主成分分析和因子分析法

谭忠

目录
CONTENTS

Part 1

源头问题与当今应用

案例引入：假定你是一个公司的财务经理，掌握公司的所有重要的数据，比如固定资产、资金流动、每一笔借贷的数额和期限、各种税费、工资支出等等. 如果让你向上级介绍公司状况，你能把这些指标和数字都原封不动的摆出去吗？当然不能. 你必须要把各个方面进行高度概括，用一两个指标简单地把情况说清楚.

其实，每个人都会遇到有很多变量的数据. 比如全国或各个地区的经济和社会变量的数据；各个学校的研究、教学及各类学生人数及科研经费等各种变量的数据等等. 这些数据的共同特点是变量很多，在如此多的变量中，有很多是相关的，人们希望能够找出少数“代表”来对它们进行描述.

主成分分析正是其中一种方法.

Part 2 主成分分析

一、定义

主成分分析（Principal Component Analysis，PCA），是一种统计方法。在处理实际问题中，多个变量之间可能存在一定的相关性，当变量的个数较多且变量之间存在复杂的关系时，增加了问题分析的难度。

主成分分析是一种数学降维的方法，该方法主要将原来众多具有一定相关性的变量，重新组合成为一种新的相互无关的综合变量。

二、主成分分析的基本思想

设研究某个实际问题要考虑 p 个随机变量 $X_1, X_2, \cdots, X_p$，它们可以构成 p 维随机变量 $X=(X_1,\ X_2, \cdots, X_p)^{\mathrm{T}}$. 为了避免遗漏重要的信息，我们要考虑尽可能多的与研究问题有关的变量，此时，会产生以下两个问题:

(1)随机变量 $X_1, X_2, \cdots, X_p$ 的个数 p 比较大，将增大计算量和增加分析问题的复杂性;

(2)随机变量 $X_1, X_2, \cdots, X_p$ 之间存在一定的相关性，因而它们的观测样本所反映的信息在于一定程度上有重叠.

为了解决这些问题，人们希望在定量研究中利用原始变量的线性组合形成几个新变量，即对 X 做线性变换$Z = A^T X$(A 必须满足一定的条件)，Z 的各分量(称为主成分)在保留原始变量主要信息的前提下起到变量降维与简化问题的作用.

当然，这一分析过程应使得：

(1)每一个新变量都是各原始变量的线性组合；

(2)新变量的数目大大少于原始变量的数目；

(3)新变量保留了原始变量所包含的绝大部分信息；

(4)各新变量之间互不相关.

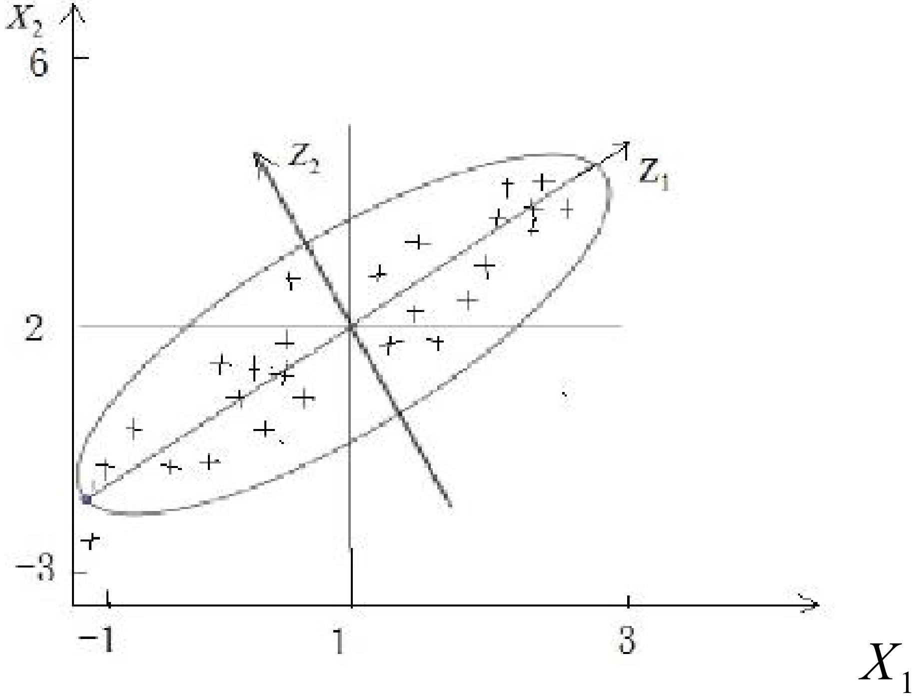
综上所述，主成分分析的基本思想是构造原始变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少量几个新变量并使它们含有足够多的原始变量带有的信息，

从而使得用这几个新变量代替原始变量分析问题和解决问题成为可能.

三、主成分的几何意义

从代数学观点看主成分就是 p 个变量的一些特殊的线性组合，而从几何上看这些线性组合正是把 $X_1, X_2, \cdots, X_p$ 构成的坐标系旋转产生的新坐标系，新坐标轴的方向就是原始数据变差最大的方向（最大的样本方差）.

设有 n 个样品，每个样品有 p 个变量(指标) $X_1, X_2, \cdots, X_p$，它们的综合指标记为 $Z_1, Z_2, \cdots, Z_p$. 当 $p = 2$ 时原变量为 X_1, X_2. 设 (X_1, X_2)服从二元正态分布，则样品点 $X_{(i)} = (x_{i1}, x_{i2})(i = 1, 2, \cdots, n)$ 的散布图在一个椭圆内分布着（当 (X_1, X_2) 的相关性越强，这个椭圆越扁）.



由图可看出，这 n 个观测无论沿 X_1 轴的方向还是 X_2 轴方向，均有较大的离散性，显然，若只考虑 X_1,X_2 中任何一个，原始数据均会有较大的损失.

现考虑若取椭圆的长轴为坐标轴 Z_1，椭圆的短轴为 Z_2，这相当于在平面上作一个坐标变换，即按逆时针方向旋转一个角度 α. 根据旋转变换公式，新老坐标之间有关系：

$$\begin{cases} Z_1 = X_1\cos\alpha + X_2\sin\alpha \\ Z_2 = -X_1\sin\alpha + X_2\cos\alpha \end{cases}$$

从图上可以看出二维平面上 n 个点的波动 (用二个变量的方差和表示)大部分可以归结为在 Z_1 方向的波动，而在 Z_2 方向上的波动很小，可以忽略.

这样一来，二维问题可以降为一维了，只取第一个综合变量即可，而 Z_1 是椭圆的长轴，容易得到.

一般情况，p 个变量组成 p 维空间，n 个样品点就是 p 维空间的 n 个点. 对于 p 元正态分布变量来说，找主成分的问题就是找 p 维空间中椭球的主轴问题.

四、主成分分析的方法原理

设

$$X = (X_1, X_2, \cdots X_p)^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

是 p 维随机向量，均值向量 $E(X) = \mu$,

协方差阵

$$D(X)=\Sigma=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{np} \end{pmatrix}$$

其中若是求样本协方差阵

$$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j}) \qquad (i,j=1,2,\cdots,p)$$

若是求总体协方差阵

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j}) \qquad i, j = 1,2,\cdots,p$$

考虑它的线性变换：

$$\begin{cases} Z_1 = a_1'X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ Z_2 = a_2'X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ \quad\vdots \\ Z_p = a_p'X = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p \end{cases} \qquad (1)$$

易见

$$\begin{aligned} Var(Z_i) &= a_i'\Sigma a_i \\ Cov(Z_i, Z_j) &= a_i'\Sigma a_j \end{aligned} \tag{2}$$

假如我们希望用 Z_1 来代替原来的 p 个变量 $X_1, X_2, \cdots, X_p$，这就要求 Z_1 尽可能多地反映原来 p 个变量的信息，这里所说的“信息”最经典的方法是用 Z_1 的方差来表达，$Var(Z_1)$ 越大，表示包含的信息越多. 由(2)式可以看出，对 a_1 必须有某种限制，否则可使 $Var(Z_1) \to \infty$.

常用的限制是：$a_1'a_1 = 1$. 若存在满足以上约束的 a_1 使 $Var(Z_1)$ 最大，Z_1 就称为**第一主成分**. 如果第一主成分不足以表达原来 p 个变量的绝大部分信息，考虑 X 的第二个线性组合 Z_2.

为了有效地代表原始变量的信息，Z_1 已反映的信息不希望在 Z_2 中出现，用统计语言来讲，就是要求

$$Cov(Z_1, Z_2) = a_1'\Sigma a_2 = 0 \tag{3}$$

于是求 Z_2 就是在约束 $a_2'a_2 = 1$和上式下，求 a_2 使 $Var(Z_2)$ 最大，所求 Z_2 称为**第二主成分**. 类似地可求得**第三主成分**，**第四主成分**等等.

定义 1 设 $X=(X_1,X_2,\cdots,X_p)^T$ 为 p 维随机向量，称 $Z_i=a_i'X$ 为 X 的第 i 主成分 $(i = 1, 2, \cdots, p)$，如果:

（1）$a_i'a_i=1$;

（2）当 $i>1$ 时，$a_i'\Sigma a_j=0\ (j=1,\cdots,i-1)$;

（3）$Var(Z_i)=\max Var(a'X)$，满足 $a_i'a_i=1$，$a_i'\Sigma a_j=0$ $(j=1,\cdots,i-1)$.

五、主成分求法

由上述定义可知，求第 i 主成分 $Z_i = a_i'X$ 的问题，即为求$a_i = (a_{1i}, a_{2i}, \cdots, a_{pi})'$，使得在 $a_i'a_i = 1$ 下， $Var(Z_i)$ 达最大. 这是条件极值问题，用拉格朗日乘子法求解.

令

$$\varphi(a_i) = Var(a_i'X) - \lambda(a_i'a_i - 1)$$

考虑

$$\begin{cases} \frac{\partial \varphi}{\partial a_i} = 2(\Sigma - \lambda I)a_i = 0 \\ \frac{\partial \varphi}{\partial \lambda} = a_i' a_i - 1 = 0 \end{cases} \tag{4}$$

因 $a_i \neq 0$，故 $|\Sigma - \lambda I| = 0$ 求解上述方程组 (4)，其实就是求 Σ 的特征值和特征向量问题.

由方程组 (4) 的 $(\Sigma - \lambda I)a_i = 0$，两边左乘 a_i' 得到

$$a_i' \Sigma a_i = \lambda \tag{5}$$

由于 X 的协差阵 Σ 为非负定的，其特征值均大于零，不妨设 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. 由 (5) 式知道 Z_i方差为 λ，那么方差最大值为 λ_1，对应的单位化正交特征向量为 a_1. 即可得出 Z_1 表达式. 同理可得 $Z_2, Z_3, \cdots, Z_P$.

主成分分析的目的之一是为了简化数据结构，故在实际应用中一般绝不用 p 个主成分，而选用 $m(m < p)$ 个主成分.

m 取多大，这是一个很实际的问题. 为此，我们引进贡献率的概念.

定义 2 我们称 $\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 为主成分 Z_k 的**贡献率**；又称 $\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 为主成分$Z_1,Z_2,\cdots,Z_m(m<p)$的**累计贡献率**.

累计贡献率的大小表达 m 个主成分提取了 $X_1,X_2,\cdots,X_p$ 的多少信息. 通常取 m，使累积贡献率达到85%以上.

但它没有表达某个变量被提取了多少信息，为此又引入第 i 个主成分 Z_i 与原始变量 X_j 的相关系数 $\rho(X_j,Z_i)=\frac{a_{ji}\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}$， $i,j=1,2,\cdots,p$. 将前 m 个主成分 $Z_1,Z_2,\cdots,Z_m$ 对原始变量 X_i 的贡献率 $v_i^{(m)}$ 定义为 X_i 与 $Z_1,Z_2,\cdots,Z_m$ 的相关系数的平方，它等于 $v_i^{(m)}=\sum_{k=1}^{m}\lambda_k a_{ik}^2/\sigma_{ii}$.

例： 已知随机向量 $X=(X_1,X_2,X_3)^T$ 的协方差阵，试求 X 的主成分及主成分对变量 X_i 的贡献率.

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

解: Σ的特征值为$\lambda_1=3+\sqrt{8},\lambda_2=2,\lambda_3=3-\sqrt{8}$. 相应单位正交

特征向量为 $a_1=\begin{pmatrix}0.383\\ -0.924\\ 0.000\end{pmatrix},a_2=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix},a_3=\begin{pmatrix}0.924\\ 0.383\\ 0.000\end{pmatrix}$,

故主成分为

$$\begin{cases} Z_1 = 0.383X_1 - 0.924X_2 \\ Z_2 = X_3 \\ Z_3 = 0.924X_1 + 0.383X_2 \end{cases}$$

下表列出 m 个主成分对变量 X_i 的贡献率 $v_i^{(m)}$

i	$\rho(X_i, Z_1)$	$\rho(X_i, Z_2)$	${v_i}^{(1)}$	${v_i}^{(2)}$
1	0.925	0	0.856	0.856
2	-0.998	0	0.996	0.996
3	0.000	1	0.000	1.000

Z_1的贡献率： $\frac{\lambda_1}{\sum_{i=1}^{3}\lambda_3}=\frac{3+\sqrt{8}}{8}=72.85\%$

Z_1,Z_2 的累计贡献率： $\frac{\sum_{k=1}^{2}\lambda_k}{\sum_{i=1}^{p}\lambda_i}=\frac{3+\sqrt{8}+2}{8}=97.85\%$

由上可见，当 $m=1$ 时，Z_1的贡献率已达 72.85%，比较理想了.
而此时 Z_1 对 X_3 的贡献率 $\nu_3^{(1)}=0$，这是因为在 Z_1 中没有包含 X_3 的任何信息.

欲使累积贡献率达到 85%以上需取 $m=2$，这时 Z_1,Z_2 的累计贡献率为 97.85%，且 Z_1,Z_2 对 X_i 的贡献率$\nu_i^{(2)}(i=1,2,3)$也较高.

六、标准化的主成分

在实际问题中，不同的变量往往有不同的量纲，而通过协方差阵 Σ 来求主成分总是优先考虑方差大的变量，有时会造成很不合理的结果，为了消除由于量纲的不同可能带来的一些不合理的影响，常采用将变量标准化.

若记 $E(X_i)=\mu_i$，$Var(X_i)=\sigma_i^2$，即令 $X_i^*=\frac{X_i-E(X_i)}{\sqrt{Var(X_i)}}=\frac{X_i-\mu_i}{\sigma_i}$ $(i=1,2,\cdots,p)$.

这时标准化后的随机向量 $X^*=\left(X_1^*,X_2^*,\cdots,X_p^*\right)$，$A$ 的协方差阵 Σ^* 就是原随机向量 X的相关阵 R，从相关阵 R 出发求主成分，记主成分向量为 $Z^*=(Z_1^*,Z_2^*,\cdots,Z_P^*)$，方法与前面一致.

七、利用主成分进行综合评价

对主成分进行加权综合，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 $Z_1, Z_2, \cdots, Z_m$ 是利用主成分分析所提取出来的前 m 个主成分，它们的特征根分别是 $\lambda_1, \lambda_2, \cdots, \lambda_p$，每个主成分 Z_i 的方差贡献率为 $\alpha_i = \frac{\lambda_i}{\sum_{i=k}^{p} \lambda_k}$，则可构造综合评价函数

$$Y = \alpha_1 Y_1 + \alpha_2 Y_2 + \cdots + \alpha_m Y_m$$

其中 Y_i 为第 i 个主成分的得分. 把 m 个主成分得分带入 Y 函数后，即可得到每个样本的综合评价函数得分，根据每个样本的综合评价函数得分的大小就可以对其进行综合排名.

八、主成分分析案例一（经管社科类）：地区居民家庭收支情况

下表是我国 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 R 软件实现主成分分析.

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	8545.1	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
内蒙古	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

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
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
河南	3	1.5	1.9	4415.8	2934
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

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
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	4.31.8	3113.2
青海	3	1.3	2.3	3971.8	3070.3
宁夏	2.9	1.3	2.2	4.78.3	3133.7
新疆	3	1.5	2.1	4.18.4	3015.1

（1）导入数据，求出相关系数矩阵.

```
> test<-read.csv("C:\\Users\\Administrator\\Desktop\\主1.csv",header=T)
> a=cor(test[,2:6])
> a
            x1         x2         x3          x4          x5
x1  1.00000000  0.5301188  0.2415472 -0.05059688  0.03818932
x2  0.53011882  1.0000000 -0.6346865  0.31947029  0.35702655
x3  0.24154721 -0.6346865  1.0000000 -0.41956430 -0.39797614
x4 -0.05059688  0.3194703 -0.4195643  1.00000000  0.96767839
x5  0.03818932  0.3570265 -0.3979761  0.96767839  1.00000000
> 
```

（2）求出相关系数矩阵的特征值和特征向量.

```
> eigen(a)
$values
[1] 2.57600426 1.38949217 0.96112276 0.04661280 0.02676801

$vectors
            [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  0.07530817  0.7873023 -0.3427597  0.4668378 -0.1976412
[2,]  0.44110005  0.5192346  0.3304279 -0.6276242  0.1809184
[3,] -0.44961650  0.1062337 -0.6836021 -0.5441940  0.1519883
[4,]  0.54395888 -0.2539892 -0.3731209 -0.2182263 -0.6728717
[5,]  0.54928423 -0.1864255 -0.4084142  0.2106667  0.6725695
```

继续求解方差贡献率，累计方差贡献率，并据此寻找主成份以及求出载荷矩阵.

```
> test.pr<-princomp(test[,2:6],cor=TRUE)
> summary(test.pr,loadings=TRUE)
Importance of components:
                          Comp.1    Comp.2    Comp.3      Comp.4      Comp.5
Standard deviation     1.6049935 1.1787672 0.9803687 0.215899986 0.163609320
Proportion of Variance 0.5152009 0.2778984 0.1922246 0.009322561 0.005353602
Cumulative Proportion  0.5152009 0.7930993 0.9853238 0.994646398 1.000000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
x1         0.787 -0.343  0.467 -0.198
x2  0.441  0.519  0.330 -0.628  0.181
x3 -0.450  0.106 -0.684 -0.544  0.152
x4  0.544 -0.254 -0.373 -0.218 -0.673
x5  0.549 -0.186 -0.408  0.211  0.673
>
```

XIAMEN UNIVERSITY

(3)主成分表达式

根据载荷矩阵可以得到主成分的表达式

pc1=0.441∗ x2−0.450∗ x3+0.544∗ x4+0.549∗ x5

$$pc2 = 0.787 * x1 + 0.519 * x2 + 0.106 * x3 - 0.254 * x4 - 0.186 * x5$$

$$pc3 = -0.343 * x1 + 0.330 * x2 - 0.684 * x3 - 0.373 * x4 - 0.408 * x5$$

(4)计算主成分得分以及综合得分，其中综合得分公式为

$$y = 0.51520 * pc1 + 0.27790 * pc2 + 0.19222 * pc3$$

便可在原数据的基础上，得到以 Y 为变量名的综合得分如下：

> zu3

	zhongwen	x1	x2	x3	x4	x5	pc1	pc2	pc3	y
1	北京	-0.5976143	0.94151885	-1.40352942	2.31217218	2.3083851	3.57192315	-1.1470997	-0.32856430	1.45831918
2	天津	-0.5976143	-0.77536847	0.02300868	0.62463748	0.7113280	0.37803045	-1.1612647	-0.58983942	-0.24133310
3	河北	-0.5976143	0.08307519	-0.69026037	-0.39588897	-0.4457447	-0.11282411	-0.3169097	1.03406503	0.05257179
4	山西	0.0000000	0.08307519	0.02300868	-0.53770288	-0.6810481	-0.64012350	0.3088064	0.49010767	-0.14976583
5	内蒙古	-0.5976143	0.08307519	-0.69026037	-0.31379955	-0.2198347	0.05585714	-0.3797797	0.91127439	0.09840198
6	辽宁	-0.5976143	-0.77536847	0.02300868	-0.44779845	-0.2452906	-0.73055831	-0.7109349	0.20047958	-0.53541627
7	吉林	0.0000000	0.08307519	-0.69026037	-0.57963710	-0.4887662	-0.23640192	0.2080868	0.91517417	0.11194782
8	黑龙江	-1.1952286	-1.63381212	1.44954678	-0.82831973	-0.8765237	-2.30462462	-1.2615148	-0.45409967	-1.62520462
9	上海	0.0000000	0.94151885	-0.69026037	2.82770895	2.6596964	3.72427395	-0.7974609	-1.35705224	1.43627897
10	江苏	-0.5976143	-0.77536847	0.73627773	0.74027427	0.4041655	-0.04866642	-1.0578977	-0.99552567	-0.51042265
11	浙江	-1.1952286	-0.77536847	-0.69026037	2.36052361	2.1939741	2.45729631	-2.4238809	-1.14938684	0.37146741
12	安徽	0.0000000	0.94151885	-0.69026037	-0.56826403	-0.6297604	0.07095289	0.6769552	1.25174404	0.46529101
13	福建	0.5976143	0.94151885	-0.69026037	0.80425572	0.4701068	1.42143072	0.5940823	0.08606665	0.91396032
14	江西	-0.5976143	0.08307519	-0.69026037	-0.53134921	-0.7953651	-0.37845610	-0.2174734	1.22723684	-0.01951698

对得分进行排名，结果如下：

```
> Y=arrange(zu3,-y)
> Y
   zhongwen         x1          x2          x3          x4         x5         pc1        pc2         pc3           y
1      广东  1.7928429  1.79996251 -0.69026037  1.66644872  2.0194457  3.11962443  1.4730854 -0.99433862  1.82546918
2      北京 -0.5976143  0.94151885 -1.40352942  2.31217218  2.3083851  3.57192315 -1.1470997 -0.32856430  1.45831918
3      上海  0.0000000  0.94151885 -0.69026037  2.82770895  2.6596964  3.72427395 -0.7974609 -1.35705224  1.43627897
4      山东 -0.5976143  1.79996251 -2.11679847  0.09639337 -0.2575959  1.65736061  0.2629064  2.31600389  1.37211614
5      西藏  2.3904572  1.79996251  0.02300868 -0.34131094  0.6343025  0.94598850  2.7866220 -0.37316358  1.19004602
6      福建  0.5976143  0.94151885 -0.69026037  0.80425572  0.4701068  1.42143072  0.5940823  0.08606665  0.91396032
7      重庆  0.5976143  0.94151885 -0.69026037  0.04943975  0.3719463  0.95692072  0.8040634  0.40766249  0.79481567
8      安徽  0.0000000  0.94151885 -0.69026037 -0.56826403 -0.6297604  0.07095289  0.6769552  1.25174404  0.46529101
9      浙江 -1.1952286 -0.77536847 -0.69026037  2.36052361  2.1939741  2.45729631 -2.4238809 -1.14938684  0.37146741
10     吉林  0.0000000  0.08307519 -0.69026037 -0.57963710 -0.4887662 -0.23640192  0.2080868  0.91517417  0.11194782
11   内蒙古 -0.5976143  0.08307519 -0.69026037 -0.31379955 -0.2198347  0.05585714 -0.3797797  0.91127439  0.09840198
12     河南  0.0000000  0.08307519 -0.69026037 -0.50205879 -0.8057917 -0.36824632  0.2473486  1.01558387  0.07423321

22     新疆  0.0000000  0.08307519  0.73627773 -0.75455363 -0.7296117 -1.10572281  0.4485259  0.10293092 -0.42523767
23     江苏 -0.5976143 -0.77536847  0.73627773  0.74027427  0.4041655 -0.04866642 -1.0578977 -0.99552567 -0.51042265
24     辽宁 -0.5976143 -0.77536847  0.02300868 -0.44779845 -0.2452906 -0.73055831 -0.7109349  0.20047958 -0.53541627
25     四川 -0.5976143 -0.77536847  0.02300868 -0.55434949 -0.4369149 -0.89372383 -0.6482288  0.31840583 -0.57938534
26     贵州  0.5976143 -0.77536847  0.73627773 -0.65353028 -0.6815177 -1.40293618  0.4387107 -0.44264123 -0.68595953
27     湖北 -0.5976143 -0.77536847  0.73627773 -0.12407897 -0.1647896 -0.83123094 -0.7325263 -0.44098822 -0.71658600
28     云南  0.0000000 -1.63381212  1.44954678 -0.37244393 -0.3535959 -1.76954086 -0.5339269 -1.24745927 -1.29983237
29     宁夏 -0.5976143 -1.63381212  1.44954678 -0.71649515 -0.6182066 -2.10197596 -0.8676428 -0.80618532 -1.47902089
30   黑龙江 -1.1952286 -1.63381212  1.44954678 -0.82831973 -0.8765237 -2.30462462 -1.2615148 -0.45409967 -1.62520462
31     青海  0.0000000 -1.63381212  2.16281582 -0.78416173 -0.6777604 -2.49245271 -0.2934495 -1.44950545 -1.64428519
>
```

结果分析：

综合主成分得分靠前的依次为广东、北京、上海、山东、西藏、福建、重庆、安徽、浙江、吉林，这是城镇居民家庭收支水平较高的地区，其中广东，北京，上海三个地区得分最高. 得分居于最后四位的是云南，宁夏，黑龙江，青海.

由上分析，进行主成分分析的主要步骤如下：

(1)指标数据标准化；

(2)求指标之间的相关阵 R；

(3)确定主成分个数 m；

(4)求主成分 Z_i 表达式；

(5)利用主成分进行综合评价.

MATLAB的操作可参考：谢中华. MATLAB统计分析与应用:40个案例分析[M]. 北京航空航天大学出版社, 2010.

主成分分析案例三（日常生活类)：在某中学随机抽取某年级 30 名学生，测量其身高 X_1、体重 X_2、胸围 X_3 和坐高 X_4，见下表. 通过这个例子，介绍如何利用 Matlab 软件实现主成分分析.

序号	X_1	X_2	X_3	X_4	序号	X_1	X_2	X_3	X_4
1	148	41	72	78	2	139	34	71	76
3	160	49	77	86	4	149	36	67	79
5	159	45	80	86	6	142	31	66	67
7	153	43	76	83	8	150	43	77	68
9	151	42	77	80	10	139	31	68	74
11	140	29	64	74	12	161	47	78	84
13	158	49	78	83	14	140	33	67	77
15	137	31	66	73	16	152	35	73	79
17	149	47	82	79	18	145	35	70	77
19	160	47	74	87	20	156	44	78	85
21	151	42	73	82	22	147	38	73	78
23	157	39	68	80	24	147	30	65	75
25	157	39	68	80	26	151	36	74	80
27	144	36	68	76	28	141	30	67	76
29	139	32	68	73	30	148	38	70	78

问题分析：

(1)读取数据，并进行标准化变换

$$[X,\ \text{textdata}\ ] = x\ \text{sread}\left({}'\,\text{examp}11l_0 3.xls'\right) \qquad XZ = 2\,\text{score}(X)$$

(2)调用 princomp 函数根据标准化后原始样本观测数据作主成分分析，返回主成分表达式的系数矩阵 COEFF，主成分得分数据 SCORE，样本相关系数矩阵的特征值向量 latent 和每个观测的霍特林 T2 统计量

```
[COEFF,SCORE,latent,tsquare] = princomp(XZ)

COEFF =

   -0.1975    0.9586   -0.2003    0.0440
    0.5372    0.2740    0.6961   -0.3897
    0.5974    0.0771   -0.0453    0.7969
    0.5617   -0.0069   -0.6880   -0.4595
```

(3)为了直观，定义元胞数组 result1，用来存放特征值、贡献率和累积贡献率等数据

```
>> explained = 100*latent/sum(latent);  %计算贡献率
[m, n] = size(X);  %求X的行数和列数
result1 = cell(n+1, 4);  %定义一个n+1行，4列的元胞数组
result1(1,:) = {'特征值', '差值', '贡献率', '累积贡献率'};
result1(2:end,1) = num2cell(latent);  %存放特征值
result1(2:end-1,2) = num2cell(-diff(latent));  %存放特征值之间的差值
result1(2:end,3:4) = num2cell([explained, cumsum(explained)])  %存放(累积)贡献率

result1 =

    '特征值'      '差值'       '贡献率'      '累积贡献率'
    [2.6591]    [1.6974]    [66.4770]    [   66.4770]
    [0.9617]    [0.6525]    [24.0415]    [   90.5185]
    [0.3092]    [0.2391]    [ 7.7290]    [   98.2474]
    [0.0701]          []    [ 1.7526]    [       100]
```

(4)调用 pcares 函数重建观测数据

```
% 通过循环计算E1(m)和E2(m)
for i = 1 : 4
    residuals = pcares(X, i);  % 返回残差
    Rate = residuals./X;  %计算相对误差
    E1(i) = sqrt(mean(residuals(:).^2));  %计算残差的均方根
    E2(i) = sqrt(mean(Rate(:).^2));  %计算相对误差的均方根
end
E1    %查看残差的均方根
E2    %查看相对误差的均方根

E1 =

    4.5232    1.8359    0.7747    0.0000


E2 =

    1.2789    0.0392    0.0171    0.0000
```

从以上结果可以看出，当只使用第一个主成分得分重建观测矩阵时，残差的均方根等于4.5232，相对误差的均方根为 1.2789. 当只使用前两个主成分得分重建观测矩阵时，E1 和 E2 的值都有所下降，当使用全部四个主成分得分重建观测矩阵时，E1 和 E2 均为 0，此时没有信息损失.

故主成分为

$$\begin{cases} Z_1 = -0.1975X_1^* + 0.5372X_2^* + 0.5974X_3^* + 0.5671X_4^* \\ Z_2 = 0.9586X_1^* + 0.2740X_2^* + 0.0771X_3^* - 0.0069X_4^* \end{cases}$$

Part3

因子分析

一、因子分析法介绍

在多变量分析中，某些变量间往往存在相关性. 是什么原因使变量间有关联呢？是否存在不能直接观测到的、但影响可观测变量变化的公共因子呢？这里我们就引进了因子分析的方法.

因子分析法目标是通过发掘隐藏在数据下的一组较少的、更为基本的无法观测的变量，来解释一组可观测变量的相关性.

例：随着年龄的增长，儿童的身高、体重会随着变化，具有一定的相关性，身高和体重之间为何会有相关性呢？因为存在着一个同时支配或影响着身高与体重的生长因子. 这就是一个潜在的无法观测的变量.

因子分析主要用于：

1、减少分析变量个数；

2、通过对变量间相关关系探测，将原始变量进行分类. 即将相关性高的变量分为一组，用共性因子代替该组变量.

二、因子分析的数学模型

模型形式为：

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + \varepsilon_i$$

式中的X_i $(i = 1, \dots, p)$ 是第 i 个可观测变量，F_j $(j = 1, \dots, m)$ 称为 m 个**公共因子**, ε_i 称为 X_i 的**特殊因子**, a_{ij}称为**因子载荷**，是第 i 个变量在第 j 个因子上的负荷.

该模型可用矩阵表示为：

$$X = AF + \varepsilon$$

该模型可用矩阵表示为：

$$X = AF + \varepsilon$$

其中

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}, F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

矩阵A称为**因子载荷矩阵。**

且满足以下四个条件：

（1）$m \le p$；

（2）$cov(F, \varepsilon) = 0$，即公共因子与特殊因子是不相关的；

（3）$D_F = D(F) = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix} = I_m$，即各个公共因子不相关且方差为 1；

（4）$D_\varepsilon = D(\varepsilon) = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_p^2 \end{bmatrix}$，即各个特殊因子不相关，方差不要求相等.

三、因子载荷阵的统计意义

1、因子载荷

对于因子模型 $X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + \varepsilon_i, (i = 1, \dots, p)$

我们可以得到，X_i 与 F_j 的协方差为

$$cov(X_i, F_j) = cov(\sum_{k=1}^{m} a_{ik}F_k + \varepsilon_i, F_j)$$
$$= cov(\sum_{k=1}^{m} a_{ik}F_k, F_j) + cov(\varepsilon_i, F_j) = a_{ij}$$

XIAMEN UNIVERSITY

如果对 X_i 作了标准化处理，X_i 的标准差为 1，且 F_j 的标准差为 1，那么

$$r_{X_i,F_j} = \frac{cov(X_i,F_j)}{\sqrt{D(X_i)}\sqrt{D(F_j)}} = cov(X_i,F_j) = a_{ij}$$

从上面的分析，我们知道对于标准化后的 X_i，a_{ij} 是 X_i 与 F_j 的相关系数，它一方面表示 X_i 对 F_j 的依赖程度，绝对值越大，密切程度越高；另一方面也反映了变量 X_i 对公共因子 F_j 的相对重要性. 了解这一点对我们理解抽象的因子含义，即因子命名，有非常重要的作用.

2、变量共同度

设因子载荷矩阵为 A，称第 i 行元素的平方和

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \qquad i = 1,2,\cdots,p$$

为变量 X_i 的共同度.

由因子模型，知

$$D(X_i) = a_{i1}^2 D(F_1) + a_{i2}^2 D(F_2) + \cdots + a_{im}^2 D(F_m) + D(\varepsilon_i)$$
$$= a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + Var(\varepsilon_i) = h_i^2 + \sigma_i^2$$

$$D(X_i) = h_i^2 + \sigma_i^2$$

上式说明，变量 X_i 的方差由两部分组成：

第一部分为共同度 h_i^2，它描述了全部公共因子对变量 X_i 的总方差所作的贡献，反映了变量 X_i 的方差中能够被全体因子解释的部分. 变量共同度越高，说明该因子分析模型的解释能力越高.

第二部分为特殊因子 ε_i 对变量 X_i 的方差的贡献，也就是变量 X_i 的方差中没有被全体因子解释的部分.

3、因子的方差贡献

设因子载荷矩阵为 A，称第 j 列元素的平方和

$$g_j^2=\sum_{i=1}^{p} a_{ij}^2 \qquad j=1,2,\cdots,m$$

为因子 F_j 对 X 的贡献，即 g_j^2 表示同一因子 F_j 对各变量所提供的方差贡献之总和，它是衡量每一个因子相对重要性的一个尺度.

四、因子载荷阵的求解

使用主成分分析法求解因子载荷阵的一般步骤：

(1)计算原始数据的协差阵 Σ 或相关矩阵 R.

(2)计算协差阵 Σ 或相关矩阵 R 的特征根为 $\lambda_1 \geq \cdots\lambda_p \geq 0$，相应的单位特征向量为 $\gamma_1, \gamma_2, \cdots\gamma_p$.

(3)利用 Σ 的特征根和特征向量计算因子载荷阵：

$$A = \left(\sqrt{\lambda_1}\gamma_1, \sqrt{\lambda_2}\gamma_2, \cdots, \sqrt{\lambda_p}\gamma_p\right)$$

由于因子分析的目的是减少变量个数，因此，因子数目 m 应小于原始变量个数 p. 所以在实际应用中，仅提取前 m 个特征根和对应的特征向量，构成仅含 m 个因子的因子载荷阵：

$$A=\left(\sqrt{\lambda_1}\gamma_1,\sqrt{\lambda_2}\gamma_2,\cdots,\sqrt{\lambda_m}\gamma_m\right)$$

可知，A 中第 j 列元素的平方和为 $(\sqrt{\lambda_j}\gamma_j)'(\sqrt{\lambda_j}\gamma_j)=\lambda_j\gamma_j'\gamma_j=\lambda_j$ （γ_j 是单位特征向量），即有

$$\lambda_j=\sum_{i=1}^{p}a_{ij}^2=g_j^2$$

这说明，第 j 个公因子的方差贡献 g_j^2 就等于样本协差阵的第 j 大特征根.

在实际应用中，有两种常用的确定因子提取个数 m 的方法:

一是仅提取方差贡献 g_j^2 大于 1的因子;

二是利用因子的累积方差贡献率 $\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 来确定公因子提取的个数，也就是寻找一个使得称 $\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 达到较大百分比的自然数 m.

五、因子命名与因子旋转

因子分析的目标之一就是要对所提取的抽象因子的实际含义进行合理解释，即对因子进行命名. 有时直接根据特征根、特征向量求得的因子载荷阵难以看出公共因子的含义.

例如，可能多个变量在同一个公共因子上的载荷大小不一，有的趋近 1，有的在 0.5 之间徘徊，有的趋于 0，说明该因子对多个变量的影响程度大小不同.

这种因子模型反而很难对因子的实际背景进行合理的解释. 这时需要通过**因子旋转**的方法，使每个因子仅在几个变量上的载荷趋于 1，在其余变量的载荷比较小趋于 0.

为了达到上述因子的要求引出了因子旋转方法，即正交旋转和斜交旋转两类，这里我们重点介绍**正交旋转**. 对公共因子作正交旋转就是对载荷矩阵 A 作一正交变换，右乘正交矩阵 Γ，使得旋转后的因子载荷阵 $B = A\Gamma$ 有更鲜明的实际意义.

旋转以后的公共因子向量为 $F^* = \Gamma'$，各个分量 $F_1^*, F_2^*, \cdots, F_m^*$ 也是互不相关的公共因子. 正交矩阵 Γ 的选取方法很多，实践中常用的方法是**最大方差旋转法**，其原理是使得旋转后因子载荷阵 B 的每一列元素的方差之和达到最大，从而实现使同一列上的载荷尽可能地向靠近 1 和靠近 0 两极分离的目的.

值得说明的是，旋转后的因子载荷阵 B 与旋转前的因子载荷阵相比，各因子的方差贡献发生了变化，已经不再等于样本协方差的第 j 大特征根，但提取出来的全部 m 个因子的总方差贡献率 $\frac{\sum_{k=1}^{m} g_k}{\sum_{i=1}^{p} g_i}$ 却不会改变，仍然等于 $\frac{\sum_{k=1}^{m} \lambda_k}{\sum_{i=1}^{p} \lambda_i}$. 另外因子旋转在改变因子载荷阵的同时也改变了**因子得分**.

六、因子得分

因子得分是因子分析的最终体现. 当因子载荷阵确定以后，便可以计算各因子在每个样本上的具体数值，称为**因子得分**. 得到了因子得分之后，就可以像主成分分析那样，用因子得分来代替原始变量，从而达到降维的效果.

七、案例分析

因子分析案例一（理工类）

问题背景：对我国各省市自治区的农业生产情况作因子分析。 从农业生产条件、生产结果和效益出发, 选取六项指标分别为:

X_1：乡村劳动力人口（万人）；X_2：人均经营耕地面积；X_3：户均生产性固定资产原值（元）；X_4：家庭基本纯收入（元）；X_5：人均农业总产值（千元/人）； X_6：增加值占总产值比重（\%）.

原始资料数据如下表：

序号	地 区	X_1	X_2	X_3	X_4	X_5	X_6
1	北 京	66.9	0.93	2972.41	3290.73	2.525	49.7
2	天 津	80.2	1.64	4803.54	2871.62	1.774	49.6
3	河 北	1621.8	2.03	4803.54	2871.81	0.8004	54
4	山 西	635.4	2.76	2257.66	1499.14	0.555	56.2
5	内蒙古	514.1	10.17	5834.94	1550.15	0.9051	66.4
6	辽 宁	605.1	2.96	3108.86	2059.35	1.4752	53.1
7	吉 林	534.2	4.73	4767.51	1940.46	1.1154	63.1
8	黑龙江	494.8	8.24	5573.02	2075.42	1.6283	57.8
9	上 海	66	1.02	1660.03	4571.81	3.0448	35.6
10	江 苏	1530.2	1.26	2826.86	2868.33	1.1921	50.6
11	浙 江	1123.1	0.94	5494.23	3289.07	0.8565	63.3
12	安 徽	1953.6	1.44	3573.62	1508.24	0.5756	59.2
13	福 建	775.8	0.82	2410.05	2295.19	1.1496	62.8
14	江 西	1103.2	1.3	2310.98	1804.93	0.6649	59.9
15	山 东	2475.1	1.44	3109.11	1989.53	0.8809	55
16	河 南	2815.8	1.5	3782.26	1508.36	0.5823	58.5
17	湖 北	1296.5	1.6	2291.6	1754.13	0.8799	62.8
18	湖 南	2089.3	1.42	2348.72	1719.18	0.587	64.7
19	广 东	1439.8	0.88	3249.61	2928.24	1.096	59.7
20	广 西	1579.9	1.43	3090.17	1590.9	0.5694	64.5
21	海 南	165.9	1.35	4454.77	1575.49	0.3535	65.2
22	四 川	3903.7	1.08	2870.45	1340.61	0.4443	64.1
23	贵 州	1376.6	1.18	2282.27	1206.25	0.2892	65.4
24	云 南	1642.2	2.42	4025.06	1096.73	0.3456	64.2
25	西 藏	88.6	2.51	11559.83	1257.71	0.4349	70.4
26	陕 西	1046.1	2.6	2228.55	1091.96	0.4383	59.7
27	甘 肃	672	5.86	2879.36	1037.12	0.4883	57.2
28	青 海	137.1	2.62	6725.11	1133.06	0.4096	70.3
29	宁 夏	139.1	4.01	5607.97	1346.89	0.4973	62.5
30	新 疆	288.5	3.96	7438.13	1161.71	1.4939	57.8

解:

第一步：将原始数据标准化

即令 $$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_{j}}{s_{j}},\quad 1\le i\le 30, 1\le j\le 6$$

其中，$$\overline{x}_{j}=\sum_{i=1}^{30}x_{ij}/30,\quad s_{j}^{2}=\sum_{i=1}^{30}\left(x_{ij}-\overline{x}_{j}\right)^{2}/29,\quad 1\le j\le 6$$

第二步：建立指标间的相关系数阵R

即对 $1 \le i, j \le 6$ 计算 $r_{ij} = \dfrac{\sum_{k=1}^{30}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{30}(x_{ki}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{30}(x_{kj}-\bar{x}_j)^2}}$

得

$$\mathbf{R} = \begin{bmatrix} 1 & -0.3325 & -0.3710 & -0.2026 & -0.3955 & 0.1413 \\ -0.3325 & 1 & 0.3492 & -0.2980 & -0.0014 & 0.1654 \\ -0.3710 & 0.3492 & 1 & -0.2481 & -0.1308 & 0.4044 \\ -0.2026 & -0.2980 & -0.2481 & 1 & 0.8145 & -0.7112 \\ -0.3955 & -0.0014 & -0.1308 & 0.8145 & 1 & -0.7967 \\ 0.1413 & 0.1654 & 0.4044 & -0.7112 & -0.7967 & 1 \end{bmatrix}$$

第三步：求 R的特征根和特征向量

序号	特征根	贡献率（%）	累计贡献率（%）
1	2.7765	46.2756	46.2756
2	1.7409	29.0160	75.2917
3	0.7116	11.8612	87.1529
4	0.4334	7.2248	94.3778
5	0.2369	3.9484	98.3263
6	0.1004	1.6736	100

第i个变量的贡献率= $\dfrac{\lambda_i}{\sum_{j=1}^{6}\lambda_j}\times\%,\quad 1\le i\le 6$

第i个变量的累计贡献率= $\dfrac{\sum_{j=1}^{i}\lambda_j}{\sum_{j=1}^{6}\lambda_j}\times\%,\quad 1\le i\le 6$

由于前三个特征根的累计贡献率已达87.15%,所以取前三个特征根所对应的特征向量，如下：

α_1	α_2	α_3
0.1460	−0.6242	−0.1854
0.1631	0.5270	0.7547
0.2421	0.5272	0.5369
−0.5463	0.0153	0.2325
−0.5455	0.2317	−0.0422
0.5453	0.0225	0.2276

第四步：计算因子载荷矩阵

令 $a_i = \sqrt{\lambda_i}\alpha_i, 1 \le i \le 3.$ 则因子载荷矩阵 $A = (a_1, a_2, a_3)$

A表示如下

因子	a_1	a_2	a_3	h_i^2
X_1	0.2433	−0.8236	−0.1564	0.7621
X_2	0.2718	0.6954	0.6366	0.9629
X_3	0.4035	0.6957	0.4529	0.8520
X_4	−0.9103	0.0202	0.1961	0.8675
X_5	−0.9089	0.3057	−0.0356	0.9210
X_6	0.9086	0.0296	0.1920	0.8634

其中 h_i^2 是公共因子对变量 x_i 的贡献，$1 \le i \le 6$

第五步：因子Varimax旋转

旋转后的因子载荷矩阵A如下

因子	$a_1\,(F_1)$	$a_2\,(F_2)$	$a_3\,(F_3)$
X_1	-0.3793	-0.7252	-0.3036
X_2	-0.1046	0.2178	0.9510
X_3	-0.2957	0.8698	0.0890
X_4	0.8862	0.0265	-0.2852
X_5	0.9499	0.1206	0.0645
X_6	-0.8976	0.2402	-0.0009

第六步：将指标按高载荷分类，并结合专业知识给出各因子的命名(解释)

在第一因子中，X4、X5、X6三项指标有较大的载荷，这些都从产出效益方面描述农业情况的，所以称为产出及效益因子。

在第二个因子中，X1、X3，有较大的载荷，这主要是人们对农业的生产工具、人力等的投入，所以称为人为投入条件因子。

在第三个因子中，X2，有较大的载荷，这主要从自然条件方面刻划农业的生产条件状况，所以称为自然条件因子。

	高载荷指标	命名
因子一	X_4：家庭基本纯收入 X_5：人均产值 X_6：增加值占总产值比重	产出及效益因子
因子二	X_1：乡村劳动力人口 X_3：户均生产性固定资产原值	人为投入条件因子
因子三	X_2：人均经营耕地面积	自然条件因子

第七步：因子得分

计算因子得分矩阵 $\mathbf{B}=\mathbf{A}'\mathbf{R}^{-1}$

$$\mathbf{B}=\begin{bmatrix} -0.1592 & 0.0271 & -0.1063 & 0.3083 & 0.3562 & -0.3336 \\ -0.4986 & -0.1906 & 0.7097 & 0.1330 & 0.0738 & 0.2086 \\ -0.0771 & 0.9657 & -0.2667 & -0.2659 & 0.0904 & -0.1590 \end{bmatrix}$$

计算因子得估计 $\hat{\mathbf{F}}=\mathbf{B}\mathbf{X}^{*}$

计算三个公共因子绝对得分之和

行政区	得分总和	行政区	得分总和	行政区	得分总和
北京	2.9342	浙江	2.5237	海南	1.9911
天津	2.5903	安徽	1.5356	四川	3.6627
河北	0.9471	福建	0.9435	贵州	1.9248
山西	1.1655	江西	1.0140	云南	1.3281
内蒙古	3.9199	山东	1.5117	西藏	5.5579
辽宁	1.4929	河南	2.1293	陕西	1.7389
吉林	1.3497	湖北	1.0150	甘肃	2.9330
黑龙江	3.4426	湖南	2.0175	青海	2.9342
上海	3.9771	广东	1.4801	宁夏	1.9522
江苏	1.9923	广西	1.5219	新疆	2.0541

综上我们可以回顾因子分析的一般步骤：

（1）求指标之间的相关阵 R；

（2）确定因子的个数；

（3）因子旋转；

（4）因子命名；

（5）因子得分.

THANK YOU