第六章 主成分分析

第一节 引言

- 多元统计分析处理的是多变量（多指标）问题。由于变量较多，增加了分析问题的复杂性。但在实际问题中，变量之间可能存在一定的相关性，因此，多变量中可能存在信息的重叠。人们自然希望通过克服相关性、重叠性，用较少的变量来代替原来较多的变量，而这种代替可以反映原来多个变量的大部分信息，这实际上是一种“降维”的思想。

多元统计

- 主成分分析也称主分量分析，是由**Hotelling**于**1933**年首先提出的。由于多个变量之间往往存在着一定程度的相关性。人们自然希望通过线性组合的方式，从这些指标中尽可能快地提取信息。当第一个线性组合不能提取更多的信息时，再考虑用第二个线性组合继续这个快速提取的过程，……，直到所提取的信息与原指标相差不多时为止。这就是主成分分析的思想。一般说来，在主成分分析适用的场合，用较少的主成分就可以得到较多的信息量。以各个主成分为分量，就得到一个更低维的随机向量；因此，通过主成分既可以降低数据“维数”又保留了原数据的大部分信息。

多元统计

- 我们知道，当一个变量只取一个数据时，这个变量（数据）提供的信息量是非常有限的，当这个变量取一系列不同数据时，我们可以从中读出最大值、最小值、平均数等信息。变量的变异性越大，说明它对各种场景的“遍历性”越强，提供的信息就更加充分，信息量就越大。主成分分析中的信息，就是指标的变异性，用标准差或方差表示它。
- 主成分分析的数学模型是，设p个变量构成的p维随机向量为$X=(X_1,\ldots,X_p)'$。对X作正交变换，令$Y=T'X$，其中T为正交阵，要求Y的各分量是不相关的，并且Y的第一个分量的方差是最大的，第二个分量的方差次之，……，等等。为了保持信息不丢失，Y的各分量方差和与X的各分量方差和相等。

第二节 主成分的几何意义及数学推导

一 主成分的几何意义

二 主成分的数学推导

一、主成分的几何意义

- 主成分分析数学模型中的正交变换，在几何上就是作一个坐标旋转。因此，主成分分析在二维空间中有明显的几何意义。假设共有n个样品，每个样品都测量了两个指标（X_1，X_2），它们大致分布在一个椭圆内如图6.1所示。事实上，散点的分布总有可能沿着某一个方向略显扩张，这个方向就把它看作椭圆的长轴方向。显然，在坐标系x_1Ox_2中，单独看这n个点的分量X_1和X_2，它们沿着x_1方向和x_2方向都具有较大的离散性，其离散的程度可以分别用的X_1方差和X_2的方差测定。如果仅考虑X_1或X_2中的任何一个分量，那么包含在另一分量中的信息将会损失，因此，直接舍弃某个分量不是“降维”的有效办法。

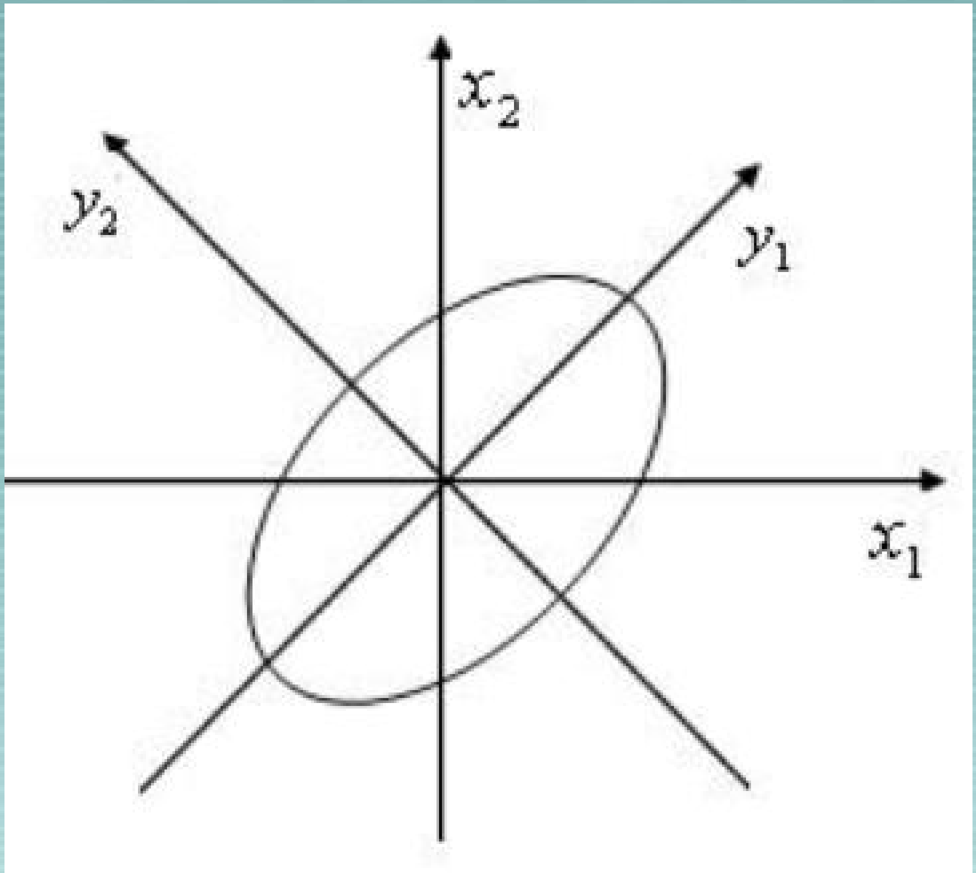


图6.1　主成分的几何意义

- 如果我们将该坐标系按逆时针方向旋转某个角度 θ 变成新坐标系 y_1Oy_2，这里 y_1 是椭圆的长轴方向，y_2 是椭圆的短轴方向。旋转公式为

$$\begin{cases} Y_1 = X_1\cos\theta + X_2\sin\theta \\ Y_2 = -X_1\sin\theta + X_2\cos\theta \end{cases} \qquad (6.1)$$

我们看到新变量 Y_1 和 Y_2 是原变量 X_1 和 X_2 的线性组合，它的矩阵表示形

式为：

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathbf{T'X} \qquad (6.2)$$

其中，$\mathbf{T'}$ 为旋转变换矩阵，它是正交矩阵，即有 $\mathbf{T'} = \mathbf{T}^{-1}$ 或 $\mathbf{T'T} = \mathbf{I}$。

多元统计

- 易见，n个点在新坐标系下的坐标Y_1和Y_2几乎不相关。称它们为原始变量X_1和X_2的综合变量，n个点y_1在轴上的方差达到最大，即在此方向上包含了有关n个样品的最大量信息。因此，欲将二维空间的点投影到某个一维方向上，则选择y_1轴方向能使信息的损失最小。我们称Y_1为第一主成分，称Y_2为第二主成分。第一主成分的效果与椭圆的形状有很大的关系，椭圆越是扁平，n个点在y_1轴上的方差就相对越大，在y_2轴上的方差就相对越小，用第一主成分代替所有样品所造成的信息损失也就越小。

- 考虑两种极端的情形：
  - 一种是椭圆的长轴与短轴的长度相等，即椭圆变成圆，第一主成分只含有二维空间点的约一半信息，若仅用这一个综合变量，则将损失约50％的信息，这显然是不可取的。造成它的原因是，原始变量X_1和X_2的相关程度几乎为零，也就是说，它们所包含的信息几乎不重迭，因此无法用一个一维的综合变量来代替。
  - 另一种是椭圆扁平到了极限，变成y_1轴上的一条线，第一主成分包含有二维空间点的全部信息，仅用这一个综合变量代替原始数据不会有任何的信息损失，此时的主成分分析效果是非常理想的，其原因是，第二主成分不包含任何信息，舍弃它当然没有信息损失。

二、主成分的数学推导

- 设 $\mathbf{X}=(X_1,\cdots,X_p)'$ 为一个 p 维随机向量，并假定存在二阶矩，其均值向量与协差阵分别记为：

$$\boldsymbol{\mu}=E(\mathbf{X}),\quad \boldsymbol{\Sigma}=D(\mathbf{X}) \qquad (6.3)$$

考虑如下的线性变换

$$\begin{cases} Y_1=t_{11}X_1+t_{12}X_2+\cdots+t_{1p}X_p=T_1'\mathbf{X} \\ Y_2=t_{21}X_1+t_{22}X_2+\cdots+t_{2p}X_p=T_2'\mathbf{X} \\ \cdots\cdots\cdots\cdots \\ Y_p=t_{p1}X_1+t_{p2}X_2+\cdots+t_{pp}X_p=T_p'\mathbf{X} \end{cases} \qquad (6.\ 4)$$

用矩阵表示为 $\mathbf{Y}=\mathbf{T}'\mathbf{X}$

其中 $\mathbf{Y}=(Y_1,Y_2,\cdots Y_p)'$，$\mathbf{T}=(T_1,T_2,\cdots,T_p)$。

- 我们希望寻找一组新的变量 $Y_1,\cdots,Y_m$（$m \le p$），这组新的变量要求充分地反映原变量 $X_1,\cdots,X_p$ 的信息，而且相互独立。这里我们应该注意到，对于 $Y_1,\cdots,Y_m$ 有

$$D(Y_i) = D(T_i'\mathbf{X}) = T_i' D(\mathbf{X}) T_i'' = T_i'\mathbf{\Sigma} T_i \qquad i = 1,2,\cdots,m$$

$$Cov(Y_i, Y_k) = Cov(T_i'\mathbf{X}, T_k'\mathbf{X}) = T_i' Cov(\mathbf{X},\mathbf{X}) T_k'' = T_i'\mathbf{\Sigma} T_k$$
$$i,k = 1,2,\cdots,m$$

这样，我们所要解决的问题就转化为，在新的变量 $Y_1,\cdots,Y_m$ 相互独立的条件下，求 T_i 使得 $D(Y_i) = T_i'\mathbf{\Sigma} T_i$，$i = 1,2,\cdots,m$,达到最大。

- 我们下面将借助投影寻踪（**Projection Pursuit**）的思想来解决这一问题。首先应该注意到，使得 $D(Y_i)$ 达到最大的线性组合，显然用常数乘以 T_i 后，$D(Y_i)$ 也随之增大，为了消除这种不确定性，不妨假设 T_i 满足 $T_i'T_i=1$ 或者 $|T|=1$。那么，问题可以更加明确。

  第一主成分为，满足 $T_1'T_1=1$，使得 $D(Y_1)=T_1'\mathbf{\Sigma}T_1$ 达到最大的

  $$Y_1=T_1'\mathbf{X}。$$

  第二主成分为，满足 $T_2'T_2=1$，且 $Cov(Y_2,Y_1)=Cov(T_2'\mathbf{X},T_1'\mathbf{X})=0$，

  使得 $D(Y_2)=T_2'\mathbf{\Sigma}T_2$ 达到最大的 $Y_2=T_2'\mathbf{X}$。

- 一般情形，第 k 主成分为，满足 $T_k'T_k=1$，

  且 $Cov(Y_k,Y_i)=Cov(T_k'\mathbf{X},T_i'\mathbf{X})=0$（$i<k$），使得 $D(Y_k)=T_k'\mathbf{\Sigma}T_k$ 达到最大的 $Y_k=T_k'\mathbf{X}$。

- 求第一主成分，构造目标函数为：

$$\varphi_1(T_1,\lambda)=T_1'\boldsymbol{\Sigma}T_1-\lambda(T_1'T_1-1) \tag{6.5}$$

对目标函数 $\varphi_1(T_1,\lambda)$ 求导数有：

$$\frac{\partial\varphi_1}{\partial T_1}=2\boldsymbol{\Sigma}T_1-2\lambda T_1=0 \tag{6.6}$$

即

$$(\boldsymbol{\Sigma}-\lambda\mathbf{I})T_1=0 \tag{6.7}$$

由 6.7 式两边左乘 T_1' 得到

$$T_1'\boldsymbol{\Sigma}T_1=\lambda \tag{6.8}$$

由于 **X** 的协差阵 **Σ** 为非负定的，其特征方程(6.7)的根均大于零，不妨设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$。由(6.8)知道 Y_1 的方差为 λ。那么，Y_1 的最大方差值为 λ_1，其相应的单位化特征向量为 T_1。

- 在求第二主成分之前，我们首先明确，由(6.6)知 $Cov(Y_2, Y_1) = T_2'\boldsymbol{\Sigma}T_1 = \lambda T_2'T_1$。那么，如果 Y_2 与 Y_1 相互独立，即有 $T_2'T_1 = 0$ 或 $T_1'T_2 = 0$。这时，我们可以构造求第二主成分的目标函数，即

$$\varphi_2(T_2, \lambda, \rho) = T_2'\boldsymbol{\Sigma}T_2 - \lambda(T_2'T_2 - 1) - 2\rho(T_1'T_2) \qquad \textbf{(6.9)}$$

对目标函数 $\varphi_2(T_2, \lambda, \rho)$ 求导数有：

$$\frac{\partial \varphi_2}{\partial T_2} = 2\boldsymbol{\Sigma}T_2 - 2\lambda T_2 - 2\rho T_1 = 0 \qquad \textbf{(6.10)}$$

用 T_1' 左乘(6.10)式有 $\quad T_1'\boldsymbol{\Sigma}T_2 - \lambda T_1'T_2 - \rho T_1'T_1 = 0$

由于 $T_1'\boldsymbol{\Sigma}T_2 = 0$，$T_1'T_2 = 0$，那么，$\rho T_1'T_1 = 0$，即有 $\rho = 0$。从而

$$(\boldsymbol{\Sigma} - \lambda \mathbf{I})T_2 = 0 \qquad \textbf{(6.11)}$$

而且 $\quad T_2'\boldsymbol{\Sigma}T_2 = \lambda \qquad \textbf{(6.12)}$

- 这样说明，如果 $\mathbf{X}$ 的协差阵 $\mathbf{\Sigma}$ 的特征根为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$。由(6.12)知道 Y_2 的最大方差值为第二大特征根 λ_2，其相应的单位化的特征向量为 T_2。

针对一般情形，第 k 主成分应该是在 $T_k'T_k = 1$ 且 $T_k'T_i = 0$ 或 $T_i'T_k = 0$（$i < k$）的条件下，使得 $D(Y_k) = T_k'\mathbf{\Sigma}T_k$ 达到最大的 $Y_k = T_k'\mathbf{X}$。这样我们构造目标函数为

$$\varphi_k(T_k, \lambda, \rho_i) = T_k'\mathbf{\Sigma}T_k - \lambda(T_k'T_k - 1) - 2\sum_{i=1}^{k-1}\rho_i(T_i'T_k) \qquad \textbf{(6.13)}$$

对目标函数 $\varphi_k(T_k, \lambda, \rho_i)$ 求导数有：

$$\frac{\partial \varphi_k}{\partial T_k} = 2\mathbf{\Sigma}T_k - 2\lambda T_k - 2\sum_{i=1}^{k-1}\rho_i T_i = 0 \qquad \textbf{(6.14)}$$

用 T_i' 左乘(6.14)式有 $$T_i'\mathbf{\Sigma}T_k-\lambda T_i'T_k-T_i'(\sum_{i=1}^{k-1}\rho_i T_i)=0$$

即有 $\rho_i T_i' T_i=0$，那么，$\rho_i=0$（$i=1,2,\cdots k-1$）。从而

$$(\mathbf{\Sigma}-\lambda\mathbf{I})T_k=0 \tag{6.15}$$

而且 $$T_k'\mathbf{\Sigma}T_k=\lambda \tag{6.16}$$

- 对于 $\mathbf{X}$ 的协差阵 $\mathbf{\Sigma}$ 的特征根 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p\ge 0$。由(6.15)和(6.16)知道 Y_k 的最大方差值为第 k 大特征根 λ_k，其相应的单位化的特征向量为 T_k。
- 综上所述，设 $\mathbf{X}=(X_1,\cdots,X_p)'$ 的协差阵为 $\mathbf{\Sigma}$，其特征根为 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p\ge 0$，相应的单位化的特征向量为 $T_1,T_2,\cdots,T_p$。那么，由此所确定的主成分为 $Y_1=T_1'\mathbf{X}$，$Y_2=T_2'\mathbf{X}$，$\cdots$，$Y_m=T_m'\mathbf{X}$，其方差分别为 $\mathbf{\Sigma}$ 的特征根。

第三节 主成分的性质

一 主成分的一般性质

二 主成分的方差贡献率

一、主成分的一般性质

- 设 $\mathbf{Y}=(Y_1,Y_2,\cdots,Y_p)'$ 是 $\mathbf{X}$ 的主成分，由 $\boldsymbol{\Sigma}$ 的所有特征根构成的对角阵为

$$\boldsymbol{\Lambda}=\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{bmatrix} \tag{6.17}$$

主成分可表示为

$$\mathbf{Y}=\mathbf{T}'\mathbf{X} \tag{6.18}$$

性质 1　主成分的协方差矩阵是对角阵。

证明：实际上，由（6.3）式知

$$E(\mathbf{Y})=E(\mathbf{T}'\mathbf{X})=\mathbf{T}'\boldsymbol{\mu}$$

$$D(\mathbf{Y})=\mathbf{T}'D(\mathbf{X})\mathbf{T}=\mathbf{T}'\boldsymbol{\Sigma}\mathbf{T}=\boldsymbol{\Lambda} \tag{6.19}$$

性质 2　主成分的总方差等于原始变量的总方差。

证明：由矩阵“迹”的性质知

$$tr(\mathbf{\Lambda}) = tr(\mathbf{T}'\mathbf{\Sigma}\mathbf{T}) = tr(\mathbf{\Sigma}\mathbf{T}\mathbf{T}') = tr(\mathbf{\Sigma})$$

所以
$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_{ii} \tag{6.20}$$

或
$$\sum_{i=1}^{p} D(Y_i) = \sum_{i=1}^{p} D(X_i) \tag{6.21}$$

性质 3　主成分 Y_k 与原始变量 X_i 的相关系数为

$$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki} \tag{6.22}$$

并称之为因子负荷量（或因子载荷量）。

证明：事实上

$$\rho(Y_k, X_i) = \frac{Cov(Y_k, X_i)}{\sqrt{D(Y_k)D(X_i)}} = \frac{Cov(T_k'\mathbf{X}, e_i\mathbf{X})}{\sqrt{\lambda_k \sigma_{ii}}}$$

其中的 $e_i = (0,\cdots,0,1,0,\cdots,0)$，它是除第 i 个元素为 **1** 外其他元素均为 **0** 的单位向量。而

$$Cov(T_k'\mathbf{X}, e_i\mathbf{X}) = T_k'\mathbf{\Sigma} e_i = e_i'(\mathbf{\Sigma} T_k) = e_i'(\lambda_k T_k) = \lambda_k e_i' T_k = \lambda_k t_{ki}$$

所以 $\rho(Y_k, X_i) = \dfrac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki}$。

性质 4　$\sum_{i=1}^{p} \rho^2(Y_k, X_i)\cdot\sigma_{ii} = \lambda_k$，（$k = 1,2,\cdots,p$）。

证明：只须将(6.22)代入左边式子整理化简即可。

二、主成分的方差贡献率

- 由主成分的性质 2 可以看出，主成分分析把 p 个原始变量 $X_1, X_2, \cdots, X_p$ 的总方差 $tr(\boldsymbol{\Sigma})$ 分解成了 p 个相互独立的变量 $Y_1, Y_2, \cdots, Y_p$ 的方差之和 $\sum_{k=1}^{p} \lambda_k$。主成分分析的目的是减少变量的个数，所以一般不会使用所有 p 个主成分的，忽略一些带有较小方差的主成分将不会给总方差带来太大的影响。这里我们称

$$\varphi_k = \lambda_k \Big/ \sum_{k=1}^{p} \lambda_k \qquad (6.23)$$

- 为第 k 个主成分 Y_k 的贡献率。第一主成分的贡献率最大，这表明 $Y_1 = T_1'\boldsymbol{X}$ 综合原始变量 $X_1, X_2, \cdots, X_p$ 的能力最强，而 $Y_2, Y_3, \cdots, Y_p$ 的综合能力依次递减。若只取 $m(< p)$ 个主成分，则称

$$\psi_m = \sum_{k=1}^{m} \lambda_k \bigg/ \sum_{k=1}^{p} \lambda_k \qquad (6.24)$$

为主成分 $Y_1, \cdots, Y_m$ 的累计贡献率，累计贡献率表明 $Y_1, \cdots, Y_m$ 综合 $X_1, X_2, \cdots, X_p$ 的能力。通常取 m ，使得累计贡献率达到一个较高的百分数（如 85%以上）。

第四节 主成分方法应用中应注意的问题

一 实际应用中主成分分析的出发点

二 如何利用主成分分析进行综合评价

一、实际应用中主成分分析的出发点

- 我们前面讨论的主成分计算是从协方差矩阵 $\mathbf{\Sigma}$ 出发的，其结果受变量单位的影响。不同的变量往往有不同的单位，对同一变量单位的改变会产生不同的主成分，主成分倾向于多归纳方差大的变量的信息，对于方差小的变量就可能体现得不够，也存在“大数吃小数”的问题。为使主成分分析能够均等地对待每一个原始变量，消除由于单位的不同可能带来的影响，我们常常将各原始变量作标准化处理，即令

$$X_i^* = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}} \qquad i = 1, \cdots, p \qquad (6.25)$$

显然，$\mathbf{X}^* = (X_1^*, \cdots, X_p^*)'$ 的协方差矩阵就是 $\mathbf{X}$ 的相关系数矩阵 $\mathbf{R}$ 。实际应用中，$\mathbf{X}$ 的相关系数矩阵 $\mathbf{R}$ 可以通过（2.13）式，利用样本数据来估计。

- 这里我们需要进一步强调的是，从相关阵求得的主成分与协差阵求得的主成分一般情况是不相同的。实际表明，这种差异有时很大。我们认为，如果各指标之间的数量级相差悬殊，特别是各指标有不同的物理量纲的话，较为合理的做法是使用R代替Σ。对于研究经济问题所涉及的变量单位大都不统一，采用R代替Σ后，可以看作是用标准化的数据做分析，这样使得主成分有现实经济意义，不仅便于剖析实际问题，又可以避免突出数值大的变量。

- 因此，在实际应用中，主成分分析的具体步骤可以归纳为：

1. 将原始数据标准化；
2. 建立变量的相关系数阵；
3. 求 **R** 的特征根为 $\lambda_1^* \geq \cdots \geq \lambda_p^* \geq 0$，相应的特征向量为 $T_1^*, T_2^*, \cdots, T_p^*$；
4. 由累积方差贡献率确定主成分的个数（m），并写出主成分为

$$Y_i = (T_i^*)' X, \qquad i = 1, 2, \cdots, m$$

二、如何利用主成分分析进行综合评价

- 人们在对某个单位或某个系统进行综合评价时都会遇到如何选择评价指标体系和如何对这些指标进行综合的困难。一般情况下，选择评价指标体系后通过对各指标加权的办法来进行综合。但是，如何对指标加权是一项具有挑战性的工作。指标加权的依据是指标的重要性，指标在评价中的重要性判断难免带有一定的主观性，这影响了综合评价的客观性和准确性。由于主成分分析能从选定的指标体系中归纳出大部分信息，根据主成分提供的信息进行综合评价，不失为一个可行的选择。这个方法是根据指标间的相对重要性进行客观加权，可以避免综合评价者的主观影响，在实际应用中越来越受到人们的重视。
- 对主成分进行加权综合。我们利用主成分进行综合评价时，主要是将原有的信息进行综合，因此，要充分的利用原始变量提供的信息。将主成分的权数根据它们的方差贡献率来确定，因为方差贡献率反映了各个主成分的信息含量多少。

- 设 $Y_1, Y_2, \cdots, Y_p$ 是所求出的 p 个主成分，它们的特征根分别是 $\lambda_1, \lambda_2, \cdots, \lambda_p$，将特征根“归一化”即有

$$w_i = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} \qquad i = 1, 2, \cdots, p$$

记为 $W = (w_1, w_2, \cdots w_p)'$，由 $\mathbf{Y} = \mathbf{T}'\mathbf{X}$，构造综合评价函数为

$$Z = w_1 Y_1 + w_2 Y_2 + \cdots + w_p Y_p = W'\mathbf{Y} = W'\mathbf{T}'\mathbf{X} = (\mathbf{T}W)'\mathbf{X} \qquad (6.26)$$

令 $\mathbf{T}W = w^*_{k\times 1}$，并代入(6.26)式，有

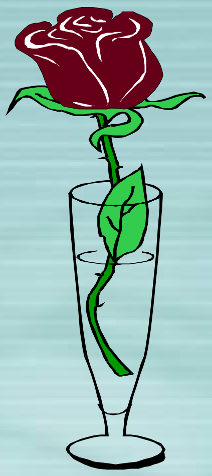
$$Z = (w^*)'\mathbf{X} \qquad (6.27)$$

这里我们应该注意，从本质上说综合评价函数是对原始指标的线性综合，从计算主成分到对之加权，经过两次线性运算后得到综合评价函数。

第五节 实例分析与计算机实现

一 主成分分析实例

二 利用SPSS进行主成分分析



一、主成分分析实例

- 表6.1是某市工业部门13个行业的8项重要经济指标的数据，这8项经济指标分别是：

  X1：年末固定资产净值，单位：万元；

  X2：职工人数据，单位：人；

  X3：工业总产值，单位：万元；

  X4：全员劳动生产率，单位：元/人年；

  X5：百元固定资产原值实现产值，单位：元；

  X6：资金利税率，单位：%；

  X7：标准燃料消费量，单位：吨；

  X8：能源利用效果，单位：万元/吨。

表6.1 某市工业部门13个行业8项指标

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
冶金	90342	52455	101091	19272	82	16.1	197435	0.172
电力	4903	1973	2035	10313	34.2	7.1	592077	0.003
煤炭	6735	21139	3767	1780	36.1	8.2	726396	0.003
化学	49454	36241	81557	22504	98.1	25.9	348226	0.985
机器	139190	203505	215898	10609	93.2	12.6	139572	0.628
建材	12215	16219	10351	6382	62.5	8.7	145818	0.066
森工	2372	6572	8103	12329	184.4	22.2	20921	0.152
食品	11062	23078	54935	23804	370.4	41	65486	0.263
纺织	17111	23907	52108	21796	221.5	21.5	63806	0.276
缝纫	1206	3930	6126	15586	330.4	29.5	1840	0.437
皮革	2150	5704	6200	10870	184.2	12	8913	0.274
造纸	5251	6155	10383	16875	146.4	27.5	78796	0.151
文教	14341	13203	19396	14691	94.6	17.8	6354	1.574

- 我们要考虑的是：如何从这些经济指标出发，对各工业部门进行综合评价与排序？
- 我们先计算这些指标的主成分，然后通过主成分的大小进行排序。表6.2和表6.3分别是特征根（累计贡献率）和特征向量的信息。
- 利用主成分得分进行综合评价时，从特征向量我们可以写出所有8个主成分的具体形式：

$$Y_1 = 0.476X_1^* + 0.473X_2^* + 0.424X_3^* - 0.213X_4^*$$
$$-0.388X_5^* - 0.352X_6^* + 0.215X_7^* + 0.055X_8^*$$
$$Y_2 = 0.296X_1^* + 0.278X_2^* + 0.378X_3^* + 0.451X_4^*$$
$$+0.331X_5^* + 0.403X_6^* - 0.377X_7^* + 0.273X_8^*$$

……

表6.2　特征根和累计贡献率

序号	特征根	方差贡献率%	累计贡献率%
1	3.1049	38.8114	38.8114
2	2.8974	36.2180	75.0294
3	0.9302	11.6277	86.6571
4	0.6421	8.0265	94.6836
5	0.3041	3.8011	98.4847
6	0.0866	1.0825	99.5672
7	0.0322	0.4023	99.9695
8	0.0024	0.0305	100.0000

表6.3 特征向量

	特征向量1	特征向量2	特征向量3	特征向量4	特征向量5	特征向量6	特征向量7	特征向量8
1	0.477	0.296	0.104	0.045	-0.184	-0.066	0.758	0.245
2	0.473	0.278	0.163	-0.174	0.305	-0.048	-0.518	0.527
3	0.424	0.378	0.156	0.059	0.017	0.099	-0.174	-0.781
4	-0.213	0.451	-0.009	0.516	-0.539	0.288	-0.249	0.220
5	-0.388	0.331	0.321	-0.199	0.450	0.582	0.233	0.031
6	-0.352	0.403	0.145	0.279	0.317	-0.714	0.056	-0.042
7	0.215	-0.377	0.140	0.758	0.418	0.194	0.053	0.041
8	0.055	0.273	-0.891	0.072	0.322	0.122	0.067	-0.003

表6.4　各行业主成分得分及排序

行业	Y_1	Y_2	……	Y_8	综合得分	排序
冶金	1.475	0.759	……	0.004	0.911	2
电力	0.498	-2.592	……	0.067	-0.654	12
煤炭	1.056	-3.226	……	-0.024	-0.629	11
化学	0.460	1.184	……	-0.052	0.618	3
机器	4.528	2.262	……	0.023	2.589	1
建材	0.330	-1.774	……	-0.067	-0.602	10
森工	-1.103	-0.318	……	-0.035	-0.573	9
食品	-2.195	2.244	……	-0.052	0.155	4
纺织	-0.841	0.896	……	-0.001	0.033	5
缝纫	-2.032	0.825	……	0.073	-0.476	8
皮革	-0.713	-0.756	……	-0.030	-0.659	13
造纸	-1.201	0.030	……	0.079	-0.437	7
文教	-0.263	0.464	……	0.015	-0.276	6

多元统计

- 我们以特征根为权，对8个主成分进行加权综合，得出各工业部门的综合得分，具体数据见表6.4。
- 综合得分的计算公式是：

$$Y=\frac{\lambda_1}{\sum_{i=1}^{8}\lambda_i}Y_1+\frac{\lambda_2}{\sum_{i=1}^{8}\lambda_i}Y_2+\cdots+\frac{\lambda_8}{\sum_{i=1}^{8}\lambda_i}Y_8$$

根据上式可计算出各工业部门的综合得分，并可据此排序。

- 从上表可以看出，机器行业在该地区的综合评价排在第一，原始数据也反映出机器行业存在明显的规模优势，另外从前两个主成分得分上看，该行业也排在第一位，同样存在效益优势；而排在最后三位的分别是皮革行业、电力行业和煤炭行业。

二、利用SPSS进行主成分分析

- SPSS没有提供主成分分析的专用功能，只有因子分析的功能。但是因子分析和主成分分析有着密切的联系。因子分析的重要步骤——因子的提取最常用的方法就是“主成分法”。利用因子分析的结果，可以很容易地实现主成分分析。具体来讲，就是利用因子载荷阵和相关系数矩阵的特征根来计算特征向量。即：

$$z_{ij} = a_{ij} \Big/ \sqrt{\lambda_j}$$

- 其中，z_{ij}为第j个特征向量的第i个元素；a_{ij}为因子载荷阵第i行第j列的元素；λ_j为第j个因子对应的特征根。然后再利用计算出的特征向量来计算主成分。
- 以下是我国2005年第1、2季度分地区城镇居民家庭收支基本情况。通过这个例子，介绍如何利用SPSS软件实现主成分分析。

表6.5 分地区城镇居民家庭收支基本情况

地区	平均每户人口（人）	平均每户就业人口（人）	平均每一就业者负担人数（人）	平均每人实际可支配收入（元）	平均每人消费性支出（元）
北京	2.9	1.6	1.8	8845.1	6249.3
天津	2.9	1.4	2	6189.1	4549.1
河北	2.9	1.5	1.9	4582.9	3317.3
山西	3	1.5	2	4359.7	3066.8
内蒙	2.9	1.5	1.9	4712.1	3557.8
辽宁	2.9	1.4	2	4501.2	3530.7
吉林	3	1.5	1.9	4293.7	3271.5
黑龙江	2.8	1.3	2.2	3902.3	2858.7
上海	3	1.6	1.9	9656.5	6623.3
江苏	2.9	1.4	2.1	6371.1	4222.1
浙江	2.8	1.4	1.9	8921.2	6127.5
安徽	3	1.6	1.9	4311.6	3121.4
福建	3.1	1.6	1.9	6471.8	4292.3
江西	2.9	1.5	1.9	4369.7	2945.1
山东	2.9	1.7	1.7	5357.7	3517.6

表6.5 分地区城镇居民家庭收支基本情况

地区	平均每户人口（人）	平均每户就业人口（人）	平均每一就业者负担人数（人）	平均每人实际可支配收入（元）	平均每人消费性支出（元）
湖南	3	1.5	2	4558.5	3338.1
湖北	2.9	1.4	2.1	5010.7	3616.4
广东	3.3	1.7	1.9	7828.8	5941.7
广西	3	1.5	2	4876.8	3508.5
海南	3.6	1.6	2.3	4323	2975.4
重庆	3.1	1.6	1.9	5283.8	4187.8
四川	2.9	1.4	2	4333.5	3326.7
贵州	3.1	1.4	2.1	4177.4	3066.3
云南	3	1.3	2.2	4619.8	3415.4
西藏	3.4	1.7	2	4668.8	4467.1
陕西	3	1.5	2	4342.7	3186.6
甘肃	2.9	1.5	1.9	4031.8	3113.2
青海	3	1.3	2.3	3971.8	3070.3
宁夏	2.9	1.3	2.2	4078.3	3133.7
新疆	3	1.5	2.1	4018.4	3015.1

（一）利用SPSS进行因子分析

- 将原始数据输入SPSS数据编辑窗口，将5个变量分别命名为X_1～X_5。在SPSS窗口中选择Analyze→Data Reduction→Factor菜单项，调出因子分析主界面，并将变量X_1～X_5移入Variables框中，其他均保持系统默认选项，单击OK按钮，执行因子分析过程（关于因子分子在SPSS中实现的详细过程，参见第7章实例）。得到如表6.6所示的特征根和方差贡献率表和表6.7所示的因子载荷阵。
- 表6.6中Total列为各因子对应的特征根，本例中共提取两个公因子；% of Variance列为各因子的方差贡献率；Cumulative %列为各因子累积方差贡献率，由表中可以看出，前两个因子已经可以解释79.31%的方差

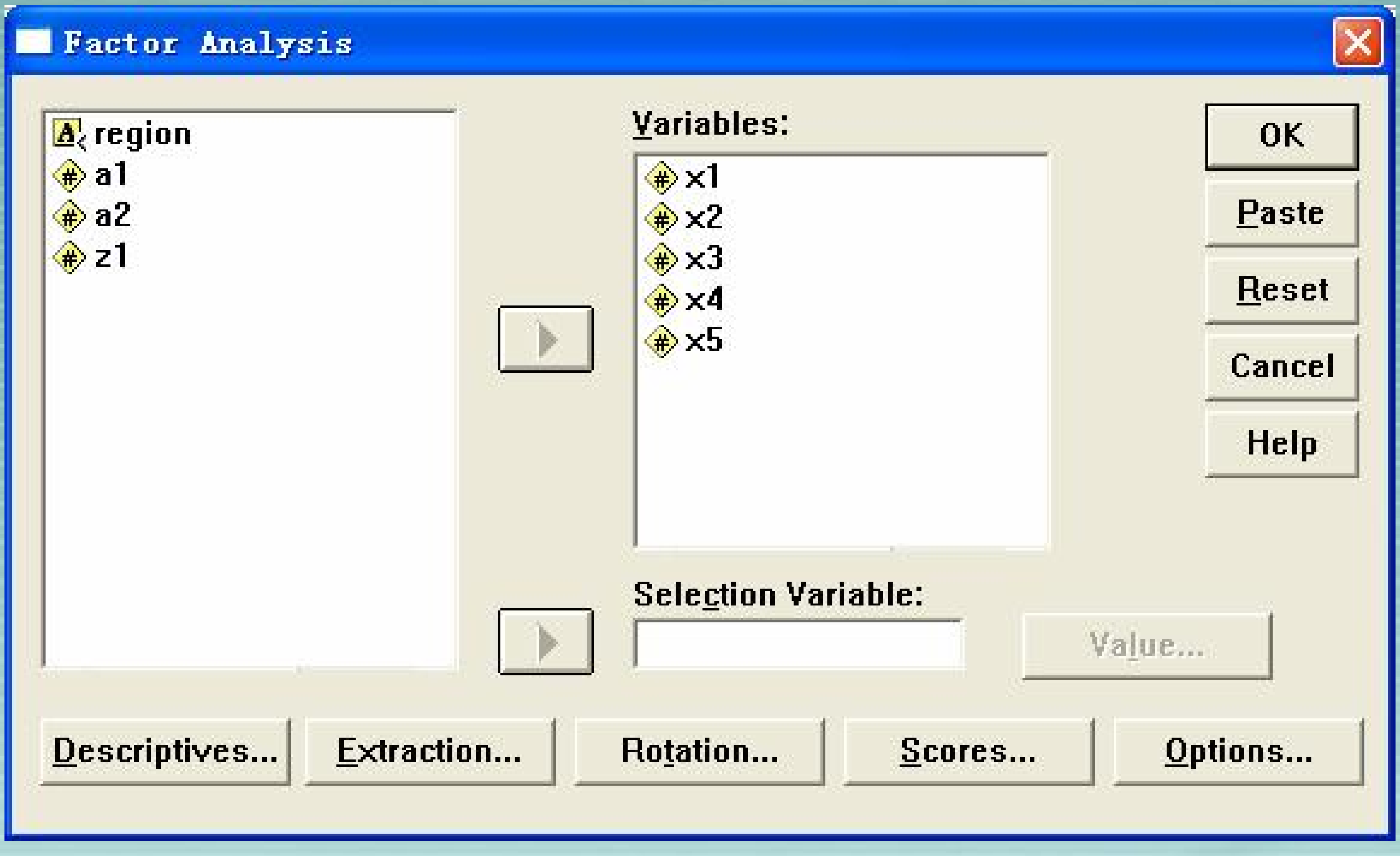

Factor Analysis
region
a1
a2
z1
Variables:
x1
x2
x3
x4
x5
OK
Paste
Reset
Cancel
Help
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...

图6.2 因子分析主界面

表6.6 特征根和方差贡献率表

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.576	51.520	51.520	2.576	51.520	51.520
2	1.389	27.790	79.310	1.389	27.790	79.310
3	.961	19.222	98.532			
4	.047	.932	99.465			
5	.027	.535	100.000			

Extraction Method: Principal Component Analysis.

（二）利用因子分析结果进行主成分分析

1. 将表6.7中因子载荷阵中的数据输入SPSS数据编辑窗口，分别命名为*a*1和*a*2。

Component Matrix [a]

	Component	
	1	2
X1	.121	.928
X2	.708	.612
X3	-.722	.125
X4	.873	-.299
X5	.882	-.220

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

表6.7 因子载荷阵

2. 为了计算第一个特征向量，点击菜单项中的Transform→Compute，调出Compute variable对话框，在对话框中输入等式：

$$z1=a1\ /\ SQRT(2.576)$$

点击OK按钮，即可在数据编辑窗口中得到以z1为变量名的第一特征向量。

再次调出Compute variable对话框，在对话框中输入等式：

$$z2=a2\ /\ SQRT(1.389)$$

点击OK按钮，得到以z2为变量名第二特征向量。这样，我们得到了如表6.8所示的特征向量矩阵。

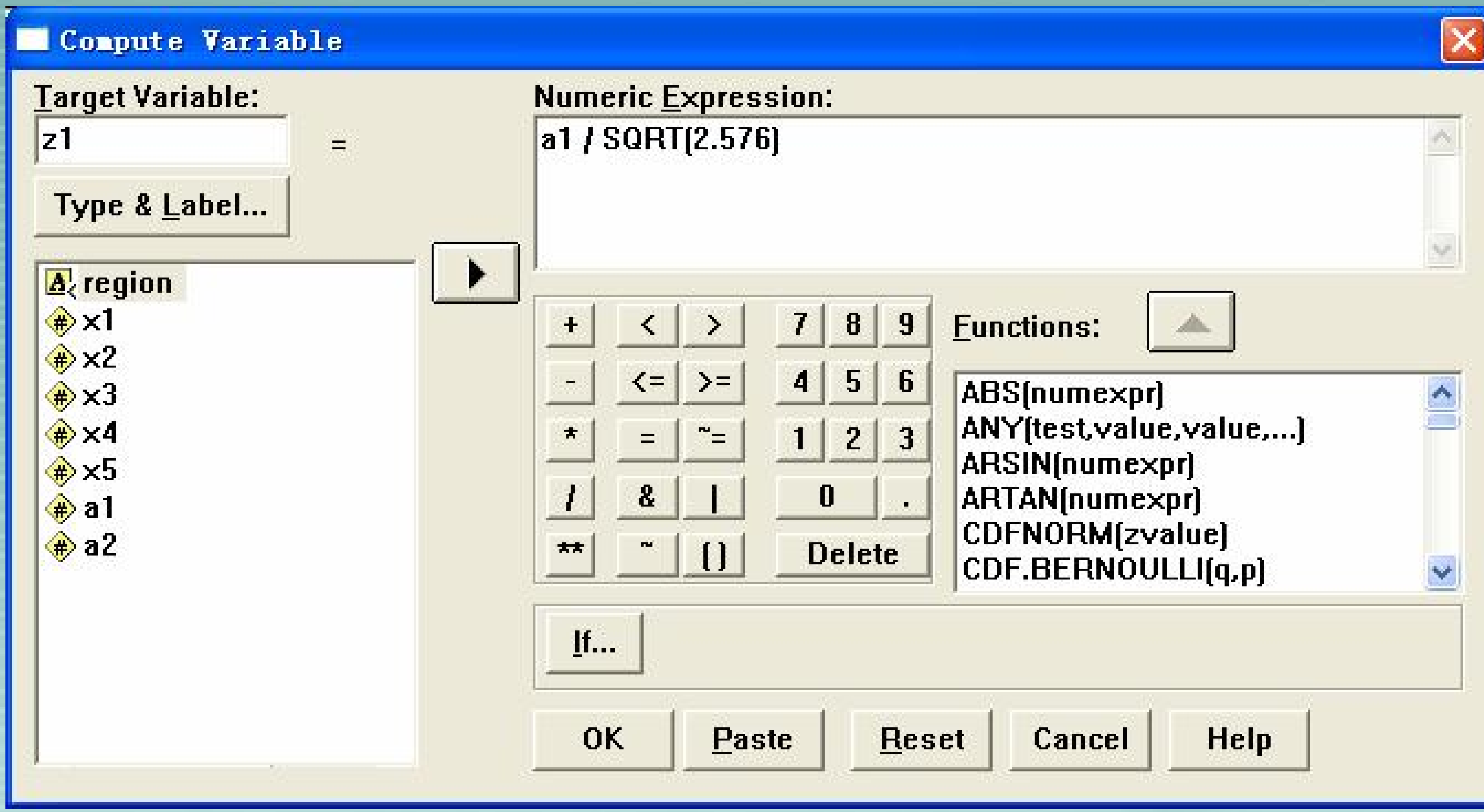


图6.3 Compute variable对话框

表6.8　特征向量矩阵

	Z1	Z2
X1	0. 075	0. 787
X2	0. 441	0. 519
X3	−0. 450	0. 106
X4	0. 544	−0. 254
X5	0. 550	−0. 187

- 根据表6.8可以得到主成分的表达式：

$$Y_1 = 0.075X_1 + 0.441X_2 - 0.450X_3 + 0.544X_4 + 0.550X_5$$

$$Y_1 = 0.787X_1 + 0.519X_2 + 0.106X_3 - 0.254X_4 - 0.178X_5$$

3. 再次使用Compute命令，就可以计算得到两个主成分。

本章结束