主成分分析

概述：在数学建模中，经常会遇到研究**多个变量的问题，而且这些变量还有一定的相关性**，例如：在研究上海世博会影响力评价时，就要考虑多个变量。当变量数较多且期间存在复杂关系时，会显著增加分析问题的复杂性。如果有一种方法可以**将多个变量总和为几个少数有代表性的变量，使这些变量可以代表以前大部分变量的信息且有互不相关**，那么这样做有利于化简问题。这就是主成分分析法。

主成分分析（PCA）基本思想

PCA 是一种**降维**的方法，**将原来众多具有相关性的变量重组为一组相互无关的变量**。通常，在数学中的方法就是将原来的变量进行线性组合，作为新的综合变量，但是这种组合如果不加以限制，可能有很多种。那么如何选择？如果将选取的第一个线性组合记为 F1，自然希望他尽可能多地反映原来变量的信息。**这里"信息"用方差来测量，于是应该让 F1 的方差尽量大**。F1 在所有线性组合中方差最大，**称为第一主成分**。如果第一主成分不足以表示原来变量的信息，则需要选取第二主成分，即 F2。F2 中可以不包含 F1 中含有的变量。**用数学语言表达即为：cov（F1，F2）=0**。以此类推，可以选取第三、四…主成分。

PCA 步骤

1、 对原始数据进行标准化处理

假设：样本观测数据矩阵为：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

那么可以按照如下方法对原始数据进行标准化处理：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{\sqrt{\mathrm{Var}(x_j)}} \qquad (i=1,2,\cdots,n;j=1,2,\cdots,p)$$

其中，$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$；$\mathrm{Var}(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ $\qquad (j=1,2,\cdots,p)$。

2、 计算样本相关系数矩阵

（2）计算样本相关系数矩阵

为了方便，假定原始数据标准化后仍用 X 表示，则经标准化处理后数据的相关系数为

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

其中，$r_{ij}=\mathrm{cov}(x_i,x_j)=\frac{\sum_{k=1}^{k=n}(x_i-\bar{x}_i)(x_j-\bar{x}_j)}{n-1}$，$n>1$。

3、 计算相关系数矩阵 R 的特征值和相应的特征向量

特征值：λ1，λ2…

特征向量：ai=（ai1，ai2…）

4、 选择重要的主成分，并写出主成分的表达式

a) 贡献率：某个主成分的方差占全部方差和的比重。实际可以说是某个特征值的比重占全部特征值的比重。一般要求累计贡献率达到 85%以上，才能保证综合变量能包括原始变量的绝大部分信息。

b) 主成分实际含义解释:

要结合具体实际问题和专业，给出恰当的解释。

一般而言，**这个解释是根据主成分表达式的系数结合定性分析来决定的**。主成分是原来变量的线性组合，在这个线性组合中系数有大有小，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映射。例如：线性组合中个变量系数绝对值相差**较大**时，应认为这一主成分为这几个**绝对值较大的变量的综合**；当系数绝对值相差**不大**时，应认为这一主成分为这**几个变量的综合**。

5、 计算主成分得分

（5）计算主成分得分

根据标准化的原始数据，按照各个样品，分别代入主成分表达式，就可以得到各主成分下的各个样品的新数据，即为主成分得分。具体形式如下：

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{bmatrix}$$

其中，$F_{ij}=a_{j1}x_{i1}+a_{j2}x_{i2}+\cdots+a_{jp}x_{ip}$，$i=1,2,\cdots,n$；$j=1,2,\cdots,k$。

6、 依据主成分得分的数据，进一步对问题进行后续的分析和建模

Matlab 程序设计实例

1. 案例问题:企业综合实力排序

为了系统地分析某 IT 类企业的经济效益,选择了 8 个不同的利润指标,对 15 家企业进行了调研,并得到如表 3-3 所示的数据。请根据这些数据对这 15 家企业进行综合实力排序。

表 3-3 企业综合实力评价表

企业序号	净利润率/%	固定资产利润率/%	总产值利润率/%	销售收入利润率/%	产品成本利润率/%	物耗利润率/%	人均利润/(千元·人$^{-1}$)	流动资金利润率/%
1	40.4	24.7	7.2	6.1	8.3	8.7	2.442	20
2	25	12.7	11.2	11	12.9	20.2	3.542	9.1
3	13.2	3.3	3.9	4.3	4.4	5.5	0.578	3.6
4	22.3	6.7	5.6	3.7	6	7.4	0.176	7.3
5	34.3	11.8	7.1	7.1	8	8.9	1.726	27.5
6	35.6	12.5	16.4	16.7	22.8	29.3	3.017	26.6
7	22	7.8	9.9	10.2	12.6	17.6	0.847	10.6
8	48.4	13.4	10.9	9.9	10.9	13.9	1.772	17.8
9	40.6	19.1	19.8	19	29.7	39.6	2.449	35.8
10	24.8	8	9.8	8.9	11.9	16.2	0.789	13.7
11	12.5	9.7	4.2	4.2	4.6	6.5	0.874	3.9
12	1.8	0.6	0.7	0.7	0.8	1.1	0.056	1
13	32.3	13.9	9.4	8.3	9.8	13.3	2.126	17.1
14	38.5	9.1	11.3	9.5	12.2	16.4	1.327	11.6
15	26.2	10.1	5.6	15.6	7.7	30.1	0.126	25.9

Matlab 程序:(其中排序和灵活运用矩阵方法要掌握)
(其中相关系数计算另加协方差知识要掌握)

```
clear all;
clc;
%从 excel 文档中读取数据
A=xlsread('E:\顾子涵专用文件夹\学习\matlab 学习\matlab 与数学模型\PCA 数
据.xlsx',1,'B2:I16');
a=size(A,1);%a 行
b=size(A,2);%b 列
%进行标准化处理
for j=1:b
    %计算每一列的平均值
    ave_x(j)=mean(A(:,j));
    %计算每一列的方差
    var_x(j)=var(A(:,j));
end
%进行标准化
for i=1:a
    for j=1:b
        SA(i,j)=(A(i,j)-ave_x(j))./sqrt(var_x(j));
```

```
    end
end
%计算样本相关系数矩阵
CM=corrcoef(SA);
%计算相关系数矩阵的特征值和特征向量
[V,D]=eig(CM);
%计算贡献率
DS(:,1)=diag(D);
for i=1:size(D)
    DS(i,2)=DS(i,1)./sum(diag(D));
end
%降序排序，取出前 85%
DS(:,3)=sort(DS(:,2),'descend');
%说明：DS 第一列为特征值升序排序
%     DS 第二列为贡献率升序排序
%      DS 第三列为贡献率降序排序
temp_sum=0;
T=0.9;
for i=1:b
    temp_sum=temp_sum+DS(i,3);
    if temp_sum>=T
        num=i;
        break;
    end
end
%提取主成分对应特征向量
for j=1:num
    PV(:,j)=V(:,b+1-j);
end
%说明：PV 第一列为最大特征值对应的特征向量
%第二列为第二大特征值对应的特征向量，以此类推
%下面计算主成分得分
new_score=SA*PV;
%计算总得分
for i=1:size(new_score,1)
    total_score(i,1)=sum(new_score(i,:));
    total_score(i,2)=i;
end
%得分数据
SCORE=[new_score,total_score];
%按照第四列降序排序，其他的随着第四列的元素移动
SCORE=sortrows(SCORE,4,"descend");


%输出结果
```

```
disp('特征值及其贡献率、累计贡献率');
DS=DS
disp('信息保留率 T 对应的主成分数与特征向量');
num=num
PV=PV
disp('主成分得分，第一二三列为各主成分得分，第四列为总得分，第五列为企业编号
');
SCORE=SCORE


%写入 excel
xlswrite('E:\顾子涵专用文件夹\学习\matlab 学习\matlab 与数学模型\PCA 分析结
果.xlsx',SCORE,1,'A1');
```