# 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。

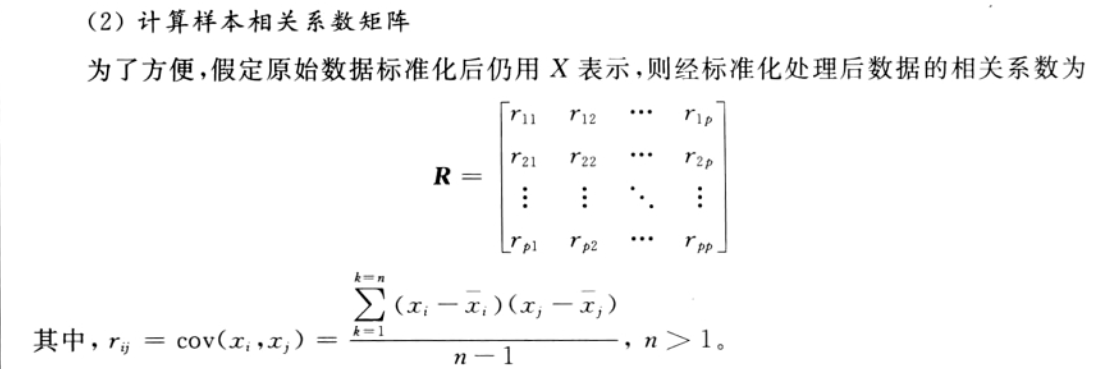
## PCA步骤

1. 对原始数据进行标准化处理

假设：样本观测数据矩阵为：



1. 计算样本相关系数矩阵



1. 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…

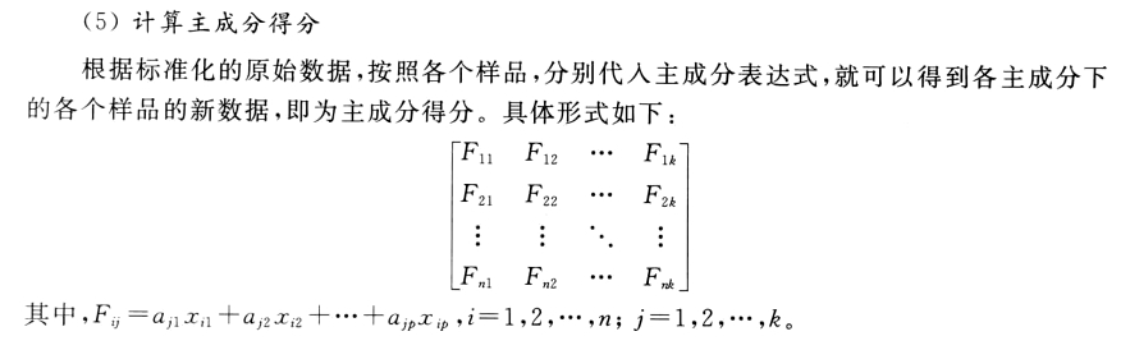
特征向量：ai=（ai1，ai2…）

1. 选择重要的主成分，并写出主成分的表达式
   1. 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。
   2. 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**。

1. 计算主成分得分



1. 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



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');