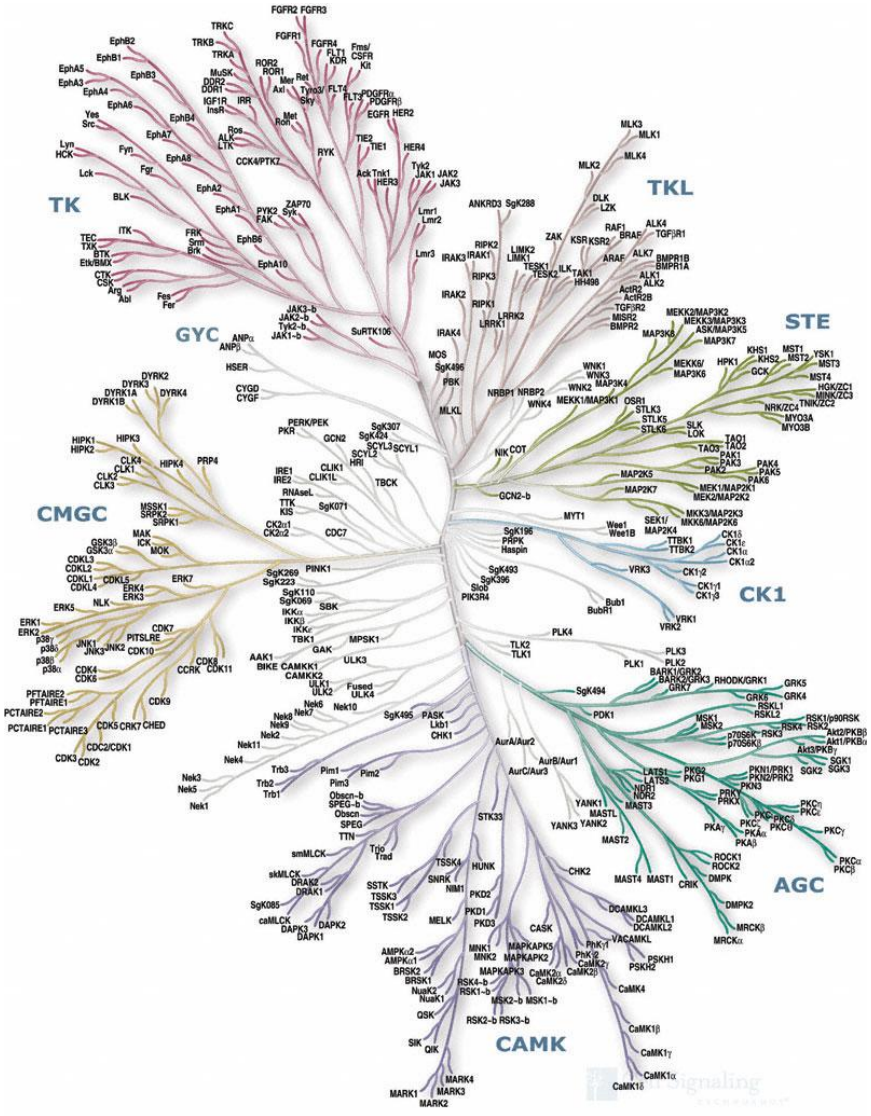
DATAGURU 炼数成金

数据分析、展现与R语言 第15周



2013.05.17

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

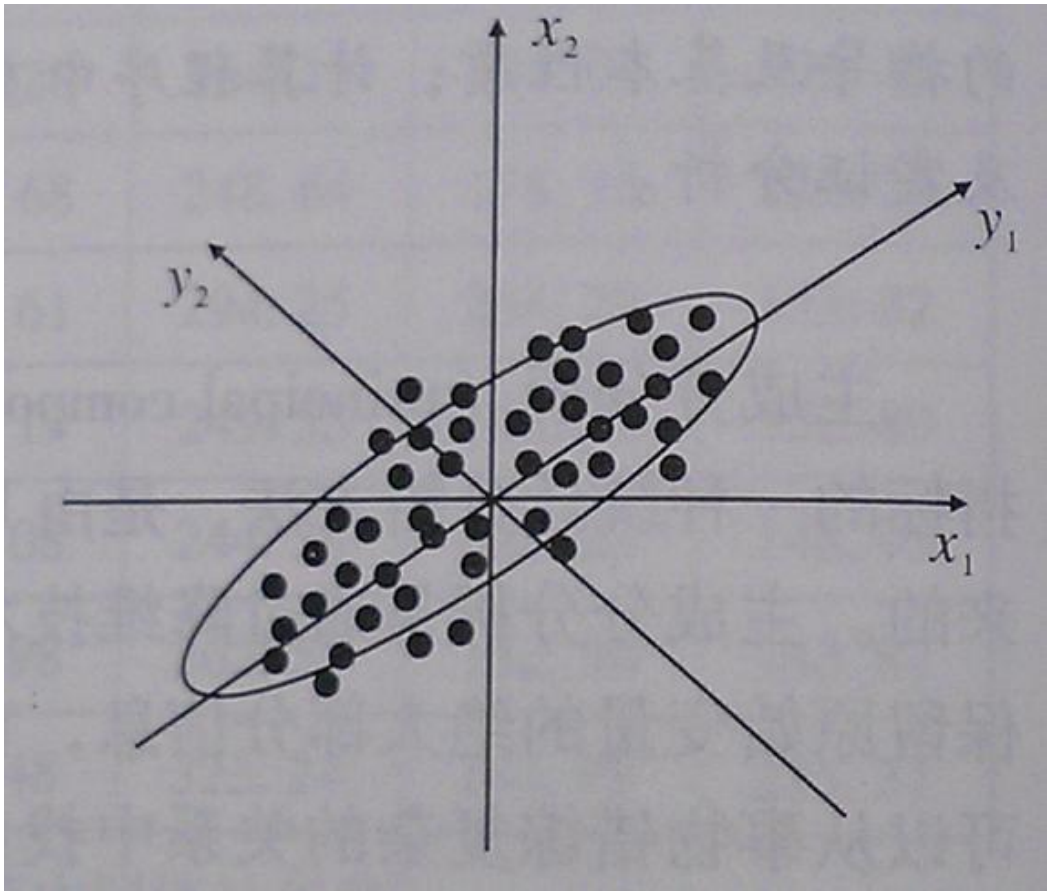
主成分分析

- Pearson于1901年提出，再由Hotelling（1933）加以发展的一种多变量统计方法
- 通过析取主成分显出最大的个别差异，也用来削减回归分析和聚类分析中变量的数目
- 可以使用样本协方差矩阵或相关系数矩阵作为出发点进行分析
- 成分的保留：Kaiser主张（1960）将特征值小于1的成分放弃，只保留特征值大于1的成分
- 如果能用不超过3-5个成分就能解释变异的80%，就算是成功

主成分分析

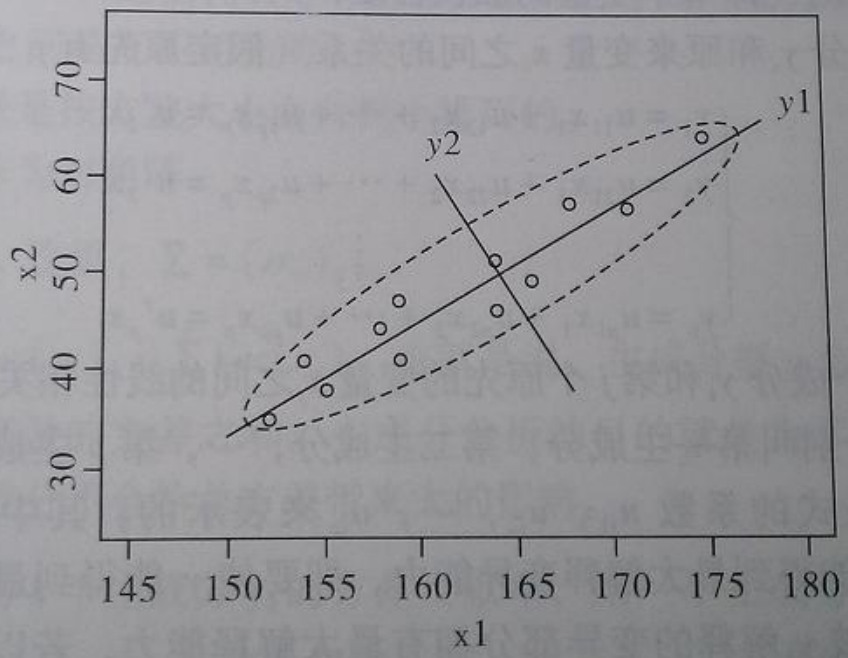
- 通过对原始变量进行线性组合，得到优化的指标
- 把原先多个指标的计算降维为少量几个经过优化指标的计算（占去绝大部分份额）
- 基本思想：**设法将原先众多具有一定相关性的指标，重新组合为一组新的互相独立的综合指标，并代替原先的指标**

主成分分析的直观几何意义



例子

```
> x1 = c(171,175,159,155,152,158,154,164,168,166,159,164)
> x2 = c(57,64,41,38,35,44,41,51,57,49,47,46)
> plot(x1,x2,xlim = c(145,180),ylim = c(25,75))
> lines(c(150,178),c(33,66));text(180,68,"y1")
> lines(c(161,168),c(60,38));text(161,63,"y2")
```



主成分分析

- 通过对原始变量进行线性组合，得到优化的指标
- 把原先多个指标的计算降维为少量几个经过优化指标的计算（占去绝大部分份额）
- 基本思想：**设法将原先众多具有一定相关性的指标，重新组合为一组新的互相独立的综合指标，并代替原先的指标**

主成分分析的数学原理

- 薛毅书p499

princomp( )函数

- 薛毅书P506

例子

- 薛毅书P508

例子：求相关矩阵特征值

- 薛毅书p487

```
> PCA=princomp(X,cor=T)
> PCA
Call:
princomp(x = X, cor = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
2.2556395 1.1632889 0.7567221 0.6376603 0.5278638 0.3502837 0.3063912
   Comp.8
0.2905094

 8  variables and  31 observations.
> PCA$loadings
```

例子：求主成分载荷

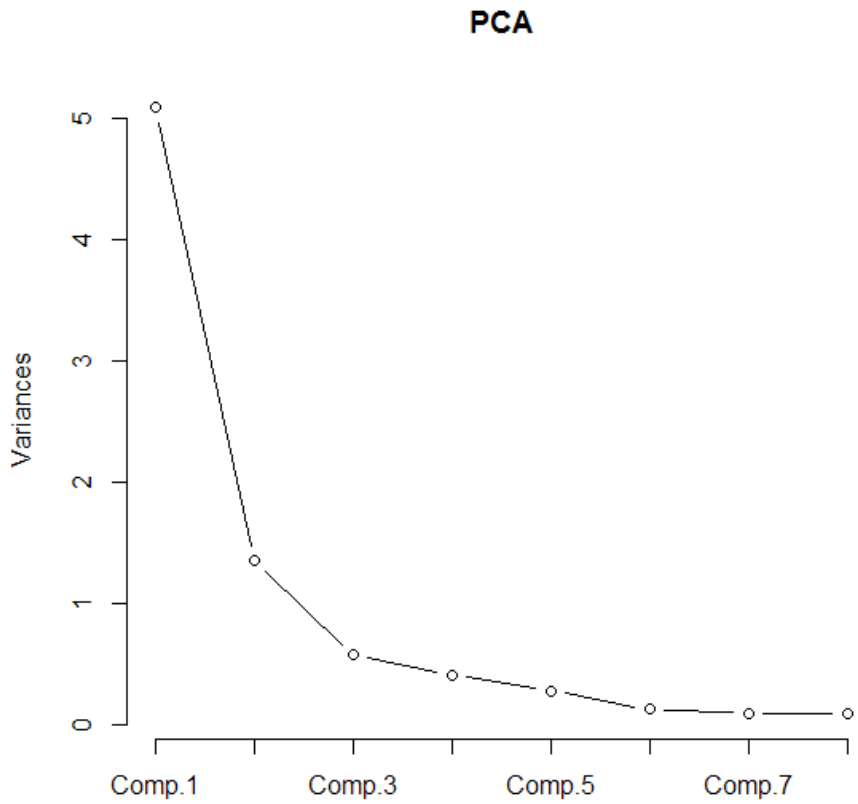
```
> PCA$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
x1 -0.399          0.416  0.214 -0.217         -0.280  0.693
x2 -0.132  0.749  0.339  0.157  0.523
x3 -0.375         -0.444  0.544         -0.562 -0.161 -0.121
x4 -0.320  0.346 -0.475 -0.657                        0.335
x5 -0.388 -0.231  0.282 -0.364  0.210 -0.109 -0.566 -0.456
x6 -0.406         -0.308  0.234          0.795        -0.229
x7 -0.327 -0.495                 0.582          0.514  0.182
x8 -0.396          0.338 -0.116 -0.538 -0.127  0.551 -0.312

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
> 
```

例子：画碎石图确定主成分

```
> screeplot(PCA,type="lines")
```



2013.05.17

例子：主成分得分-相当于predict( )

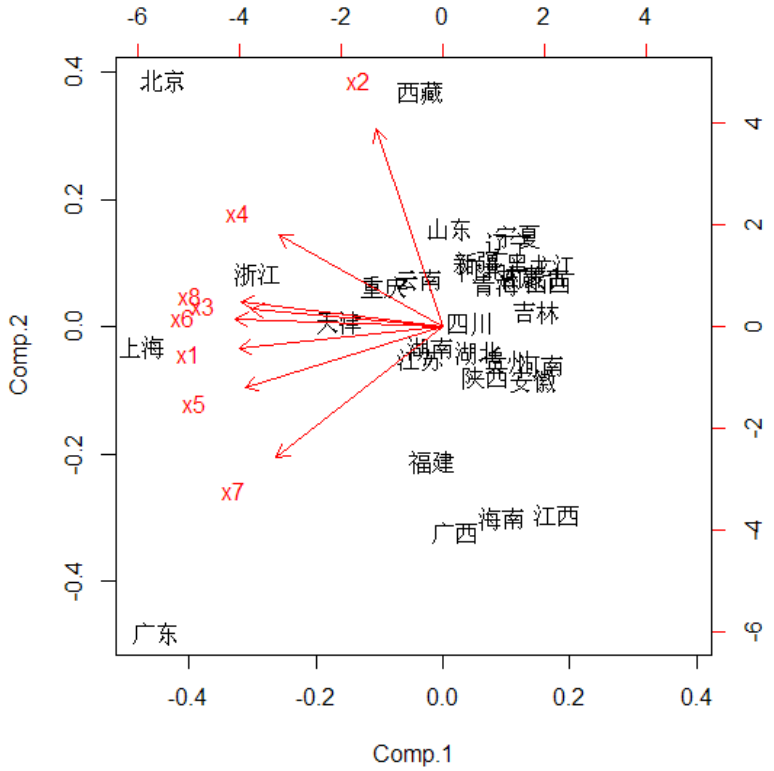
```
> PCA$score
          Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6       Comp
北京    -5.5068881  2.51368747 -0.77052784 -0.34499076 -0.48456544  0.73526042  0.142820
天津    -2.0391525  0.04696816 -0.83866069  0.84294280 -0.23905123 -0.36965072  0.438523
河北     0.7647412  0.58939950 -0.63809135 -0.40004970  0.32727289  0.02069393 -0.108875
山西     2.1042564  0.45779593 -0.29703426 -0.21190291 -0.16277216 -0.21169100  0.366478
内蒙古   1.8368141  0.51548336  0.14950198 -0.09308007  0.19160016  0.13617218 -0.010774
辽宁     1.3232250  0.85489639 -0.05242441 -0.56123733  0.43320901  0.10274050 -0.199007
吉林     1.8750798  0.14967842 -0.02016675 -0.28215689  0.45133137  0.36714488 -0.038957
黑龙江   1.9411347  0.64393452 -0.25831381 -0.84845435  0.37526772 -0.08315897 -0.086928
上海    -5.9397413 -0.19531943  0.09487298  1.07297060 -0.60041434 -0.09156896  0.065314
江苏    -0.4173225 -0.31874237 -0.21558331  0.85952388 -0.39145266 -0.42795347 -0.199799
浙江    -3.6407775  0.54489693 -0.77999195 -0.68115276  0.19016696 -0.41219749 -0.509992
安徽     1.8169295 -0.53363884  0.33919645  0.64984975 -0.04126297  0.49854622 -0.528359
福建    -0.1976522 -1.36531052  1.29563886  0.23492502  0.12124119 -0.19422385 -0.489680
江西     2.2557443 -1.90231267  0.08063848  0.33710287  0.09292676  0.00724231  0.403240
山东     0.1360728  0.99920233 -0.34711211  0.92327895  0.53080961 -0.29793692 -0.123394
河南     1.9613045 -0.39761168 -0.20088982 -0.23566368  0.30206294 -0.49375497  0.224554
湖北     0.7167909 -0.25396283 -0.03587219  0.29134913  0.81888494  0.66366667  0.443813
湖南    -0.2318682 -0.20807224 -0.01570997  0.47810304  0.47020168  0.52874605  0.065600
广东    -5.6676807 -3.11520051  0.51838684 -1.53211943  0.90023275 -0.21946848  0.129630
广西     0.2480444 -2.09427753 -0.03594804  0.29165788 -0.04979176  0.44518529  0.146873
海南     1.1715466 -1.94839070  0.44408295 -0.60362333 -1.85888240  0.34575391 -0.284233
重庆    -1.1363085  0.41532157  0.13949690  0.63934241  0.56936685  0.28511495 -0.703780
四川     0.5349560  0.03922716  0.17181794  0.42545284  0.12711946  0.30779276  0.254154
```

例子：结果解释

- Z1：日常必需消费开支
- Z2：衣着和居住

例子：成分图

```
> biplot(PCA,choices=1:2,scale=1)
```



例子：聚类

```
> kmeans(PCA$score[,1:2],5)
K-means clustering with 5 clusters of sizes 7, 4, 10, 6, 4

Cluster means:
      Comp.1      Comp.2
1  0.6787254  0.27889640
2 -5.1887719 -0.06298388
3  1.7232375  0.27928061
4 -0.7843413  0.46952434
5  0.8694208 -1.82757285

Clustering vector:
  北京   天津   河北   山西 内蒙古   辽宁   吉林 黑龙江   上海   江苏
     2      4      1      3      3      3      3      3      2      4
  浙江   安徽   福建   江西   山东   河南   湖北   湖南   广东   广西
     2      3      5      5      1      3      1      4      2      5
  海南   重庆   四川   贵州   云南   西藏   陕西   甘肃   青海   宁夏
     5      4      1      3      4      4      1      3      1      3
  新疆
     1
```

主成分回归

- 薛毅书P516

因子分析

- 降维的一种方法，是主成分分析的推广和发展
- 是用于分析隐藏在表面现象背后的因子作用的统计模型。试图用最少个数的不可测的公共因子的线性函数与特殊因子之和来描述原来观测的每一分量
- 例子：各科学习成绩（数学能力，语言能力，运动能力等）
- 例子：生活满意度（工作满意度，家庭满意度）
- 例子：薛毅书P522

因子分析的主要用途

- 减少分析变量个数
- 通过对变量间相关关系的探测，将原始变量分组，即将相关性高的变量分为一组，用共性因子来代替该变量
- 使问题背后的业务因素的意义更加清晰呈现

与主成分分析的区别

- 主成分分析侧重“变异量”，通过转换原始变量为新的组合变量使到数据的“变异量”最大，从而能把样本个体之间的差异最大化，但得出来的主成分往往从业务场景的角度难以解释
- 因子分析更重视相关变量的“共变异量”，组合的是相关性较强的原始变量，目的是找到在背后起作用的少量关键因子，因子分析的结果往往更容易用业务知识去加以解释

因子分析使用了复杂的数学手段

- 比主成分分析更加复杂的数学模型
- 求解模型的方法：主成分法，主因子法，极大似然法
- 结果还可以通过因子旋转，使到业务意义更加明显

数学模型

1．数学模型

设 $X=(X_1,X_2,\cdots,X_p)^T$ 是可观测的随机向量，且

$$E(X)=\mu=(\mu_1,\mu_2,\cdots,\mu_p)^T,\qquad \mathrm{Var}(X)=\Sigma=(\sigma_{ij})_{p\times p}.$$

因子分析的一般模型为

$$\begin{cases} X_1-\mu_1=a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1\\ X_2-\mu_2=a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2\\ \quad\vdots\\ X_p-\mu_p=a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases}$$

数学模型

$$X = \mu + AF + \varepsilon, \tag{9.22}$$

$$E(F) = 0, \quad \mathrm{Var}(F) = I_m, \tag{9.23}$$

$$E(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2), \tag{9.24}$$

$$\mathrm{Cov}(F, \varepsilon) = 0. \tag{9.25}$$

因子模型性质

2. 因子模型的性质

(1) Σ 的分解

$$\Sigma = AA^T + D. \tag{9.26}$$

(2) 模型不受单位的影响. 若 $X^* = CX$, 则有

$$X^* = \mu^* + A^*F^* + \varepsilon^*,$$

其中 $\mu^* = C\mu$, $A^* = CA$, $F^* = F$, $\varepsilon^* = C\varepsilon$.

(3) 因子载荷不是惟一的. 设 T 是一 m 阶正交矩阵, 令 $A^* = AT$, $F^* = T^TF$, 则模型 (9.22) 可表示为

$$X = \mu + A^*F^* + \varepsilon. \tag{9.27}$$

统计意义

- 因子载荷的意义
- 共同度
- 特殊方差
- 总方差贡献

因子载荷矩阵和特殊方差矩阵的估计

- 主成分法
- 主因子法
- 极大似然法

主成分法

- 通过样本估算期望和协方差阵
- 求协方差阵的特征值和特征向量
- 省去特征值较小的部分，求出A、D
- 程序
- 例子

主因子法

- 首先对变量标准化
- 给出m和特殊方差的估计（初始）值
- 求出简约相关阵R*（p阶方阵）
- 计算R*的特征值和特征向量，取其前m个，略去其它部分
- 求出A*和D*，再迭代计算

极大似然法

- 似然函数
- 极大似然函数
- 算法描述（薛毅书p533）

Jöreskog 和 Lawley 等人 (1967) 提出了一种较为实用的迭代法，使极大似然法逐步被人们采用. 其基本思想是，先取一个初始矩阵

$$D_0 = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2),$$

现计算 A_0, 计算 A_0 的办法是先求 $D_0^{-1/2}\hat{\Sigma}D_0^{-1/2}$ 的特征值 $\theta_1 \geq \theta_2 \geq \theta_p$, 及相应的特征向量 $l_1, l_2, \cdots, l_p$. 令 $\Theta = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_m)$, $L = (l_1, l_2, \cdots, l_m)$ 且令

$$A_0 = D_0^{1/2} L \left(\Theta - I_m\right)^{1/2}. \tag{9.43}$$

再由式 (9.41) 得到 D_1, 然后再按上述方法得到 A_1, 直到满足方程 (9.40) 为止.

方差最大的正交旋转

- 由于因子载荷矩阵不是唯一，有时因子的实际意义会变得难以解释。
- 因子载荷矩阵的正交旋转
- 因子载荷方差
- 载荷值趋于1或趋于0，公共因子具有简单化的结构
- varimax( ) 函数

因子分析函数factanal( )

函数 factanal() 采用极大似然法估计参数，其使用格式为

```
factanal(x, factors, data = NULL, covmat = NULL, n.obs = NA,
         subset, na.action, start = NULL,
         scores = c("none", "regression", "Bartlett"),
         rotation = "varimax", control = NULL, ...)
```

其中 x 是数据的公式，或者是由数据 (每个样本按行输入) 构成的矩阵，或者是数据框. factors 是因子的个数. data 是数据框，当 x 由公式形式给出时使用. covmat 是样本的协方差矩阵或样本的相关矩阵，此时不必输入变量 x. scores 表示因子得分的方法，scores="regression", 表示用回归方法计算因子得分，当参数为 scores="Bartlett", 表示用 Bartlett 方法计算因子得分 (具体意义见下小节), 缺省值为 "none"", 即不计算因子得分. rotation 表示旋转，缺省值为方差最大旋转，当 rotation="none" 时，不作旋转变换.

因子得分

- 薛毅书p543

知识点

1 数据分析体系的多层模型。数据挖掘与统计分析有什么区别？

2 ETL是什么？ETL层负责哪些功能？

3 OLAP是什么？DW=ETL+OLAP

4 什么是BI？BI系统主要由哪些部分构成？

5 R语言的历史和特点

6 R中与向量和矩阵运算有关的函数和运算符

7 R中用于求基本统计量的函数

8 R中数据框的操作，及怎样从外部数据文件读入数据

9 R中产生各种分布随机数的函数

10 R中涉及下标操作及定位、筛选有关的函数和写法

知识点

11 直方图、散点图（多种）、箱型图、柱状图、饼图、星相图、脸谱图、茎叶图、向日葵散点图、热力图、密度图、三维图等常见统计图的画法和意义

12 熟悉R常用的内置数据集

13 R的条件判别语句与循环语句

14 R的判别函数

15 R的集合运算函数

16 协方差与相关系数的意义与计算

17 怎样使用R进行线性回归分析，及有关建模和计算函数

18 线性回归模型结果的解释，及各项指标的意义

19 多元线性回归应该怎样选择合适的变量？

20 logistic回归模型

知识点

21 怎样用apriori算法做购物篮分析？

22 线性分类法的原理及线性判别函数

23 距离判别法的原理。有哪些距离（点与点之间，点集与点集之间）？

24 贝叶斯分类器的原理

25 怎样利用决策树算法进行分类？

26 knn分类算法的细节

27 层次聚类法的原理与有关实现函数

28 k-means聚类法的原理与实现函数

29 k中心聚类法的原理

30 dbscan聚类法的原理

知识点

31 主成分分析的原理和计算

32 因子分析的原理和计算

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间