437	3.642	96.329			
6	.229	1.909	98.238			
7	.076	.631	98.869			
8	.073	.612	99.482			
9	.032	.264	99.746			
10	.021	.179	99.925			
11	.007	.056	99.981			
12	.002	.019	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix[a]

	Component		
	1	2	3
x1	.878	−.325	.143
x2	.854	.254	.265
x3	.830	−.210	.330
x4	.789	−.203	−.403
x5	.956	.062	.130
x6	.984	−.032	−.058
x7	.933	−.207	−.143
x8	.971	−.039	−.163
x9	.059	.465	.727
x10	.207	.898	−.131
x11	.243	.927	−.052
x12	.241	.698	−.359

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

此时得到的未旋转的公共因子的实际意义不好解释，因此，对公共因子进行方差最大化正交旋转。在 Factor Analysis 对话框中，点击 Rotation 按钮，进入 Rotation 对话框，选中 Varimax 进行方差最大化正交旋转（若做主成分分析就选择 none），得到输出结果 6—10。

输出结果 6—10　Rotated Component Matrix[a]

	Component		
	1	2	3
x1	.929	−.183	.039
x2	.806	.309	.344
x3	.870	−.147	.253
x4	.791	.091	−.437
x5	.934	.194	.155
x6	.970	.174	−.053
x7	.947	.030	−.191
x8	.952	.199	−.155
x9	.010	.205	.840
x10	.034	.914	.175
x11	.068	.921	.259
x12	.092	.809	−.106

Rotation Sums of Squared Loadings		
Total	% of Variance	Cumulative %
6.526	54.381	54.381
2.649	22.077	76.458
1.278	10.647	87.105

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

由输出结果 6—10，原变量 x_1 可由各因子表示为：

$$x_1=0.929\times F_1-0.183\times F_2+0.039\times F_3$$

原变量 x_2 可由各因子表示为：

$$x_2=0.806\times F_1+0.309\times F_2+0.344\times F_3$$

其余以此类推。

为便于得出结论，在 Factor Analysis 主对话框中点击 Options 按钮进入 Options 对话框，在 Coefficient Display Format 框中选中 Sorted by size，使输出的载荷矩阵中各列按载荷系数大小排列，在同一个公共因子上具有较高载荷的变量排在一起。然后点击 Continue，OK 运行。

得到输出结果 6—11。

最后，计算因子得分，以各因子的方差贡献率占三个因子总方差贡献率的比重作为权重进行加权汇总，得出各城市的综合得分 F（这种综合评价方法目前应用较多，但也有较大争议，故应慎用），即

$$F=(54.381\times F_1+22.077\times F_2+10.647\times F_3)/87.105$$

在 Factor Analysis 主对话框中点击按钮 Scores 进入 Factor Scores 对话框，选中 Save as variables，在 Method 中选择 Regression 计算因子得分。

输出结果 6—11 Rotated Component Matrix[a]

	Component		
	1	2	3
x6	.970	.174	−.053
x8	.952	.199	−.155
x7	.947	.030	−.191
x5	.934	.194	.155
x1	.929	−.183	.039
x3	.870	−.147	.253
x2	.806	.309	.344
x4	.791	.091	−.437
x11	.068	.921	.259
x10	.034	.914	.175
x12	.092	.809	−.106
x9	.010	.205	.840

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.

得到运行结果并计算综合得分，结果如表 6—2 所示。

表 6—2

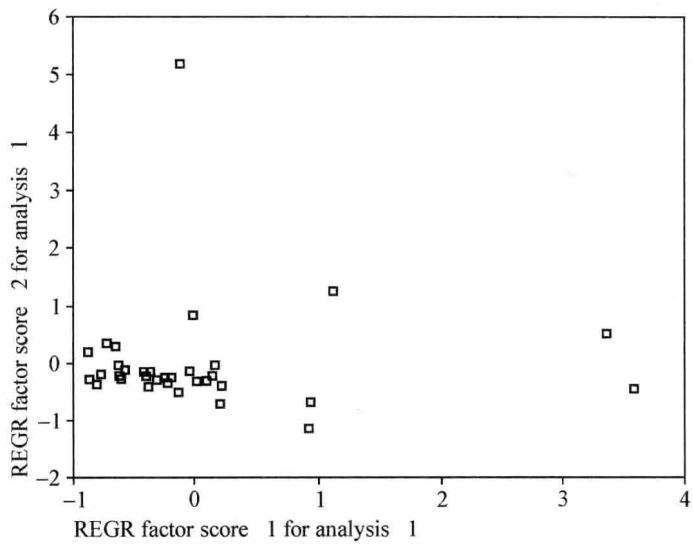
城市	F_1	F_2	F_3	F
北京	3.374 369	0.489 283	−3.040 15	1.859 078
天津	0.955 132	−0.653 44	0.963 646	0.548 475
石家庄	−0.216 24	−0.330 6	0.349 232	−0.176 11
太原	−0.388 36	−0.386 52	−0.250 67	−0.371 06
呼和浩特	−0.792 28	−0.193 65	0.258 952	−0.512 06
沈阳	0.007 808	−0.328 34	−0.679 84	−0.161 44
长春	−0.231 53	−0.240 23	−0.650 31	−0.284 93
哈尔滨	0.151 193	−0.222 19	−1.587 07	−0.155 91
上海	3.581 377	−0.456 2	2.452 754	2.420 089
南京	−0.004 38	0.854 253	−0.615 98	0.138 488
杭州	0.094 03	−0.305 61	0.358 684	0.025 089
合肥	−0.724 84	0.350 884	−0.009 4	−0.364 75
福州	−0.379 2	−0.135 31	0.326 439	−0.231 13
南昌	−0.617 37	−0.217 6	−0.112 94	−0.454 39
济南	−0.203 55	−0.283 92	0.883 371	−0.091 06
郑州	−0.295 25	−0.288 81	0.605 291	−0.183 54
武汉	0.213 515	−0.728 49	0.005 27	−0.050 69
长沙	−0.439 22	−0.130 93	0.297 917	−0.270 98
广州	1.128 811	1.255 558	−0.577 46	0.952 374
南宁	−0.638 35	−0.024 52	0.172 766	−0.383 63
海口	−0.812 98	−0.376 04	0.969 04	−0.484 42
成都	0.212 641	−0.364 64	0.254 51	0.071 445

续前表

城市	F_1	F_2	F_3	F
贵阳	−0.664 83	0.312 044	−0.961 59	−0.453 51
昆明	−0.389 96	−0.231 11	−0.776 92	−0.397
西安	−0.132 92	−0.495 72	−0.922 19	−0.321 35
兰州	−0.611 91	−0.275 55	−0.760 09	−0.544 77
西宁	−0.859 88	−0.299 11	−0.097 39	−0.624 55
银川	−0.892 65	0.206 994	−0.914 78	−0.616 65
乌鲁木齐	−0.571 59	−0.114 2	0.736 093	−0.295 82
大连	−0.040 28	−0.124 16	−0.883 3	−0.164 59
宁波	−0.170 49	−0.271 04	0.345 12	−0.132 95
厦门	−0.613 29	0.010 6	0.998 044	−0.258 21
青岛	0.160 302	−0.037 48	−0.152 74	0.071 91
深圳	−0.117 2	5.194 94	1.266 561	1.398 318
重庆	0.929 383	−1.159 16	1.749 143	0.500 237

以 F_1 因子得分为 x 轴，F_2 因子得分为 y 轴，画出各城市的因子得分图（见输出结果 6—12），其操作步骤参见例 6—1。

输出结果 6—12



（3）结果分析。由旋转后的因子载荷矩阵可以看出，公共因子 F_1 在 x_1（非农业人口数），x_2（工业总产值），x_3（货运总量），x_4（批发零售住宿餐饮业从业人数），x_5（地方政府预算内收入），x_6（城乡居民年底储蓄余额），x_7（在岗职工人数），x_8（在岗职工工资总额）上的载荷值都很大。x_1，x_7，x_8 是反映城市规模的指标；x_2，x_3 反映城市工业发展规模；x_4 反映城市第三产业的发展规模；x_5 是政府作为国家的管理者和国有资产的所有者而获得的收入，在一定程度上反映了居民的收入水平，而在我国现今的收入分配格局下，政府和居民是再分配收入的获得大

户，因而，x_5，x_6 在一定程度上反映了城市的国民收入水平，因而 F_1 为反映城市规模及经济发展水平的公共因子，在这个因子上的得分越高，城市经济发展水平越高，城市规模越大。公共因子 F_2 由于在 x_{10}（每万人拥有公共汽车数），x_{11}（人均拥有铺装道路面积），x_{12}（人均公共绿地面积）上的载荷较大，是反映城市的基础设施水平的公共因子，在此因子上的得分则反映了一个城市的基础设施水平。公共因子 F_3 仅在 x_9（人均居住面积）上有较大的载荷，是反映城市居民住房条件的公共因子。

有了对各个公共因子合理的解释，结合各个城市在三个公共因子上的得分和综合得分，就可对各中心城市的综合发展水平进行评价了。在城市经济规模因子 F_1 上得分最高的前五个城市依次是上海、北京、广州、天津和重庆，其中，上海的得分为 3.58，北京为 3.37，远高于其他城市，这就是说，就城市经济发展规模而言，上海、北京是我国最大的城市，且其规模远大于其他城市。城市规模较小、经济发展相对较慢的城市有西宁和银川，而海口由于城市规模小，在 F_1 上的得分也较低。深圳、广州和南京在 F_2 上的得分较高，而重庆、武汉得分较低，说明深圳、广州、南京的城市基础设施在全国是较好的，而重庆等城市的基础设施相对较差，还需要下大力气进行改善。上海、重庆、深圳等城市在 F_3 上的得分比较高，说明居民在居住条件上比别的城市好，北京、哈尔滨等则需要改善。

将各城市在三个因子上的得分进行加权综合，就得到了综合得分。根据综合得分就可综合评价城市的发展水平。综合得分前五名的城市依次是上海、北京、深圳、广州和天津；综合得分最低的五个城市依次是西宁、银川、兰州、呼和浩特和海口。再结合各因子得分进行分析，北京在城市规模及经济发展水平、基础设施建设方面均位于前列，但是在居民住房面积上的得分较低，因此，需在这方面加大改善力度。上海在城市规模、经济发展水平及居民住房上得分最高，在基础设施方面得分不太理想，这可能是因为上海人口较多。而综合得分较低的城市在经济发展水平上的得分都较低，在城市发展战略上应把经济发展放在首位，只有经济发展了，城市基础设施水平及其他方面才能搞上去。另外，因子得分图分析表明，就城市规模而言，历史悠久的城市大于新兴城市；就城市基础设施水平而言，南方城市普遍好于北方城市，新兴城市好于老城市；综合来讲，东部地区城市发展水平高于西部地区城市。上海、北京、深圳三个城市综合发展水平较接近，上海规模大，但基础设施水平较低；北京规模大，基础设施水平较高，但是居民人均居住面积较小；深圳规模不大，但是基础设施水平较高，人均居住面积较大。此外，综合得分值大于零的城市还有广州、天津、重庆、南京、青岛、成都、杭州等，但是这些城市与上海、北京及深圳有一定的差距。其他城市综合得分都小于零，在因子得分图中大致位于原点附近，城市综合发展水平都还较低，发展格局也较相近，其中有 18 个城市位于因子得分图的第三象限，这些城市多位于中西部地区。因而，如何加快这些城市的发展以带动周边地区的进步，是影响我国整体经济发展的重要课题。

这种综合评价方法应用非常普遍，但有些文献提出不同看法，主要是认为产生主因子的特征向量的各级分量符号不一致，很难进行排序评价。因此，认为这种综

合评价方法不严谨。我们认为这与其他的统计方法一样，其实很多理论问题并没有解决，但似乎并不影响人们使用的热情。统计学应用中许多问题的完善需要人们去实践、去探讨，这个问题当然也在其中。

□ 参考文献

[1] 方开泰，张尧庭. 多元统计分析引论. 北京：科学出版社，1982

[2] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993

[3] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[4] M. 肯德尔. 多元分析. 北京：科学出版社，1999

[5] Bryan F. J. Manly. *Multivariate Statistical Methods: A Primer*. Chapman and Hall，1986

□ 思考与练习

1. 因子分析与主成分分析有什么本质不同？
2. 因子载荷 a_{ij} 的统计定义是什么？它在实际问题分析中的作用是什么？
3. 试用 SPSS 软件对一个实际问题的研究应用因子分析。

第7章 Chapter 7 对应分析

学习目标

1. 理解列联表分析及对应分析的基本思想；
2. 了解对应分析的基本理论；
3. 掌握对应分析的方法；
4. 能够用 SPSS 软件进行对应分析并正确理解输出结果。

对应分析是 R 型因子分析与 Q 型因子分析的结合，它也是利用降维的思想来达到简化数据结构的目的，不过，与因子分析不同的是，它同时对数据表中的行与列进行处理，寻求以低维图形表示数据表中行与列之间的关系。对应分析的思想首先由理查森（Richardson）和库德（Kuder）在 1933 年提出，后来法国统计学家让-保罗·贝内泽（Jean-Paul Benzécri）和日本统计学家林知己夫（Hayashi Chikio）对该方法进行了详细的论述而使其得到了发展。对应分析方法广泛应用于对由属性变量构成的列联表数据的研究，利用对应分析可以在一张二维图上同时画出属性变量不同取值的情况，列联表的每一行及每一列均以二维图上的一个点来表示，从而以直观、简洁的形式描述属性变量各种状态之间的相互关系及不同属性变量之间的相互关系。本章主要讲述对应分析的基本思想、对应分析的基本理论与方法及如何用 SPSS 软件进行对应分析。

7.1 列联表及列联表分析

在讨论对应分析之前，我们先简要回顾一下列联表及列联表分析的有关内容。在实际研究工作中，人们常常用列联表的形式来描述属性变量（定类尺度或定序尺

度）的各种状态或相关关系，这在某些调查研究项目中运用得尤为普遍。比如，公司的管理者为了解消费者对自己产品的满意情况，需要针对不同职业的消费者进行调查，而调查数据很自然地就以列联表的形式呈现出来（见表 7—1）。

表 7—1

评价 / 职业	非常满意	比较满意	一般	不太满意	不满意	汇总
一般工人						
管理者						
行政官员						
⋮						
汇总						

以上是两变量列联表的一般形式，横栏与纵列交叉位置的数字是相应的频数。这样从表中数据就可以清楚地看到不同职业的人对该公司产品的评价，以及所有被调查者对该公司产品的整体评价、被调查者的职业构成情况等信息。通过这张列联表，还可以看出职业分布与各种评价之间的相关关系，如管理者与比较满意交叉单元格的数字相对较大（“相对”指应抵消不同职业在总的被调查者中的比例的影响），则说明职业栏的管理者这一部分与评价栏的比较满意这一部分有较强的相关性。由此可以看到，借助列联表，可以得到很多有价值的信息。

在研究经济问题的时候，研究者也往往用列联表的形式把数据呈现出来。比如说横栏是不同规模的企业，纵列是不同水平的获利能力，通过这样的形式，可以研究企业规模与获利能力之间的关系。更为一般地，可以对企业进行更广泛的分类，如按上市与非上市分类，按企业所属的行业分类，按不同所有制关系分类等。同时，用列联表的格式来研究企业的各种指标，如企业的盈利能力、企业的偿债能力、企业的发展能力等。这些指标既可以是简单的，也可以是综合的，甚至可以是用因子分析或主成分分析提取的公共因子。把这些指标按一定的取值范围进行分类，就可以很方便地用列联表来研究。

一般，假设按两个特性对事物进行研究，特性 A 有 n 类，特性 B 有 p 类，属于 A_i 和 B_j 的个体数目为 n_{ij}（$i=1, 2, \cdots, n$；$j=1, 2, \cdots, p$），则可以得到形如表 7—2 的列联表。

表 7—2

		特性 B						合计
		B_1	B_2	$\cdots$	B_j	$\cdots$	B_p	
特性 A	A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
	A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2.}$
	⋮	⋮	⋮		⋮		⋮	⋮
	A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i.}$
	⋮	⋮	⋮		⋮		⋮	⋮
	A_n	n_{n1}	n_{n2}	$\cdots$	n_{nj}	$\cdots$	n_{np}	$n_{n.}$
合计		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	n

在表 7—2 中，$n_{i.}=n_{i1}+n_{i2}+\cdots+n_{ip}$，$n_{.j}=n_{1j}+n_{2j}+\cdots+n_{nj}$，右下角元素 n 是所有频数的和，有 $n=n_{1.}+n_{2.}+\cdots+n_{n.}=n_{.1}+n_{.2}+\cdots+n_{.p}$。为了更为方便地表示各频数之间的关系，人们往往用频率来代替频数，即将列联表中每一个元素都除以元素的总和 n，令 $p_{ij}=\frac{n_{ij}}{n}$，于是得到如下频率意义上的列联表（见表 7—3）。

表 7—3

		特性 B						合计
		B_1	B_2	$\cdots$	B_j	$\cdots$	B_p	
特性 A	A_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1p}	$p_{1.}$
	A_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2p}	$p_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	A_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{ip}	$p_{i.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	A_n	p_{n1}	p_{n2}	$\cdots$	p_{nj}	$\cdots$	p_{np}	$p_{n.}$
合计		$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.p}$	1

上表中，令

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{np} \end{bmatrix}$$

$$\boldsymbol{P}_I'=(p_{1.},\ p_{2.},\ \cdots,\ p_{n.}),\qquad \boldsymbol{P}_J'=(p_{.1},\ p_{.2},\ \cdots,\ p_{.p})$$

$$\mathbf{1}'=(1,\ 1,\ \cdots,\ 1)$$

则由上表的定义知，下列各式成立：

$$\mathbf{1}'\boldsymbol{P}\mathbf{1}=\boldsymbol{P}_I'\mathbf{1}=\boldsymbol{P}_J'\mathbf{1}=1,\ \boldsymbol{P}\mathbf{1}=\boldsymbol{P}_I,\ \boldsymbol{P}'\mathbf{1}=\boldsymbol{P}_J$$

对于研究对象的总体，表 7—3 中的元素有概率的含义，p_{ij} 是特性 A 第 i 状态与特性 B 第 j 状态出现的概率，而 $p_{.j}$ 与 $p_{i.}$ 则表示边缘概率。考察各种特性之间的相关关系，可以通过研究各种状态出现的概率入手。如果特性 A 与特性 B 之间是相互独立的，则对任意的 i 与 j，有下式成立：

$$p_{ij}=p_{i.}\times p_{.j} \tag{7.1}$$

式（7.1）表示，如果特性 A 与特性 B 之间相互独立，特性 A 第 i 状态与特性 B 第 j 状态同时出现的概率则应该等于总体中第 i 状态出现的概率乘以第 j 状态出现的概率。由此令 $\hat{p}_{ij}=p_{i.}\times p_{.j}$ 表示由样本数据得到的特性 A 第 i 状态与特性 B 第 j 状态出现的期望概率的估计值。我们可以通过研究特性 A 第 i 状态和特性 B 第 j 状态同时出现的实际概率 p_{ij} 与特性 A 第 i 状态和特性 B 第 j 状态同时出现的期望概率 $\hat{p}_{ij}$ 的差别大小，来判断特性 A 与特性 B 是否独立。此处 A 与 B 为属性变量，在实

际研究中，根据实际问题它们可以有不同的意义，它实质上是列联表的横栏与纵列按某种规则的分类。我们关心的是属性变量 A 与 B 是否独立，由此提出以下假设：

H_0：属性变量 A 与 B 相互独立

H_1：属性变量 A 与 B 不独立

由上面的假设构建如下 χ^2 统计量：

$$\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{[n_{ij}-\hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}=n\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-p_{i\cdot}\,p_{\cdot j})^2}{p_{i\cdot}\,p_{\cdot j}} \tag{7.2}$$

注意到，除了常数项 n 外，χ^2 统计量实际上反映了矩阵 $\boldsymbol{P}$ 中所有元素的观察值与理论值经过某种加权的总离差情况。可以证明，在 n 足够大的条件下，当原假设为 H_0 时，χ^2 遵从自由度为 $(n-1)(p-1)$ 的 χ^2 分布。拒绝域为：

$$\chi^2>\chi_a^2[(n-1)(p-1)]$$

通过上面的方法，可以判断两个分类变量是否独立，而当拒绝原假设后，我们进一步想了解两个分类变量及分类变量各个状态（取值）之间的相关关系，用对应分析方法可以解决这一问题。

7.2　对应分析的基本理论

当 A 与 B 的取值较少时，把所得到的数据放到一张列联表中，就可以很直观地对 A 与 B 之间及它们的各种取值之间的相关性做出判断。当 p_{ij} 比较大时，说明属性变量 A 的第 i 状态与 B 的第 j 状态之间有较强的依赖关系。但是，当 A 或者 B 的取值比较多时，就很难正确地做出判断，此时需要利用降维的思想来简化列联表的结构。由前面的讨论知道，因子分析（或主成分分析）是用少数综合变量索取原始变量大部分信息的有效方法。但因子分析也有不足之处，当我们要研究属性变量 A 的各种状态时，需要做 Q 型因子分析，即要分析一个 $n\times n$ 阶矩阵的结构，而当我们要研究属性变量 B 的各种状态时，就是进行 R 型因子分析，需要分析一个 $p\times p$ 阶矩阵的结构。由于因子分析的局限性，无法使 R 型因子分析与 Q 型因子分析同时进行，而当 n 或者 p 比较大时，单独进行因子分析就会加大计算量。对应分析可以弥补上述不足，同时对两个（或多个）属性变量进行分析。

如前所述，对应分析利用降维思想分析原始数据结构，旨在以简洁、明了的方式揭示属性变量之间及属性变量各种状态之间的相关关系。对应分析的一大特点就是可以在一张二维图上同时表示出两类属性变量的各种状态，以直观地描述原始数据结构。

假定下面讨论的都是形如表 7—3 的规格化的列联表数据。为了论述方便，先对有关概念进行说明。

7.2.1 有关概念

1. 行剖面与列剖面

在表7—3中，p_{ij}表示变量A的第i状态与变量B的第j状态同时出现的概率，相应的$p_{i.}$与$p_{.j}$就有边缘概率的含义。所谓行剖面，是指当变量A的取值固定为i时（$i=1, 2, \cdots, n$），变量B的各个状态相对出现的概率情况，也就是把矩阵$\boldsymbol{P}$中第i行的每一个元素均除以$p_{i.}$，这样，就可以方便地把第i行表示成p维欧氏空间中的一个点，其坐标为：

$$\boldsymbol{p}_i^{r\prime} = \left(\frac{p_{i1}}{p_{i.}}, \frac{p_{i2}}{p_{i.}}, \cdots, \frac{p_{ip}}{p_{i.}}\right), \quad i = 1,2,\cdots,n \tag{7.3}$$

其中$\boldsymbol{p}_i^r$中的分量$\dfrac{p_{ij}}{p_{i.}}$表示条件概率$P(B=j|A=i)$，可知

$$\boldsymbol{p}_i^{r\prime}\mathbf{1} = 1 \tag{7.4}$$

形象地，第i个行剖面$\boldsymbol{p}_i^r$就是把矩阵$\boldsymbol{P}$中第i行剖裂开来，单独研究第i行的各个取值在p维超平面$x_1+x_2+\cdots+x_p=1$上的分布情况。记n个行剖面的集合为$n(r)$。

由于列联表的行与列的地位是对等的，由上面定义行剖面的方法可以很容易地定义列剖面。对矩阵$\boldsymbol{P}$第j列的每一个元素p_{ij}均除以该列各元素的和$p_{.j}$，则第j个列剖面：

$$\boldsymbol{p}_j^{c\prime} = \left(\frac{p_{1j}}{p_{.j}}, \frac{p_{2j}}{p_{.j}}, \cdots, \frac{p_{nj}}{p_{.j}}\right), \quad j = 1,2,\cdots,p \tag{7.5}$$

表示当属性变量B的取值为j时，属性变量A的不同取值的条件概率，它是n维超平面$x_1+x_2+\cdots+x_n=1$上的一个点。有$\boldsymbol{p}_j^{c\prime}\mathbf{1}=1$，记$p$个列剖面的集合为$p(c)$。

在定义了行剖面与列剖面之后，我们看到，属性变量A的各个取值的情况可以用p维空间上的n个点来表示，而B的不同取值情况可以用n维空间上的p个点来表示。而对应分析就是利用降维的思想，既把A的各个状态表现在一张二维图上，又把B的各个状态表现在一张二维图上，且通过后面的分析可以看到，这两张二维图的坐标轴有相同的含义，即可以把A的各个取值与B的各个取值同时在一张二维图上表示出来。

2. 距离与总惯量

通过上面行剖面与列剖面的定义，A的不同取值就可以用p维空间中的不同点来表示，各个点的坐标分别为$\boldsymbol{p}_i^r(i=1, 2, \cdots, n)$；$B$的不同取值可以用$n$维空间中的不同点来表示，各个点的坐标分别为$\boldsymbol{p}_j^c(j=1, 2, \cdots, p)$。对此，可以引入距离的概念来分别描述$A$的各个状态之间与$B$的各个状态之间的接近程度。因为对

列联表行与列的研究是对等的，此处只对行做详细论述。

变量 A 的第 k 状态与第 l 状态的普通欧氏距离为：

$$d^2(k,l)=(\boldsymbol{p}_k^r-\boldsymbol{p}_l^r)'(\boldsymbol{p}_k^r-\boldsymbol{p}_l^r)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{lj}}{p_{l.}}\right)^2 \tag{7.6}$$

如此定义的距离有一个缺点，即受到变量 B 的各个状态边缘概率的影响，当变量 B 的第 j 个状态出现的概率特别大时，式（7.6）所定义距离的 $\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{lj}}{p_{l.}}\right)^2$ 部分的作用就被高估了，因此，用 $\frac{1}{p_{.j}}$ 作权重，得到如下加权的距离公式：

$$\begin{aligned}D^2(k,l)&=\sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{lj}}{p_{l.}}\right)^2\Big/p_{.j}\\&=\sum_{j=1}^{p}\left(\frac{p_{kj}}{\sqrt{p_{.j}}p_{k.}}-\frac{p_{lj}}{\sqrt{p_{.j}}p_{l.}}\right)^2\end{aligned} \tag{7.7}$$

因此，式（7.7）定义的距离也可以看作坐标为：

$$\left(\frac{p_{i1}}{\sqrt{p_{.1}}p_{i.}},\frac{p_{i2}}{\sqrt{p_{.2}}p_{i.}},\cdots,\frac{p_{ip}}{\sqrt{p_{.p}}p_{i.}}\right),\quad i=1,2,\cdots,n \tag{7.8}$$

的任意两点之间的普通欧氏距离。

类似地，定义属性变量 B 的两个状态 s，t 之间的加权距离为：

$$D^2(s,t)=\sum_{i=1}^{n}\left(\frac{p_{is}}{\sqrt{p_{i.}}p_{.s}}-\frac{p_{it}}{\sqrt{p_{i.}}p_{.t}}\right)^2 \tag{7.9}$$

式（7.8）是行剖面消除了变量 B 的各个状态概率影响的相对坐标，下面给出式（7.8）定义的各点的平均坐标，即重心的表达式。由行剖面的定义，$\boldsymbol{p}_i^r$ 的各分量是当 A 取 i 时变量 B 各个状态出现的条件概率，也就是说，式（7.8）的坐标也同时消除了变量 A 的各个状态出现的概率影响。然而，当我们研究由式（7.8）定义的 n 个点的平均坐标时，这 n 个点的地位不是完全平等的，出现概率较大的状态应当占有较高的权重。因此，我们定义如下按 $p_{i.}$ 加权的 n 个点的平均坐标，其第 j 个分量为：

$$\sum_{i=1}^{n}\frac{p_{ij}}{\sqrt{p_{.j}}p_{i.}}p_{i.}=\frac{1}{\sqrt{p_{.j}}}\sum_{i=1}^{n}p_{ij}=\sqrt{p_{.j}},\quad j=1,2,\cdots,p \tag{7.10}$$

因此，由式（7.8）定义的 n 个点的重心为：

$$\boldsymbol{p}_J^{1/2\prime}=(\sqrt{p_{.1}},\sqrt{p_{.2}},\cdots,\sqrt{p_{.p}})$$

其中，每一分量恰恰是矩阵 $\boldsymbol{P}$ 每一列边缘概率的平方根。根据上面的准备，可以给

出如下行剖面集合 $n(r)$的总惯量的定义：由式（7.8）定义的 n 个点与其重心的加权欧氏距离之和称为行剖面集合 $n(r)$的总惯量，记为 I_I。有

$$I_I = \sum_{i=1}^{n} D^2(\boldsymbol{p}_i^r, \boldsymbol{p}_J^{1/2}) \tag{7.11}$$

令 $\boldsymbol{D}_p^{1/2}=\text{diag}(\boldsymbol{p}_J^{1/2})$ 表示由向量 $\boldsymbol{p}_J^{1/2}$的各个分量为对角线元素构成的对角阵，则总惯量式（7.11）可写为：

$$\begin{aligned} I_I &= \sum_{i=1}^{n} d^2[\boldsymbol{p}_i^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}, \boldsymbol{p}_J^{1/2\prime}] = \sum_{i=1}^{n}\sum_{j=1}^{p} p_{i.}\left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}} - \sqrt{p_{.j}}\right)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \frac{1}{n}\chi^2 \end{aligned} \tag{7.12}$$

由式（7.12）可以看到，总惯量不仅反映了行剖面集在式（7.8）意义上定义的各点与其重心加权距离的总和，同时与 χ^2 统计量仅相差一个常数，而由前面列联表的分析我们知道，χ^2 统计量反映了列联表横栏与纵列的相关关系，因此，此处总惯量也反映了两个属性变量各状态之间的相关关系。对应分析就是在总惯量信息损失最小的前提下，简化数据结构以反映两属性变量之间的相关关系。实际上，总惯量的概念类似于主成分分析或因子分析中方差总和的概念，在 SPSS 软件中进行对应分析时，系统会给出对总惯量信息的提取情况。

完全对应地，可以得到对列联表的列进行分析的相应结论，列剖面 p 个点经 $p_{.j}$加权后的平均坐标，即重心为：

$$\boldsymbol{p}_I^{1/2\prime} = (\sqrt{p_{1.}}, \sqrt{p_{2.}}, \cdots, \sqrt{p_{n.}}) \tag{7.13}$$

列剖面集合 $p(c)$的总惯量为：

$$I_J = I_I = \frac{1}{n}\chi^2 \tag{7.14}$$

7.2.2 R 型与 Q 型因子分析的对等关系

经过以上数据变换，在引入加权距离函数之后，或者对行剖面集的各点进行式（7.8）的变换，对列剖面的各点进行类似变换之后，可以直接计算属性变量各状态之间的距离，通过距离的大小来反映各状态之间的接近程度，同类型的状态之间距离应当较短，而不同类型的状态之间距离应当较长，据此可以对各种状态进行分类以简化数据结构。但是，这样做不能对两个属性变量同时进行分析，因此不计算距离，而代之求协方差矩阵，进行因子分析，提取主因子，用主因子所定义的坐标轴作为参照系，对两个变量的各状态进行分析。

先对行剖面进行分析，即 Q 型因子分析。假定各个行剖面的坐标均经过了形如式（7.8）的变换，以消除变量 B 的各个状态发生的边缘概率的影响。即变换后的

行剖面为：

$$\boldsymbol{p}_i^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}, \qquad i=1,2,\cdots,n$$

则变换后的 n 个行剖面所构成的矩阵为：

$$\boldsymbol{p}_r=\begin{bmatrix}\boldsymbol{p}_1^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}\\ \boldsymbol{p}_2^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}\\ \vdots\\ \boldsymbol{p}_n^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}\end{bmatrix} \tag{7.15}$$

进行 Q 型因子分析就是从矩阵 $\boldsymbol{p}_r$ 出发，分析其协方差矩阵，提取公共因子（主成分）的分析。设 $\boldsymbol{p}_r$ 的加权协方差阵为 $\boldsymbol{\Sigma}_r$，则有

$$\boldsymbol{\Sigma}_r=\sum_{i=1}^{n}p_{i\cdot}[(\boldsymbol{D}_p^{1/2})^{-1}\boldsymbol{p}_i^r-\boldsymbol{p}_J^{1/2}][\boldsymbol{p}_i^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}-\boldsymbol{p}_J^{1/2\prime}] \tag{7.16}$$

因为对任意的 i（$i=1$，2，…，n），有

$$\begin{aligned}&[\boldsymbol{p}_i^{r\prime}(\boldsymbol{D}_p^{1/2})^{-1}-\boldsymbol{p}_J^{1/2\prime}]\boldsymbol{p}_J^{1/2}\\&=\begin{bmatrix}\dfrac{p_{i1}-p_{i\cdot}p_{\cdot1}}{\sqrt{p_{\cdot1}}p_{i\cdot}} & \dfrac{p_{i2}-p_{i\cdot}p_{\cdot2}}{\sqrt{p_{\cdot2}}p_{i\cdot}} & \cdots & \dfrac{p_{ip}-p_{i\cdot}p_{\cdot p}}{\sqrt{p_{\cdot p}}p_{i\cdot}}\end{bmatrix}\begin{bmatrix}\sqrt{p_{\cdot1}}\\ \sqrt{p_{\cdot2}}\\ \vdots\\ \sqrt{p_{\cdot p}}\end{bmatrix}=0\end{aligned} \tag{7.17}$$

所以，$\boldsymbol{\Sigma}_r\boldsymbol{p}_J^{1/2}=0$。也就是说，变换后行剖面点集的重心 $\boldsymbol{p}_J^{1/2}$ 是 $\boldsymbol{\Sigma}_r$ 的一个特征向量，且其对应的特征根为零。因此，该因子轴对公共因子的解释而言是无用的，在对应分析中，总是不考虑该轴。实际上，在对列剖面进行分析时，也存在类似的情况，$\boldsymbol{p}_I^{1/2}$ 是变换后列剖面集所构成矩阵的协方差矩阵的一个特征向量，且其对应的特征根也为零。因此，因子轴 $\boldsymbol{p}_I^{1/2}$ 也是无用的。

为了更清楚地了解对应分析的具体计算过程，我们看一下 $\boldsymbol{\Sigma}_r$ 中的元素。设

$$\boldsymbol{\Sigma}_r=(a_{ij})_{p\times p}$$

则有

$$\begin{aligned}a_{ij}&=\sum_{a=1}^{n}\left(\frac{p_{ai}}{\sqrt{p_{\cdot i}}p_{a\cdot}}-\sqrt{p_{\cdot i}}\right)\left(\frac{p_{aj}}{\sqrt{p_{\cdot j}}p_{a\cdot}}-\sqrt{p_{\cdot j}}\right)p_{a\cdot}\\&=\sum_{a=1}^{n}\left(\frac{p_{ai}}{\sqrt{p_{\cdot i}}\sqrt{p_{a\cdot}}}-\sqrt{p_{\cdot i}}\sqrt{p_{a\cdot}}\right)\left(\frac{p_{aj}}{\sqrt{p_{\cdot j}}\sqrt{p_{a\cdot}}}-\sqrt{p_{\cdot j}}\sqrt{p_{a\cdot}}\right)\\&=\sum_{a=1}^{n}\left(\frac{p_{ai}-p_{\cdot i}p_{a\cdot}}{\sqrt{p_{\cdot i}p_{a\cdot}}}\right)\left(\frac{p_{aj}-p_{\cdot j}p_{a\cdot}}{\sqrt{p_{\cdot j}p_{a\cdot}}}\right)\\&=\sum_{a=1}^{n}z_{ai}z_{aj}\end{aligned} \tag{7.18}$$

其中 $z_{ij}=\dfrac{p_{ij}-p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}$，$i=1,2,\cdots,n$；$j=1,2,\cdots,p$

若令 $\boldsymbol{Z}=(z_{ij})$，则有

$$\boldsymbol{\Sigma}_r=\boldsymbol{Z}\boldsymbol{Z}' \tag{7.19}$$

依照上述方法，可以对列剖面进行分析，设变换后的列剖面集所构成矩阵的协方差矩阵为 $\boldsymbol{\Sigma}_c$，则可以得到

$$\boldsymbol{\Sigma}_c=\boldsymbol{Z}'\boldsymbol{Z} \tag{7.20}$$

其中，矩阵 $\boldsymbol{Z}$ 的定义与上面完全一致。这样，对应分析的过程就转化为基于矩阵 $\boldsymbol{Z}$ 的分析过程，由式（7.19）和式（7.20）可以看出，矩阵 $\boldsymbol{\Sigma}_r$ 与 $\boldsymbol{\Sigma}_c$ 存在简单的对等关系，如果把原始列联表中的数据 n_{ij} 变换成 z_{ij}，则 z_{ij} 对两个属性变量有对等性。

由矩阵的知识可知，$\boldsymbol{\Sigma}_r=\boldsymbol{Z}\boldsymbol{Z}'$ 与 $\boldsymbol{\Sigma}_c=\boldsymbol{Z}'\boldsymbol{Z}$ 有完全相同的非零特征根，记作 λ_1，λ_2，$\cdots$，λ_r（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r$），而经过上面的分析可知，$\boldsymbol{\Sigma}_r$ 与 $\boldsymbol{\Sigma}_c$ 均有一个特征根为零，且其所对应的特征向量分别为 $\boldsymbol{p}_j^{1/2}$，$\boldsymbol{p}_i^{1/2}$，由这两个特征向量构成的因子轴为无用轴。因此，在对应分析中，公共因子轴的最大维数为 $\min(n,p)-1$。所以有 $0<r\leqslant\min(n,p)-1$。设 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$，$\cdots$，$\boldsymbol{u}_r$ 为相对于特征根 λ_1，λ_2，$\cdots$，λ_r 的 $\boldsymbol{\Sigma}_r$ 的特征向量，则有

$$\boldsymbol{\Sigma}_r\boldsymbol{u}_j=\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{u}_j=\lambda_j\boldsymbol{u}_j \tag{7.21}$$

对上式两边左乘矩阵 $\boldsymbol{Z}'$，有

$$\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{u}_j)=\lambda_j(\boldsymbol{Z}'\boldsymbol{u}_j)$$

即

$$\boldsymbol{\Sigma}_c(\boldsymbol{Z}'\boldsymbol{u}_j)=\lambda_j(\boldsymbol{Z}'\boldsymbol{u}_j) \tag{7.22}$$

表明 $(\boldsymbol{Z}'\boldsymbol{u}_j)$ 即为相对于特征根 λ_j 的 $\boldsymbol{\Sigma}_c$ 的特征向量，这就建立了对应分析中 R 型因子分析与 Q 型因子分析的关系，这样，就可以由 R 型因子分析的结果很方便地得到 Q 型因子分析的结果，从而大大减少了计算量，特别是克服了当某一属性变量的状态特别多时计算上的困难。又由于 $\boldsymbol{\Sigma}_r$ 与 $\boldsymbol{\Sigma}_c$ 具有相同的非零特征根，而这些特征根正是各个公共因子所解释的方差，或提取的总惯量的份额，即有 $\sum_{i=1}^{r}\lambda_i=I_I=I_J$。那么，在变量 B 的 p 维空间 R^p 中的第一主因子、第二主因子……直到第 r 个主因子与变量 A 的 n 维空间 R^n 中相对应的各个主因子在总方差中所占的百分比完全相同。这样就可用相同的因子轴同时表示两个属性变量的各个状态，把两个变量的各个状态同时反映在具有相同坐标轴的因子平面上，以直观地反映两个属性变量及各个状态之间的相关关系。一般情况下，我们取两个公共因子，这样，就可以在一张二维图上同时画出两个变量的各个状态。

7.2.3 对应分析应用于定量变量的情况

上面对对应分析方法的描述都是以属性变量数据为例展开的，这是因为在实际中，对应分析广泛地应用于对属性变量列联表数据的研究。实际上，对应分析方法

也适用于定距尺度与定比尺度的数据。假设要分析的数据为 $n\times p$ 的表格形式（n 个观测，p 个变量），沿用上面的思想，同样可以对数据进行规格化处理，再进行 R 型因子分析与 Q 型因子分析，进而把观测与变量在同一张低维图形上表示出来，分析各观测与各变量之间的接近程度。

其实，对于定距尺度与定比尺度的情况，完全可以把每一个观测都分别看成一类，这也是对原始数据进行的最细的分类；同时把每一个变量都看成一类。这样，对定距尺度数据与定比尺度数据的处理问题就变成与上面分析属性变量相同的问题了，自然可以运用对应分析来研究行与列之间的相关关系。但是应当注意，对应分析要求数据阵中每一个数据都是大于或等于零的，当用对应分析研究普通的 $n\times p$ 的表格形式的数据时，若有小于零的数据，则应当先对数据进行加工，比如将该变量的各个取值都加上一个常数。有的研究人员将对应分析方法用于对经济问题截面数据的研究，得到了比较深刻的结论。

本章第 4 节将通过例 7—2 给出一个具体的对应分析应用于分类汇总数据的例子，对该问题以及多重对应分析有兴趣的读者请参阅参考文献［2］和［6］。

7.2.4　需要注意的问题

需要注意的是，用对应分析生成的二维图上的各状态点，实际上是两个多维空间上的点的二维投影，在某些特殊的情况下，在多维空间中相隔较远的点，在二维平面上的投影却很接近。此时，我们需要对二维图上的各点做更深入的了解，即哪些状态对公共因子的贡献较大，这与因子分析在判断原始变量对公共因子贡献的方法类似，不同的是，因为对应分析中 $\boldsymbol{\Sigma}_r$ 与 $\boldsymbol{\Sigma}_c$ 存在简单的对等关系，我们可以任选一个变量，分析其各个状态对公共因子的贡献，不妨以变量 A 的各个状态为例进行说明。由于

$$\mathrm{var}(F_k)=\sum_{i=1}^{n}p_{i\cdot}\,a_{ik}^2=\lambda_k$$

式中，a_{ik} 为因子载荷。设状态 i 对公共因子的贡献为 $CTR(i)$，于是有 $CTR(i)=p_{i\cdot}\,a_{ik}^2/\lambda_k$，$CTR(i)$ 的值越大，说明状态 i 对第 k 个公共因子的贡献越大。同时，如有需要，我们可以仿照因子分析的方法分析每一个公共因子的贡献的大小，在此不再详述。

另外还需注意的是，对应分析只能用图形的方式提示变量之间的关系，但不能给出具体的统计量来度量这种相关程度，这容易使研究者在运用对应分析时得出主观性较强的结论。

7.3　对应分析的步骤及逻辑框图

7.3.1　对应分析的步骤

由前面的分析可知，对来源于实际问题的列联表数据，运用对应分析方法进行

研究的过程可以最终转化为进行 R 型因子分析与 Q 型因子分析的过程。一般来说，对应分析应包括如下几个步骤：

（1）由原始列联表数据计算规格化的概率意义上的列联表。

（2）计算 $\boldsymbol{Z}$ 矩阵。

（3）由 $\boldsymbol{\Sigma}_r$ 或 $\boldsymbol{\Sigma}_c$ 出发进行 R 型因子分析或 Q 型因子分析，并由 R(或 Q) 型因子分析的结果推导出 Q(或 R) 型因子分析的结果。

（4）在二维图上画出原始变量各个状态，并对原始变量相关性进行分析。

7.3.2 对应分析的逻辑框图

对应分析的逻辑框图如图 7—1 所示。

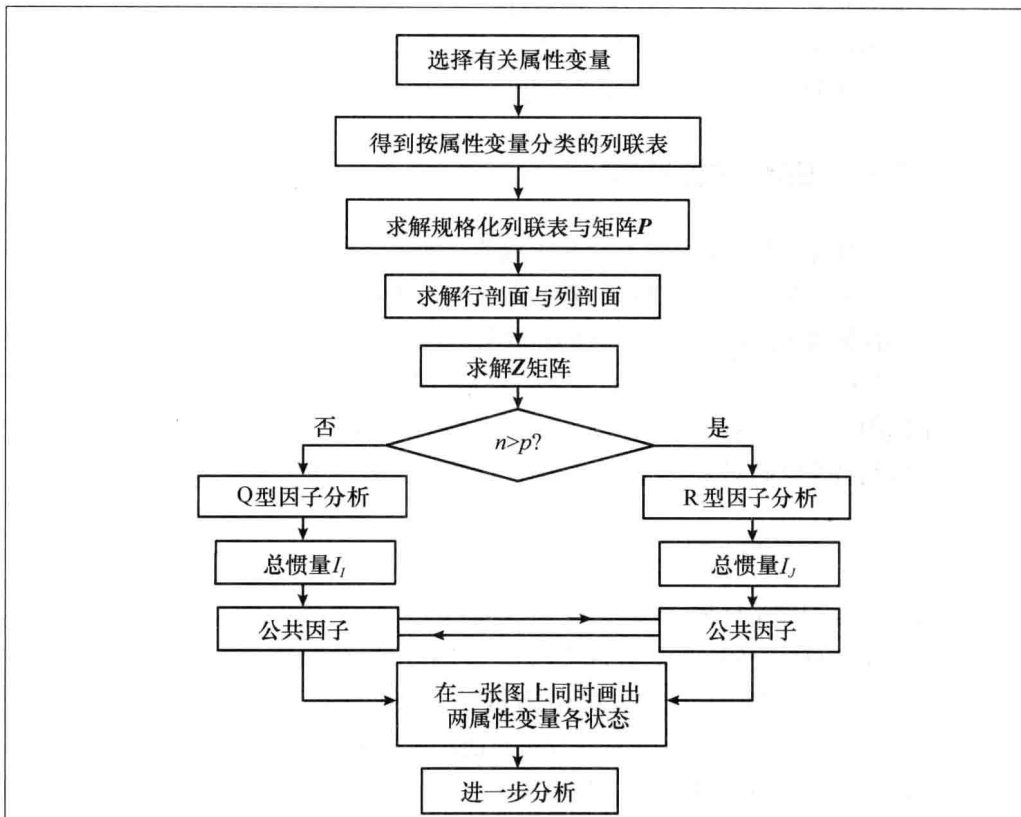


图 7—1 对应分析逻辑框图

7.4 对应分析的上机实现

SPSS 软件的 Correspondence Analysis 模块是专门进行对应分析的模块。下面举例说明用 Correspondence Analysis 模块进行对应分析的方法。

例 7—1

选用 GSS93 subset. sav 数据。该数据共包括 1 500 个观测，67 个变量。我们仅借助它来说明 Correspondence Analysis 模块的使用方法，不对其具体意义做过多的分析。选用该数据集中 Degree（学历）与 Race（人种）变量为例来说明。其中，Degree 变量是定序尺度的，其各个取值的含义如下：0——中学以下（Less than high school）；1——中学（High school）；2——专科（Junior college）；3——本科（Bachelor）；4——研究生（Graduate）；7，8，9——缺失。Race 变量是定类尺度的，其各个取值的含义如下：1——白人（white）；2——黑人（black）；3——其他（other）。

打开 GSS93 subset. sav 数据，对变量 Degree 与变量 Race 进行对应分析，依次点选 Analyze→Dimension Reduction→Correspondence Analysis…进入 Correspondence Analysis 对话框。数据集中所有的变量名（标签）均已出现在左边的窗口中，将 Degree 变量选入右侧行变量（Row）的小窗口中，此时该窗口显示的 Degree 变量形如 Degree(? ?)，同时，其下方的 Define Range 按钮被激活，点击该按钮，进入 Define Row Range 对话框，在该对话框中需要确定 Degree 变量的取值范围，此处我们不研究缺失值，最小值（Minimum value）与最大值（Maximum value）处分别填上 0 和 4，按右侧的 Update（更新）按钮，可以看到 Degree 的取值 0～4 已出现在 Category Constraints 框架左侧的窗口中，该框架的作用是对 Degree 的各状态加以限定，保持默认值 None 不变，即对 Degree 的取值不加以限定。点击 Continue 继续，回到 Correspondence Analysis 对话框，可以看到，此时行变量 Degree 的显示变为 Degree(0 4)。按照同样的方法把 Race 选为列变量且设定其取值范围为 1～3，点击 OK 按钮运行，则可以得到输出结果 7—1。

输出结果 7—1　　**Correspondence Table**

RS Highest Degree	Race of Respondent			
	white	black	other	Active Margin
Less than HS	214	48	17	279
High school	658	92	30	780
Junior college	74	13	3	90
Bachelor	209	7	18	234
Graduate	99	7	7	113
Active Margin	1 254	167	75	1 496

(1)

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	0. 144	0. 021			0. 852	0. 852	0. 021	0. 065
2	0. 060	0. 004			0. 148	1. 000	0. 026	
Total		0. 024	36. 482	0. 000[a]	1. 000	1. 000		

a. 8 degrees of freedom.

(2)

Overview Row Points*

RS Highest Degree	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
Less than HS	0.186	−0.462	−0.414	0.008	0.276	0.531	0.750	0.250	1.000
High school	0.521	−0.078	0.192	0.002	0.022	0.322	0.285	0.715	1.000
Junior college	0.060	−0.304	0.193	0.001	0.039	0.037	0.857	0.143	1.000
Bachelor	0.156	0.723	−0.203	0.012	0.566	0.107	0.968	0.032	1.000
Graduate	0.076	0.429	−0.041	0.002	0.096	0.002	0.996	0.004	1.000
Active Total	1.000			0.024	1.000	1.000			

* Symmetrical normalization.

(3)

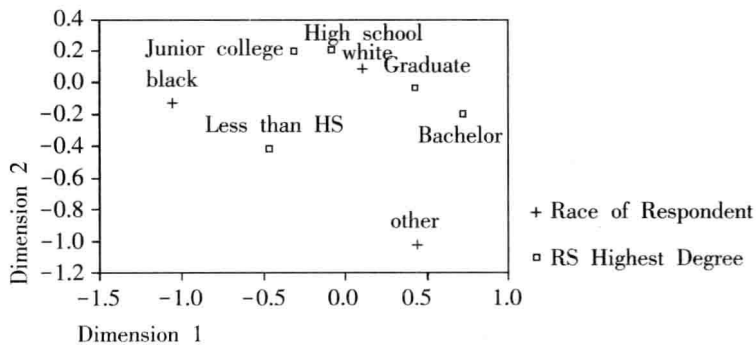
Overview Column Points*

Race of Respondent	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
white	0.838	0.113	0.079	0.002	0.074	0.088	0.830	0.170	1.000
black	0.112	−1.051	−0.134	0.018	0.855	0.033	0.993	0.007	1.000
other	0.050	0.452	−1.026	0.005	0.071	0.879	0.318	0.682	1.000
Active Total	1.000			0.024	1.000	1.000			

* Symmetrical normalization.

(4)

Row and Column Points
Symmetrical Normalization



(5)

其中，输出的第一部分 Correspondence Table 表是由原始数据按 Degree 与 Race 分类的列联表，可以看到观测总数 $n=1\ 496$ 而不是原始数据观测个数 1 500，这是因为原始数据中有 4 条记录有缺失。

第二部分 Summary 表给出了总惯量、χ^2 值及每一维度（公共因子）所解释的总惯量的百分比的信息。可知总惯量为 0.024，χ^2 值为 36.482，有关系式：$36.482=0.024\times1\ 496$[①]，由此可以清楚地看到总惯量与 χ^2 值的关系，同时说明总惯量描述了列联表行与列之间总的相关关系。Singular Value 反映的是行与列各状态在二维图中分值的相关程度，实际上是对行与列进行因子分析产生的新的综合变量的典型相关系数，其在取值上等于特征根的平方根。Sig. 是假设 χ^2 值为 0 成立的概率，表注表明自由度为 $(5-1)\times(3-1)=8$，Sig. 值很小说明列联表的行与列之间有较

① 存在误差。

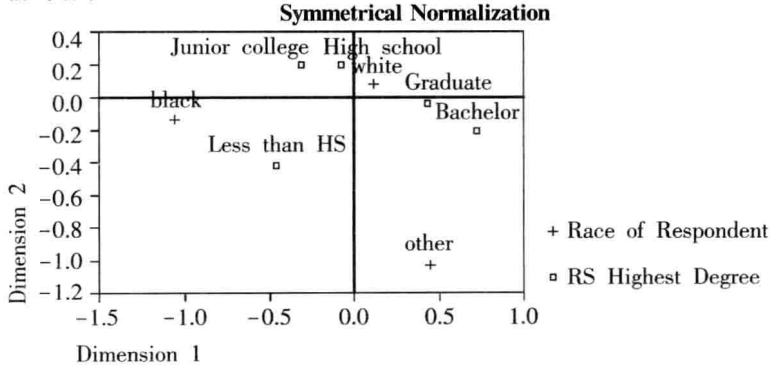
强的相关性。Proportion of Inertia 部分是各维度（公共因子）分别解释总惯量的比例及累计百分比，类似于因子分析中公共因子解释能力的说明。

第三部分和第四部分是对列联表行与列各状态有关信息的概括。其中，Mass 部分分别指列联表中行与列的边缘概率，也就是 $\boldsymbol{P}_I$ 与 $\boldsymbol{P}_J$。Score in Dimension 是各维度的分值，也就是行与列各状态在二维图中的坐标值。Inertia 是惯量，是每一行（列）与其重心的加权距离的平方，可以看到 $I_I=I_J=0.024$，即行剖面的总惯量等于列剖面的总惯量。Contribution 部分是指行（列）的每一状态对每一维度（公共因子）特征根的贡献及每一维度对行（列）各个状态的特征根的贡献。由此可以更好地理解维度的来源及意义，如第一维度中，Bachelor 对应的数值最大，为 0.566，说明 Bachelor 这一状态对第一维度的贡献最大。在表的最后部分维度对各状态特征根的贡献部分，看到除 High school 外，其余各最高学历的特征根的分布大部分集中在第一维度上，说明第一维度反映了最高学历各状态大部分的差异，这实际上相当于因子分析中对共同度的分解。

输出的最后一部分是 Degree（学历）各状态与 Race（人种）的各状态同时在一张二维图上的投影。在图上既可以看到每一变量内部各状态之间的相关关系，又可以同时考察两变量之间的相关关系。为了更清楚地显示各状态之间的距离，我们可以给上图画上 X 轴与 Y 轴的参考线，方法如下：在 SPSS 的结果输出窗口中，双击该图形，进入图形编辑窗口，可以看到顶部的菜单相应发生了变化，点击 Options，选择 X Axis Reference line，进入 Properties 对话框，在 Reference Line 的 Position 文本框中看到默认为 0，点击 Add 按钮，则 0 出现在下方的主窗口中，表明画出参考线 $X=0$，输入别的值再按 Add 可以画出其他的参考线，而点击 Remove 按钮可以移去相应的参考线。此处我们只画出 $X=0$ 的参考线，按 OK 继续，可以看到 $X=0$ 的参考线已经出现在图形中。用同样的方法，画出 Y 的参考线，然后关闭图形编辑窗口，则输出窗口的图形也发生了变化，上面的二维图变为输出结果 7—2 所示的形式。

输出结果 7—2

Row and Column Points
Symmetrical Normalization



在同一变量内部，最高学历为 High school 及以上的各状态之间距离相近，而 Less than high school 可以单独归为一类；对于人种，black，white，other 之间的距离均很大，很明显形成三大类。同时考察两变量各状态，可以看到白人（white）受教育程度一般较高，其与学历较高的点比较接近，而黑人明显学历较低，与 Less

than high school 比较靠近。other 的最高学历没有显著特点。

以上是由 SPSS 默认设置得到的结果。实际研究时，可以根据不同的研究目的对有关设置进行修改。下面对 SPSS 提供的有关选项进行简要说明。在 Correspondence Analysis 对话框中点击右侧的 Model 按钮进入 Model 对话框，在该对话框中，可以设定进行对应分析的有关方法：在上部 Dimensions in solution 处可以规定对应分析的最大维数，默认维数是 2。由本章的论述知，最大维数应该是 $\min(n, p)-1$，此处保留默认值即可。Distance Measure 对话框中可以规定距离量度方法，默认为卡方距离，也就是加权的欧氏距离，还可以规定用欧氏距离（Euclidean）。在 Standardization Method 对话框中可以规定标准化方法，若距离的量度使用卡方距离，则应使用默认的标准化方法，即对行与列均进行中心化处理；若选择欧氏距离，则有不同的标准化方法可以选择，此处不再详述。最下方 Normalization Method 框架中可以规定不同的正态化方法，默认为 Symmetrical 方法，当我们的分析目的是考察两变量各状态之间的差异性或相似性时，应选择此方法。当我们的目的是考察两个属性变量之间各状态及同一变量内部各状态之间的差异性时，则应当选择 Principal 方法。当我们的目的是考察不同行（列）之间的差异性或相似性时，则应当选择 Row principal(Column principal)，而选中 Custom 并自己设定一个－1～1 之间的值，则可能输出更容易解释的二维图。

在 Correspondence Analysis 对话框中点击 Statistics 按钮，进入 Statistics 对话框，选中 Row profiles 和 Column profiles 交由程序运行，则除上面的结果外，还可以输出行剖面与列剖面，如输出结果 7—3 所示。

输出结果 7—3

Row Profiles

RS Highest Degree	Race of Respondent			
	white	black	other	Active Margin
Less than HS	0.767	0.172	0.061	1.000
High school	0.844	0.118	0.038	1.000
Junior college	0.822	0.144	0.033	1.000
Bachelor	0.893	0.030	0.077	1.000
Graduate	0.876	0.062	0.062	1.000
Mass	0.838	0.112	0.050	

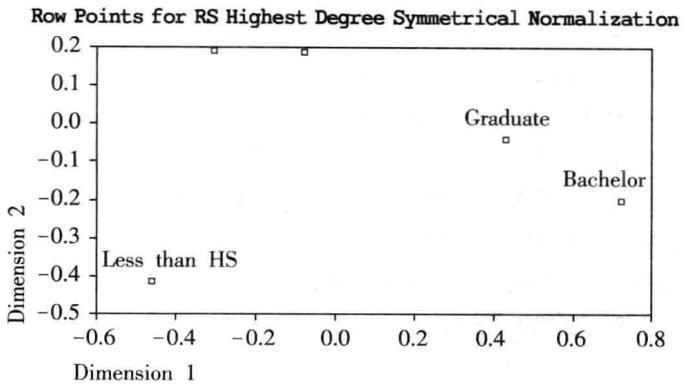
Column Profiles

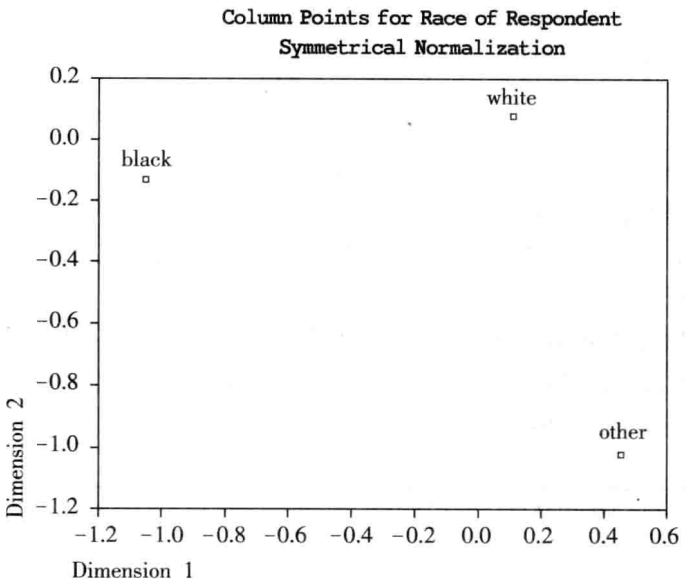
RS Highest Degree	Race of Respondent			
	white	black	other	Mass
Less than HS	0.171	0.287	0.227	0.186
High school	0.525	0.551	0.400	0.521
Junior college	0.059	0.078	0.040	0.060
Bachelor	0.167	0.042	0.240	0.156
Graduate	0.079	0.042	0.093	0.076
Active Margin	1.000	1.000	1.000	

在 Statistics 对话框中选择其他选项，可以输出一些有用的统计量，这些统计量有助于检验对应分析的效果，此处不再详述。在 Correspondence Analysis 对话框中，点击 Plots 按钮进入 Plots 对话框，看到在 Scatterplots 框架中系统默认输出 Biplot，即在同一张二维图上同时输出两个属性变量的各个状态。为了考察列联表

各行（列）之间的相关性，有时候有必要输出仅包括一个变量各种状态参数的二维图，选择 Row points 及 Column points 可以实现。同时选中 Row points 与 Column points 并交由程序运行，则可以得到输出结果 7—4。

输出结果 7—4





这样可以更清楚地考察每一变量各个状态之间的距离或接近程度。SPSS 软件还提供了许多其他有用的选项，可以针对不同的研究问题及研究目的，选择这些选项以得到更多的结果，此处不再详细说明。

因为对应分析所需的数据是原始列联表数据，所以 SPSS 软件也提供了直接读入列联表数据的功能，这样对上例，就不用从 1 500 条原始观测开始进行分析，从而提高了分析的效率。由输出结果 7—1 对列联表作如下变换，如表 7—4 所示。

表 7—4

row	column	freq	row	column	freq
0	1	214	2	3	3
0	2	48	3	1	209
0	3	17	3	2	7
1	1	658	3	3	18

续前表

row	column	freq	row	column	freq
1	2	92	4	1	99
1	3	30	4	2	7
2	1	74	4	3	7
2	2	13			

在 SPSS 的数据窗口输入以上数据，依次点选 Data→Weight Cases…进入 Weight Cases 对话框，系统默认是对观测不使用权重，选中 Weight Cases by 选项，此时下面的 Frequency variable 被激活，选中 Freq 并点击箭头，使变量 Freq 充当权数的作用，点击 OK。然后按上述方法选择变量，设定取值范围并进行分析（此处行变量为 Row，取值范围为 0～4；列变量为 Column，取值范围为 1～3），可以得到与上面一致的结果。为了比较，此处仅给出所输出的列联表，如输出结果 7—5 所示。

输出结果 7—5　　Correspondence Table

ROW	COLUMN			
	1	2	3	Active Margin
0	214	48	17	279
1	658	92	30	780
2	74	13	3	90
3	209	7	18	234
4	99	7	7	113
Active Margin	1 254	167	75	1 496

可以看到，列联表与输出结果 7—1 是相同的，但此处由于没有对 Row 与 Column 的取值设定标签，所以显示的是其实际取值，这可以在 Variable View 窗口进行设定，此处不再进行说明。

这样读入数据仍然有些麻烦，利用 SPSS 的语法可以读入列联表形式的数据。实际上，利用 SPSS 语法可以灵活地读入以上各种格式的数据。这里给出上例的情况下所用的程序（见表 7—5），但不再详细说明。

表 7—5

```
DATA LIST FREE/ROWCAT _ COL1 COL2 COL3.
BEGIN DATA.
0  214    48  17
1  658    92  30
2   74    13   3
3  209     7  18
4   99     7   7
END DATA.
VARIABLE LABELS
COL1'white'COL2'black'COL3'other'.
VALUE LABELS ROWCAT _ 0'less than high school'1'high school'2'junior
college'3'Bachelor'4'Graduate'.
CORRESPONDENCE TABLE = ALL (5, 3)
  /DIMENSIONS = 2
  /MEASURE = CHISQ
  /STANDARDIZE = RCMEAN
  /NORMALIZATION = SYMMETRICAL
  /PRINT = TABLE RPOINTS CPOINTS
  /PLOT = NDIM (1, MAX) BIPLOT (20).
```

前面的例子是关于利用对应分析对交叉表数据进行分析的，下面通过例题讲述如何利用对应分析对分类汇总数据进行对应分析。分类汇总的数据单元格内不再是频数，而是相应的统计指标，如均数等。

对汇总数据，由于单元格内不再是频数，不存在行、列合计频数，也就不能再像交叉表时一样基于无效假设计算标准化残差，而是使用欧氏距离来代表相应单元格数值偏离无关联假设的程度。由于指标量纲以及量级的差异，对应分析中针对欧氏距离提供了 5 种标准化方法，含义如下：

（1）Row and Column Means Removed：为缺省设置，在数据标准化时将行合计均数以及列合计均数的影响都移去，这样行、列类别间均数的差异不再对结果产生影响，在结果中呈现的只是行、列变量类别间的交互作用。

（2）Row/Column Means Removed：在数据标准化时只移除行/列变量合计均数差异的影响，这样行/列均数的差异不再对结果产生影响，在结果中呈现的只是列/行变量类别间的差异。

（3）Row/Column Totals are Equalized and Row/Column Means Removed：在数据标准化时首先将原始数据除以行/列合计，然后再移除行、列均数的影响。

距离测量方式以及相应的距离标准化方法均在 Model 子对话框中选择，在对欧氏距离进行标准化后，剩余的步骤与普通的对应分析完全相同。

对一个具体问题，如何选择以上 5 种标准化方法取决于具体的研究目的，一般是在对问题进行定性分析的基础上，选择合适的方法以便对定性分析的结论进行实证分析。下面通过一个具体问题说明该方法的应用。

例 7—2

按现行统计报表制度，农民家庭人均纯收入主要由四部分构成，即工资性收入、家庭经营纯收入、财产性收入、转移性收入。表 7—6 列出了 2012 年全国 31 个省、直辖市、自治区农民家庭人均纯收入的数据。试进行对应分析，揭示全国农民家庭人均纯收入的特征以及各省、直辖市、自治区与各收入指标间的关系。

表 7—6　　2012 年各地区按来源分农村居民家庭人均纯收入

省区	工资性收入	家庭经营纯收入	财产性收入	转移性收入
北京	10 843.48	1 318.105	1 716.357	2 597.795
天津	7 922.257	4 126.286	920.999 4	1 055.995
河北	4 005.282	3 254.567	218.302 2	603.234 4
山西	3 175.504	2 334.408	140.799 9	705.914 8
内蒙古	1 459.055	4 689.107	322.977 1	1 140.171
辽宁	3 630.236	4 783.35	246.173	723.956 5
吉林	1 792.023	5 617.627	392.957 1	795.560 7
黑龙江	1 816.841	5 433.686	580.337 5	772.984 7
上海	11 477.71	902.606 3	1 381.832	4 041.533

续前表

省区	工资性收入	家庭经营纯收入	财产性收入	转移性收入
江苏	6 775.886	3 873.896	458.456 6	1 093.714
浙江	7 678.216	5 291.36	588.528 3	993.815
安徽	3 243.468	3 265.642	111.807 9	539.539 9
福建	4 474.487	4 570.445	319.801 6	602.436 1
江西	3 532.722	3 742.427	120.923 9	433.356 7
山东	4 383.221	4 234.554	257.196 1	571.569 3
河南	2 989.356	3 973.434	135.490 7	426.662 3
湖北	3 189.844	4 123.488	65.872 06	472.508 5
湖南	3 847.59	2 903.212	112.773 9	576.593 8
广东	6 804.428	2 566.102	556.471 5	615.842 9
广西	2 245.953	3 234.553	53.870 26	473.174 2
海南	2 475.565	4 182.73	173.304 9	576.402 2
重庆	3 400.774	2 975.306	175.563 5	831.631 1
四川	3 088.858	3 004.923	166.554	741.091 8
贵州	1 977.732	2 249.205	71.535 05	454.526 3
云南	1 435.871	3 328.097	234.187	418.384 5
西藏	1 201.932	3 678.656	127.712 4	711.080 2
陕西	2 727.852	2 294.43	200.052 7	540.181 2
甘肃	1 787.715	2 114.751	112.081 7	492.116 1
青海	1 989.692	2 221.922	95.259 12	1 057.508
宁夏	2 510.528	3 071.519	101.550 4	496.725 6
新疆	1 008.019	4 238.978	170.729 6	975.951 5

资料来源：中华人民共和国国家统计局：《中国统计年鉴（2013）》，北京，中国统计出版社，2014。

软件 SPSS 的实际操作和分析如下。

操作步骤：

（1）打开 SPSS 文件，在表格下方有两个选项，分别是 Data View 和 Variable View，点击 Variable View 选项，将各选项改为如下形式（见图 7—2）。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	province	Numeric	12	1	省区	{1.0, 北京}...	None	12	Right	Nominal	Input
2	income	Numeric	12	1	收入类别	{1.0, 工资}...	None	12	Right	Nominal	Input
3	money	Numeric	12	2		None	None	12	Right	Scale	Input
4											

图 7—2

其中 Values 项需要作如下设置：在弹出的对话框里，对北京至新疆的 31 省区以及工资等 4 项收入进行数字赋值（见图 7—3）。

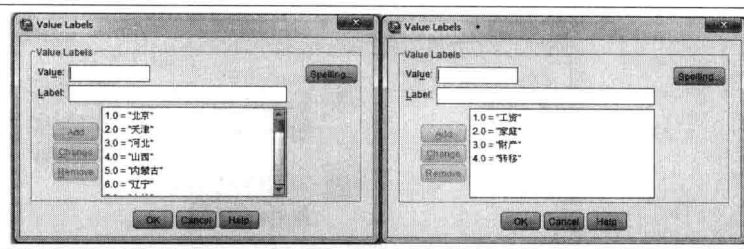


图 7—3

然后点击 Data View 进行如下数据的输入（见图 7—4）。

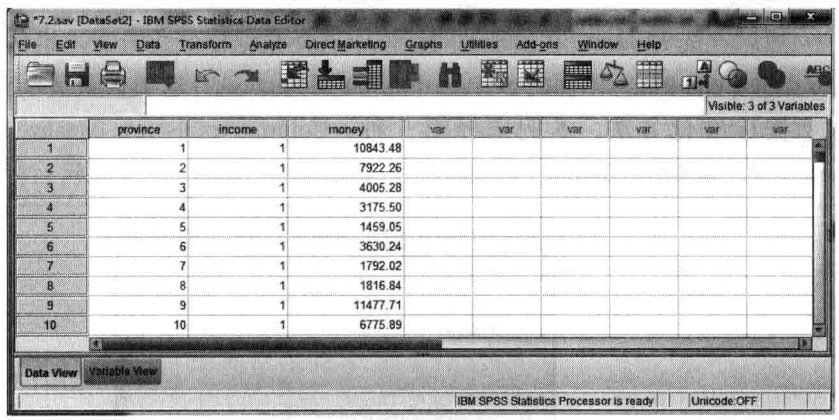


图 7—4

在 SPSS 的数据窗口输入以上数据，然后依次点选 Data→Weight Cases…进入 Weight Cases 对话框，系统默认是对观测不使用权重，选中 Weight cases by 选项，此时下面的 Frequency Variable 被激活，选中 money 并点击箭头，使变量money 充当权数的作用，点击 OK（见图 7—5）。

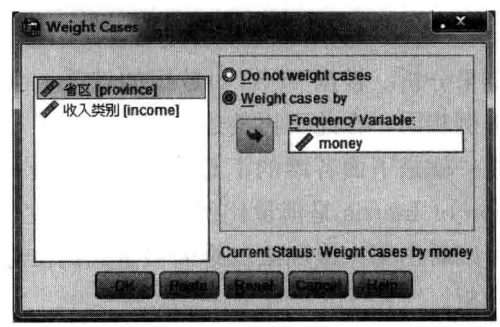


图 7—5

（2）数据输入完成后，选择 Analyze→Dimension Reduction→Correspondence Analysis，然后把“省区”选入“Row”，再点击 Define Range 来定义范围为 1（Minimum value）到 31（Maximum value），之后点击 Update，再点击 Continue。之后同样地，把“收入类别”选入 Column，并定义其范围为 1～4（见图 7—6）。

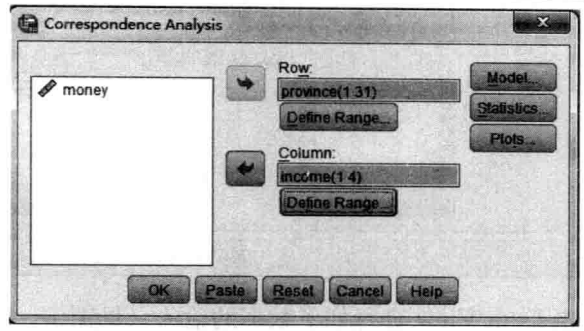


图 7—6

然后点选 Model，在出现的对话框中选择数据标准化方法，本例 Distance Measure 点选 Euclidean，下面的 Standardization Method 选择选项被激活，有 5 种可供选择的数据标准化方法，本例选择第 5 种：Column totals are equalized and means are removed，读者也可尝试使用其他方法。其余选项为默认，点击 OK 运行（见图 7—7）。

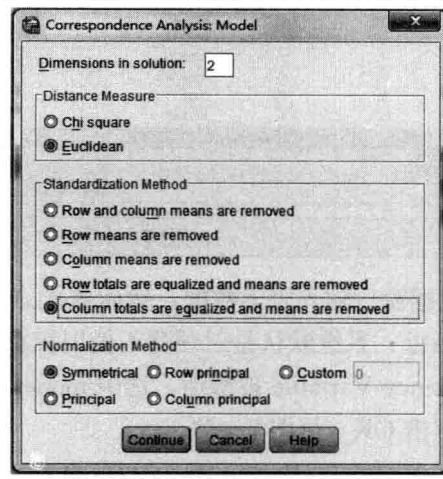


图 7—7

（3）输出结果分析。根据 SPSS 对数据的计算，会得到一系列的表格，大多数表格的解释与上例相同，在此不再赘述。

输出表格之一就是下面各维的汇总表。输出结果 7—6 中给出了行和列记分的关系。Proportion of Inertia 是惯量比例，代表各维度分别解释总惯量的比例及累计百分比，从中可以看出第一维和第二维的惯量比例占总惯量的 95.5%，因此可以选取两维来进行分析。

输出结果 7—6 **Summary**

Dimension	Singular Value	Inertia	Proportion of Inertia		Confidence Singular Value	
			Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.736	.542	.883	.883	.001	−.059
2	.210	.044	.072	.955	.002	
3	.167	.028	.045	1.000		
Total		.614	1.000	1.000		

在 SPSS 的输出结果中还给出了绘制最后叠加的散点图所需的两套坐标。首先是关于行变量（省区）的点坐标表，例如北京（－2.886，0.413），天津（－1.088，1.003）等（见输出结果 7—7）。

输出结果 7—7　　**Overview Row Points[a]**

省区	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
北京	.032	－2.886	.413	.202	.365	.026	.978	.006	.983
天津	.032	－1.088	1.003	.035	.052	.155	.805	.195	.999
河北	.032	.225	.084	.002	.002	.001	.701	.028	.729
山西	.032	.331	－.309	.004	.005	.015	.643	.159	.803
内蒙古	.032	.081	－.336	.005	.000	.017	.031	.153	.185
辽宁	.032	.190	.171	.002	.002	.005	.446	.103	.549
吉林	.032	.100	.239	.006	.000	.009	.042	.068	.110
黑龙江	.032	－.141	.479	.009	.001	.035	.051	.166	.217
上海	.032	－2.996	－1.414	.228	.393	.308	.935	.059	.995
江苏	.032	－.427	.240	.007	.008	.009	.657	.059	.716
浙江	.032	－.602	.769	.015	.016	.091	.563	.262	.826
安徽	.032	.434	－.056	.005	.008	.000	.918	.004	.922
福建	.032	.079	.429	.002	.000	.028	.086	.711	.796
江西	.032	.447	.147	.005	.009	.003	.890	.027	.917
山东	.032	.173	.327	.002	.001	.016	.383	.387	.771
河南	.032	.468	.155	.005	.010	.004	.958	.030	.988
湖北	.032	.532	.059	.007	.012	.001	.919	.003	.922
湖南	.032	.377	－.085	.005	.006	.001	.734	.011	.745
广东	.032	－.408	.659	.010	.007	.067	.411	.305	.716
广西	.032	.596	－.157	.009	.016	.004	.969	.019	.989
海南	.032	.402	.041	.004	.007	.000	.942	.003	.944
重庆	.032	.236	－.282	.002	.002	.012	.634	.258	.893
四川	.032	.299	－.229	.003	.004	.008	.813	.136	.949
贵州	.032	.584	－.270	.010	.015	.011	.839	.051	.890
云南	.032	.432	.068	.006	.008	.001	.739	.005	.745
西藏	.032	.489	－.326	.007	.010	.016	.795	.101	.896
陕西	.032	.339	－.116	.004	.005	.002	.684	.023	.707
甘肃	.032	.528	－.288	.009	.012	.013	.775	.066	.841
青海	.032	.342	－.826	.007	.005	.105	.371	.618	.989
宁夏	.032	.507	－.117	.006	.011	.002	.954	.015	.969
新疆	.032	.360	－.471	.007	.006	.034	.449	.219	.668
Active Total	1.000			.614	1.000	1.000			

a. Symmetrical normalization.

同样，列变量（收入类别）的点坐标表见输出结果 7—8，例如工资性收入（－0.716，0.282），家庭性收入（0.133，0.383）等。

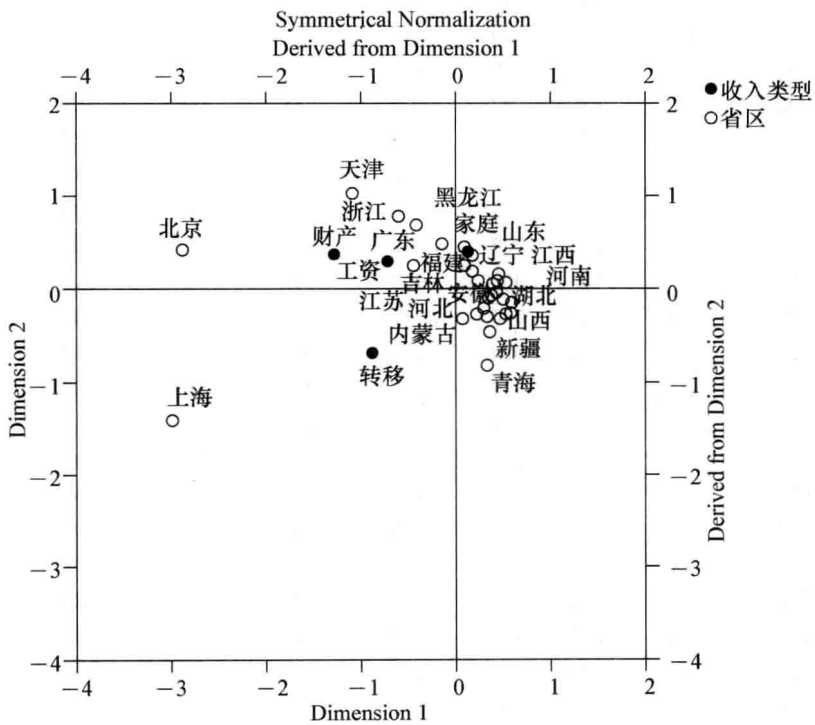
输出结果 7—8 **Overview Column Points[a]**

收入类型	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
工资	.250	−.716	.282	.117	.174	.095	.808	.036	.843
家庭	.250	.133	.383	.027	.006	.175	.121	.288	.409
财产	.250	−1.283	.358	.316	.559	.153	.960	.021	.981
转移	.250	−.877	−.696	.170	.261	.577	.834	.150	.984
Active Total	1.000			.614	1.000	1.000			

a. Symmetrical normalization.

由以上两张坐标表可以得出如下的叠加散点图（见输出结果 7—9）。从输出结果 7—9 中不难看出，我国经济发达省区，如浙江、广东、江苏、福建、天津等，主要以工资性收入和财产性收入为主；而山东、辽宁和江西等多依靠家庭经营性收入；个别省区，如上海、北京，经济发展迅速，依靠工资性收入的同时，也会有相当部分的转移性收入和财产性收入。

输出结果 7—9 **Row and Column Points**



从我国目前的经济发展状况来看，大部分农民仍是以工资性收入和家庭经营性收入为主要的收入来源。在经济发达地区，农民外出打工较多，因此以工资性收入为主；而在经济不发达地区，大部分农民还是以农业生产为主，因此以家庭经营性收入为主。随着我国社会经济的不断发展和进步，这种格局也必然会发生一定的变化，转移性收入和财产性收入也会有所表现。

综上所述，对应分析方法较好地揭示了指标与指标、样品与样品、指标与样品之间的内在联系。因此，这种方法能够以较小的代价从原始数据中提取较多的信息。

□ 参考文献

[1] 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982

[2] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993

[3] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[4] 任若恩，王惠文. 多元统计数据分析——理论、方法、实践. 北京：国防工业出版社，1999

[5] 郭志刚. 社会统计分析方法——SPSS 软件应用. 北京：中国人民大学出版社，1999

[6] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, William C. Black. *Multivariate Data Analysis*. Fifth Edition. Prentice-Hall, 1998

□ 思考与练习

1. 试述对应分析的思想方法及特点。
2. 试述对应分析中总惯量的意义。
3. 试对一个实际问题运用 SPSS 软件进行对应分析。

第 8 章

Chapter 8 典型相关分析

学习目标

1. 理解典型相关分析的思想；
2. 了解典型相关分析的基本理论及分析方法；
3. 掌握利用 SPSS 软件或 SAS 软件实现典型相关分析的方法并能正确理解、解释各种输出结果。

典型相关分析（canonical correlation analysis）是研究两组变量之间相关关系的多元分析方法。它借用主成分分析降维的思想，分别对两组变量提取主成分，且使从两组变量提取的主成分之间的相关程度达到最大，而从同一组内部提取的各主成分之间互不相关，用从两组分别提取的主成分的相关性来描述两组变量整体的线性相关关系。典型相关分析的思想首先由霍特林于 1936 年提出，计算机的发展解决了典型相关分析在应用中计算方面的困难，目前它已成为普遍应用的两组变量之间相关性分析的技术。本章主要介绍典型相关分析的思想、基本理论及分析方法，并介绍利用 SAS 和 SPSS 软件进行典型相关分析的方法。

8.1 典型相关分析的基本理论及方法

8.1.1 典型相关分析的统计思想

典型相关分析研究两组变量间整体的线性相关关系，它是将每一组变量作为一个整体来进行研究，而不是分析每一组变量内部的各个变量。所研究的两组变量可以是一组变量为自变量，而另一组变量为因变量的情况，也可以处于同等的地位，

但典型相关分析要求两组变量都至少是间隔尺度的。

典型相关分析借助主成分分析的思想，对每一组变量分别寻找线性组合，使生成的新的综合变量能代表原始变量大部分的信息，同时，与由另一组变量生成的新的综合变量的相关程度最大，这样一组新的综合变量称为第一对典型相关变量，同样的方法可以找到第二对、第三对……使各对典型相关变量之间互不相关，典型相关变量之间的简单相关系数称为典型相关系数。典型相关分析就是用典型相关系数衡量两组变量之间的相关性。

一般，设 $\boldsymbol{x}=(X_1,X_2,\cdots,X_p)'$，$\boldsymbol{y}=(Y_1,Y_2,\cdots,Y_q)'$是两个相互关联的随机向量，利用主成分分析的思想，分别在两组变量中选取若干有代表性的综合变量 U_i，V_i，使每一综合变量都是原变量的一个线性组合，即

$$\begin{aligned} U_i&=a_{i1}X_1+a_{i2}X_2+\cdots+a_{ip}X_p\equiv\boldsymbol{a}'\boldsymbol{x}\\ V_i&=b_{i1}Y_1+b_{i2}Y_2+\cdots+b_{iq}Y_q\equiv\boldsymbol{b}'\boldsymbol{y}\end{aligned}\tag{8.1}$$

我们可以只考虑方差为 1 的 $\boldsymbol{x}$，$\boldsymbol{y}$ 的线性函数 $\boldsymbol{a}'\boldsymbol{x}$ 与 $\boldsymbol{b}'\boldsymbol{y}$，求使它们相关系数达到最大的这一组。若存在常向量 $\boldsymbol{a}_1$，$\boldsymbol{b}_1$，使得

$$\begin{aligned} &\rho(\boldsymbol{a}_1{}'\boldsymbol{x},\boldsymbol{b}_1{}'\boldsymbol{y})=\max\rho(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})\\ &\mathrm{var}(\boldsymbol{a}'\boldsymbol{x})=\mathrm{var}(\boldsymbol{b}'\boldsymbol{y})=1\end{aligned}\tag{8.2}$$

则称 $\boldsymbol{a}_1{}'\boldsymbol{x}$，$\boldsymbol{b}_1{}'\boldsymbol{y}$ 是 $\boldsymbol{x}$，$\boldsymbol{y}$ 的第一对典型相关变量。求出第一对典型相关变量之后，可以类似地去求第二对、第三对……使得各对之间互不相关。这些典型相关变量就反映了 $\boldsymbol{x}$，$\boldsymbol{y}$ 之间的线性相关情况。也可以按照相关系数绝对值的大小来排列各对典型相关变量之间的先后次序，使得第一对典型相关变量相关系数的绝对值最大，第二对次之……更重要的是，我们可以检验各对典型相关变量相关系数的绝对值是否显著大于零。如果是，这一对综合变量就真的具有代表性；如果不是，这一对变量就不具有代表性，不具有代表性的变量可以忽略。这样就可通过对少数典型相关变量的研究，代替原来两组变量之间的相关关系的研究，从而容易抓住问题的本质。在研究实际问题时，可以通过典型相关分析找出几对主要的典型相关变量，根据典型相关变量相关程度及各典型相关变量线性组合中原变量系数的大小，结合对所研究实际问题的定性分析，尽可能给出较为深刻的分析结果。

8.1.2　典型相关分析的基本理论及方法

1. 总体典型相关和典型变量

设随机向量 $\boldsymbol{x}=(X_1,X_2,\cdots,X_p)'$，$\boldsymbol{y}=(Y_1,Y_2,\cdots,Y_q)'$，$\boldsymbol{x}$，$\boldsymbol{y}$ 的协方差矩阵为：

$$\mathrm{cov}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}=\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\tag{8.3}$$

不失一般性，设 $p<q$，$\boldsymbol{\Sigma}_{11}$是 $p\times p$ 阶矩阵，它是第一组变量的协方差阵；$\boldsymbol{\Sigma}_{22}$是 $q\times q$ 阶矩阵，它是第二组变量的协方差阵。而 $\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}'_{21}$是两组变量之间的协方差阵。而且当 $\boldsymbol{\Sigma}$ 是正定阵时，$\boldsymbol{\Sigma}_{12}$与 $\boldsymbol{\Sigma}_{21}$也是正定的。

为了研究两组变量 $\boldsymbol{x}$，$\boldsymbol{y}$ 之间的相关关系，我们考虑它们的线性组合：

$$\begin{cases}U_1=\boldsymbol{a}'\boldsymbol{x}=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p\\ V_1=\boldsymbol{b}'\boldsymbol{y}=b_{11}Y_1+b_{12}Y_2+\cdots+b_{1q}Y_q\end{cases}\tag{8.4}$$

式中，$\boldsymbol{a}=(a_{11},a_{12},\cdots,a_{1p})'$，$\boldsymbol{b}=(b_{11},b_{12},\cdots,b_{1q})'$，是任意非零常数向量。我们希望在 $\boldsymbol{x}$，$\boldsymbol{y}$ 及 $\boldsymbol{\Sigma}$ 给定的条件下，选取 $\boldsymbol{a}$，$\boldsymbol{b}$ 使 U_1 与 V_1 之间的相关系数

$$\rho=\frac{\operatorname{cov}(U_1,V_1)}{\sqrt{\operatorname{var}(U_1)\operatorname{var}(V_1)}}=\frac{\operatorname{cov}(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})}{\sqrt{\operatorname{var}(\boldsymbol{a}'\boldsymbol{x})\operatorname{var}(\boldsymbol{b}'\boldsymbol{y})}}\tag{8.5}$$

达到最大。

由于随机变量 U_1，V_1 乘以任意常数并不改变它们之间的相关关系，不妨限定取标准化的随机变量 U_1 与 V_1，即规定 U_1 及 V_1 的方差为 1，也即

$$\begin{cases}\operatorname{var}(U_1)=\operatorname{var}(\boldsymbol{a}'\boldsymbol{x})=\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}=1\\ \operatorname{var}(V_1)=\operatorname{var}(\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}=1\end{cases}\tag{8.6}$$

所以
$$\rho=\operatorname{cov}(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{a}'\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{b}=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}\tag{8.7}$$

于是，我们的问题是，在式（8.6）的约束下，求 $\boldsymbol{a}\in R^p$，$\boldsymbol{b}\in R^q$，使得式（8.7）达到最大。由拉格朗日乘数法，这一问题等价于求 $\boldsymbol{a}$，$\boldsymbol{b}$，使

$$G=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}-\frac{\lambda}{2}(\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}-1)-\frac{\mu}{2}(\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}-1)\tag{8.8}$$

达到最大。式中，λ，μ 为拉格朗日乘数因子。将 G 分别对 $\boldsymbol{a}$ 及 $\boldsymbol{b}$ 求偏导并令其为 0，得方程组：

$$\begin{cases}\dfrac{\partial G}{\partial \boldsymbol{a}}=\boldsymbol{\Sigma}_{12}\boldsymbol{b}-\lambda\boldsymbol{\Sigma}_{11}\boldsymbol{a}=0\\ \dfrac{\partial G}{\partial \boldsymbol{b}}=\boldsymbol{\Sigma}_{21}\boldsymbol{a}-\mu\boldsymbol{\Sigma}_{22}\boldsymbol{b}=0\end{cases}\tag{8.9}$$

用 $\boldsymbol{a}'$，$\boldsymbol{b}'$分别左乘式（8.9）的两式，有

$$\begin{cases}\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}=\lambda\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}=\lambda\\ \boldsymbol{b}'\boldsymbol{\Sigma}_{21}\boldsymbol{a}=\mu\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}=\mu\end{cases}$$

又
$$(\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b})'=\boldsymbol{b}'\boldsymbol{\Sigma}_{21}\boldsymbol{a}$$

所以有
$$\mu=\boldsymbol{b}'\boldsymbol{\Sigma}_{21}\boldsymbol{a}=(\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b})'=\lambda\tag{8.10}$$

也就是说，λ 恰好等于线性组合 U 与 V 之间的相关系数，于是式（8.9）可写为：

$$\begin{cases}\boldsymbol{\Sigma}_{12}\boldsymbol{b}-\lambda\boldsymbol{\Sigma}_{11}\boldsymbol{a}=0\\ \boldsymbol{\Sigma}_{21}\boldsymbol{a}-\lambda\boldsymbol{\Sigma}_{22}\boldsymbol{b}=0\end{cases} \tag{8.11}$$

或者可以写为：

$$\begin{bmatrix}-\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}=\boldsymbol{0} \tag{8.12}$$

而式（8.12）有非零解的充要条件是：

$$\begin{vmatrix}-\lambda\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & -\lambda\boldsymbol{\Sigma}_{22}\end{vmatrix}=0 \tag{8.13}$$

式（8.13）左端为 λ 的 $p+q$ 次多项式，因此有 $p+q$ 个根，设为 λ_1，λ_2，…，λ_{p+q}（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_{p+q}$），再以 $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ 左乘式(8.11) 中第二式，则有

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}-\lambda\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{b}=0$$

即

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}=\lambda\boldsymbol{\Sigma}_{12}\boldsymbol{b} \tag{8.14}$$

又由式（8.11）中第一式，得

$$\boldsymbol{\Sigma}_{12}\boldsymbol{b}=\lambda\boldsymbol{\Sigma}_{11}\boldsymbol{a}$$

代入式（8.14），得

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}-\lambda^2\boldsymbol{\Sigma}_{11}\boldsymbol{a}=0$$

$$(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\lambda^2\boldsymbol{\Sigma}_{11})\boldsymbol{a}=0 \tag{8.15}$$

再用 $\boldsymbol{\Sigma}_{11}^{-1}$ 左乘式(8.15)，得

$$(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\lambda^2\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11})\boldsymbol{a}=0$$

$$(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\lambda^2\boldsymbol{I}_p)\boldsymbol{a}=0 \tag{8.16}$$

因此，对 λ^2 有 p 个解，设为 λ_1^2，λ_2^2，…，λ_p^2（$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2$），对 $\boldsymbol{a}$ 也有 p 个解。

类似地，用 $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ 左乘式(8.11) 中第一式，则有

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b}-\lambda\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{a}=0 \tag{8.17}$$

又由式（8.11）中第二式，得

$$\boldsymbol{\Sigma}_{21}\boldsymbol{a}=\lambda\boldsymbol{\Sigma}_{22}\boldsymbol{b}$$

代入式(8.17)，得

$$(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}-\lambda^2\boldsymbol{\Sigma}_{22})\boldsymbol{b}=0 \tag{8.18}$$

再以 $\boldsymbol{\Sigma}_{22}^{-1}$ 左乘式(8.18)，得

$$(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}-\lambda^2\boldsymbol{I}_q)\boldsymbol{b}=0 \tag{8.19}$$

因此对 λ^2 有 q 个解，对 $\boldsymbol{b}$ 也有 q 个解，λ^2 为 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 的特征根，$\boldsymbol{a}$ 是对应于 λ^2 的特征向量。同时 λ^2 也是 $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ 的特征根，$\boldsymbol{b}$ 为相应的特征向量。而式(8.16)、式(8.19)有非零解的充分必要条件为：

$$|\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\lambda^2\boldsymbol{I}_p|=0 \tag{8.20}$$

$$|\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}-\lambda^2\boldsymbol{I}_q|=0 \tag{8.21}$$

对式(8.20)，由于 $\boldsymbol{\Sigma}_{11}>0$，$\boldsymbol{\Sigma}_{22}>0$，故 $\boldsymbol{\Sigma}_{11}^{-1}>0$，$\boldsymbol{\Sigma}_{22}^{-1}>0$，所以

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}=\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}$$

而 $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}$ 与 $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ 有相同的特征根。如果记

$$\boldsymbol{T}=\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$$

则

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}=\boldsymbol{T}\boldsymbol{T}'$$

类似地，对式(8.21)，可得

$$\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}=\boldsymbol{T}'\boldsymbol{T}$$

而 $\boldsymbol{T}\boldsymbol{T}'$ 与 $\boldsymbol{T}'\boldsymbol{T}$ 有相同的非零特征根，从而推知式(8.16)、式(8.19)的非零特征根是相同的。设已求得 $\boldsymbol{T}\boldsymbol{T}'$ 的 p 个特征根依次为：

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2>0$$

则 $\boldsymbol{T}'\boldsymbol{T}$ 的 q 个特征根中，除了上面的 p 个之外，其余的 $q-p$ 个都是零。故 p 个特征根排列是 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2>0$，因此，只要取最大的 λ_1，U_1 与 V_1 即具有最大的相关系数。令 $\boldsymbol{a}_1$，$\boldsymbol{b}_1$ 为式（8.12）的解，且按式（8.6）进行了正规化，这时 $U_1=\boldsymbol{a}_1'\boldsymbol{x}$ 与 $V_1=\boldsymbol{b}_1'\boldsymbol{y}$ 即分别为 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的正规化的线性组合，且具有最大的相关系数 λ_1。

综上所述，有如下定义。

定义 8.1 在一切使方差为 1 的线性组合 $\boldsymbol{a}'\boldsymbol{x}$ 与 $\boldsymbol{b}'\boldsymbol{y}$ 中，其中两者相关系数最大的 $U_1=\boldsymbol{a}_1'\boldsymbol{x}$ 与 $V_1=\boldsymbol{b}_1'\boldsymbol{y}$ 称为第一对典型相关变量，它们的相关系数 λ_1 称为第一典型相关系数。

更一般地，在定义了 $i-1$ 对典型相关变量后，在一切使方差为 1 且与前 $i-1$对典型相关变量都不相关的线性组合 $U_i=\boldsymbol{a}_i'\boldsymbol{x}$ 与 $V_i=\boldsymbol{b}_i'\boldsymbol{y}$ 中，其两者相关系数最大者称为第 i 对典型相关变量，其相关系数称为第 i 对典型相关系数。

由上述推导，进一步有：求 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的第 i 个典型相关系数就是求方程（8.13）的第 i 个最大根 λ_i，而第 i 对典型变量即为 $U_i=\boldsymbol{a}_i'\boldsymbol{x}$ 与 $V_i=\boldsymbol{b}_i'\boldsymbol{y}$，其中 $\boldsymbol{a}_i$ 与 $\boldsymbol{b}_i$ 为方程（8.12）当 $\lambda=\lambda_i$ 时所求得的解。

我们不加证明地给出典型变量的以下两个性质。

（1）由 X_1，X_2，…，X_p 所组成的典型变量 U_1，U_2，…，U_p 互不相关，同样，由 Y_1，Y_2，…，Y_q 所组成的典型变量 V_1，V_2，…，V_p 也互不相关，且它们的方差均等于 1。即

$$\operatorname{cov}(U_i,U_j)=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

$$\operatorname{cov}(V_i,V_j)=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

（2）同一对典型变量 U_i 及 V_i 之间的相关系数为 λ_i，不同对的典型变量 U_i 及 $V_i(i\neq j)$ 间互不相关。即

$$\operatorname{cov}(U_i,V_i)=\lambda_i\neq 0,\quad i=1,2,\cdots,p$$

$$\operatorname{cov}(U_i,V_j)=0,\quad i\neq j$$

2. 样本典型相关和典型变量

上面的讨论都是基于总体情况已知的情形进行的，而实际研究中总体协方差阵 $\boldsymbol{\Sigma}$ 常常是未知的，我们只获得了样本数据，必须根据样本数据对 $\boldsymbol{\Sigma}$ 进行估计。

设 $\begin{bmatrix}\boldsymbol{x}_i\\ \boldsymbol{y}_i\end{bmatrix}(i=1,2,\cdots,n)$ 是来自正态总体 $N_{p+q}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的容量为 n 的样本，则总体协方差阵 $\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$，$\boldsymbol{\Sigma}_{(p+q)\times(p+q)}$（$\boldsymbol{\Sigma}>0$）的极大似然估计为：

$$\hat{\boldsymbol{\Sigma}}=\boldsymbol{A}=\frac{1}{n}\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix} \tag{8.22}$$

其中

$$\boldsymbol{A}_{11}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})' \tag{8.23}$$

$$\boldsymbol{A}_{22}=\sum_{i=1}^{n}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})' \tag{8.24}$$

$$\boldsymbol{A}_{12}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})'=\boldsymbol{A}_{21}' \tag{8.25}$$

当 $n>p+q$ 时，在正态情况下，$P(\hat{\boldsymbol{\Sigma}}>0)=1$，且由 $\boldsymbol{\Sigma}$ 所定义的 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 和 $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ 的非零特征根以概率 1 互不相同，故由极大似然估计的性质得，$\hat{\boldsymbol{\Sigma}}$ 所产生的

$$\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}=\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} \tag{8.26}$$

是 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 的极大似然估计。$\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ 是 $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ 的极大似然估计。$\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{21}^{-1}\boldsymbol{A}_{21}$ 和 $\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ 的非零特征根 $\hat{\lambda}_1^2$，$\hat{\lambda}_2^2$，$\cdots$，$\hat{\lambda}_k^2$（$k=\operatorname{rank}(\boldsymbol{A})$）是 λ_1^2，λ_2^2，$\cdots$，λ_k^2 的极大似然估计，相应的特征向量 $\hat{\boldsymbol{a}}_1$，$\cdots$，$\hat{\boldsymbol{a}}_k$ 为 $\boldsymbol{a}_1$，$\cdots$，$\boldsymbol{a}_k$ 的极大似然估计，$\hat{\boldsymbol{b}}_1$，$\cdots$，$\hat{\boldsymbol{b}}_k$ 是 $\boldsymbol{b}_1$，$\cdots$，$\boldsymbol{b}_k$ 的极大似然估计。所以平行于总体的讨论，$\hat{\lambda}_1$，$\cdots$，$\hat{\lambda}_k$ 称为样本的典型相关系数，$(\hat{\boldsymbol{a}}_1'\boldsymbol{x},\ \hat{\boldsymbol{b}}_1'\boldsymbol{y})$，$\cdots$，$(\hat{\boldsymbol{a}}_k'\boldsymbol{x},\ \hat{\boldsymbol{b}}_k'\boldsymbol{y})$ 称为典型

相关变量。

如果将样本（$\boldsymbol{x}_i$，$\boldsymbol{y}_i$）（$i=1$，2，…，n）代入典型变量 $\hat{U}_i$ 及 $\hat{V}_i$ 中，求得的值称为第 i 对典型变量的得分。利用典型变量的得分可以绘出样本的典型变量的散点图，类似因子分析可以对样品进行分类研究。

3. 典型相关系数的显著性检验

典型相关系数的显著性检验可以用巴特莱特提出的大样本的 χ^2 检验来完成。

如果随机向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 之间互不相关，则协方差矩阵 $\boldsymbol{\Sigma}_{12}$ 仅包含零，因而典型相关系数

$$\lambda_i=\boldsymbol{a}_i'\boldsymbol{\Sigma}_{12}\boldsymbol{b}_i$$

都变为零。

这样，检验典型相关系数的显著性问题即变为进行如下检验：

$$H_0:\lambda_1=0,\quad H_1:\lambda_1\neq 0$$

求出 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 的 p 个特征根，并按大小顺序排列：

$$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2$$

做乘积：

$$\Lambda_1=(1-\lambda_1^2)(1-\lambda_2^2)\cdots(1-\lambda_p^2)=\prod_{i=1}^{p}(1-\lambda_i^2)$$

则对于大的 n（这里要求 $n>\dfrac{p+q+1}{2}+k$，k 为非零特征根个数），计算统计量

$$Q_1=-\left[n-1-\frac{1}{2}(p+q+1)\right]\ln\Lambda_1$$

Q_1 近似服从 $\chi^2(pq)$。因此在检验水平 α 下，若 $Q_1>\chi_\alpha^2(pq)$，则拒绝原假设 H_0，说明至少有第一对典型变量显著相关，或说相关性系数 λ_1 在显著性水平 α 下是显著的。

在去掉第一典型相关系数后，继续检验余下的 $p-1$ 个典型相关系数的显著性。更一般地，若前 $j-1$ 个典型相关系数在水平 α 下是显著的，则当检验第 j 个典型相关系数的显著性时，计算

$$\Lambda_j=(1-\lambda_j^2)(1-\lambda_{j+1}^2)\cdots(1-\lambda_p^2)=\prod_{i=j}^{p}(1-\lambda_i^2)$$

并计算统计量

$$Q_j=-\left[n-j-\frac{1}{2}(p+q+1)\right]\ln\Lambda_j$$

则 Q_j 服从自由度为 $(p-j+1)(q-j+1)$ 的 χ^2 分布。在检验水平 α 下，若 $Q_j>\chi_\alpha^2[(p-j+1)(q-j+1)]$，则拒绝 H_0，接受 H_1，即认为第 j 个典型相关系数在显