

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第五章 主成分分析

主讲教师 周仕君

为什么要进行主成份分析？

- 由于变量个数太多，并且彼此之间存在着一定的相关性，因而使得所观测到的数据在一定程度上反映的信息有所重叠。
- 例如：某人要做一件上衣要测量很多尺寸，如身长、袖长、胸围、腰围、肩宽、肩厚等十几项指标，但事实上服装厂做衣服时一般只考虑三项指标，即反映长度的指标、反映胖瘦的指标和反映特体的指标。



- 另外，当变量较多时，在高维空间中研究样本的分布规律比较复杂，势必增加分析问题的复杂性。
- 因此，我们希望用较少的综合变量来代替原来较多的变量；而这几个综合变量又能够尽可能多地反映原来变量的信息，并且彼此之间互不相关。
- 利用这种降维的思想，产生了主成份分析、因子分析、典型相关分析、偏最小二乘回归等统计方法。



什么是主成份分析？

- 主成份分析就是设法将原来指标重新组合成一组新的互相无关的几个综合指标来代替原来指标，同时根据实际需要从中可取几个较少的综合指标尽可能多地反映原来指标的信息。
- 这种将多个指标化为少数互相无关的综合指标的统计方法叫做主成份分析或称主分量分析。



主成份分析的经典案例

美国的统计学家斯通在1947年所做的关于国民经济的研究中，曾利用美国1929年-1938年各年的数据，得到17个反映国民收入与支出的变量要素，如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等。

在他进行主成份分析后，发现能够以97.4%的精度，用3个新变量取代原来的17个变量。这3个新变量可以分别命名为总收入、总收入变化率 and 经济发展或衰退的趋势。

更有趣的是，这3个变量其实都是可以直接测量的。



主成份分析

- 第一节 主成份的几何意义及数学推导
- 第二节 主成份的性质
- 第三节 主成份方法应用中应注意的问题
- 第四节 主成份回归



第一节 主成份的几何意义及数学推导

- 主成份的几何意义
- 主成份的数学推导

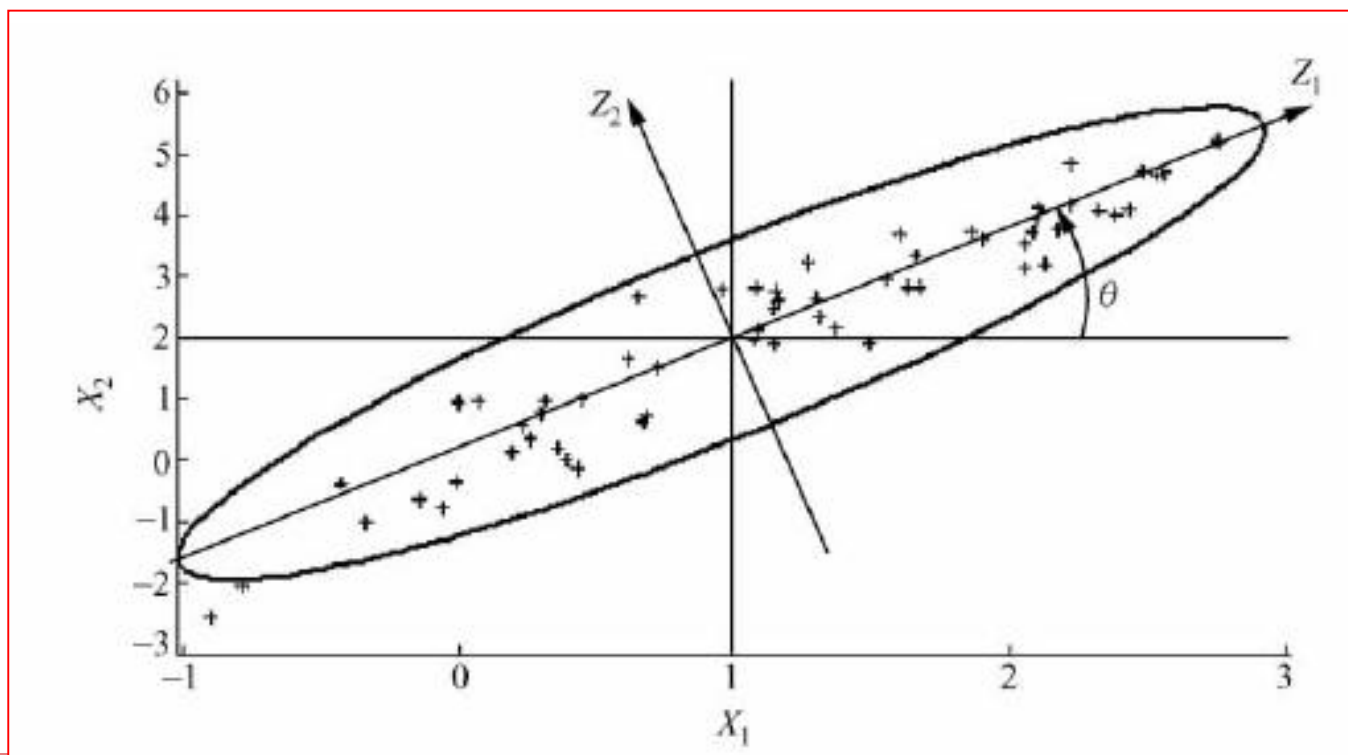


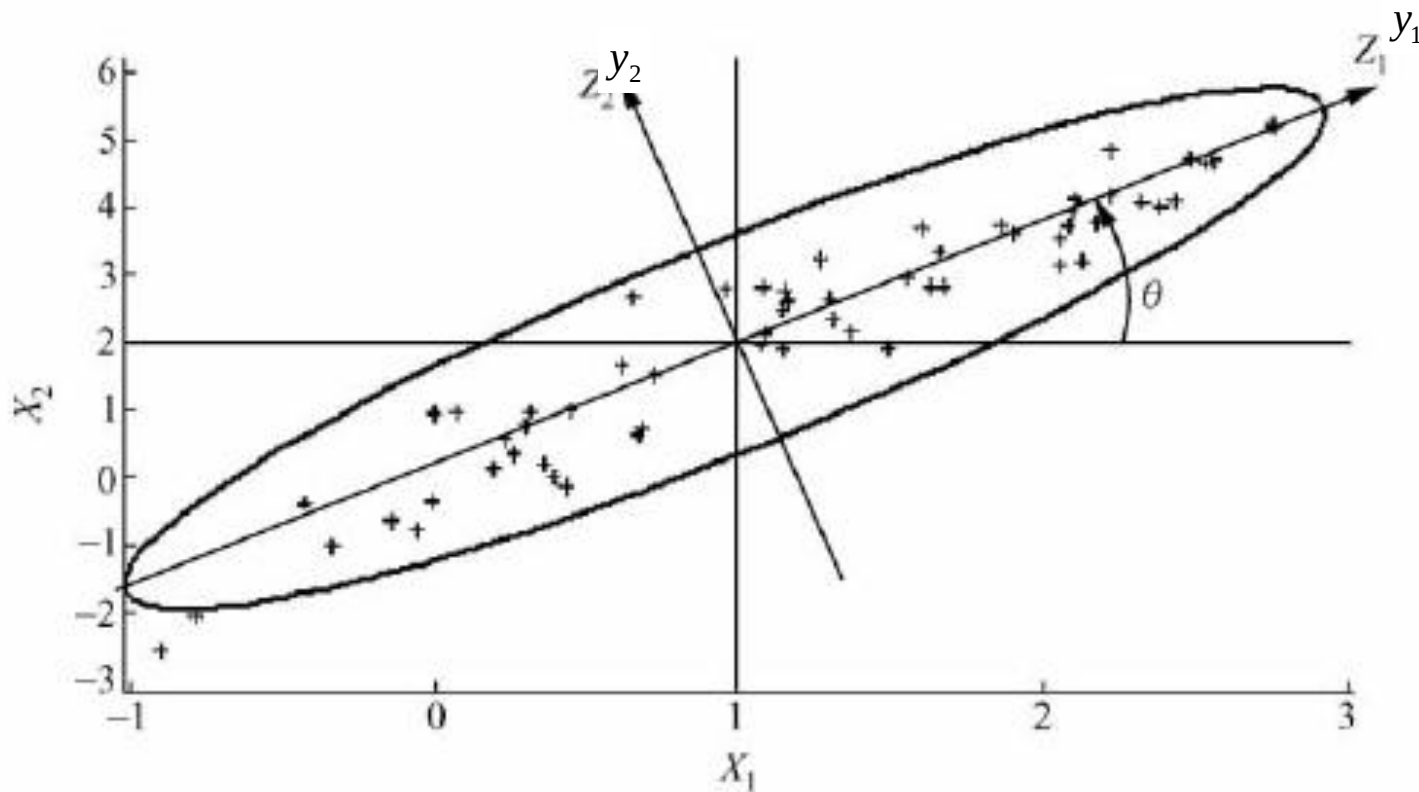
一、主成份的几何意义

- 主成份分析从代数观点看就是 p 个原始变量的一些特殊的线性组合；
- 而主成份分析从几何意义上看，相当于这些线性组合通过把由 $X_1, X_2, \dots, X_p$ 构成的坐标系旋转而产生的新坐标系。这样的新坐标轴使其通过样本变差最大的方向（或者说具有最大的样本方差）。



- 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所示。
- 当原始变量 X_1, X_2 相关性越强，这个椭圆就越扁。





从图上可以看出二维平面上 n 个点波动（用两个变量的方差和表示），大部分波动归为第一主成份，称 y_1 为第一主成份，第二主成份的效果与椭圆的形状有很大的关系，椭圆越扁， y_1 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，用第一主成份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，第一主成份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，第一主成份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，此时的主成份分析效果是非常理想的，其原因是，第二主成份不包含任何信息，舍弃它当然没有信息损失。



- 如果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

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

- 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:

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

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 $\mathbf{T}' \mathbf{T} = \mathbf{I}$ 。



二、主成份的数学推导

□ 设 $\mathbf{X}=(X_1, \dots, X_p)'$ 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：

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

□ 考虑如下的线性变换

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

用矩阵表示为 $\mathbf{Y}=(Y_1, Y_2, \dots, Y_p)'$ ，其中 $\mathbf{Y} = \mathbf{T}'\mathbf{X}$, $\mathbf{T}=(T_1, T_2, \dots, T_p)$ 。



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

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

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

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



- 由主成份定义可知，要找出 p 个原始变量的第一主成份，就是要找出使 $D(Y_i)$ 达到最大的线性组合。由于 $D(Y_i) = T_i' \Sigma T_i$ 是向量 T 的增函数，若对 T 不加任何限制，则 $T_i' \Sigma T_i$ 可以任意大。为了消除这种不确定性，通常将线性组合 $Y = T'X$ 的系数标准化即单位化，令 $T_i' T_i = 1$ 或者 $|T| = 1$ ，即 T 的长度为1。那么，问题就变成了在约束条件下，求最大的问题了。



- 第一主成份满足 $T_1' T_1 = 1$ 使得 $D(Y_1) = T_1' \Sigma T_1$ 达到最大的 $Y_1 = T_1' X$ 。
- 第二主成份满足 $T_2' T_2 = 1, \text{Cov}(Y_2, Y_1) = \text{Cov}(T_2' X, T_1' X) = 0$ 使得 $D(Y_2) = T_2' \Sigma T_2$ 达到最大的 $Y_2 = T_2' X$ 。
- 第 k 主成份满足 $T_k' T_k = 1, \text{Cov}(Y_k, Y_i) = \text{Cov}(T_k' X, T_i' X) = 0 \ (i < k)$ 使得 $D(Y_k) = T_k' \Sigma T_k$ 达到最大的 $Y_k = T_k' X$ 。



- 求第一主成份，构造目标函数为：

$$\varphi_1(T_1, \lambda) = T_1' \Sigma T_1 - \lambda(T_1' T_1 - 1) \quad (6.5)$$

- 对目标函数 $\varphi_1(T_1, \lambda)$ 求导数有：

$$\frac{\partial \varphi_1}{\partial T_1} = 2\Sigma T_1 - 2\lambda T_1 = 0 \quad (6.6)$$

$$\text{即 } (\Sigma - \lambda \mathbf{I})T_1 = 0 \quad (6.7)$$

- 由6.7式两边左乘 T_1' 得到

$$T_1' \Sigma T_1 = \lambda \quad (6.8)$$

- 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 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 。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_{1^0}



□ 在求第二主成份之前，首先明确，由(6.6)知

$\text{Cov}(Y_2, Y_1) = T_2' \Sigma T_1 = \lambda T_2' T_1$ 。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，可以构造求第二主成份
的目标函数，即

$$\varphi_2(T_2, \lambda, \rho) = T_2' \Sigma T_2 - \lambda(T_2' T_2 - 1) - 2\rho(T_1' T_2) \quad (6.9)$$

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

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

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

□ 由于 $T_1' T_2 = 0$, $T_1' \Sigma T_2 = 0$ ，那么， $\rho T_1' T_1 = 0$ ，即有 $\rho = 0$ 。

□ 从而 $(\Sigma - \lambda I)T_2 = 0$ (6.11) 而且 $T_2' \Sigma T_2 = \lambda$ (6.12)



□ 这样说明，如果 X 的协差阵 Σ 的特征根为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 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 主成份应该是在 $T_k' T_k = 1$ 且 $T_i' T_k = 0$ 或 $T_k' T_i = 0$ ($i < k$) 的条件下，使得 $D(Y_k) = T_k' \Sigma T_k$ 达到最大的 $Y_k = T_k' X$ 。这样我们构造目标函数为：

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

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

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



□ 用 T_i' 左乘(6.14)式有 $T_i'\Sigma T_k - \lambda T_i' T_k - T_i'(\sum_{i=1}^{k-1} \rho_i T_i) = 0$ 即 $\rho_i T_i' T_i = 0$
那么 $\rho_i = 0 (i=1, 2, \dots, k-1)$ 。从而 $(\Sigma - \lambda \mathbf{I})T_k = 0$ (6.15)

而且 $T_k'\Sigma T_k = \lambda$ (6.16)

□ 对于 X 的协差阵 Σ 的特征根 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 。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 $X = (X_1, \dots, X_p)'$ 的协差阵为 Σ ，其特征根为 $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ ，相应的单位化的特征向量为 $T_1, T_2, \dots, T_p$ 。那么，由此所确定的主成份为 $Y_1 = T_1' X, \dots, Y_m = T_m' X$ ，其方差分别为 Σ 的特征根。



例题

- 例1.假设市场上肉类 X_1 ，鸡蛋 X_2 ，水果 X_3 三种商品价格的月份资料的协方差矩阵为：

$$\Sigma = \begin{pmatrix} 2 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix}$$

- 试求三种商品月份价格的所有主成份。



□ 解：根据协方差阵，写出其特征多项式：

$$|\lambda I - \Sigma| = \begin{vmatrix} \lambda - 2 & -2 & 2 \\ -2 & \lambda - 5 & 4 \\ 2 & 4 & \lambda - 5 \end{vmatrix} = (\lambda - 1)^2 (\lambda - 10)$$

令此多项式等于0，得到特征根为： $\lambda_1 = 10, \lambda_2 = \lambda_3 = 1$

$$T_1 = (1, 2, -2)', T_2 = (0, \sqrt{2}, \sqrt{2}), T_3 = (4\sqrt{2}, -\sqrt{2}, -\sqrt{2})' \quad \text{化}$$
$$T_1 = \left(\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}\right) \quad T_2 = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \quad T_3 = \left(\frac{2\sqrt{2}}{3}, -\frac{\sqrt{2}}{6}, -\frac{\sqrt{2}}{6}\right)'$$



□ 于是，三种商品价格的三个主成份分别为：

$$Y_1 = T_1' X = \frac{1}{3} X_1 + \frac{2}{3} X_2 - \frac{2}{3} X_3$$

$$Y_2 = T_2' X = \frac{\sqrt{2}}{2} X_2 + \frac{\sqrt{2}}{2} X_3$$

$$Y_3 = T_3' X = \frac{2\sqrt{2}}{3} X_1 - \frac{\sqrt{2}}{6} X_2 + \frac{\sqrt{2}}{6} X_3$$

□ 其中，第一个主成份的方差占了原始变量的总方差的绝大部分，所以第一主成份综合反映了三种商品价格的绝大部分变动。



第二节 主成份的性质

- 主成份的一般性质
- 主成份的方差贡献率



一、主成份的一般性质

□ 设 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)$ 是 $\mathbf{X}$ 的主成份, 由 Σ 的所有特征根

构成的对角阵为

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \quad (6.17)$$

主成份可表示为 $\mathbf{Y} = \mathbf{T}'\mathbf{X}$ (6.18)

□ 性质1 主成份的协方差矩阵是对角阵。

证明: 实际上, 由 (6.3) 式知 $E(\mathbf{Y}) = E(\mathbf{T}'\mathbf{X}) = \mathbf{T}'\boldsymbol{\mu}$

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



□ 性质2 主成份的总方差等于原始变量的总方差。

(所谓总方差指的是p维数据的波动情况, 即数据阵X的差异信息, 也可称为总惯量)

证明: 由矩阵“迹”的性质知

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

$$\text{所以 } \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \sigma_{ii} \quad \text{或} \quad \sum_{i=1}^p D(Y_i) = \sum_{i=1}^p D(X_i)$$

□ 上式表明原始变量的全部主成份解释了原始变量的全部变动, 并且反映了对差异信息的不同分配。



□ 性质3 主成份 Y_k 与原始变量 X_i 的相关系数为：

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

并称之为因子负荷量（或因子载荷量）。



证明：事实上

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

其中的 $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ ，它是除第 i 个元素为1外

其他元素均为0的单位向量。而

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

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

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

证明：只须将(6.22) 代入左边式子整理化简即可。



二、主成份的方差贡献率

- 由主成份的性质2可以看出，主成份分析把 p 个原始变量 $X_1, X_2, \dots, X_p$ 的总方差 $tr(\Sigma)$ 分解成了 p 个相互独立的变量 $Y_1, Y_2, \dots, Y_p$ 的方差之和 $\sum_{k=1}^p \lambda_k$ 。主成份分析的目的是减少变量的个数，所以一般不会使用所有 p 个主成份的。但是，那么究竟应该选取几个主成份呢？该怎么选取呢？



□ 这里称
$$\varphi_k = \lambda_k / \sum_{k=1}^p \lambda_k$$

为第 k 个主成份 Y_k 的贡献率。第一主成份的贡献率最大，这表明 $Y_1 = T_1'X$ 综合原始变量 $X_1, X_2, \dots, X_p$ 的能力最强，而 $Y_2, Y_3, \dots, Y_p$ 的综合能力依次递减。

□ 若只取 $m(< p)$ 个主成份，则称

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

为主成份 $Y_1, \dots, Y_m$ 的累计贡献率，累计贡献率表明

$Y_1, \dots, Y_m$ 综合 $X_1, X_2, \dots, X_p$ 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）。



注意：信息提取率

- 虽然主成份的贡献率给出了选取主成份的一个标准，但是累积贡献率只是表达了前 k 个主成份提取了 $X_1, X_2, \dots, X_p$ 的多少信息，并没有表达某个变量被提取了多少信息，因此仅仅使用累积贡献率这一标准，不能保证每一个变量都被提取了足够的信息，也可能85%的累积贡献率只是提取了 p 个原始变量的一个子集的信息，而有些变量根本没有考虑到，因此还需要另一个辅助的标准——信息提取率。



-
- 由于任一原始变量 X_i 可用主成份表示为：

$$X_i = t_{1i}Y_1 + t_{2i}Y_2 + \cdots + t_{pi}Y_p$$

- 其中各个主成份之间互不相关，所以原始变量的 X_i 的方差为： $Var(X_i) = t_{1i}^2\lambda_1 + t_{2i}^2\lambda_2 + \cdots + t_{pi}^2\lambda_p$

$$= \sum_{j=1}^p t_{ji}^2\lambda_j = \sigma_{ii}$$

- 比率 $t_{ji}^2\lambda_j / \sigma_{ii}$ 反映了第 j 个主成份从第 i 个原始变量中所提取的信息的大小。



□ 如果只选取前 m 个主成份，则这 m 个主成份从原始变量

中所提取的总信息量为： $\sum_{j=1}^m t_{ji}^2 \lambda_j / \sigma_{ii}$

即这 m 个主成份能够解释第 i 个变量变动的程度为

$\sum_{j=1}^m t_{ji}^2 \lambda_j / \sigma_{ii}$ ，所以可以定义比率：

$$\Omega_i = \sum_{j=1}^m t_{ji}^2 \lambda_j / \sigma_{ii}$$

为原始变量 X_i 的信息提取率。

➤ 因此，在选取主成份时，不仅要考虑累积贡献率，还应考虑信息提取率。



例题

- 例2.假设市场上棉鞋 X_1 ，凉鞋 X_2 ，布鞋 X_3 三种商品销售量的协方差矩阵为：

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

- 试求各主成份，并计算各主成份的贡献率和各个原始变量的信息提取率。



□ 解：根据协方差阵，写出其特征多项式：

$$|\lambda I - \Sigma| = \begin{vmatrix} \lambda - 1 & 2 & 0 \\ 2 & \lambda - 5 & 0 \\ 0 & 0 & \lambda - 2 \end{vmatrix} = (\lambda - 2)(\lambda^2 - 6\lambda + 1)$$

令此多项式等于0，得到特征根为： $\lambda_1 = 5.83, \lambda_2 = 2, \lambda_3 = 0.17$
将特征根代入特征方程，得到相应的特征向量，其单位化的结果为：

$$T_1 = (0.383, -0.924, 0)' \quad T_2 = (0, 0, 1)' \quad T_3 = (0.924, 0.383, 0)'$$



□ 于是，三种商品价格的三个主成份分别为：

$$Y_1 = T_1' X = 0.383X_1 - 0.924X_2$$

$$Y_2 = T_2' X = X_3$$

$$Y_3 = T_3' X = 0.924X_1 + 0.383X_2$$

□ 若只取第一个主成份，则累积贡献率为：

$$\frac{\lambda_1}{\sum \lambda} = \frac{5.83}{5.83 + 2 + 0.17} = 72.875\%$$



□ 若进一步计算每个变量的信息提取率，则：

$$\Omega_1 = \frac{\lambda_1 t_{11}^2}{\sigma_{11}} = \frac{5.83 \times (0.383)^2}{1} = 0.855$$

$$\Omega_2 = \frac{\lambda_1 t_{12}^2}{\sigma_{22}} = \frac{5.83 \times (-0.924)^2}{5} = 0.996$$

$$\Omega_3 = \frac{\lambda_1 t_{13}^2}{\sigma_{33}} = \frac{5.83 \times 0^2}{2} = 0$$

□ 可见，第三个变量的信息提取率为0，说明Y1和与x3都不相关，在第一个主成份中一点都没有包含x3的信息，因此还需再取第二个主成份，此时累积贡献率为：

$$\frac{\lambda_1 + \lambda_2}{\sum \lambda} = \frac{5.83 + 2}{5.83 + 2 + 0.17} = 97.875\%$$



□ 此时，三个变量的信息提取率为：

$$\Omega_1 = \frac{\sum_{j=1}^2 \lambda_j t_{j1}^2}{\sigma_{11}} = \frac{5.83 \times (0.383)^2 + 2 \times 0}{1} = 0.855$$

$$\Omega_2 = \frac{\sum_{j=1}^2 \lambda_j t_{j2}^2}{\sigma_{22}} = \frac{5.83 \times (-0.924)^2 + 2 \times 0}{1} = 0.996$$

$$\Omega_3 = \frac{\sum_{j=1}^2 \lambda_j t_{j3}^2}{\sigma_{33}} = \frac{5.83 \times 0 + 2 \times 1^2}{2} = 1$$

□ 可见，三个变量的信息提取率都比较高，应取两个主成份来进行分析。



第三节 主成份方法应用中应注意的问题

- 实际应用中主成份分析的出发点
- 如何利用主成份分析进行综合评价



一、实际应用中主成份分析的出发点

- 在实际问题中，不同的变量往往有不同的量纲，利用协差矩阵来求主成份，需要计算各变量方差的总和，而不同量纲的方差相加是没有意义的，并且通过协差阵来求主成份，首先优先照顾方差大的变量，有时会造成很不合理的结果。为了消除量纲的不同可能带来的一些不合理的影响，在进行主成份之前，常常需要将变量标准化。



- 标准化：为使主成份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

$$X_i^* = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}} \quad i = 1, \dots, p$$

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



【例 5—1】 假定我们研究某一经济问题共涉及两个指标：产值和利税。其中产值以百万元计，利税以万元计，得原始资料矩阵如下：

$$\mathbf{X} = \begin{bmatrix} 12.5 & 586 \\ 24 & 754 \\ 15.3 & 850 \\ 18 & 667 \\ 31.2 & 750 \end{bmatrix}$$

可以得到，原始变量的协方差阵与相关阵分别为：

$$\mathbf{\Sigma} = \begin{bmatrix} 55.90 & 242.65 \\ 242.65 & 9\,927.80 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & 0.325\,7 \\ 0.325\,7 & 1 \end{bmatrix}$$



由协方差阵出发求解主成分，得到结果见表 5—1。

表 5—1

特征值	解释方差比例	累积比例
9 933.760 7	0.995 0	0.995 0
49.934 3	0.005 0	1.000 0

对应两特征值的标准正交特征向量见表 5—2。

表 5—2

特征值 1	特征值 2
0.024 6	0.999 7
0.999 7	-0.024 6

因此，所得的主成分的表达式为：

$$Y_1 = 0.024\ 6(X_1 - \bar{X}_1) + 0.999\ 7(X_2 - \bar{X}_2)$$

$$Y_2 = 0.999\ 7(X_1 - \bar{X}_1) - 0.024\ 6(X_2 - \bar{X}_2)$$



对产值与利税均以万元计，原始数据资料阵变为以下形式：

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 250 & 586 \\ 2 & 400 & 754 \\ 1 & 530 & 850 \\ 1 & 800 & 667 \\ 3 & 120 & 750 \end{bmatrix}$$

相关矩阵没有变化，协方差矩阵变为：

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 558.950 & 24.265 \\ 24.265 & 9.927.8 \end{bmatrix}$$



由此协方差矩阵出发重新求主成分，结果见表 5—5。

表 5—5

特征值	解释方差比例	累积比例
560 020.348	0.984 4	0.984 4
8 857.452	0.015 6	1.000 0

对应两特征值的标准正交特征向量见表 5—6。

表 5—6

特征值 1	特征值 2
0.999 029	-0.044 068
0.044 068	0.999 029

此时所得主成分的表达式为：

$$Y_1 = 0.999\ 029(X_1 - \bar{X}_1) - 0.044\ 068(X_2 - \bar{X}_2)$$

$$Y_2 = 0.044\ 068(X_1 - \bar{X}_1) + 0.999\ 029(X_2 - \bar{X}_2)$$



由相关矩阵求解主成分的结果见表 5—3。

表 5—3

特征值	解释方差比例	累积比例
1.325 7	0.662 9	0.662 9
0.674 2	0.337 1	1.000 0

对应两特征值的标准正交特征向量见表 5—4。

表 5—4

特征值 1	特征值 2
0.707 1	0.707 1
0.707 1	-0.707 1

此时，所得主成分的表达式为：

$$Y_1 = 0.707\ 1Z_1 + 0.707\ 1Z_2 = 0.707\ 1\left(\frac{X_1 - \bar{X}_1}{\sqrt{55.9}}\right) + 0.707\left(\frac{X_2 - \bar{X}_2}{\sqrt{9\ 927.80}}\right)$$

$$Y_2 = 0.707\ 1Z_1 - 0.707\ 1Z_2 = 0.707\ 1\left(\frac{X_1 - \bar{X}_1}{\sqrt{55.9}}\right) - 0.707\left(\frac{X_2 - \bar{X}_2}{\sqrt{9\ 927.80}}\right)$$



标准化的实质

- 对原始数据做标准化处理之后倾向于各个指标在主成分的构成中相等。其原因就在于对数据进行标准化的过程实际上抹杀了数据本身离散度的信息。标准化后各原始指标的方差均为1，但实际上方差也是对数据信息的重要概括形式。因此对于同度量或相似量级的变量还是用原始的协方差阵求解主成分为宜。



-
- **注意：**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：
 - 使用原始变量的 R 代替 Σ 做主成份分析；
 - 使用标准化变量的相关阵或者协差阵做主成份分析，二者结果是相同的。

 - 需要进一步强调的是，从原始变量的相关阵求得的主成份与用原始变量的协差阵求得的主成份一般情况是不相同的。实际表明，这种差异有时很大。



例题

□ 例1.假设市场上肉类 X_1 ，鸡蛋 X_2 ，水果 X_3 三种商品价格的月份资料的相关矩阵为：

$$R = \begin{pmatrix} 1 & \frac{2}{\sqrt{10}} & -\frac{2}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} & 1 & -\frac{4}{5} \\ -\frac{2}{\sqrt{10}} & -\frac{4}{5} & 1 \end{pmatrix}$$

□ 试求三种商品月份价格的所有主成份。



□ 解：根据相关阵，写出其特征多项式：

$$|\lambda I - R| = \begin{vmatrix} \lambda - 1 & -\frac{2}{\sqrt{10}} & \frac{2}{\sqrt{10}} \\ -\frac{2}{\sqrt{10}} & \lambda - 1 & \frac{4}{5} \\ \frac{2}{\sqrt{10}} & \frac{4}{5} & \lambda - 1 \end{vmatrix} = \left(\lambda - \frac{1}{5}\right) \left(\lambda^2 - \frac{14}{5}\lambda + 1\right)$$

令此多项式等于0，得到特征根为： $\lambda_1 = 2.38, \lambda_2 = 0.42, \lambda_3 = 0.2$

将特征根代入特征方程，得到相应的特征向量，其单位化的结果为：

$$T_1 = (0.54, 0.59, -0.59)' \quad T_2 = (0.84, -0.39, 0.39)' \quad T_3 = (0, 0.71, 0.71)'$$



□ 于是，三种商品价格的三个主成份分别为：

$$Y_1 = T_1' X = 0.54X_1 + 0.59X_2 - 0.59X_3$$

$$Y_2 = T_2' X = 0.84X_1 - 0.39X_2 + 0.39X_3$$

$$Y_3 = T_3' X = 0.71X_2 + 0.71X_3$$

□ 其中，第一个主成份的方差占了原始变量的总方差的绝大部分，所以第一主成份综合反映了三种商品价格的绝大部分变动。



在实际应用中，主成份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 $\lambda_1^* \geq \dots \geq \lambda_p^* \geq 0$ ，相应的特征向量为 $T_1^*, T_2^*, \dots, T_p^*$ ；
- ④ 由累积方差贡献率确定主成份的个数 (m)，并写出主成份为： $Y_i = (T_i^*)' X \ (i=1, 2, \dots, m)$



二、如何利用主成份分析进行综合评价

- 一般在进行综合评价时，需要给各指标赋予权重，再进行综合评价。
- 如果是用主成份分析进行综合评价，则在评价前需要给主成份赋予权重，此时，可以方差贡献率作为参考来确定权重。



注意：主成份得分

□ 主成份得分是指第 i 个主成份 $Y_i = T_i' X$ 。将第 t 个样品的值：

$$X_{(t)} = (X_{t1}, X_{t2}, \dots, X_{tp})'$$

代入 $Y_i = T_i' X$ ，经计算得到的值称为第 t 个样品在第 i 个主成份的得分。

□ **注意：**此时的样品数据是标准化后的数据。



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

$$w_i = \lambda_i / \sum_{i=1}^m \lambda_i \quad i=1, 2, \dots, p$$

- 记为 $W = (w_1, w_2, \dots, w_p)'$ ，由 $Y = T'X$ ，构造综合评价函数为：

$$Z = w_1 Y_1 + w_2 Y_2 + \dots + w_p Y_p = W'Y = W'T'X = (TW)'X \quad (6.26)$$

- 令 $TW = w_{k \times 1}^*$ ，并代入(6.26)式，有

$$Z = (w^*)'X \quad (6.27)$$

- 从本质上说综合评价函数是对原始指标的线性综合，从计算主成份到对之加权，经过两次线性运算后得到综合评价函数。



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



-
- 综合评价即是从这些经济指标出发，对各工业部门进行综合评价与排序。
 - 首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。



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



-
- 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：

$$Y_1 = 0.476X_1^* + 0.473X_2^* + 0.424X_3^* - 0.213X_4^* \\ - 0.388X_5^* - 0.352X_6^* + 0.215X_7^* + 0.055X_8^*$$

$$Y_2 = 0.296X_1^* + 0.278X_2^* + 0.378X_3^* + 0.451X_4^* \\ + 0.331X_5^* + 0.403X_6^* - 0.377X_7^* + 0.273X_8^*$$

.....



□ 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。

□ 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。具体数据见表6.4。



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

➤从左表可以看出，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。



案例2

例 对全国 30 个省市自治区经济发展基本情况的八项指标作主成分分析，原始数据如下：

省份	GDP X_1	居民消 费水平 X_2	固定资 产投资 X_3	职工平 均工资 X_4	货物周 转 量 X_5	居民消费 价格指数 X_6	商品零售 价格指数 X_7	工业总 产 值 X_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.31
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1849.55
吉林	1129.2	1872	320.45	4430	197.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64



浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	174	4609	908.3	114.8	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.11	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	6002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69
河南	3002.74	1034	670.35	4344	1571.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1403	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1311	382.59	5105	556	118.4	116.4	554.97
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.81	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.58	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.55	5348	339	119.7	116.7	428.76



第一步 将原始数据标准化。

第二步 建立指标之间的相关系数阵 R 如下：

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	.267	.951	.191	.617	-.274	-.264	.874
X_2	.267	1.000	.426	.718	-.151	-.234	-.593	.363
X_3	.951	.426	1.000	.400	.431	-.282	-.359	.792
X_4	.191	.718	.400	1.000	-.356	-.134	-.539	.104
X_5	.617	-.151	.431	-.356	1.000	-.255	.022	.659
X_6	-.274	-.234	-.282	-.134	-.255	1.000	.760	-.126
X_7	-.264	-.593	-.359	-.539	.022	.760	1.000	-.192
X_8	.874	.363	.792	-.104	.659	-.126	-.192	1.000



第三步 求 R 的特征值和特征向量。

主成分	特征值	方差贡献率	累计贡献率
1	3.755	46.943	46.943
2	2.195	27.443	74.386
3	1.214	15.178	89.564
4	.403	5.033	94.596
5	.213	2.660	97.256
6	.139	1.737	98.993
7	6.594E-02	.824	99.817
8	1.462E-02	.183	100.000

从上表看，前 3 个特征值累计贡献率已达 89.564%，说明前 3 个主成分基本包含了全部指标具有的信息，我们取前 3 个特征值，并计算出相应的特征向量：



第一特征向量 a_1	第二特征向量 a_2	第三特征向量 a_3
0.470641	0.107995	0.19241
0.456708	0.258512	0.109819
0.424712	0.287536	0.19241
-0.31944	0.400931	0.397525
0.312729	-0.40431	0.24505
0.250802	0.498801	-0.24777
0.240481	-0.48868	0.332179
-0.26267	0.167392	0.723351



因而前三个主成分为：

$$\begin{aligned} \text{第一主成分: } F_1 = & 0.470641X_1 + 0.456708X_2 + 0.424712X_3 \\ & - 0.31944X_4 + 0.312729X_5 + 0.250802X_6 \end{aligned}$$

$$\begin{aligned} \text{第二主成分: } F_2 = & 0.107995X_1 + 0.258512X_2 + 0.287536X_3 \\ & + 0.400931X_4 - 0.40431X_5 + 0.498801X_6 \\ & - 0.48868X_7 + 0.167392X_8 \end{aligned}$$

$$\begin{aligned} \text{第三主成分: } F_3 = & 0.19241X_1 + 0.109819X_2 + 0.19241X_3 \\ & + 0.397525X_4 + 0.24505X_5 - 0.24777X_6 \\ & + 0.332179X_7 + 0.723351X_8 \end{aligned}$$



在第一主成分的表达式中第一、二、三项指标的系数较大，这三个指标起主要作用，我们可以把第一主成分看成是由国内生产总值、固定资产投资和居民消费水平所该划的反映经济发展状况的综合指标；

在第二主成分中，第四、五、六、七项指标的影响大，且第六、七项指标的影响尤其大，可将之看成是反映物价指数、职工工资和货物周转量的综合指标；

在第三主成分中，第八项指数影响最大，远超过其它指标的影响，可单独看成是工业总产值的影响。



第四节 主成份回归

- 多元线性回归
- 主成份回归的原理
- 主成份回归的应用



一、多元线性回归

□ 样本回归模型：

$$Y = (1_n \ X) \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} + \mu = \beta_0 1_n + X \beta + \mu$$

□ 模型系数向量估计量为：

$$\begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} = \left[(1_n \ X)' (1_n \ X) \right]^{-1} (1_n \ X)' Y$$



二、主成份回归的原理

- 最小二乘估计的假定前提之一：自变量不相关，即自变量之间不存在共线性问题。但在经济管理中，许多经济变量之间都存在相关关系。
- 解决此问题的一个方法就是：
 - 先对模型中的自变量进行主成份分析，并提取一定的主成份；
 - 然后用因变量对所取的主成份进行回归；
 - 最后，再将因变量对所提取的主成份的回归方程转化为对原自变量的回归方程。

这就是主成份回归。



□ 实际上，对于上述回归模型中的自变量 $X_1, X_2, \dots, X_p$ 的样本观测数据，若进行主成份分析，则可得 k 个主成份分别为：

$$\left\{ \begin{array}{l} Z_1 = t_{11}X_1 + t_{12}X_2 + \cdots + t_{1p}X_p \\ Z_2 = t_{21}X_1 + t_{22}X_2 + \cdots + t_{2p}X_p \\ \dots\dots\dots \\ Z_k = t_{k1}X_1 + t_{k2}X_2 + \cdots + t_{kk}X_k \end{array} \right.$$



- 记第*i*个样品在第*j*个主成份上的得分为 Z_{ij} ，并记全部样本的主成份得分矩阵为 $Z = (z_{ij})_{n \times k}$ ，记主成份系数矩阵为 $T = (t_{ij})_{n \times k}$ ，即有：

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix} \quad T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1k} \\ t_{21} & t_{22} & \cdots & t_{2k} \\ \vdots & \vdots & \vdots & \\ t_{n1} & t_{n2} & \cdots & t_{nk} \end{bmatrix}$$

- 若主成份分析是根据样本协方差 S 做出的，则主成份得分矩阵为： $Z = X_0 T$
- 若主成份分析是根据样本相关阵 R 做出的，则主成份得分矩阵为： $Z^* = X^* T^*$



-
- 若将主成份得分矩阵看作是自变量的样本观测矩阵，建立因变量 Y 对 k 个主成份的线性回归方程：

$$Y = r_0 + r_1 z_1 + \cdots + r_k z_k + \mu$$

- 或者： $Y = r_0^* + r_1^* z_1^* + \cdots + r_k^* z_k^* + \mu$

- 由于此回归方程的因变量与原回归方程的因变量相同，所以随机误差项也相同，并且 $E(\mu) = 0, E(\mu^2) = \sigma^2$ ，即可使用最小二乘法进行系数的估计。



在计算主成份得分值时，所用的变量数据都是经过中心化或标准化变换的。

□ 则模型系数的估计量为：

$$\begin{pmatrix} \hat{r}_0 \\ \hat{r} \end{pmatrix} = \begin{pmatrix} n & 1_n' & Z' \\ Z'1_n & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} 1_n Y \\ Z'Y \end{pmatrix}$$

□ 或者：
$$\begin{pmatrix} \hat{r}_0^* \\ \hat{r}^* \end{pmatrix} = \begin{pmatrix} n & 1_n' & Z^{*'} \\ Z^{*'}1_n & Z^{*'}Z^* \end{pmatrix}^{-1} \begin{pmatrix} 1_n Y \\ Z^{*'}Y \end{pmatrix}$$

□ 主成份回归的系数向量估计值与相应的原自变量回归系数向量估计值之间的关系：

$$\hat{\beta} = T\hat{r} \text{ 或者 } \hat{\beta}^* = T^*\hat{r}^*$$

□ 根据上述系数值即可得出因变量与原始变量的回归模型。



三、主成份回归实例

美国经济数据

联邦政府 雇员人数 (千人)	国民总产 出隐含平 减指数(%)	国民总产 出(百万 美元)	失业 人数 (千人)	武装力 量人数 (千人)	14 岁及以上非 慈善机构人口 数(千人)	时间 (年份)
1892	83.0	234289	2356	1590	107608	1947
1863	88.5	259426	2325	1456	108632	1948
1908	88.2	258054	3682	1616	109773	1949
1828	89.5	284599	3351	1650	110929	1950
2302	96.2	328975	2099	3099	112075	1951
2420	98.1	346999	1932	3594	113270	1952
2305	99.0	365385	1870	3547	115094	1953
2188	100.0	363112	3578	3350	116219	1954
2187	101.2	397469	2904	3048	117388	1955
2209	104.6	419180	2822	2857	118734	1956
2217	108.4	442769	2936	2798	120445	1957
2191	110.8	444546	4681	2637	121950	1958
2233	112.6	482704	3813	2552	123366	1959
2270	1142	502601	3931	2514	125368	1960
2279	115.7	518173	4806	2572	127852	1961
2340	116.9	554894	4007	2827	130081	1962



用各变量的均值与标准差将各个 x 变量标准化, 计算相关矩阵 R , 得到:

$$R = \begin{bmatrix} 1 & 0.9916 & 0.6206 & 0.4647 & 0.9792 & 0.9911 \\ 0.9916 & 1 & 0.6043 & 0.4464 & 0.9911 & 0.9953 \\ 0.6206 & 0.6043 & 1 & -0.1774 & 0.6866 & 0.6683 \\ 0.4647 & 0.4464 & -0.1774 & 1 & 0.3644 & 0.4172 \\ 0.9792 & 0.9911 & 0.6866 & 0.3644 & 1 & 0.9940 \\ 0.9911 & 0.9953 & 0.6683 & 0.4172 & 0.9940 & 1 \end{bmatrix}$$

- 可以看出, 变量 X_1, X_2, X_5, X_6 两两之间有很高的相关性。
- 因此, 应进行主成份回归。



- 先对自变量进行主成份分析。相关阵的特征根和特征向量如下：

自变量相关阵的特征根和特征向量

特征向量	$\alpha^*(1)$	$\alpha^*(2)$	$\alpha^*(3)$	$\alpha^*(4)$	$\alpha^*(5)$	$\alpha^*(6)$
x_1^*	0.4618	0.0578	-0.1491	-0.7929	-0.3379	-0.1352
x_2^*	0.4615	0.0532	-0.2777	-0.1216	0.1496	0.8158
x_3^*	0.3213	-0.5955	0.7283	-0.0077	-0.0092	0.1075
x_4^*	0.2015	0.7982	0.5616	-0.0773	-0.0243	0.0180
x_5^*	0.4623	-0.0456	-0.1960	0.5897	-0.5486	-0.3116
x_6^*	0.4649	0.0006	-0.1281	0.0524	-0.7495	-0.4504
特征根 λ_i^*	4.6034	1.1753	0.2035	0.0149	0.0025	0.0004

- 取前三个主成份即可。



□ 这三个主成份为：

$$z^{*(1)} = 0.4618x_1^* + 0.4615x_2^* + 0.3213x_3^* + 0.2015x_4^* \\ + 0.4623x_5^* + 0.4649x_6^* .$$

$$z^{*(2)} = 0.0578x_1^* + 0.0532x_2^* - 0.5955x_3^* + 0.7982x_4^* \\ - 0.0456x_5^* + 0.0006x_6^*$$

$$z^{*(3)} = -0.1491x_1^* - 0.2777x_2^* + 0.7283x_3^* + 0.5616x_4^* \\ - 0.1960x_5^* - 0.1281x_6^*$$

□ 第一主成份反映经济发展的综合水平；第二主成份是反映人力资源利用状况的综合指数；第三主成份是反映人力资源与经济发展的其他方面结构协调程度的一个综合指数。



- 将各年份每一变量的标准化数值代入这3个主成份，得到各年份的3个主成份得分。

年份	第一主成份 $z^{*(1)}$	第二主成份 $z^{*(2)}$	第三主成份 $z^{*(3)}$
1947	-3.4788	-0.7515	-0.3079
1948	-3.0105	-0.8490	-0.6422
1949	-2.3433	-1.5400	0.4934
1950	-2.0939	-1.2763	0.1113
1951	-1.4382	1.2358	0.0291
1952	-1.0103	1.9221	0.1612
1953	-0.7024	1.9106	-0.0671
1954	0.0325	0.5930	1.0390
1955	0.0995	0.6935	0.0976
1956	0.4494	0.5478	-0.2930
1957	0.9551	0.4294	-0.4452
1958	1.8171	-0.8632	0.6774
1959	1.9400	-0.3866	-0.2660
1960	2.3611	-0.4991	-0.3657
1961	3.0780	-0.9100	0.2020
1962	3.3448	-0.1767	-0.4239



- 用这3个主成份作自变量，用上表的得分值作为样本数据，建立回归模型，使用最小二乘法得到：

$$\hat{Y} = 2170.75 + 55.6130z_1^* + 108.7449z_2^* + 19.8311z_3^*$$

- 则由原自变量表示的回归方程的系数向量估计值为：

$$\begin{aligned} \hat{\beta}^* = A_{(n)}^+ \hat{y}_{(n)}^* &= \begin{bmatrix} 0.4618 & 0.0578 & -0.1491 \\ 0.4615 & 0.0532 & -0.2777 \\ 0.3213 & -0.5955 & 0.7283 \\ 0.2015 & 0.7982 & 0.5616 \\ 0.4623 & -0.0456 & -0.1960 \\ 0.4649 & 0.0006 & -0.1281 \end{bmatrix} \begin{bmatrix} 55.6130 \\ 108.7449 \\ 19.8311 \end{bmatrix} \\ &= \begin{bmatrix} 29.0164 \\ 25.9450 \\ -32.4459 \\ 109.1430 \\ 16.8700 \\ 23.3834 \end{bmatrix} \end{aligned}$$



□ 由此，可得出用原标准化自变量表示的回归方程为：

$$\begin{aligned}\hat{Y} = & 2170.75 + 29.0164X_1^* + 25.9450X_2^* - 32.4459X_3^* \\ & + 109.1430X_4^* + 16.8700X_5^* + 23.3834X_6^*\end{aligned}$$



疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回复问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

