浅谈主成分分析与因子分析

yiliyiyi

主成分分析与因子分析有相似之处，也有明显区别，以下是对主成分分析和因子分析进行的一个简单比较。

主成分分析
基本思想
主成分几何意义及求解
主成分分析优缺点
分析步骤
因子分析
基本思想
与主成分分析的区别和联系
分析步骤

【1】主成分分析——基本思想

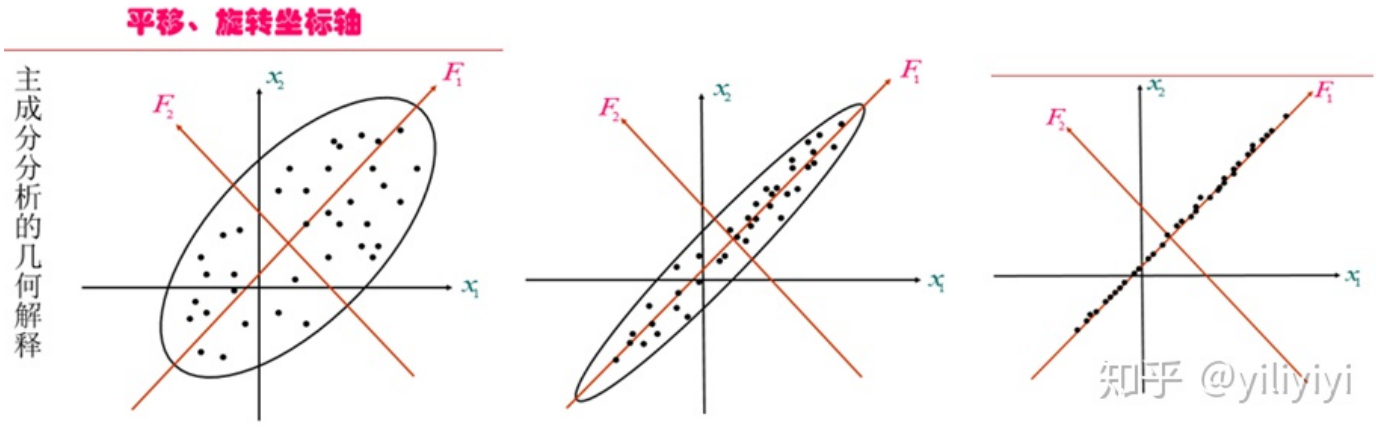
主成分分析（Principal Component Analysis，PCA）通过将原始变量转换为原始变量的线性组合（主成分），在保留主要信息的基础上，达到简化和降维的目的。

主成分与原始变量之间的关系：

主成分是原始变量的线性组合
主成分的数量相对于原始数量更少
主成分保留了原始变量的大部分信息
主成分之间相互独立

【2】主成分分析——几何意义及求解

通过旋转变换，将分布在x1，x2坐标轴上的原始数据，转换到F1,F2坐标轴表示的坐标系上，使得数据在F1轴上离散程度最大，此时，可以忽略F2轴，仅通过F1轴就可以表示数据的大部分信息，从而达到降维的目的。



不同的线性变换，得到的Fi统计特性不同，为得到较好的效果，我们希望主成分之间相互独立，同时方差尽可能得大，即

$$\max \mathrm{Var}(F_i)=\mathrm{Var}(u_i^T X)=u_i^T Var(X)u_i, \text{记为 } u_i^T \sum u_i$$

$$\text{s.t } u^T u = 1, \text{即} u_{i1}^2 + u_{i2}^2 + \cdots u_{ip}^2 (i = 1,2,\cdots p)$$

$$F_i \text{与} F_j \text{相互无关} \ (i \neq j, i, j = 1,2,\cdots p)$$

求解以上公式，得

$$\sum u_i = \lambda u_i, \text{ 带入公式, } \mathrm{Var}(F_i)=u_i^T \sum u_i = u_i^T \lambda u_i = \lambda$$

因此，我们只需要**对协方差矩阵∑求特征值λ及特征向量ui**，即可构成主成分分析的解。

由此可知，主成分分析是把p个随机变量的方差分解为p个不相关的随机变量的方差和，使得第一个主成分的方差达到最大，其贡献率等于其方差在全部主成分方差中的占比。

主成分分析的一个关键问题是：**主成分的个数选多少个比较合适?**

有3个主要的衡量标准：

保留的主成分使得方差贡献率达到**80%**以上
保留的主成分的**方差（特征值）大于1**
Cattell**碎石检验**绘制了关于各主成分及其特征值的图形，我们只需要保留图形中变化最大之处以上的主成分即可

【3】主成分分析——优缺点

优点

不要求数据呈正态分布，主成分就是按数据离散程度最大的方向对基组进行旋转，这特性扩展了其应用范围，比如，用于人脸识别

通过对原始变量进行综合与简化，可以客观地确定各个指标的权重，避免主观判断的随意性

缺点

主成分分析适用于变量间有较强相关性的数据，若原始数据相关性弱，则起不到很好的降维作用

降维后，存在少量信息丢失，不可能包含100%原始数据

原始数据经过标准化处理之后，含义会发生变化，且主成分的解释含义较原始数据比较模糊

假设标准化后的原始变量间存在多重共线性，即原始变量之间存在不可忽视的信息重叠，主成分分析不能有效剔除信息重叠

【4】主成分分析——分析步骤

主成分分析步骤：

选取初始变量

根据初始变量特性选择使用协方差矩阵还是相关矩阵来求主成分

计算协方差矩阵或相关矩阵的特征值和特征向量

确定主成分个数

对主成分做经济解释，主成分的经济意义由各线性组合中权重较大的几个指标来确定

案例：对中国各个地区的经济水平影响因素的分析。

分析前，先进行相关性检验，变量之间存在较强相关性，才能使用主成分分析方法。

	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值
GDP	1.0000000	0.2667649	0.9505842	0.1903841	0.61723824	-0.2725597	-0.26363059	0.8737437
居民消费水平	0.2667649	1.0000000	0.4261367	0.7177550	-0.15101233	-0.2351387	-0.59272738	0.3630990
固定资产投资	0.9505842	0.4261367	1.0000000	0.3993680	0.43062284	-0.2804858	-0.35905170	0.7918605
职工平均工资	0.1903841	0.7177550	0.3993680	1.0000000	-0.35597525	-0.1343476	-0.53838854	0.1037891
货物周转量	0.6172382	-0.1510123	0.4306228	-0.3559752	1.00000000	-0.2531754	0.02172214	0.6585765
居民消费价格指数	-0.2725597	-0.2351387	-0.2804858	-0.1343476	-0.25317541	1.0000000	0.76283800	-0.1252172
商品零售价格指数	-0.2636306	-0.5927274	-0.3590517	-0.5383885	0.02172214	0.7628380	1.0000000	-0.1920742
工业总产值	0.8737437	0.3630990	0.7918605	0.1037891	0.65857654	-0.1252172	-0.19207416	1.0000000

具体步骤：

1）获取初始数据，统一量纲，将数据进行标准化处理

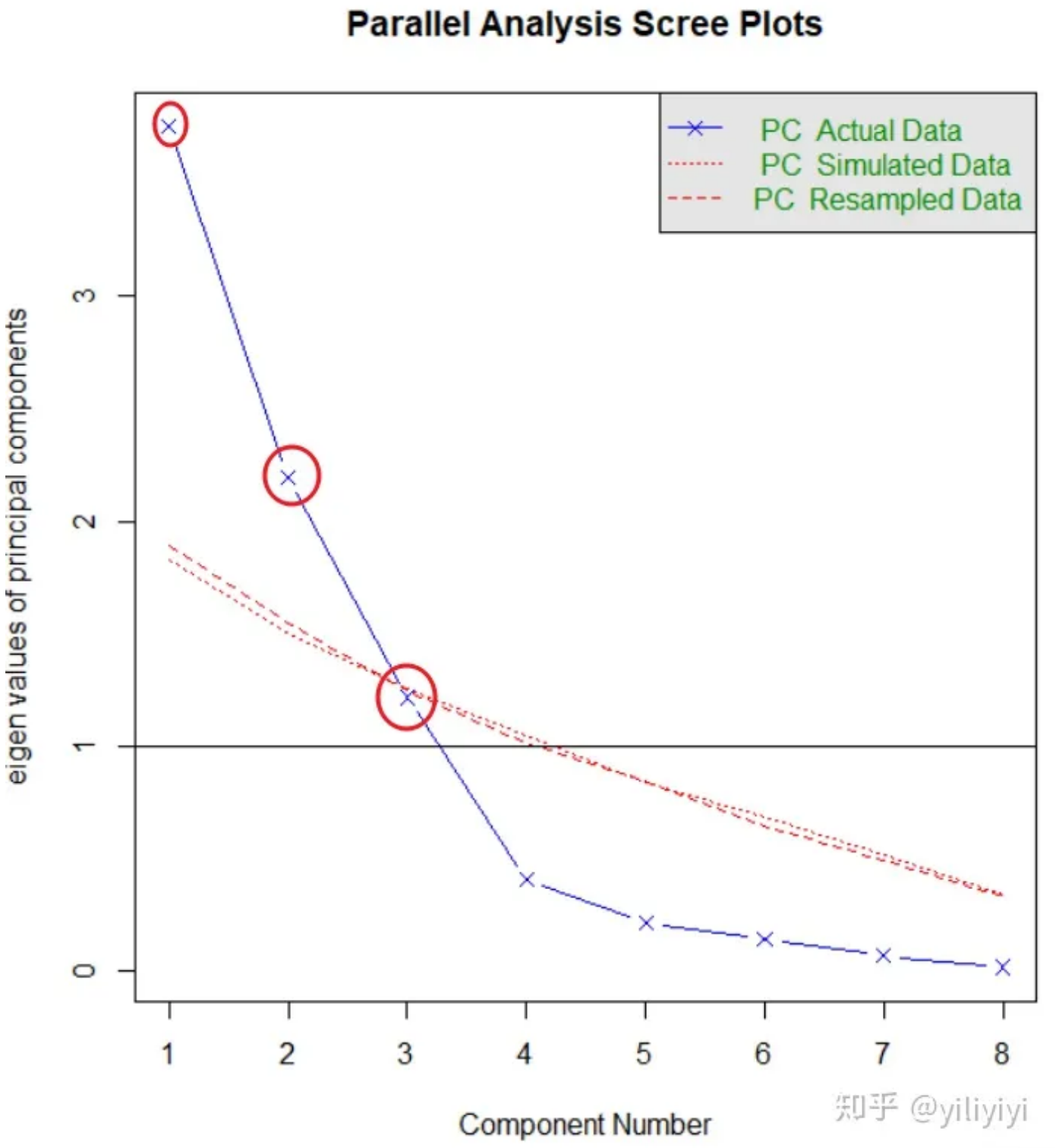
```
#将数据进行标准化
data1_PCA<-scale(data1_PCA)
```

2）计算相关系数矩阵，求得特征值和特征向量

```
sigma<-cor(data1_PCA)#求相关系数矩阵
e<-eigen(sigma) #计算特征值及特征向量
e$values #特征值
e$vectors #特征向量
```


3）确定主成分个数

```
library(psych)
fa.parallel(data1_PCA,fa="pc") #画碎石图
abline(h=1)
```


Parallel Analysis Scree Plots
PC Actual Data
PC Simulated Data
PC Resampled Data
eigen values of principal components
0
1
2
3
1
2
3
4
5
6
7
8
Component Number
知乎 @yiliyiyi

4）提取主成分

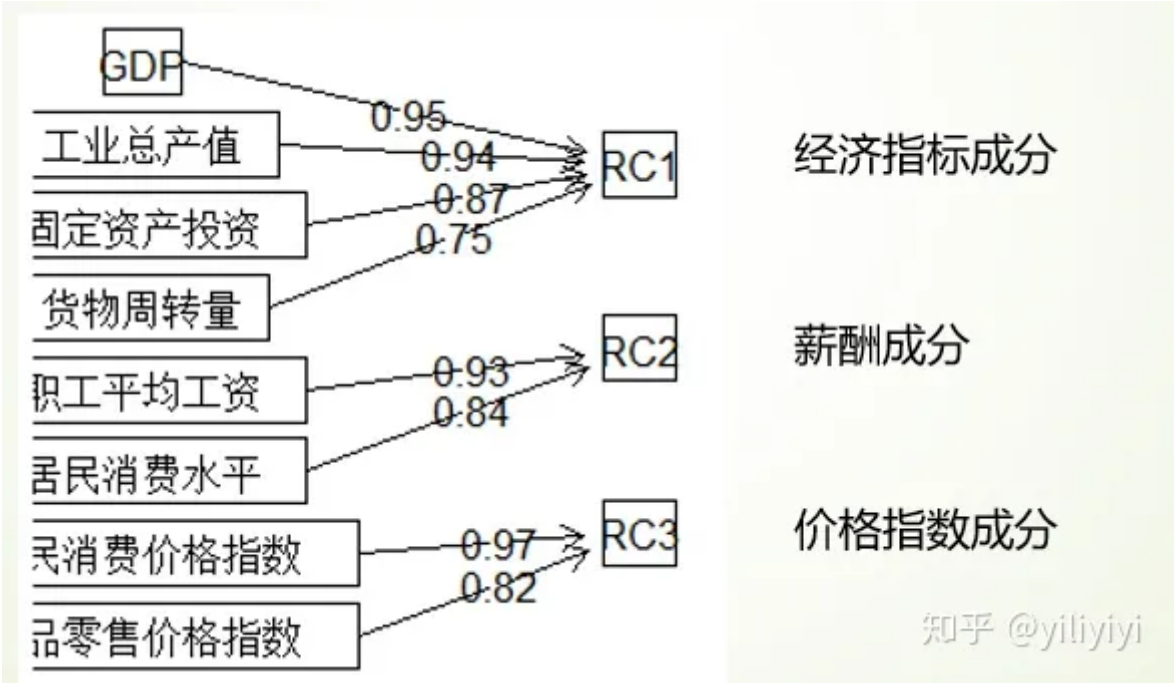
```
> fit<-principal(data1_PCA,
+                nfactors=3,
+                rotate="varimax",# 方差最大旋转
+                scores=T)
> fit
Principal Components Analysis
Call: principal(r = data1_PCA, nfactors = 3, rotate = "varimax",
    scores = T)
Standardized loadings (pattern matrix) based upon correlation matrix
     RC1   RC2   RC3   h2    u2 com
x1  0.95  0.13 -0.13 0.94 0.055 1.1
x2  0.22  0.84 -0.21 0.80 0.201 1.3
x3  0.87  0.35 -0.14 0.90 0.098 1.4
x4  0.05  0.93 -0.11 0.87 0.125 1.0
x5  0.75 -0.51 -0.19 0.86 0.143 1.9
x6 -0.13 -0.01  0.97 0.96 0.043 1.0
x7 -0.10 -0.49  0.82 0.93 0.071 1.7
x8  0.94  0.11 -0.02 0.90 0.097 1.0

                       RC1  RC2  RC3
SS loadings           3.21 2.22 1.74
Proportion Var        0.40 0.28 0.22
Cumulative Var        0.40 0.68 0.90
Proportion Explained  0.45 0.31 0.24
Cumulative Proportion 0.45 0.76 1.00
```


$$
\begin{aligned}
&RC1 \\
&= 0.95 \times GDP + 0.22 \times \text{居民消费水平} + 0.87 \times \text{固定资产投资} + 0.05 \\
&\times \text{职工平均水平} + 0.75 \times \text{货物周转率} - 0.13 \times \text{居民消费价格指数} - 0.1 \\
&\times \text{商品零售价格指数} + 0.94 \times \text{工业总产值} \\
&RC2 \\
&= 0.13 \times GDP + 0.84 \times \text{居民消费水平} + 0.35 \times \text{固定资产投资} + 0.93 \\
&\times \text{职工平均水平} - 0.51 \times \text{货物周转率} - 0.01 \times \text{居民消费价格指数} - 0.49 \\
&\times \text{商品零售价格指数} + 0.11 \times \text{工业总产值} \\
&RC3 \\
&= -0.13 \times GDP - 0.21 \times \text{居民消费水平} - 0.14 \times \text{固定资产投资} - 0.11 \\
&\times \text{职工平均水平} - 0.19 \times \text{货物周转率} + 0.97 \times \text{居民消费价格指数} + 0.82 \\
&\times \text{商品零售价格指数} - 0.02 \times \text{工业总产值}
\end{aligned}
$$

得到主成分的式子


```
# 绘制主成分分析的载荷矩阵，查看各个主成分的综合构成变量
fa.diagram(fit,digits=2)
```



5）将原数据分别按第一，第二，第三主成分得分排序，观察各地区主要受哪个主成分影响

```
data_PCA<-data_PCA[order(data_PCA$RC1,decreasing=T),]
View(head(data_PCA))
```

第一主成分

	地区	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值	RC1
15	山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69	2.1179261
10	江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64	2.0342283
19	广东	5381.72	2699	1639.83	8250	656.5	114.0	111.6	1396.35	1.4815138
3	河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85	1.2278905
22	四川	3534.00	1261	822.54	4645	902.3	118.5	117.0	1431.81	1.1093449
16	河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92	1.0710606
6	辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114.0	1840.55	0.9590441

```
data_PCA<-data_PCA[order(data_PCA$RC2,decreasing=T),]
View(head(data_PCA))
```

第二主成分

	地区	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值	RC1	RC2
9	上海	2462.57	5343	996.48	9279	207.4	118.7	113.0	1642.95	0.6054180	3.660729136
19	广东	5381.72	2699	1639.83	8250	656.5	114.0	111.6	1396.35	1.4815138	1.685617126
1	北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43	-0.3975205	1.586773980
2	天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51	-0.8887078	0.983433231
11	浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59	0.6513345	0.748748291
25	西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57	-1.5933403	0.629936109
13	福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67	-0.2701307	0.271317984
10	江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64	2.0342283	0.2385663


```
data_PCA<-data_PCA[order(data_PCA$RC3,decreasing=T),]
View(head(data_PCA))
```

第三主成分

	地区	GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品零售价格指数	工业总产值	RC1	RC2	RC3
24	云南	1206.68	1261	334.00	5149	310.4	121.3	118.1	716.65	-0.2766098	0.043729737	2.04581025
23	贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72	-0.7656986	-0.349818271	1.66359533
17	湖北	2391.42	1527	571.68	4685	849.0	120.0	116.6	1220.72	0.6025424	-0.300458005	1.26538081
30	新疆	834.57	1469	376.96	5348	339.0	119.7	116.7	428.76	-0.5763933	0.114782634	1.14055989
27	甘肃	553.35	1007	114.81	5493	507.0	119.8	116.5	468.79	-0.7173669	-0.222264910	1.01881286
22	四川	3534.00	1261	822.54	4645	902.3	118.5	117.0	1431.81	1.1093449	-0.525800739	0.97724298
26	陕西	1000.03	1208	300.27	4396	500.9	119.0	117.0	600.98	-0.4133576	-0.535857292	0.88581687
9	上海	2462.57	5343	996.48	9279	207.4	118.7	113.0	1642.95	0.6054180	3.660729136	0.82507775

【5】因子分析——基本思想

因子分析（Factor Analysis，FA）是一种数据简化技术，通过研究众多变量之间的内部依赖关系，探求观测数据的基本结构，并用少数几个假想变量（因子）来表示原始数据。

因子能够反映众多原始变量的主要信息。

因子的特点：

因子个数远远少于原始变量个数
因子并非原始变量的简单取舍，而是一种新的综合

因子之间没有线性关系
因子具有明确解释性，可以最大限度地发挥专业分析的作用

【6】因子分析——例子

在市场调查中我们收集了食品的五项指标（x1~x5）:味道、价格、风味、是否快餐、能量，经过因子分析，我们发现了：

x1 = 0.02 * z1 + 0.99 * z2 + e1
x2 = 0.94 * z1 - 0.01 * z2 + e2
x3 = 0.13* z1 + 0.98 * z2 + e3
x4 = 0.84 * z1 + 0.42 * z2 + e4
x5 = 0.97 * z1 - 0.02 * z2 + e1

（数字代表实际变量间的相关系数，值越大，相关性越大）

第一个公因子**z1**主要与价格、是否快餐、能量有关，代表“价格与营养”；

第二个公因子**z2**主要与味道、风味有关，代表“口味”；

e1-5是特殊因子，是公因子中无法解释的，在分析中一般略去。

【7】因子分析——分析步骤

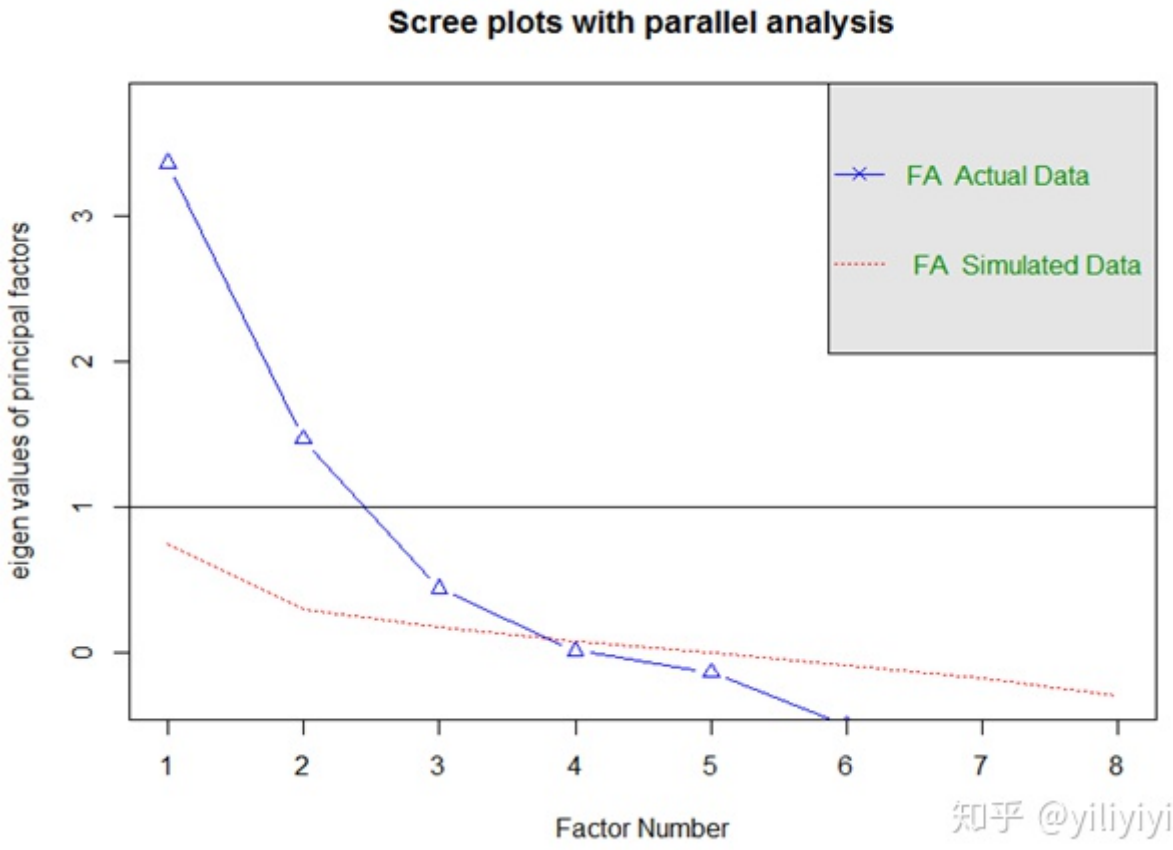
因子分析步骤：

选择分析变量
计算原始变量的相关系数矩阵
提取公因子
取方差（特征值）大于0的因子
因子的累积方差贡献率达到80%
因子旋转
因子的实际意义更容易解释
计算因子得分

案例分析步骤：

1）数据标准化处理，计算相关系数矩阵，求特征值及特征向量

2）确定因子个数



3）提取公共因子，使用fa()函数获得相应结果

```
fa<-fa(sigma,nfactors=3,rotate="varimax",fm="pa")
fa
```

使用正交旋转提取公共因子

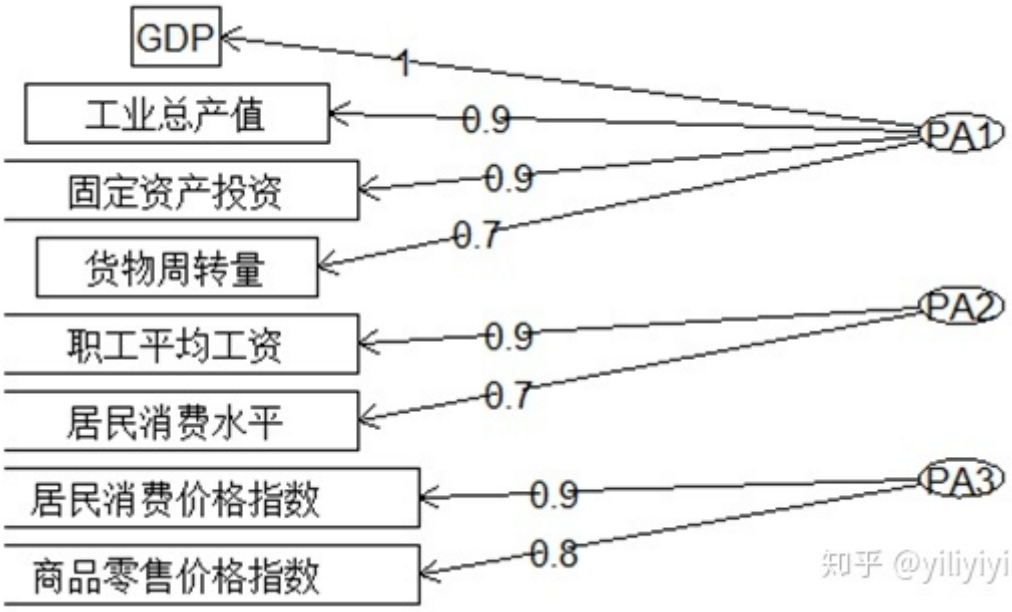
```
Factor Analysis using method =  pa
Call: fa(r = sigma, nfactors = 3, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1   PA2   PA3   h2    u2 com
x1  0.96  0.14 -0.13 0.97 0.035 1.1
x2  0.20  0.75 -0.21 0.65 0.353 1.3
x3  0.85  0.36 -0.14 0.87 0.126 1.4
x4  0.04  0.91 -0.11 0.84 0.160 1.0
x5  0.71 -0.45 -0.17 0.73 0.269 1.8
x6 -0.14 -0.03  0.93 0.88 0.119 1.0
x7 -0.10 -0.51  0.79 0.90 0.100 1.7
x8  0.91  0.11 -0.03 0.85 0.150 1.0

                       PA1  PA2  PA3
SS loadings           3.06 2.01 1.61
Proportion Var        0.38 0.25 0.20
Cumulative Var        0.38 0.63 0.84
Proportion Explained  0.46 0.30 0.24
Cumulative Proportion 0.46 0.76 1.00
```

结果显示因子变得更好解释了，**x1（GDP），x3（固定资产投资），x5（货物周转量），x8（工业总产值）在第一因子上载荷较大，x2（居民消费水平），x4（职工平均工资）在第二因子上载荷较大，x6（居民消费价格指数），x7（商品零售价格指数）在第三因子上载荷较大。**

4）使用fa.diagram()函数绘制正交旋转结果的图形

```
fa.diagram(fa,simple=TRUE)
```



【8】因子分析与主成分分析的比较

区别:

（1）因子分析需要构造因子模型，着重要求新变量**具有实际的意义**，能解释原始变量间的内在结构。

（2）主成分分析仅仅是变量变换，是原始变量的线性组合表示新的综合变量，**强调新变量贡献了多大比例的方差**，不关心新变量是否有明确的实际意义。

联系:

两者都是**降维和信息浓缩**的方法。

生成的新变量均代表了原始变量的大部分信息且互相独立，都可以用于后续的回归分析、判别分析、聚类分析等等。