信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第12章 MATLAB多元统计分析

讲授人：牛言涛

日期：2020年4月25日

目录

CONTENTS

因子分析简介

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 作个比喻，对面来了一群女生，我们一眼就能够分辨出孰美孰丑，这是判别分析；并且我们的脑海中会迅速将这群女生分为两类：美的一类，丑的一类，这是聚类分析。我们之所以认为某个女孩漂亮，是因为她具有漂亮女孩所具有的一些共同点，比如漂亮的脸蛋、高挑的身材、白皙的皮肤等等。其实这种从研究对象中寻找公共因子的办法就是因子分析（Factor Analysis）。

长相	五官	颜值	性格	脾气	腿长	身高
5	6	11	30	25	69	174
7	8	15	25	20	59	164
5	6	11	14	9	55	160
4	5	9	22	17	54	159
9	10	19	30	25	52	157
5	6	11	28	23	51	156
6	7	13	13	8	53	158

颜值	性格	身材
7	28	122
10	23	112
7	12	108
6	20	107
13	28	105
7	26	104
9	11	106

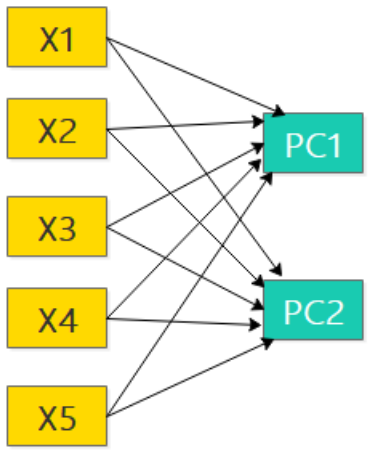
因子分析简介

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

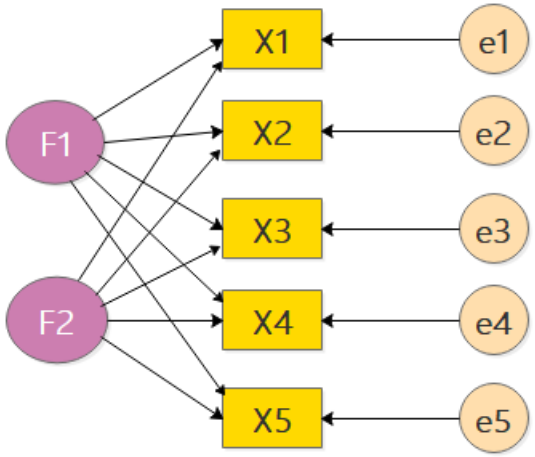
- 例1 为了解学生的知识和能力，对学生进行了抽样命题考试，考题包括的面很广， 但可归结为学生的语文水平、数学推导、艺术修养、历史知识、生活知识等五个方面，把每一个方面称为一个（公共）因子，显然每个学生的成绩均可由这五个因子来确定，即可设想每个学生考试的分数能用这五个公共因子的线性组合表示出来。
- 例2 诊断时，医生检测了病人的五个生理指标：收缩压、舒张压、心跳间隔、呼吸间隔和舌下温度，但依据生理学知识，这五个指标是受植物神经支配的，植物神经又分为交感神经和副交感神经，因此这五个指标可用交感神经和副交感神经两个公共因子来确定，从而也构成了因子模型。
- 例3 Holjinger和Swineford在芝加哥郊区对145名七、八年级学生进行了24个心理测验，通过因子分析，这24个心理指标被归结为4个公共因子，即词语因子、速度因子、 推理因子和记忆因子。

因子分析简介

- 因子分析的思想源于1904年查尔斯·斯皮曼（charles spearman）对学生考试成绩的研究，目前因子分析已经在很多领域得到广泛应用。
- 因子分析也是利用降维的思想，把每一个原始变量分解成两部分，一部分是少数几个公共因子的线性组合，另一部分是该变量所独有的特殊因子，其中公共因子和特殊因子都是不可观测的隐变量，我们需要对公共因子作出具有实际意义的合理解释。



主成分分析模型



因子分析模型

- 主成分分析：把主成分表示成各原始变量的线性组合。解释原始变量的总方差。几个原始变量，就有几个主成分。给定的协方差矩阵或相关矩阵特征值唯一时，主成分也是唯一的。
- 因子分析：原始变量表示成各因子的线性组合。解释原始变量的协方差。因子个数可以根据业务场景的需要人为指定。因子不是唯一的。并且通过旋转可以得到不同因子。
- 主成分分析和因子分析用途：数据处理，降维，变量间关系的探索等。主成分分析是因子分析的一个特例。

1. 因子分析原理分析

- 因子分析是将多个实测变量转换为少数几个不相关的综合指标的多元统计方法。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。假想变量是不可观测的潜在变量，称为因子。
- 假定这p个有相关关系的随机变量含有m个彼此独立的因子，可表示为：

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \qquad \vdots \qquad \vdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

或用矩阵表示为$X = AF + \varepsilon$。$F_1, F_2, \ldots, F_m$称为公共因子，是不可观测的变量，它们的系数称为因子载荷，A称为因子载荷矩阵，它反映了公共因子对变量的重要程度，对解释公共因子具有重要的作用。ε是独有的特殊因子，是不能包含在公共因子的部分。

1. 因子分析原理分析

- 需要满足：

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \quad \vdots \quad \vdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

  - $m \leq p$，即公共因子数不超过原变量个数；
  - 公共因子之间互不相关，且每个F_i方差为1，即F的协方差矩阵为单位矩阵I；
  - 公共因子和特殊因子之间彼此互不相关，即$Cov(F, \varepsilon) = 0$；
  - 特殊因子之间彼此互不相关，但方差不一定相同，记ε_I的方差为：

$$\text{var}(\varepsilon) = D = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

  - 理想的情况是，对于每个原始变量而言，其在因子载荷矩阵中，在一个公共因子上的载荷较大，在其他的因子上载荷较小。可以通过因子旋转方法调整因子载荷矩阵。

1. 因子分析原理分析

- 因子载荷阵的求解方法有极大似然估计法、主成分分析法、主因子法。这里仅介绍最为常用的主成分分析法，且不加证明地给出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$\boldsymbol{\Sigma}$。

  2. 计算协差阵$\boldsymbol{\Sigma}$的特征根为$\lambda_1 > \lambda_2 > \cdots \lambda_p > 0$，相应的单位特征向量为$u_1, u_2, \cdots, u_p$。

  3. 利用$\boldsymbol{\Sigma}$的特征根和特征向量计算因子载荷阵：

$$A = \left(\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \cdots, \sqrt{\lambda_p}u_p\right)$$

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，构成仅包含m个因子的因子载荷阵：

$$A = \left(\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \cdots, \sqrt{\lambda_m}u_m\right)$$

2. 因子载荷阵的统计意义

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

1. 对于因子模型

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{ij}F_j + \cdots a_{im}F_m + \varepsilon_i \qquad i = 1, 2, \cdots, p$$

协方差 $Cov\left(X_i, F_j\right) = Cov\left(\sum_{k=1}^{m} a_{ik}F_k + \varepsilon_i, F_j\right) = Cov\left(\sum_{k=1}^{m} a_{ik}F_k, F_j\right) + Cov\left(\varepsilon_i, F_j\right) = a_{ij}$

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，因此

$$\rho_{X_i, F_j} = \frac{Cov\left(X_i, F_j\right)}{\sqrt{D\left(X_i\right)}\sqrt{D\left(F_j\right)}} = Cov\left(X_i, F_j\right) = a_{ij}$$

那么，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。了解这一点对理解抽象的因子含义，即因子命名，有非常重要的作用。

2. 因子载荷阵的统计意义

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

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 $h_i^2=\sum_{j=1}^{m}a_{ij}^2$　　$i=1,2,\cdots,p$ 为变量 X_i 的共同度。

由因子模型，知

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

- 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。
  - 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。
- 变量共同度越高，说明该因子分析模型的解释能力越高。

2. 因子载荷阵的统计意义

信阳师范学院 数学与统计学院 SCHOOL OF MATHEMATICS AND STATISTICS

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

由因子载荷阵的表达式 $A=\left(\sqrt{\lambda_1}u_1,\sqrt{\lambda_2}u_2,\cdots,\sqrt{\lambda_p}u_p\right)$，可知$A$中第$j$列元素的平方和为

$\left(\sqrt{\lambda_j}u_j\right)'\left(\sqrt{\lambda_j}u_j\right)=\lambda_j u_j'u_j=\lambda_j$（$u_j$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（λ_j）大于1的因子；二是利用因子的累积方差贡献率 $\sum_{j=1}^{m}\lambda_j\Big/\sum_{j=1}^{p}\lambda_j$ 来确定公因子提取的个数，也就是寻找一个使得 $\sum_{j=1}^{m}\lambda_j\Big/\sum_{j=1}^{p}\lambda_j$ 达到较大百分比的自然数m 。

3. 因子旋转

- 因子分析的主要目的是对公共因子给出符合实际意义的合理的解释，解释的主要依据就是因子载荷阵的各列元素的取值。当因子载荷阵某一列上各元素的绝对值差距比较大，并且绝对值大的元素较少时，则该公共因子就易于解释，反之，公共因子的解释就变得比较困难。
- 此时可以考虑对因子和因子载荷阵进行旋转（例如正交旋转），使得旋转后的因子载荷阵的各列元素的绝对值尽可能两极分化，这样就使得因子的解释变得容易。这就好比一个女孩，正面看上去可能不觉得漂亮，可女孩不经意的一个转身，或许看到她楚楚动人的某个侧面。
- 因子旋转的方法有正交旋转和斜正交旋转两种，这里只介绍一种普遍使用的正交旋转法：最大方差因子旋转法(Varimax)。它是由Kaiser于1958年提出的，这种旋转方法的目的是使因子载荷阵每列上的各元素的绝对值（或平方值）尽可能地向两极分化，即少数元素的绝对值（或平方值）取尽可能大的值，而其它元素尽量接近于0。

3. 因子旋转

- 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$F^* = \Gamma' F$，它的各个分量F_1^*，$F_2^*, \cdots$，F_m^*也是互不相关的公共因子。根据正交矩阵Γ的不同选取方式，将构造出不同的正交旋转的方法。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A相比，各因子的方差贡献g_j^2发生了变化，已经不再等于样本协差阵的第j大特征根，但提取出的全部m个因子的总方差贡献率 $\sum_{j=1}^{m} g_j^2 \Big/ \sum_{j=1}^{p} g_j^2$ 却不会改变，仍然等于 $\sum_{j=1}^{m} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j$ 。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。

4. 因子得分

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 在对公共因子作出合理的解释之后，有时还需要求出各观测所对应的各个公共因子的得分，比如我们知道某个女孩是一个美女，可能很多人更关心该给她的脸蛋，身材等各打多少分。
- 在因子分析模型$\boldsymbol{X} = \boldsymbol{AF} + \boldsymbol{\varepsilon}$中，如果不考虑特殊因子的影响，当$\boldsymbol{m} = \boldsymbol{p}$且$\boldsymbol{A}$可逆时，可以非常方便地从每个样品的指标取值$\boldsymbol{X}$计算出其在因子$\boldsymbol{F}$上的相应取值：$\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{X}$，即该样品在因子$\boldsymbol{F}$上的“得分”情况，简称为该样品的因子得分。
- 但是因子分析模型在实际应用中要求$\boldsymbol{m} < \boldsymbol{p}$，因此，不能精确计算出因子的得分情况，只能对因子得分进行估计。估计因子得分的方法也有很多，常用的方法包括回归法（Regression）、巴特莱特法（Bartlett）、安德森 - 鲁宾法（Anderson-Rubin）等。
- 可以证明，如果使用回归法，则因子得分可以由下面的式子给出：$\boldsymbol{F} = \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$，其中，$\boldsymbol{\Sigma}$为样本协差阵。称$\boldsymbol{m} \times \boldsymbol{p}$的矩阵$\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}$为因子得分系数矩阵。
- 应该注意，如果因子载荷阵经过了旋转，则上式中的因子载荷阵 $\boldsymbol{A}$ 应该是旋转后的因子载荷阵。

因子分析中的Heywood（海伍德）现象

信阳师范学院
数学与统计学院

- 如果X的各个分量都已经标准化了，则其方差=1。即共性方差与特殊方差的和为1。也就是说共性方差与特殊方差均大于0，并且小于1。但在实际进行参数估计的时候，共性方差的估计可能会等于或超过1，如果等于1，就称之为海伍德现象，如果超过1，称之为超海伍德线性。超海伍德现象意味着某些特殊因子的方差为负，表明肯定存在问题。造成这种现象的可能原因包括：
  - 共性方差本身估计的问题；
  - 太多的共性因子，出现了过拟合；
  - 太少的共性因子，造成拟合不足；
  - 数据太少，不能提供稳定的估计；
  - 因子模型不适合这些数据。
- 当出现海伍德现象或超海伍德现象时，应对估计结果保持谨慎态度。可以尝试增加数据量，或改变公共因子数目，让公共因子数目在一个允许的范围内变动，观察估计结果是否有改观；还可以尝试用其他多元统计方法进行分析，比如主成分析。

5. 因子分析与主成分分析的区别

1. 因子分析中是把变量表示成各因子的线性组合，而主成分分析中则是把主成分表示成各个变量的线性组合。主成分分析仅仅是变量变换：用原始变量的线性组合表示新的综合变量，即主成分。因子分析需要构造因子模型：用潜在的假想变量和随机影响变量的线性组合表示原始变量。因子模型估计出来后，需要对所得的公共因子进行解释。

2. 主成分分析的重点在于解释个变量的总方差，而因子分析则把重点放在解释各变量之间的协方差。

3. 主成分分析中不需要有假设(assumptions)，因子分析则需要一些假设。因子分析的假设包括：各个共同因子之间不相关，特殊因子（specific factor）之间也不相关，共同因子和特殊因子之间也不相关。

4. 主成分分析中，当给定的协方差矩阵或者相关矩阵的特征值是唯一时，主成分一般是独特的；而因子分析中因子不是独特的，可以旋转得到不到的因子。

5. 在因子分析中，因子个数需要分析者指定，根据一定的条件自动设定，只要是特征值大于1的因子进入分析，而指定的因子数量不同而结果不同。在主成分分析中，成分的数量是一定的，一般有几个变量就有几个主成分。

5. 因子分析与主成分分析的区别

和主成分分析相比，由于因子分析可以使用旋转技术帮助解释因子，在解释方面更加有优势。大致说来，当需要寻找潜在的因子，并对这些因子进行解释的时候，更加倾向于使用因子分析，并且借助旋转技术帮助更好解释。而如果想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进入后续的分析，则可以使用主成分分析。当然，这中情况也可以使用因子得分做到。所以这种区分不是绝对的。

总得来说，主成分分析主要是作为一种探索性的技术，在分析者进行多元数据分析之前，用主成分分析来分析数据，让自己对数据有一个大致的了解是非常重要的。

主成分分析一般很少单独使用（不一定，可单独用）：①了解数据 (screening the data)；②和cluster analysis 一起使用；③和判别分析一起使用，比如当变量很多，个案数不多，直接使用判别分析可能无解，这时候可以使用主成份发对变量简化（reduce dimensionality）；④在多元回归中，主成分分析可以帮助判断是否存在共线性（条件指数），还可以用来处理共线性。

在算法上，主成分分析和因子分析很类似，不过，在因子分析中所采用的协方差矩阵的对角元素不再是变量的方差，而是和变量对应的共同度（变量方差中被各因子所解释的部分）。

6. MATLAB求解因子分析函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- matlab函数主要有rotatefactors 和factoran，其中factoran调用了rotatefactors函数。这里主要介绍factoran函数。factoran用来根据原始数据样本观测数据，样本协方差矩阵或样本相关系数矩阵，计算因子模型中因子载荷阵A的最大似然估计，求特殊方差的估计，因子旋转矩阵和因子得分，还能对因子模型进行检验。factoran函数的调用格式如下：
- lambda=factoran(X,m)：返回包含m个公共因子的因子模型的载荷阵lambda。
  - 输入参数X是n行d列的矩阵，每行对应一个观测，每列对应一个变量。
  - m是一个正整数，表示模型中公共因子的个数。
  - 输出参数lambda是一个d行m列的矩阵，第i行第j列元素表示第i个变量在第j个公共因子的载荷。
  - 默认情况下，factoran函数调用用rotatefactors 函数，并用'varimax'选项（rotatefactors函数的可用选项）来计算旋转后因子载荷阵的估计。

6. MATLAB求解因子分析函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- [lambda,psi,T,stats]= factoran(X,m)：返回特殊方差的最大似然估计psi，psi是包含d个元素的列变量，分别对应d个特殊方差的最大似然估计；返回m行m列的旋转矩阵T。返回一个包含模型检验信息的结构体变量stats，模型检验的原假设是H0：因子数 = m 。参数stats包括4个字段：
  - 其中stats.loglike表示对数似然函数最大值，stats.dfe表示误差自由度，误差自由度的取值为 $\left[(\boldsymbol{d}-\boldsymbol{m})^2-(\boldsymbol{d}+\boldsymbol{m})\right]/\boldsymbol{2}$，stats.chisq 表示近视卡方检验统计量，stats.p表示检验p值。
  - 对于给定的显著性水平$\boldsymbol{\alpha}$，若检验的p值大于显著性水平$\boldsymbol{\alpha}$，则接受原假设H0，说明用含有m个公共因子的模型拟合原始数据是合适的，否则，拒绝原假设，说明拟合不合适。
  - 注意 只有当stats.def是正的，并且psi中特殊方差的估计都是正数时，factoran函数才计算，stats.chisq和stats.p。当输入参数X是协方差矩阵或相关系数矩阵时，若要计算stats.chisq和stats.p必须指定'nobs'参数。

6. MATLAB求解因子分析函数

- [lambda,psi,T,stats,F]=factoran(X,m)：F是一个n行m列的矩阵，每一行对应一个观测的m个公共因子的得分。若果X是一个协方差矩阵或相关系数矩阵，则factoran函数不能计算因子得分。factoran函数用相同的旋转矩阵计算因子载荷阵lambda和因子得分F。
- [...] = factoran(..., param1, val1, param2, val2)：允许用户指定可选的成对出现的参数名誉参数值，用来控制模型的拟合和输出，可用的参数名与参数值如表所列：

参数名	参数值	说明
'xtype'	指定输入参数x的类型，可以是下列2者之一	
	'data'	原始数据（默认情况）
	'covariance'	正定的协方差矩阵或相关系数矩阵
'scores'	预测因子得分的方法。若X不是原始数据，'scores'将被忽略	
	'wls' 'Bartlett'	加权最小二乘估计（默认情况）
	'regression' 'thomson'	最小均方误差法，相当于岭回归法
'coeff'	一个介于0和1之间的数	经常记为 γ，不同的值对应不同的orthomax旋转。若取值为0，对应quartimax旋转，若取值为1（默认情况），对应varimax旋转（最大方差旋转）

6. MATLAB求解因子分析函数

'rotate'	指定因子载荷阵和因子得分的旋转方法。'rotate'与'rotatefactors'函数的'method'参数有相同值	
	'none'	不进行旋转
	'equamax'	orthomax旋转的特殊情况。用'normalize','reltol'和'maxit'参数来控制旋转
	'orthomax'	最大方差旋转法（一种正交旋转方法）。用'coeff','normalize','reltol'和'maxit'参数来控制旋转
	'parsimax'	orthomax旋转的一个特殊情况（默认情况）。用'normalize','reltol'和'maxit'参数来控制旋转
	'pattern'	执行斜交旋转（默认）或正交旋转，以便和一个指定的模式矩阵（即目标矩阵）达到最佳匹配。用'type'参数选择旋转类型，用'target'指定模式矩阵
	'procrustes'	执行斜交旋转（默认）或正交旋转，以便和一个指定的模式矩阵（即目标矩阵）在最小二乘意义上达到最佳匹配，用'type'参数选择旋转类型，用'target'指定模式矩阵
	'promax'	执行一次斜交procrustes旋转，与一个目标矩阵相匹配，这个目标矩阵是orthomax解经过一定运算后得到的。用'power'参数指定生成目标矩阵的幂指数，由于'promax'旋转的内部用到了orthomax旋转，此时可以指定orthomax旋转的参数
	'quartimax'	orthomax旋转的一个特殊情况（默认情况）。用'normalize','reltol'和'maxit'参数来控制旋转
	'varimax'	orthomax旋转的一个特殊情况（默认情况）。用'normalize','reltol'和'maxit'参数来控制旋转
	函数句柄	用户自定义的旋转函数的句柄，旋转函数形如 [B,T]=myrotation(A,…)：这里的A是一个d行m列的未经旋转的因子载荷阵，B是经过旋转的d行m列的因子载荷阵，T是相应的m行m列旋转矩阵，此时可用factoran函数的'userargs'参数传递额外的输入参数给自定义旋转函数

6. MATLAB求解因子分析函数

信阳师范学院
数学与统计学院

参数名	参数值	说明
'start'	最大似然估计中特殊方差psi的初始值，可如下设置	
	'random'	选取d在[0,1]区间上服从均匀分布的随机数
	'Rsquared'	用一个尺度因子乘以diag(inv(corrcoef(x)))作为初始点（默认情况）
	正整数	指定最大似然拟合的次数，每次拟合随机选择初始点，返回对数
		似然函数取最大值时的拟合结果
	矩阵	用一个d行多列的矩阵指定最大似然法的初始点，矩阵每一列对
		应一个初始点，也对应一次拟合，返回对数似然函数取最大值时的拟合结果
'reltol'	正标量	指定'orthomax'或'varimax'旋转的收敛容限，默认值为sqrt(eps)
'maxit'	正整数	指定'orthomax'或'varimax'旋转的最大迭代次数，默认值为250
'target'	矩阵	指定'procrustes'旋转所需的目标因子载荷阵，没有默认值
'type'	'oblique'或'orthogonal'	指定'procrustes'旋转的类型，默认值为'oblique'

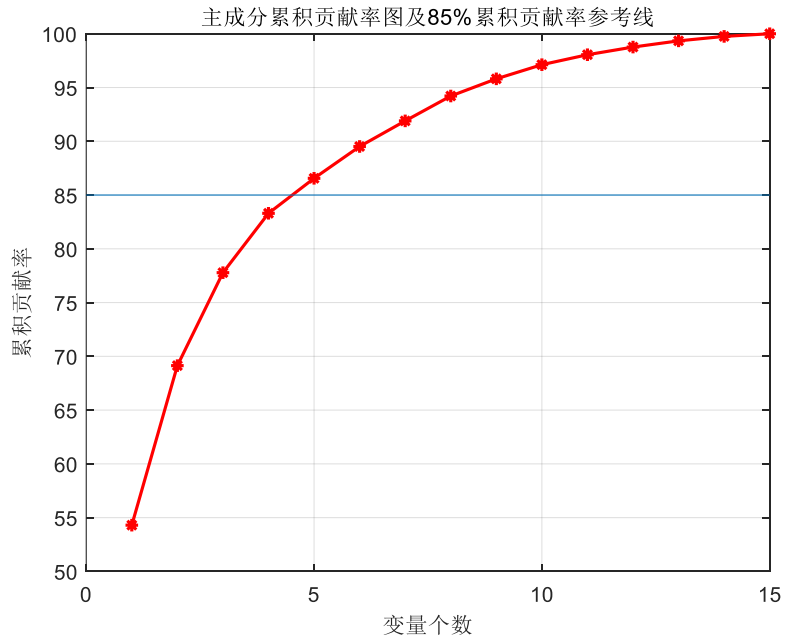
6. MATLAB求解因子分析函数

参数名	参数值	说明
'normalize'	'on'或1，'off'或0	对于'orthomax'或'varimax'旋转，用来指示是否对因子载荷阵按行单位化的标识，若为'on'或1（默认情况），则单位化，若为'off'或0，不进行单位化
'power'	大于或等于1的标量	指定'promax'旋转中生成目标矩阵的幂指数，默认值为4
'userargs'	自定义旋转函数的额外参数	一个标记开始位置的参数，在'userargs'参数的后面开始自定义旋转函数传递额外的输入参数
'nobs'	正整数	如果输入参数X是协方差矩阵或相关系数矩阵，用来指定实际观测的个数，它被用来进行模型检验。也就是说即使没有原始观测数据，指定了该参数的取值，同样可以进行模型检验。若X为原始观测数据，则'nobs'将被忽略，'nobs'参数没有默认值
'delta'	[0,1）内取值的标量	设定最大似然估计中特殊方差psi下界，默认值为0.005
'optimopts'	由命令statset('factoran')生成的结构体变量	指定用来计算最大似然估计的迭代算法的控制参数。由statset('factoran')命令可以查看默认值

7. 典型案例分析

例1：48名应聘者应聘某公司的某职位，公司为这些应聘者的15项指标打分，试用因子分析的方法对15项指标做因子分析，在因子分析中选择5个因子。15项指标说明：求职信形式：FL，外貌：APP，专业能力：AA，讨人喜欢：LA，自信心：SC，洞察力：LC，诚实：HON，推销能力：SMS，经验：EXP，驾驭水平：DRV，事业心：AMB，理解能力：GSP，潜在能力:POT，交际能力:KJ，适应性:SUIT。

```
>> [candidata,text] = xlsread('candidate.xlsx');
>> [~,~,latent,~,explained] = pca(candidata);
>> cs = cumsum(explained);  %累积贡献率
>> plot(cs,'r-*','LineWidth',1.5)
>> refline(0,85)  %添加参考线
>> xlabel('变量个数');  ylabel('累积贡献率')
>> title('主成分累积贡献率图及85%累积贡献率参考线')
```



从图中可以判断至少提取5个公共因子可使得解释程度达到85%以上。当摇摆不定时，高估因子数通常比低估因子数的结果好，因为高估因子数一般较少曲解“真实”情况。

7. 典型案例分析

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

```
% 因子分析，取五个公共因子计算
>> [lambda,psi,T,stats,F]=factoran(candidata,5);
>> n = size(candidata,2);  %取变量个数
>> res = cell(n,5+1);  %组合载荷矩阵
>> res(:,1) = text(2,2:end);  %第一列存变量名称
>> res(:,2:end) = num2cell(lambda);
```

```
res =
  15×6 cell 数组
    '求职信形式'    [0.1275]    [ 0.7216]    [ 0.1019]    [    -0.1173]    [ 0.0101]
    '外貌'          [0.4517]    [ 0.1339]    [ 0.2698]    [     0.2058]    [-0.2562]
    '专业能力'      [0.0594]    [ 0.1288]    [ 0.0022]    [     0.6863]    [-0.0156]
    '讨人喜欢'      [0.2213]    [ 0.2458]    [ 0.8273]    [    -0.0562]    [ 0.0800]
    '自信心'        [0.9166]    [-0.0933]    [ 0.1670]    [    -0.0720]    [-0.0118]
    '洞察力'        [0.8497]    [ 0.1247]    [ 0.2785]    [     0.0246]    [ 0.4235]
    '诚实'          [0.2284]    [-0.2198]    [ 0.7771]    [-3.2541e-04]    [-0.0625]
    '推销能力'      [0.8801]    [ 0.2660]    [ 0.1111]    [    -0.0474]    [ 0.0157]
    '经验'          [0.0805]    [ 0.7726]    [-0.0499]    [     0.1706]    [-0.0178]
    '驾驶水平'      [0.7546]    [ 0.3927]    [ 0.1990]    [    -0.0395]    [-0.1109]
    '事业心'        [0.9094]    [ 0.1871]    [ 0.1127]    [    -0.0363]    [-0.1623]
    '理解能力'      [0.7827]    [ 0.2945]    [ 0.3544]    [     0.1480]    [ 0.1846]
    '潜在能力'      [0.7169]    [ 0.3625]    [ 0.4456]    [     0.2673]    [-0.0176]
    '交际能力'      [0.4179]    [ 0.3987]    [ 0.5629]    [    -0.5850]    [-0.0479]
    '适应性'        [0.3506]    [ 0.7645]    [ 0.0582]    [     0.1479]    [ 0.0068]
```

- 在得到的结果中，公共因子还有比较鲜明的实际意义：
  - 第一公共因子中，系数绝对值大的变量主要是： 自信心SC，洞察力LC，推销能力SMS，驾驭水平DRV，事业心AMB，理解能力GSP，潜在能力POT，这些主要表现求职者工作能力和内在素质，可以命名为基本素质；

7. 典型案例分析

- 在得到的结果中，公共因子还有比较鲜明的实际意义：
  - 第一公共因子中，系数绝对值大的变量主要是： 自信心SC，洞察力LC，推销能力SMS，驾驭水平DRV，事业心AMB，理解能力GSP，潜在能力POT，这些主要表现求职者工作能力和内在素质，可以命名为基本素质；
  - 第二公共因子中，系数绝对值大的变量主要是：求职信的形式FL，经验EXP，适应性SUIT，这些主要反映了求职者的经验和对工作的适应能力，可以命名为工作经验素质；
  - 第三公共因子中，系数绝对值大的变量主要是：讨人喜欢LA，诚实HON，这些主要反映了求职者的人品人格；
- 第四、第五公共因子系数绝对值较小，这说明这两个公共因子相对次要一些；
  - 第四公共因子中，系数绝对值大的变量主要是：专业能力AA，交际能力KJ，这些主要反映了求职者的专业能力；
  - 第五公共因子中，系数绝对值较大的变量主要是：外貌APP，洞察力LC，这些主要反映了求职者的外貌。

7. 典型案例分析

基于以上分析，五个公共因子中各变量的表示形式为：

$$FL = 0.127\,f_1 + 0.722\,f_2 + 0.102\,f_3 - 0.117\,f_4$$

$$APP = 0.451\,f_1 + 0.134\,f_2 + 0.270\,f_3 + 0.206\,f_4 + 0.258\,f_5$$

$$AA = 0.129\,f_2 + 0.686\,f_4$$

…

可从输出结果psi中查看变量的独特性（Uniquenesses），独特性为1减去共同度后得到的结果。如果所有因子联合解释某变量的大部分方差，则该变量具有高共同度，因而具有较低的独特性。

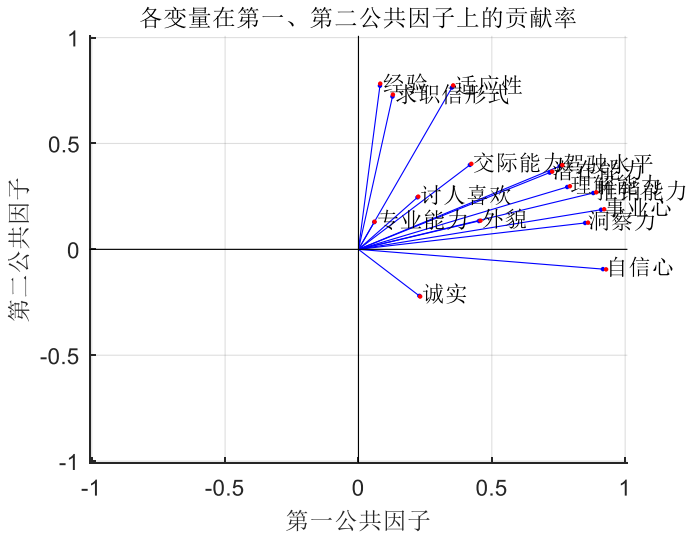
```
>> [psis,ind] = sort(psi,'descend');    %对特殊方差进行降序排列
>> varnames = text(2,2:end)';  %提取变量名称
>> res_psi = cell(n,2);  %组合数据
>> res_psi(:,1) = varnames;
>> res_psi(:,2) = num2cell(psis);
```

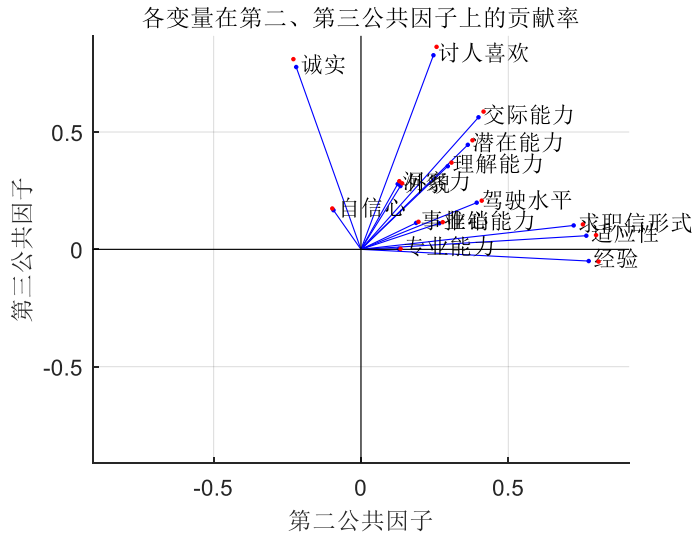
```
res_psi =
  15×2 cell 数组
    '求职信形式'      [0.5972]
    '外貌'          [0.5086]
    '专业能力'        [0.4387]
    '讨人喜欢'        [0.3646]
    '自信心'         [0.2918]
    '洞察力'         [0.2674]
    '诚实'          [0.2229]
    '推销能力'        [0.1966]
    '经验'          [0.1398]
    '驾驶水平'        [0.1191]
    '事业心'         [0.1179]
    '理解能力'        [0.0976]
    '潜在能力'        [0.0843]
    '交际能力'        [0.0050]
    '适应性'         [0.0050]
```

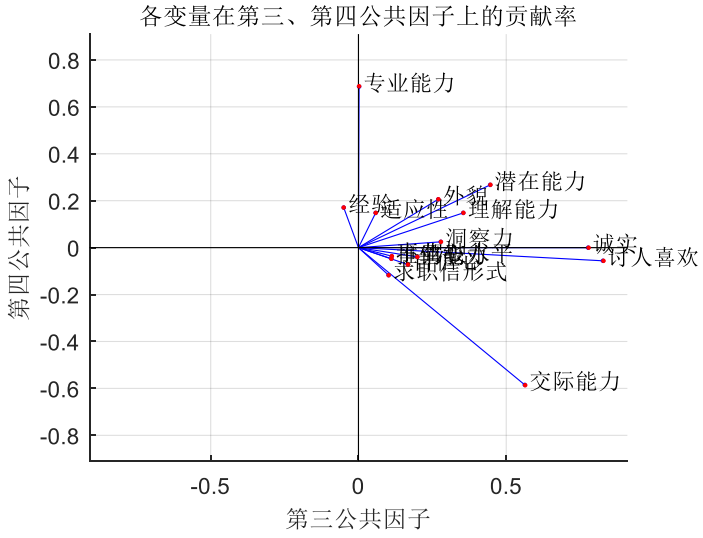
从结果可以看出，求职信形式、外貌、专业能力、讨人喜欢和自信心为前五个包含最大特殊方差的变量。反之，共同度最低。

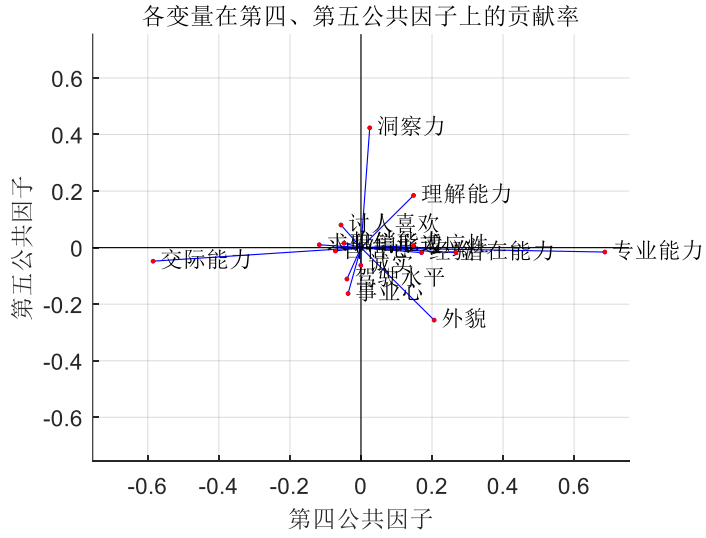
7. 典型案例分析

```
subplot(2,2,1)
biplot(lambda(:,1:2),'Score',lambda(:,1:2),'VarLabels',varnames)
subplot(2,2,2)
biplot(lambda(:,2:3),'Score',lambda(:,2:3),'VarLabels',varnames)
subplot(2,2,3)
biplot(lambda(:,3:4),'Score',lambda(:,3:4),'VarLabels',varnames)
subplot(2,2,4)
biplot(lambda(:,4:5),'Score',lambda(:,4:5),'VarLabels',varnames)
```



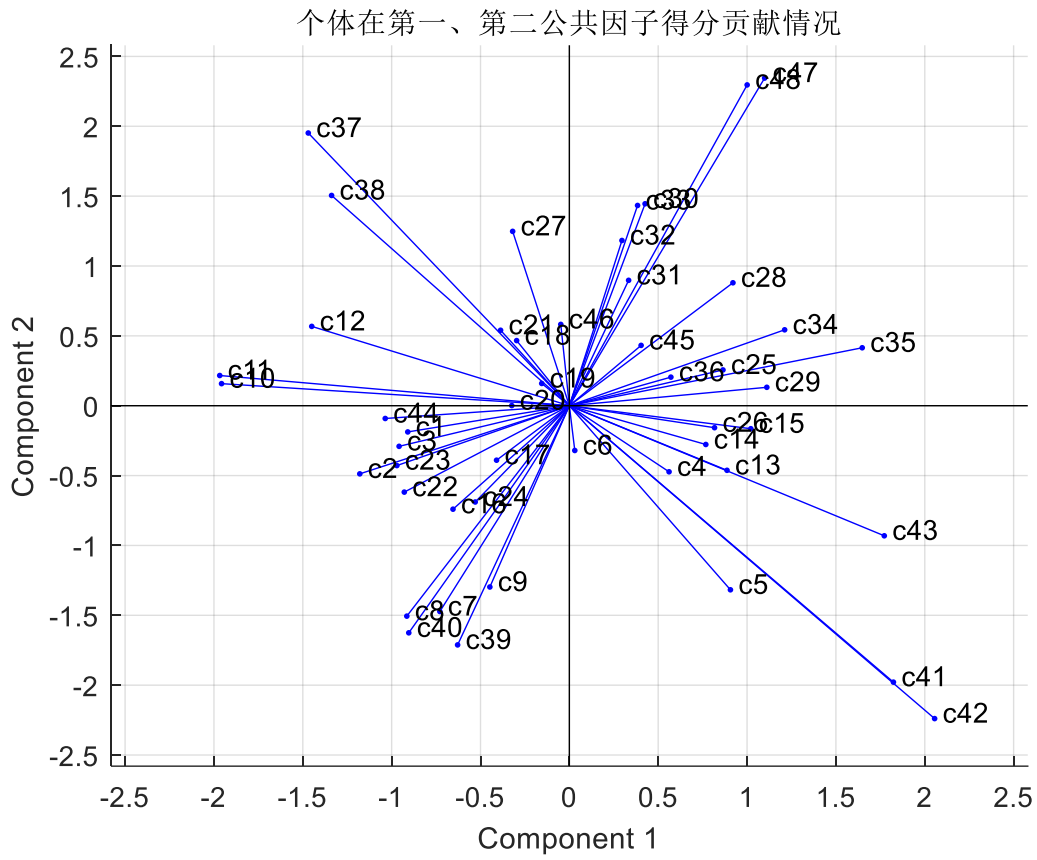






7. 典型案例分析

```
>> rownames = text(3:end,1);
>> biplot(F(:,1:2),'VarLabels',rownames)
>> title('个体在第一、第二公共因子得分贡献情况')
```



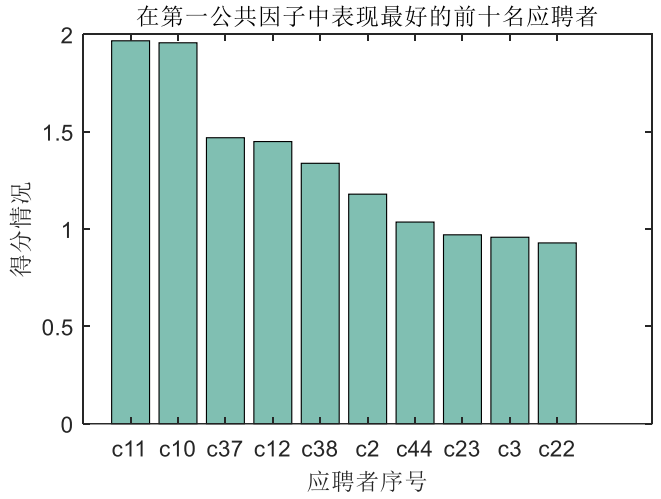
- 此图反看，一三象限和二四象限调换。
- 第一因子主要是求职者的工作能力和内在素质，第二公共因子主要表现求职者的经验和适应性，公司可以选择两者得分都比较高的应聘者。
- 对于第一因子来说，显然10、11号得分较高，但在第二因子上得分最低。
- 8、39、40号应聘者对第一因子和第二因子都得分较多，综合能力较强。而47、48号最差。
- 41、42号在求职者经验和适应性上较好，而37、38号在工作能力和内在素质方面较优秀。

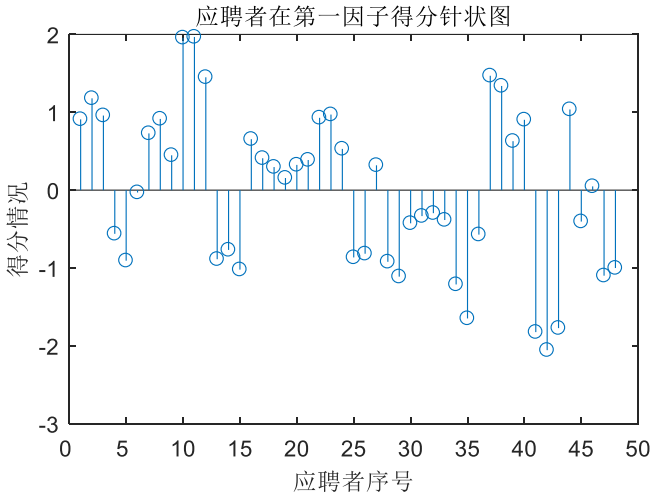
```
>> plot(F(:,1),F(:,2),'+')
>> gname(rownames)   %通过鼠标标记名称
```

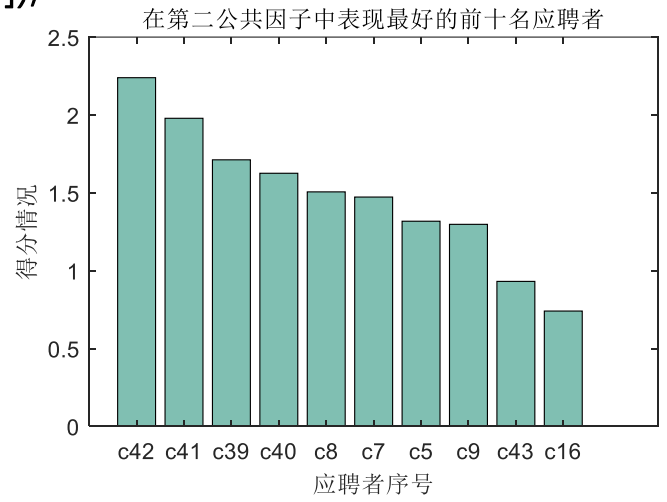
7. 典型案例分析

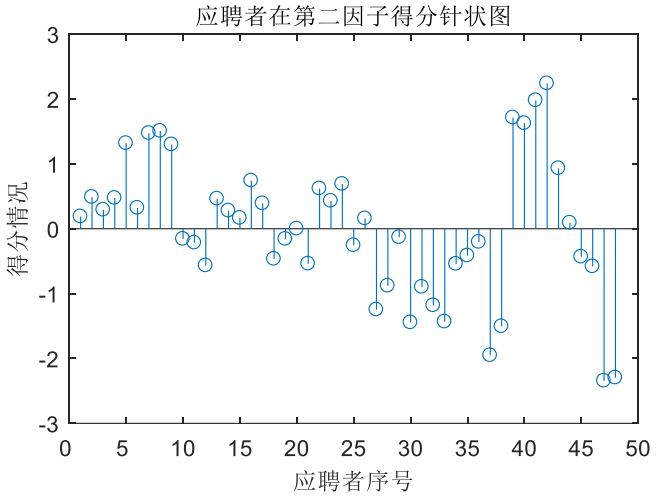
```
>> [F1,ind] = sort(F(:,1),'descend');
>> Fname = rownames(ind);
>> bar(F1(1:10))  %前十名
>> set(gca,'xticklabel',Fname(1:10))
>> colormap(summer)
>> alpha(0.5)
>> stem(F(:,1))  %绘制针状图
>> set(gca,'XTick',[0:5:50],'XTickLabel',[0:5:50]);
```

- 从图中可以看出11、10号在第一公共因子中表现最好；
- 而42、41号在第二公共因子中表现最好；
- 公司可根据公司要求或岗位要求选额应聘者。









7. 典型案例分析

例2：各主要城市家庭消费情况，进行因子分析。其中变量为Food、Clothing、Residence、Household、Transport、Education、HealthCare、Others。

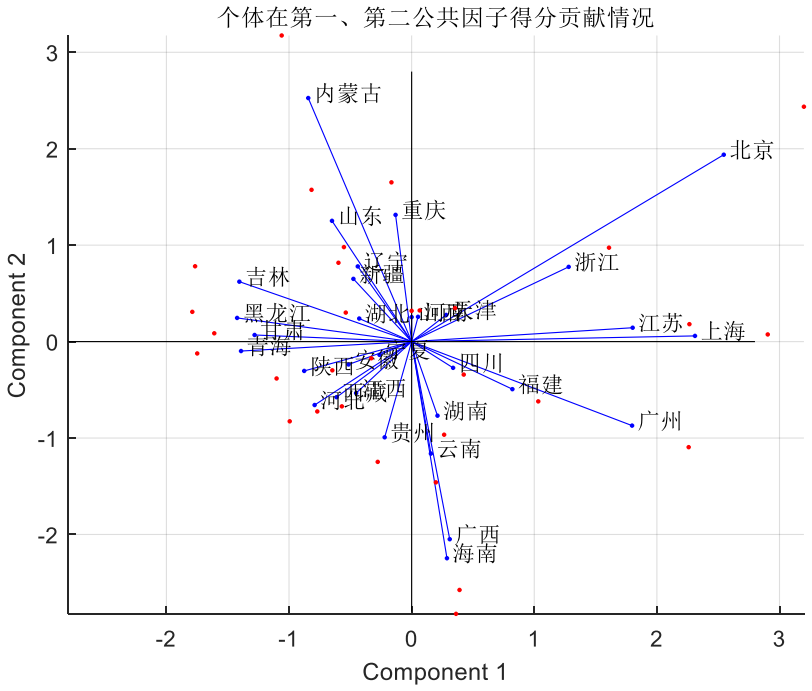
```
[consume,text] = xlsread('fendiq2013e.xlsx');
[lambda,psi,T,stats,F]=factoran(consume,3);
n = size(consume,2);  %取变量个数
res = cell(n,3+1);  %组合载荷矩阵
res(:,1) = text(1,2:end);  %第一列存变量名称
res(:,2:end) = num2cell(lambda);
```

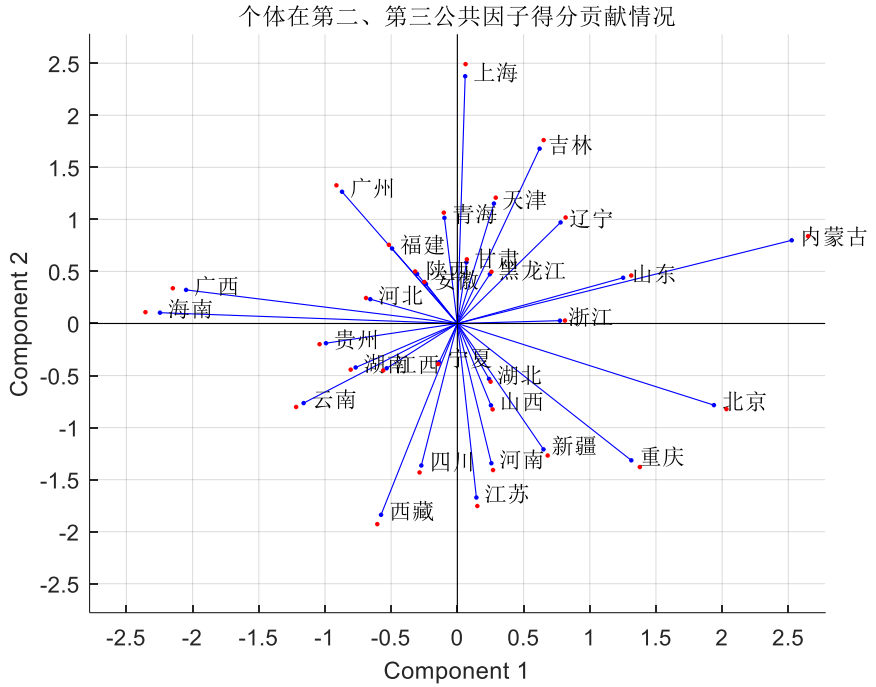
```
res =
  8×4 cell 数组
    'Food'          [0.8061]    [0.1343]    [0.2246]
    'Clothing'      [0.1909]    [0.9790]    [0.0097]
    'Residence'     [0.5533]    [0.2275]    [0.7982]
    'Household'     [0.8219]    [0.3648]    [0.1720]
    'Transport'     [0.7927]    [0.2159]    [0.4520]
    'Education'     [0.8864]    [0.2058]    [0.2665]
    'HealthCare'    [0.2163]    [0.6644]    [0.2943]
    'Others'        [0.6833]    [0.5107]    [0.3465]
```

- 第一公共因子与Food、Household、 Transport、Education和Others载荷因子系数较大，可归结于家庭日常消费。
- 第二公共因子与Clothing和HealthCare载荷因子系数较大，可归结于家庭健康支出；
- 第三公共因子与Residence载荷因子系数较大，可归结于住房支出。

7. 典型案例分析

```
rownames = text(2:end,1);
biplot(F(:,1:2),'Score',F(:,1:2),'VarLabels',rownames)
title('个体在第一、第二公共因子得分贡献情况')
biplot(F(:,2:3),'Score',F(:,2:3),'VarLabels',rownames)
title('个体在第二、第三公共因子得分贡献情况')
```

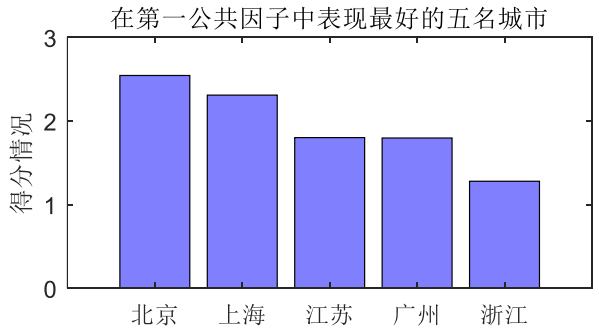


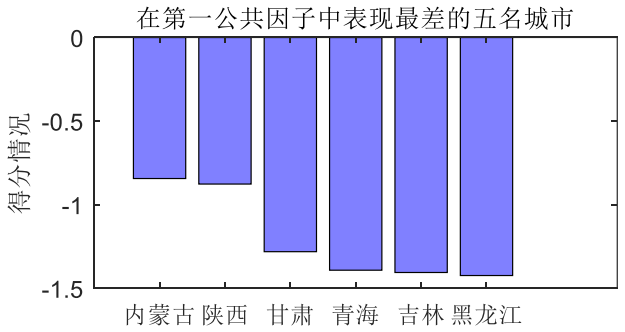


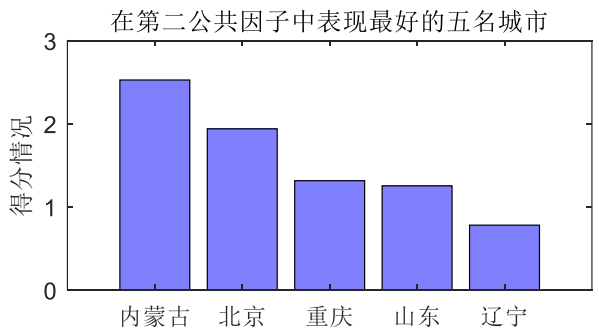
从图中可以看出：北京、上海、江苏、浙江、在第一因子上贡献较大，可以理解家庭一般消费较高；而内蒙古、北京、重庆在第二因子上贡献较大，家庭健康支出较大；而上海、吉林、广州、天津等地住房支出较大。

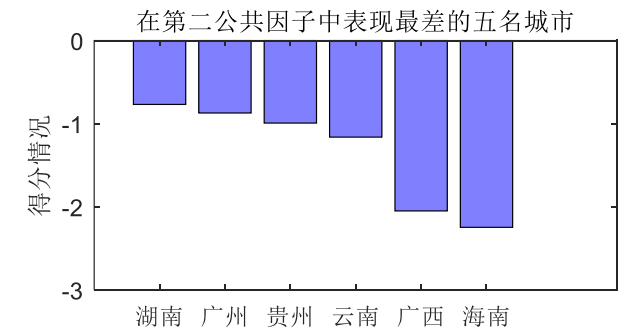
7. 典型案例分析

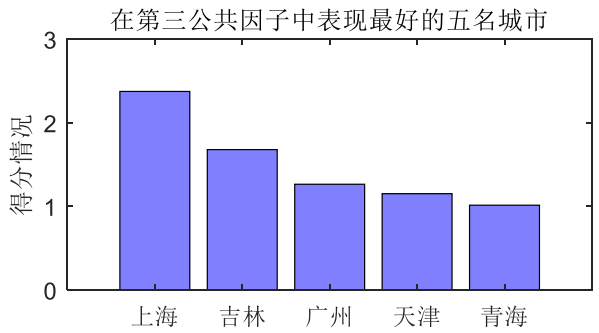
```
subplot(3,2,1)
[F1,ind] = sort(F(:,1),'descend');
Fname = rownames(ind);
bar(F1(1:5))
set(gca,'xticklabel',Fname(1:5))
colormap(spring)
alpha(0.5)
title('在第一公共因子中表现最好的五名城市')
subplot(3,2,2)
bar(F1(31-5:end))
set(gca,'xticklabel',Fname(31-5:end))
alpha(0.5)
title('在第一公共因子中表现最差的五名城市')
```

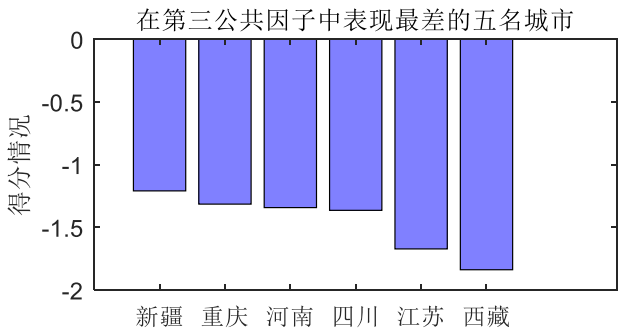













01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听