第一章 主成分分析法

为了全面系统地分析某一个实际问题，经常会考虑众多指标，这些指标难免有一定的相关性，即所反映的信息在一定程度上会有重叠。为了简化问题，通常对这些变量加以“改造”，用较少的互不相关的综合变量来反映原变量所提供的绝大部分信息。主成分分析法（Principal Component Analysis，PCA）就是满足上述要求的一种统计方法。

主成分分析是以最少的信息丢失为前提，将原有变量通过线性组合的方式综合成少数几个新变量，利用新变量进行数据建模，大大减少了计算量。由于所选取的新变量之间互不相关，故能有效地解决变量信息重叠、多重共线性等诸多问题。下面给出两个案例。

1947 年，美国的统计学家斯通 (stone) 在关于国民经济的研究问题中，曾利用美国 1929 一 1938 年各年的数据，得到了 17 个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。在进行主成分分析后，竟以 97.4% 的精度，用三新变量就取代了原来的 17 个变量。根据经济学知识，斯通给这三个新变量分别命名为总收入、总收入变化率和经济发展或衰退的趋势，这大大简化了所研究的问题。

1961 年，英国统计学家 Scott 对英国 157 个城镇发展水平进行研究问题中，调查得到影响城镇发展水平的 57 个原始变量，由于计算和研究非常烦琐，他经过主成分分析发现，用原来变量的线性组合构造 5 个新变量，可以以 95% 的精确度概括原始数据的信息，达到了简化数据的目的。

1.1 主成分分析法的原理

假设有 m 个对象，每个对象都由 n 个指标 $x_1, x_2, \ldots, x_n$ 构成。降维处理后，它们的综合指标（新变量）为：$y_1, y_2, \ldots, y_p (p \leq n)$。

设变换为：

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{cases}$$

其中，系数满足：

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2 = 1,\ i = 1, 2, \ldots, n.$$

变换后的新变量满足如下关系：

1. y_i 与 y_j 相互无关；

2. y_1 为 $x_1, x_2, \ldots, x_n$ 的所有线性组合中方差最大者；

y_2 为与 y_1 不相关的 $x_1, x_2, \ldots, x_n$ 的所有线性组合中方差最大者；

...；

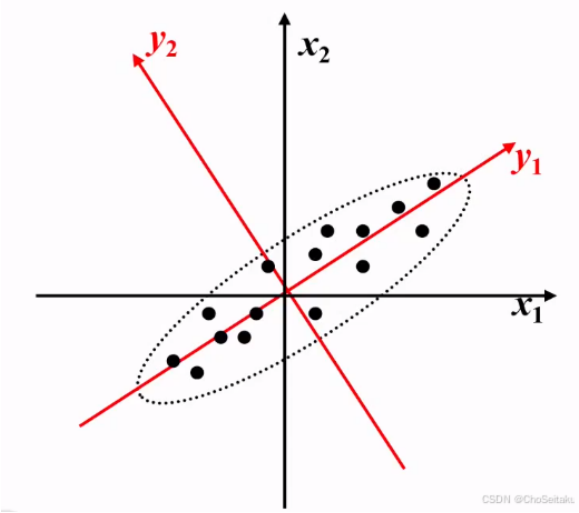
y_n 是与 $y_1, y_2, \ldots, y_{n-1}$ 不相关的 $x_1, x_2, \ldots, x_n$ 的所有线性组合中方差最大者。

称 y_1 为原变量 $x_1, x_2, \ldots, x_n$ 的第一主成分；称 y_2 为原变量 $x_1, x_2, \ldots, x_n$ 的第二主成分；...；称 y_n 为原变量 $x_1, x_2, \ldots, x_n$ 的第 n 主成分。

找主成分 y_i 就是要确定系数 a_{ij}。由矩阵的知识可知，它们分别是 $x_1, x_2, \ldots, x_n$ 的相关系数矩阵的 i 个较大的特征值所对应的特征向量。

1.2 几何意义

以二位平面为例。如果两个 2 维的向量线性无关，则这两向量在平面上相交。特别地，若两向量正交，则表示这两个向量互相垂直，如 $e_1 = (1,0)^\top$ 和 $e_2 = (0,1)^\top$ 为两垂直的向量，分别在 x 轴和 y 轴上。主成分分析是一种通过降维技术将多个变量化为少数几个主成分的统计分析方法。这些主成分通常表示为原始变量的某种线性组合，能够反映原始变量的绝大部分信息通过坐标变换后，可静思为更小维度的数据。如图所示，经过坐标变换后，原来的在二维数据可以近似为一维数据。



1.3 计算步骤

步骤 1. 计算相关系数：

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)}}$$

步骤 2. 构建相关系数矩阵：

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}$$

步骤 3. 求出矩阵 R 的特征值：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$$

及所对应的特征向量 $u_1, u_2, \dots, u_n$。并计算由特征向量组成的新指标变量：

$$y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{n1}x_n$$

$$y_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{n2}x_n$$

$$\dots$$

$$y_n = u_{1n}x_1 + u_{2n}x_2 + \cdots + u_{nn}x_n$$

步骤 4. 计算主成分的贡献率及累计贡献率。

主成分 y_j 的贡献率为：

$$b_j = \frac{\lambda_j}{\sum_{i=1}^{n}\lambda_i}$$

前 p 个主成分的累积贡献率为：

$$b = b_1 + b_2 + \cdots + b_p.$$

步骤 5. 当前 p 个主成分累计贡献率达到一定值，如 85%，就取前 p 个主成分 $y_1, y_2, \dots, y_p$ 作为新变量。

步骤 6. 构建主成分综合评价模型：

$$y = b_1y_1 + b_2y_2 + \cdots + b_py_p.$$

例题 1.1 利用主成分分析对 Excel 表 “Data_PCA” 给出的投资效益进行分析和排序。

符号说明：

x_1：	投资效果系数（无时滞）
x_2：	投资效果系数（时滞一年）
x_3：	全社会固定资产交付使用率
x_4：	建设项目投产率
x_5：	基建房屋竣工率
a_{ij}：	第 i 年的 x_j 指标值

计算过程如下。

(1) 对原始数据进行标准化处理。

将各指标值 a_{ij} 转换成标准化指标 $\widetilde{a}_{ij}$：

$$\widetilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j},\ i = 1, 2, \ldots, 17; j = 1, 2, \ldots, 5$$

其中，μ_j, s_j 分别为第 j 个指标的样本均值和样本标准差：

$$\mu_j = \frac{1}{17}\sum_{i=1}^{17} a_{ij},\ s_j = \sqrt{\frac{1}{17-1}\sum_{i=1}^{17}(a_{ij} - \mu_j)^2}$$

进而可得标准化指标变量：

$$\widetilde{x}_j = \frac{x_j - \mu_j}{s_j},\ j = 1, 2, \ldots, 5$$

(2) 计算相关系数矩阵 R。

相关系数矩阵 $R = (r_{ij})_{5\times 5}$，其中，r_{ij} 表示第 i 个指标与第 j 个指标的相关系数，计算公式为：

$$r_{ij} = \frac{\sum\limits_{k=1}^{17} \widetilde{a}_{ki} \cdot \widetilde{a}_{kj}}{17-1},\ i, j = 1, 2, \ldots, 5$$

显然，r_{ij} 满足：

$$r_{ii} = 1,\ r_{ij} = r_{ji}.$$

(3) 计算 R 的特征值和特征向量。

记 R 的特征值为：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_5 \geq 0$$

及对应的标准化特征向量：$u_1, u_2, \ldots, u_5$。其中，$u_j = \left(u_{1j}, u_{2j}, \ldots, u_{5j}\right)^T$。由特征向量组成 5 个新的指标变量：

$$\begin{aligned}
y_1 &= u_{11}\widetilde{x}_1 + u_{21}\widetilde{x}_2 + \cdots + u_{51}\widetilde{x}_5 \\
y_2 &= u_{12}\widetilde{x}_1 + u_{22}\widetilde{x}_2 + \cdots + u_{52}\widetilde{x}_5 \\
&\cdots \\
y_5 &= u_{15}\widetilde{x}_1 + u_{25}\widetilde{x}_2 + \cdots + u_{55}\widetilde{x}_5
\end{aligned}$$

其中：$y_1, y_2, \ldots, y_5$ 分别为第 $1, 2, \ldots, 5$ 主成分。

(4) 计算主成分 y_j 的贡献率：

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{5}\lambda_k}$$

前 p 个主成分的累积贡献率为：

$$b = b_1 + b_2 + \cdots + b_p.$$

(5) 求解当前 p 个主成分累计贡献率是否达到一定值，如 85%，就取前 p 个主成分 $y_1, y_2, \ldots, y_p$ 作为新变量。

(6) 构建主成分综合评价模型：

$$y = b_1 y_1 + b_2 y_2 + \cdots + b_p y_p.$$

可以求出，相关系数矩阵的 5 个特征值及其贡献率如下：

序号	特征值	贡献率	累计贡献率
1	3.1343	62.6866	62.6866
2	1.1683	23.3670	86.0536
3	0.3502	7.0036	93.0572
4	0.2258	4.5162	97.5734
5	0.1213	2.4266	100.0000

可以看出，前三个特征值的累计贡献率就达到 93% 以上，主成分分析效果很好。前三个特征矩阵对应得特征向量为：

	第 1 特征向量	第 2 特征向量	第 3 特征向量
$\widetilde{x}_1$	0.4905	-0.2934	0.5109
$\widetilde{x}_2$	0.5254	0.0490	0.4337
$\widetilde{x}_3$	-0.4871	-0.2812	0.3714
$\widetilde{x}_4$	0.0671	0.8981	0.1477
$\widetilde{x}_5$	-0.4916	0.1606	0.6255

选取前三个主成分进行综合评价：

$$y_1 = 0.4905\widetilde{x}_1 + 0.5254\widetilde{x}_2 - 0.4871\widetilde{x}_3 + 0.0671\widetilde{x}_5 - 0.4916\widetilde{x}_5$$

$$y_2 = -0.2934\widetilde{x}_1 + 0.0490\widetilde{x}_2 - 0.2812\widetilde{x}_3 + 0.8981\widetilde{x}_4 + 0.1606\widetilde{x}_5$$

$$y_3 = 0.5109\widetilde{x}_1 + 0.4337\widetilde{x}_2 + 0.3714\widetilde{x}_3 + 0.1477\widetilde{x}_4 + 0.6255\widetilde{x}_5$$

分别以三个主成分的贡献率为权重，构建主成分综合评价模型：

$$y = 0.6269y_1 + 0.2337y_2 + 0.0700y_3$$

代码如下。

```
clc
clear
A = xlsread('Data_PCA');
A = zscore(A);      % 数据标准化
R = corrcoef(A);      % 计算相关系数矩阵
[x, y, z] = pcacov(R); %x 为特征向量，y 为特征值，z 为贡献率
f = sign(sum(x));       % 构造元素为 ±1 的行向量
x = x.* f;                % 修改特征向量的正负号
n = 3;                      % 选取的主成分的个数
s = A * x(:, [1:n]);       % 计算各个主成分的得分
sz = s * z(1:n)/100;       % 计算综合得分
[szd, ind] = sort(sz, 'descend'); % 得分从高到低排列
```