2010-6-15

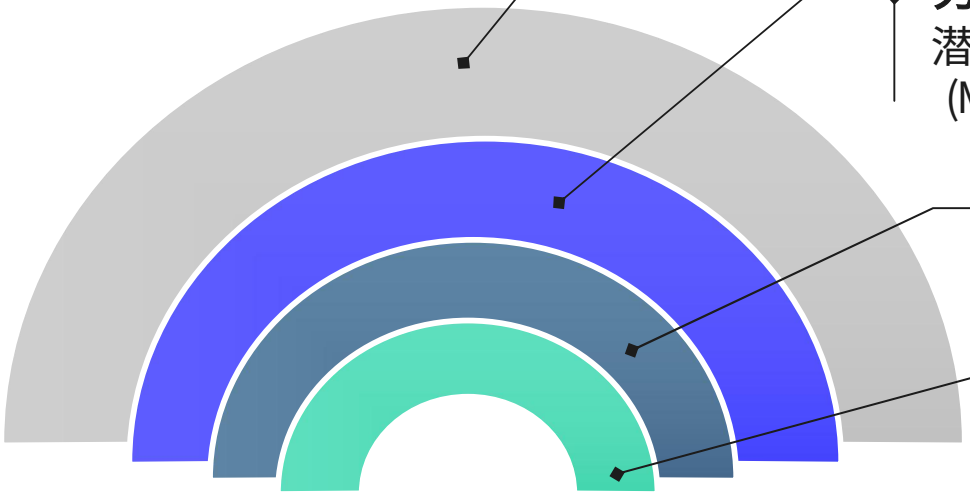
R与因子分析的新发展

主讲：祝迎春

时间：2010-06-15

联系方式：ereree@126.com

R在市场研究中的架构（第二届中国R语言会议）



商业业务应用 精确营销分析、偏好研究、设计可用性研究、市场细分等等

分析技术 结构方程、多水平线性模型、潜类模型、Maximum Difference Scaling (MDS)

软件套件 svm、pls、OpenMx、poLCA、survey

算法与R代码 函数 类 包 群集

变量选择 → 样本量 → 抽样方法 → 分析技术 → 输出形式

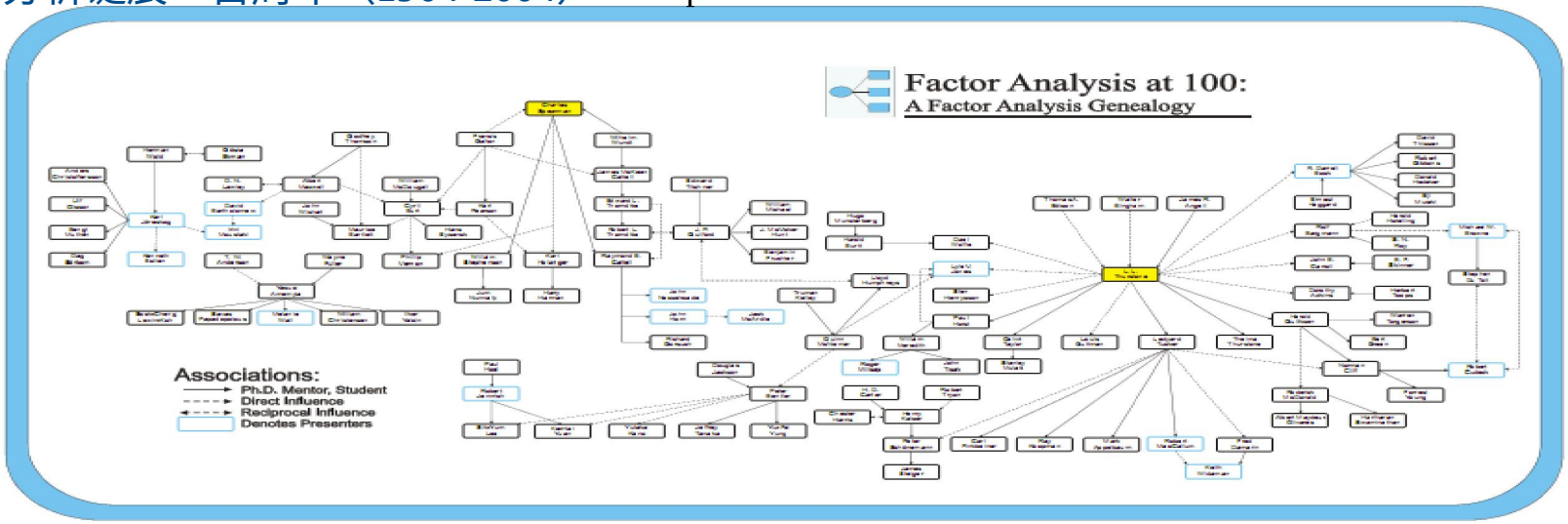
- 因子分析的过去
- 最新的技术发展
- 各个独立R包或者代码介绍
- R中因子分析应用
- 总结与展望

因子分析的过去

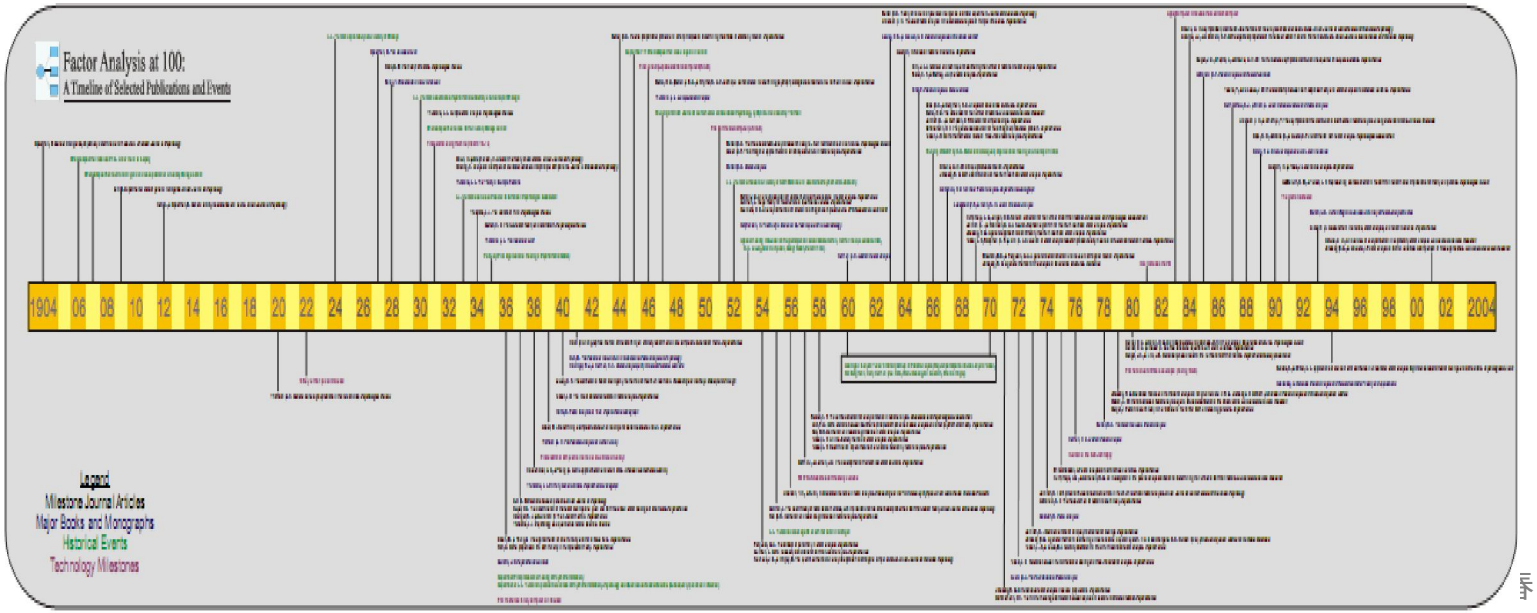
因子分析发展历程：纪念因子分析诞辰 106周年

纪念因子分析诞辰一百周年（1904-2004）：http://www.fa100.info/down.htm

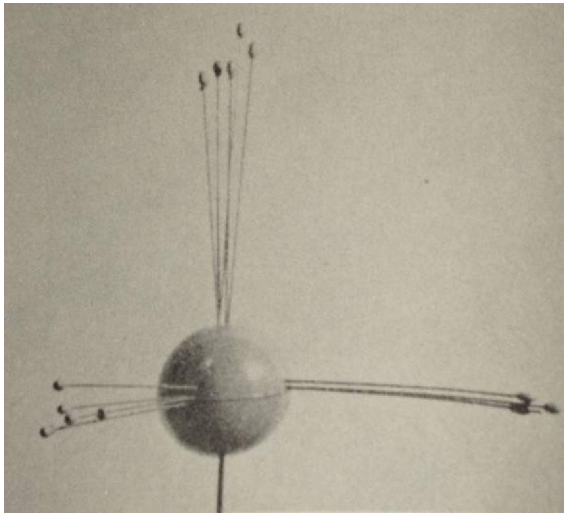
因子分析谱系



因子分析时间轴



传统主成分/因子分析（过时与错误的做法）



因子分析(Factor Analysis)是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(因子)来表示基本的数据结构的方法。

国内各种统计软件的书籍和论文、研究项目中的错误。
使用SPSS默认设置的主成分分析或因子分析导致错误

✗ 主成分+正交旋转方法

✗ 验证因子分析或结构方程时不区分形成型变量和反映型变量

√ 极大似然法+斜交旋转方法

传统主成分 /因子分析

主成分分析

线形性假设
正态分布
方差大的变量有较大重要性
主成分之间正交

注意：R中默认的prcomp是基于SVD分解（svd()函数，princomp是基于特征向量eigen()函数的问题使得求得的数值绝对值正确，符号却出现错误

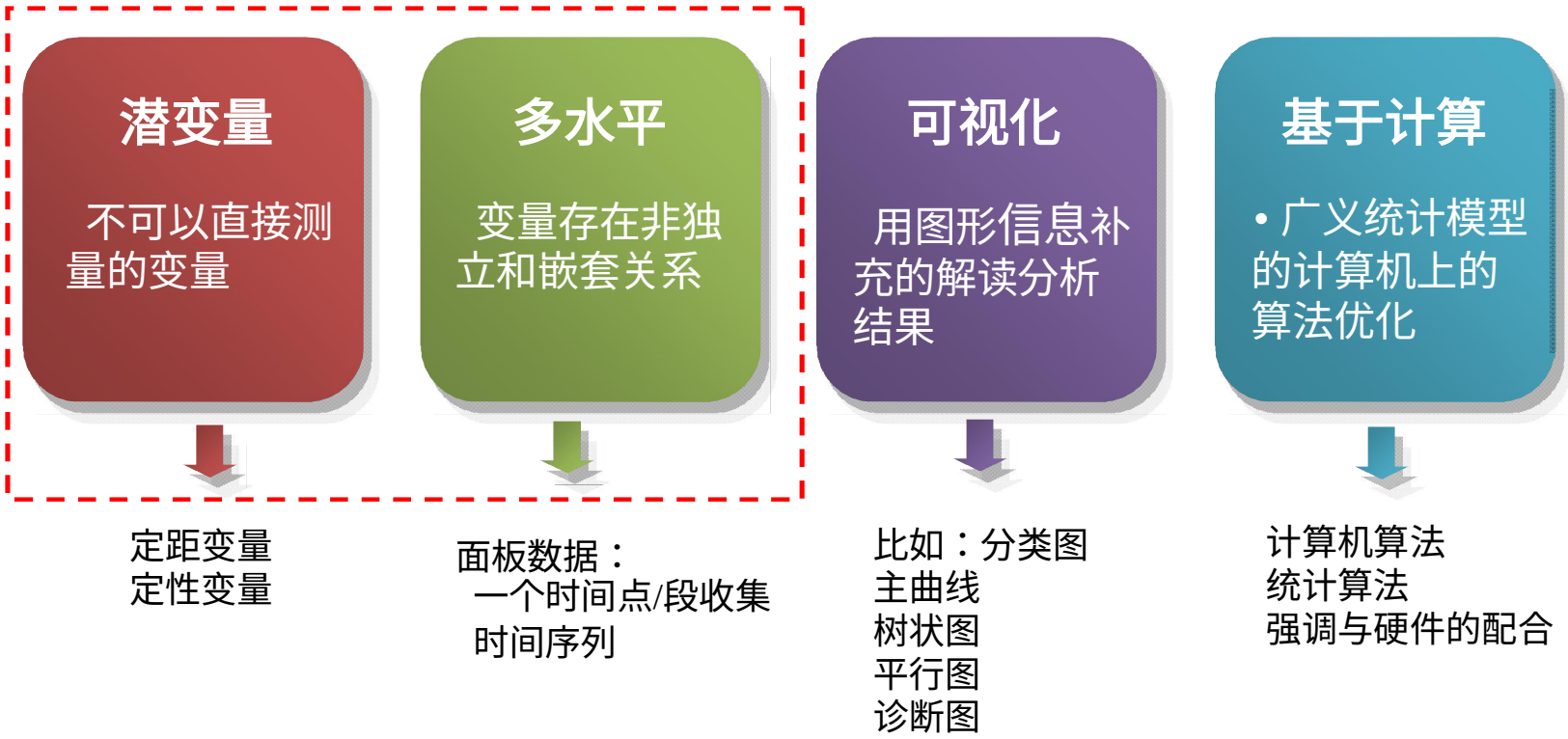
主成分回归的无聊和无奈：
1.主成分不能消除共线性的作用
2.经常出现是第1主成分,常常不是对y的最好解释变量.

探索性因子分析

线形性假设
正态分布
1.假定所有的因子(旋转后) 都会影响测度项：
在实际研究中，我们往往会假定一个因子之间没有因果关系，所以可能不会影响另外一个因子的测度项。
2.探索性因子分析假定测度项残差之间是相互独立的：
实际上，测度项的残差之间可以因为共同方法偏差、子因子等因素而相关。
3.探索性因子分析强制所有的因子为独立的：
这虽然是求解因子个数时不得不采用的机宜之计，却与大部分的研究模型不符。最明显的是，自变量与因变量之间是应该相关的，而不是独立的。这些局限性就要求有一种更加灵活的建模方法，使研究者不但可以更细致地描述测度项与因子之间的关系，而且并对这个关系直接进行测试。
4.多于一个因素的因子分析不存在唯一解,研究者必须从众多恒等的合适解中选择一个单一解。

定义：“新发展”

总体的方向


潜变量
不可以直接测量的变量
多水平
变量存在非独立和嵌套关系
可视化
用图形信息补充的解读分析结果
基于计算
• 广义统计模型的计算机上的算法优化
定距变量
定性变量
面板数据：
一个时间点/段收集
时间序列
比如：分类图
主曲线
树状图
平行图
诊断图
计算机算法
统计算法
强调与硬件的配合

与因子分析有关新的统计技术

(全部已经可以用R实现的，是我见过统计软件中最全的)：

- Ø基于变量选择Variable Selection
- Ø样本量Sample Size
- Ø非线性Nonlinear：
  - 核估计Kernel Estimate
  - 基于投影追踪Projection Pursuit
  - 神经网络Neural Network
- Ø基于贝叶斯Bayes Method
- Ø主曲线Principal Curves
- Ø独立主成分Independent Component Analysis
- Ø独立因子分析Independent Factor Analysis
- Ø基于稳健估计Robust Estimate
- Ø非独立变量 Non-independent Variables
- Ø重复测量Repeated Measures
- Ø成分数据Compositional Data
- Ø跨变量水平Ordinal Data定序的
- Ø多水平Multilevel/Hierarchical Models
- Ø缺失值Missing Values

主成分 /因子分析技术关键点

研究设计　样本量和抽样

完整步骤 Steps:

1.Missing Values（缺失值）
2.Outlier（离群值）
3.Mvnormality Test（多元正态分布检验）
4.Communalities（估计共同性*）
5.Number of Factors（决定因子数目）
6.Statistical Power（统计功效）
7.Estimation（估计因子负荷 ）
8.Rotations（选择因子旋转 ）
9.Factor Name（因子命名 ）
10.Factor Score（因子得分 ）

什么是好的因子分析？
ü稳健
ü因子结构清晰
ü因子结构稳定（超越方法本身）
ü因子得分与研究的其他变量显著相关

*共同性(又叫分解共性):所有的因子加起来，对于观察变量的方差的总解释能力。

主成分 /因子分析样本量与功效

determination of sample size for covariance structure modeling

因子分析的样本量=3500

Sample Size for RMSEA.R

```
vn <-13 #变量数
m  <-8  #因子数
mseaO <- 0.08 #原假设 RMSEA取0.07or0.08
mseaa <- 0.05 #对立假设 RMSEA
d <- ((vn-m)^2-(vn+m))/2 #自由度
alpha <- 0.05 #显著性水平
desired <- 0.8 #统计功效
```

因子分析的统计功效=0.8

Power for RMSEA.R

```
vn <-13 #变量数
m  <-8  #因子数
n <- 3500  #样本量
mseaO <- 0.08 #原假设 RMSEA取0.07or0.08
mseaa <- 0.05 #对立假设RMSEA
d <- ((vn-m)^2-(vn+m))/2 #自由度
alpha <- 0.05 #显著性水平
```

R中相关包介绍

主成分 /因子分析检验

determination of sample size for covariance structure modeling

Bartlett球面性检验

library(psych)
cortest.bartlett(data)

```
vn <-13 #变量数
m  <-8  #因子数
rmsea0 <- 0.08 #原假设 RMSEA取0.7or0.8
rmseaa <- 0.05 #对立假设 RMSEA
d <- ((vn-m)^2-(vn+m))/2 #自由度
alpha <- 0.05 #显著性水平
desired <- 0.8 #统计功效
```

因子分析的 Kaiser-Meyer-Olkin检验

kmo.R

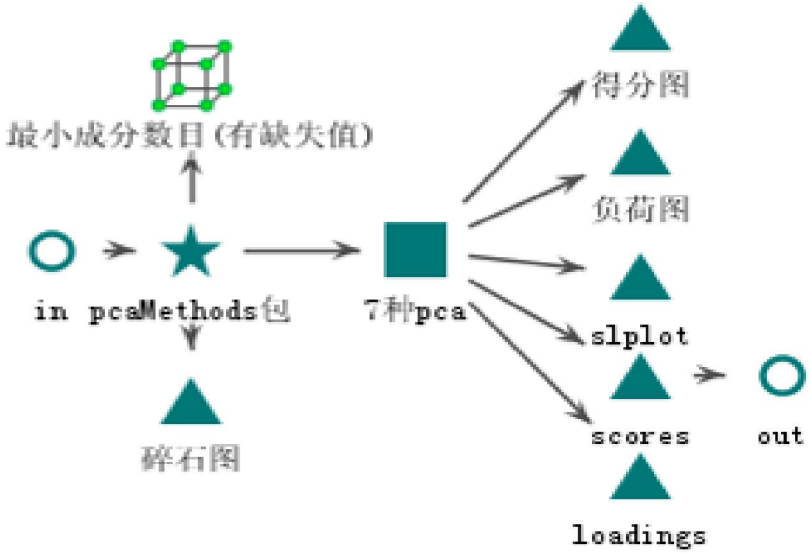
```
vn <-13 #变量数
m  <-8  #因子数
n <- 3500  #样本量
rmsea0 <- 0.08 #原假设 RMSEA取0.7or0.8
rmseaa <- 0.05 #对立假设RMSEA
d <- ((vn-m)^2-(vn+m))/2 #自由度
alpha <- 0.05 #显著性水平
```

主成分分析

Principal Component Analysis

pcaMethods	bpca(pcaMethods)	Bayesian PCA Estimator
pcaMethods	nipalsPca(pcaMethods)	PCA by non-linear iterative partial least squares
pcaMethods	nlpca(pcaMethods)	Neural network based non-linear PCA
pcaMethods	ppca(pcaMethods)	Probabilistic PCA
pcaMethods	RnipalsPca(pcaMethods)	PCA by non-linear iterative partial least squares
pcaMethods	robustPca(pcaMethods)	PCA implementation based on robustSvd
pcaMethods	svdPca(pcaMethods)	Perform PCA using singular value decomposition
pcaMethods	svdImpute(pcaMethods)	SVDimpute algorithm

bpca，ppca和nipalspca可用于执行有缺失值的数据主成分分析以及准确估计成分数目。



因子分析

Factor analysis

fa(psych)

```
fa(r, nfactors=1, residuals = FALSE, rotate = "varimax",n.obs = NA, scores = FALSE,SMC=TRUE,
missing=FALSE,impute="median",min.err = 0.001, max.iter =
50,symmetric=TRUE,warnings=TRUE,fm="minres", ...)
```

fm="ml"最大似然法
fm="minres"最小残差法 (默认)
m="pa"主轴因子
fm="wls"加权最小二乘法
fm="gls"广义最小二乘法

rotate="none"不旋转
rotate="varimax"直交解(默认)
rotate="quartimax"直交解
rotate="bentlerT"直交解
rotate="geominT"直交解
rotate="promax"斜交解
rotate="oblimin"斜交解
rotate="simplimax"斜交解
rotate="bentlerQ"斜交解
rotate="geominQ"斜交解
rotate="cluster""斜交解

主成分 /因子分析的因子数目选择

传统的方法来确定因子数目：

Lederman's condition (1937): $k_{max} \le \frac{1}{2}\left(2p+1-\sqrt{8p+1}\right)$

· Kaiser 方法 - 仅保留特征值大于 1 的成份，几乎是所有统计软件的默认选项，只适用于主成分法所求解的因子,容易提取过多的因子Fabrigar等人(1999)。

· 碎石检验 - 在碎石图中找到特征值平滑递减达到稳定的位置。保留陡曲线中在开始线趋势的第一个点之前的那些成份。

· 解释的百分比变异 - 保留按累积方式解释一定变异百分比的成份。一般为80%方差。

· 若数据属于多变量正态分布，可利用最大似然法进行共同因子的选取，也可利用赤池信息准则（AIC）及Schwarz‘ sbayesian准则（SBC）的，做为判断指标

· 若选取到第m个因子时，其AIC的值或是SBC值最小，则m为最适的共同因子个数

平行检验图（因子数目选择）

parallel analysis（Horn，1965）

是确定探索性因素分析中保留因子个数的最精确的方法。

一个玩笑：国内喜欢把好东西藏起来，相关论文极少。

我可以实现的：R（推荐） SPSS SAS STATA（推荐）。

平行分析可以通过以下几个步骤来完成:首先,生成一组随机数据的矩阵,这些矩阵和真实的数据矩阵要有相同的变量个数和被试的个数。然后,求出这组随机数据矩阵的特征值,并计算这些特征值的平均值。最后,通过比较真实数据中特征值的碎石图和这组随机矩阵的平均特征值的曲线,我们可以找到两条特征值曲线的交点,根据交点的位置,我们就可以确定要抽取因子的绝对最大数目。如果真实数据的特征值落在随机矩阵的平均特征值曲线之上,则保留这些因子;反之,则舍弃这些因子。《平行分析在探索性因素分析中的应用》（孔明，2007）

小提示：

随机矩阵的特征根取第 95 百分位数为比较好：避免高估因子数目

使用 nFactors包

```
library(nFactors)
ev <- eigen(cor(mydata))
ap <- parallel(subject=nrow(mydata),var=ncol(mydata),
rep=100,cent=.05)
nS <- nScree(ev$values, ap$eigen$qevpea)
plotnScree(nS) # 彩色有具体因子数目
```

稳健主成分分析

Robust Principal Components

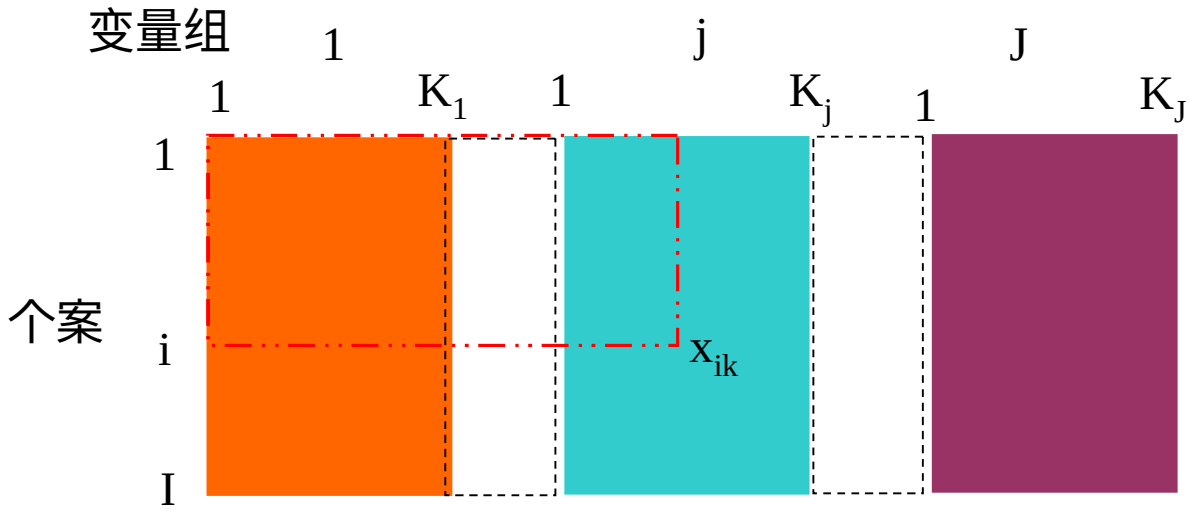
amap	acprob(amap)	
pcaPP	PCAgrid(pcaPP)	Robust Principal Components using the Grid search algorithm
pcaPP	PCAproj(pcaPP)	Robust Principal Components using the algorithm of Croux and Ruiz-Gazen
rrcov	PcaGrid(rrcov)	Robust Principal Components based on Projection Pursuit (PP): GRID search Algorithm
rrcov	PcaHubert(rrcov)	ROBPCA - ROBust method for Principal Components Analysis
rrcov	PcaLocantore(rrcov)	Spherical Principal Components
rrcov	PcaProj(rrcov)	Robust Principal Components based on Projection Pursuit (PP): Croux and Ruiz-Gazen (2005) algorithm

小提示：
pcaPP可以做2种诊断图（离群值）和双协方差矩阵图（好看）,其Croux and Ruiz-Gazen参数比rrcov的详细
Rrcov中多了一个ROBPCA稳健pca，且可以做诊断图
Rrcov中可以做球面主成分

多元因子分析 (MFA)

Multiple Factor Analysis FactoMineR包

有I个个案，有J个变量，其中变量分为k组,组内变量水平必须一样，变量可以是定类的

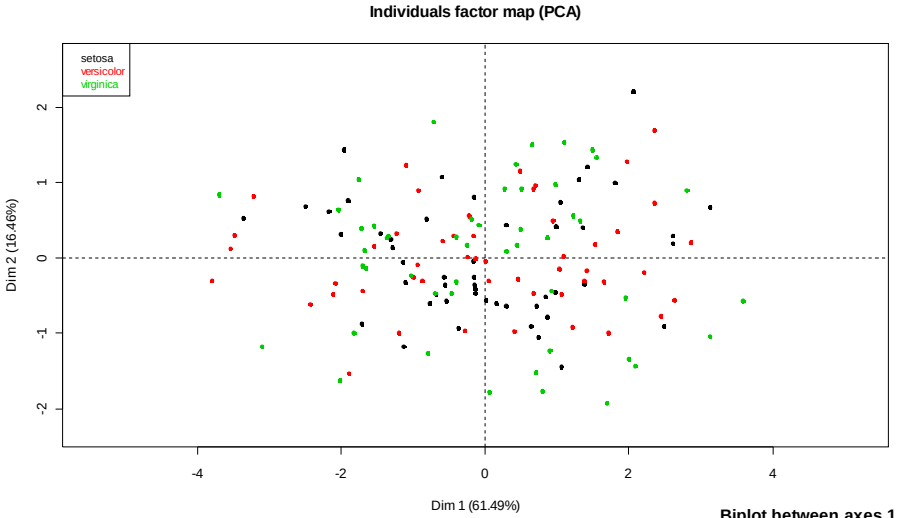


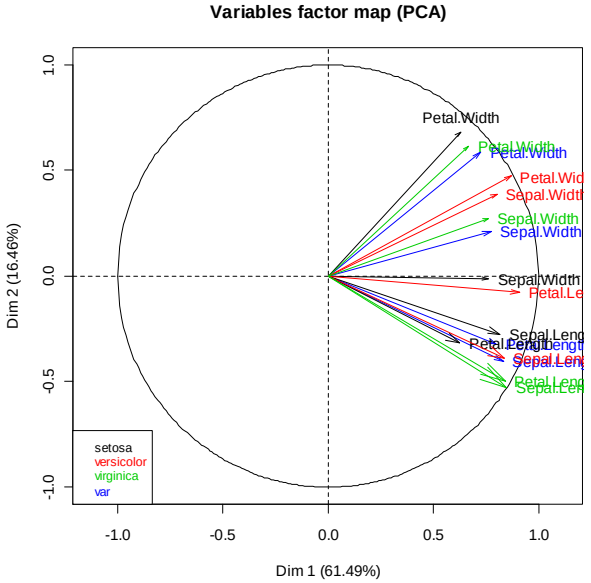
应用场景：
问卷（市场研究方法论的发展）：一组行为问题，一组态度问题
感官分析 （化工）：一组化学性质，一组感觉变量
生态（生物学）：一组土壤性质，一组植物属性
时间序列（计量经济学） ：不同时间上收集的一系列变量

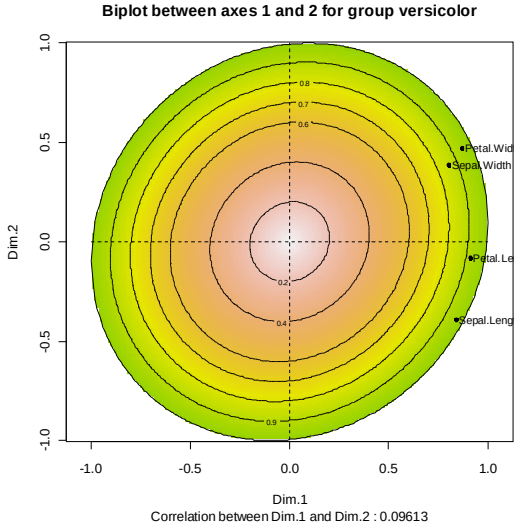
双重多元因子分析 (DMFA)

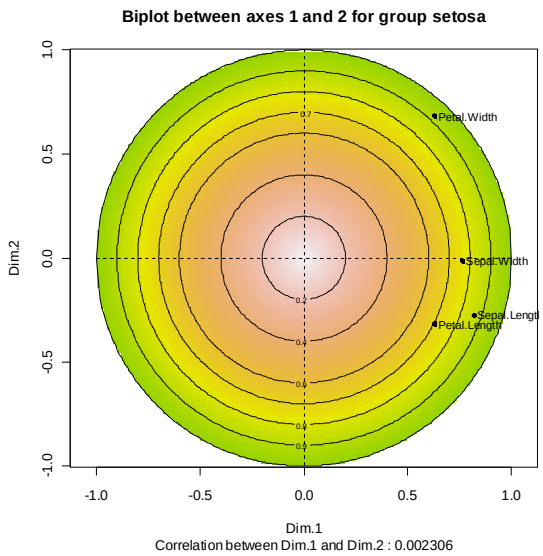
Dual Multiple Factor Analysis FactoMineR包

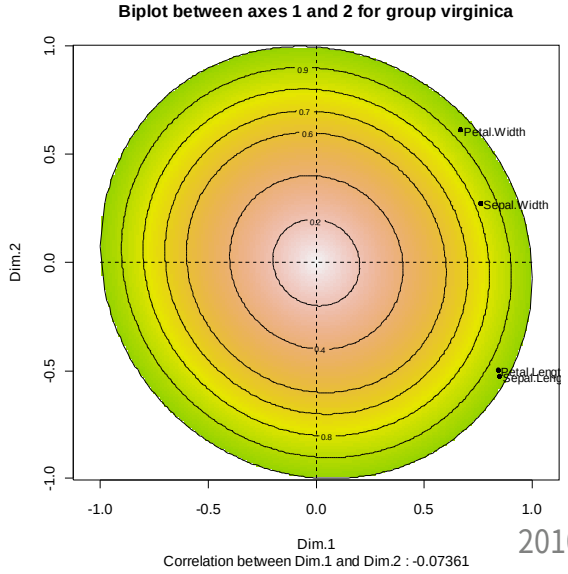
应用场景：一组变量来自不同总体





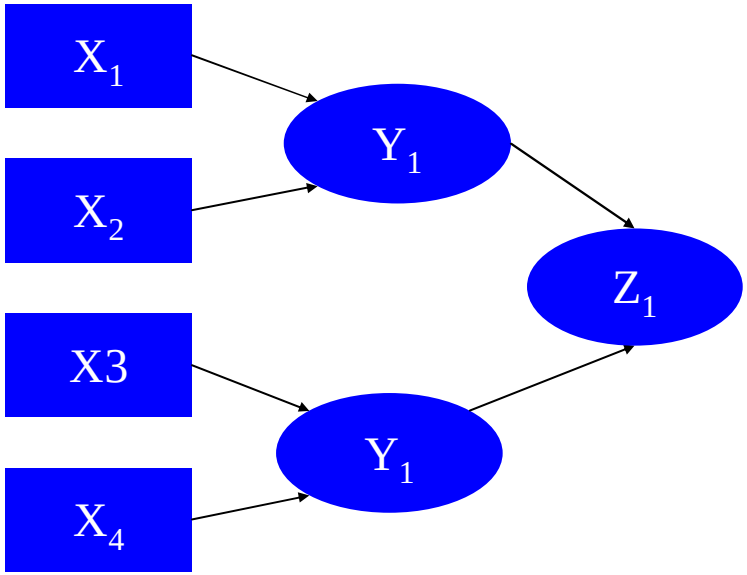


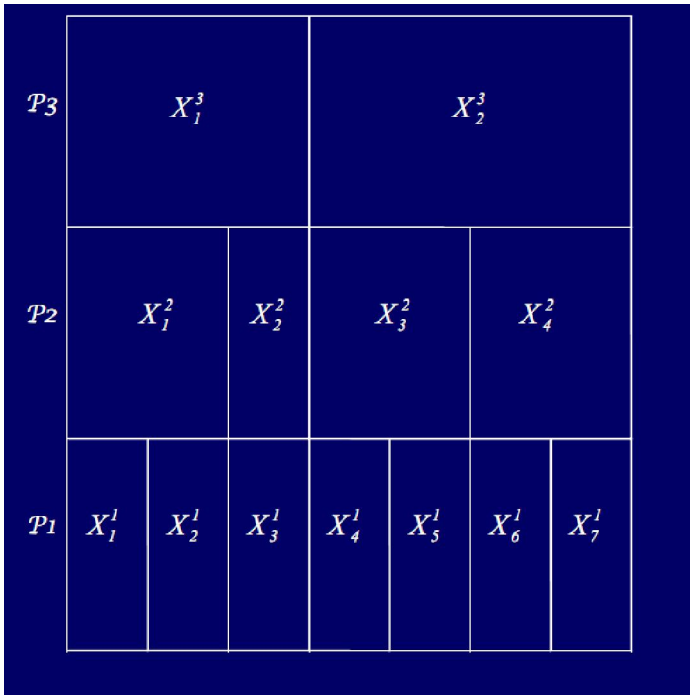




高阶因子分析 (HMFA)

Hierarchical Multiple Factorial Analysis FactoMineR包



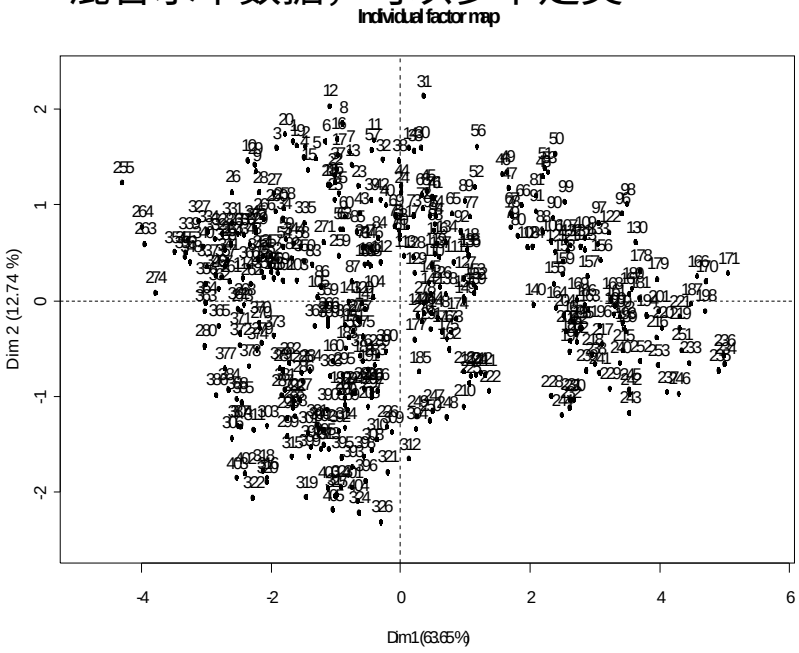


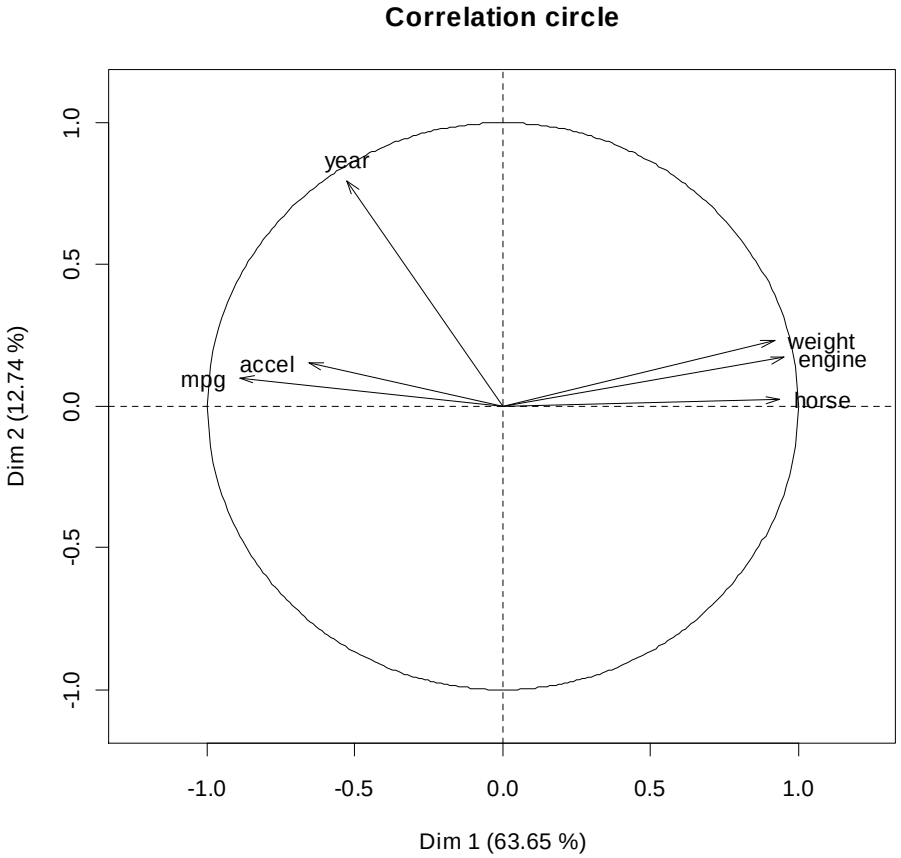
值得推荐的软件：
不推荐使用RcmdrPlugin.FactoMineR版本低
验证性的多水平协方差模型用Mplus
Stata包：gllamm(Generalized Linear Latent And Mixed Models)

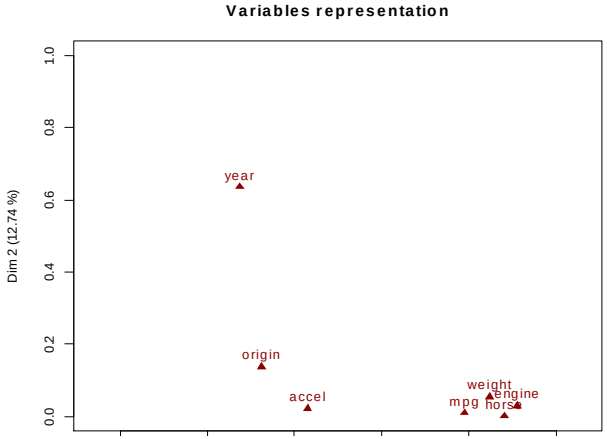
混合数据主成分/因子分析 (AFDM)

Mixed Data Multiple Factor Analysis　FactoMineR包

混合水平数据，可以多个定类







高斯混合因子分析

Factor Mixture Analysis model

数据是来自一个异质且不可观测的总体，保证测量不变性

小提示: 包'fma'是在R版本2.10.0之前建的：你得重新组装

```
fma(iris[,c(1:4)],k=3,r=2,it=50,eps=0.0001,scaling=TRUE)
```

k The number of the mixture components

r The number of factors

判别结果

	setosa	versicolor	virginica
1	50	0	0
2	0	0	6
3	0	50	44

负荷

	[,1]	[,2]
[1,]	-0.5193324	0.7587328
[2,]	-0.6956906	-0.6317622
[3,]	-0.2606269	0.9594145
[4,]	-0.2944360	0.9163729

跨变量水平主成分分析

定序变量：nlPCA.R

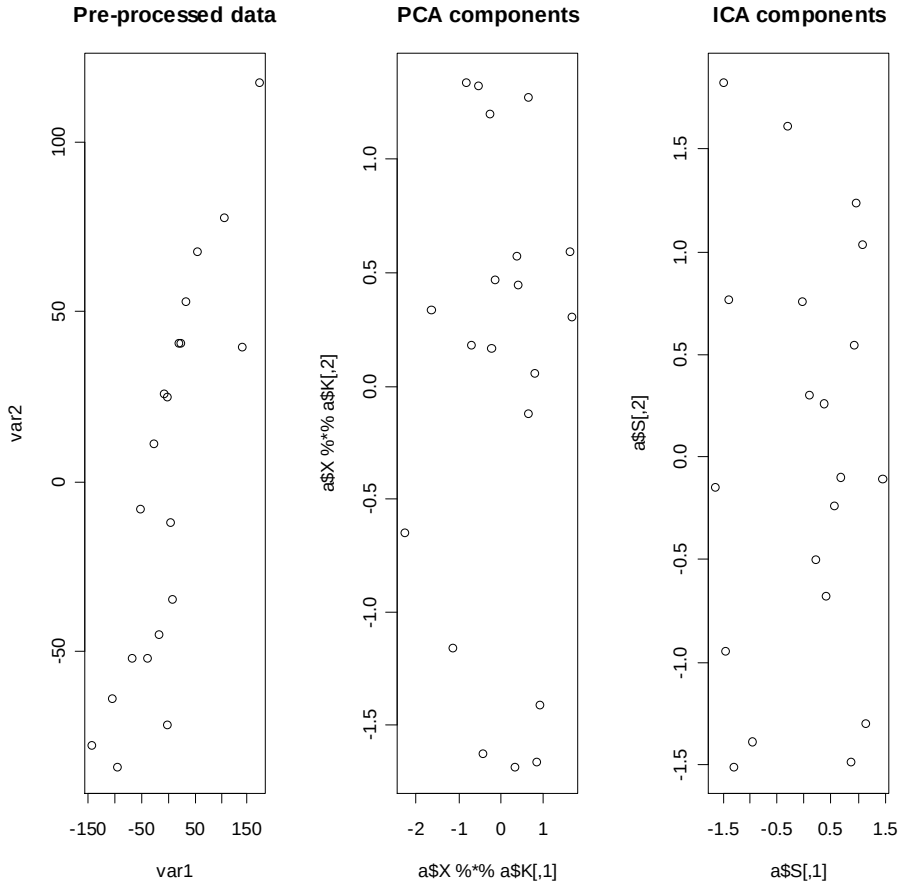
```
source("D:\\Program Files\\R\\pca\\nlPCA.R")
 ordPCA(mydata)
```

独立主成分 /独立因子分析

Independent Component Analysis/Independent Factor Analysis

独立成分分析或独立分量分析是一个线性变换。这个变换把数据或信号分离成统计独立的非高斯的信号源的线性组合。 其最重要的假设就是信号源(独立成分)统计独立*。这个假设在大多数盲信号分离的情况中符合实际情况。即使当该假设不满足时，仍然可以用独立成分分析来把观察信号统计独立化，从而进一步分析数据的特性。

其应用领域与应用前景都是非常广阔的，目前主要应用于盲源分离、图像处理、语言识别、通信、生物医学信号处理、脑功能成像研究、故障诊断、特征提取、金融时间序列分析和数据挖掘等。



独立因子分析：因子分析+独立主成分

```
library(ifa)
Dataset <- read.table("C:/Documents and
Settings/Administrator/桌面/Dataset.txt",
  header=TRUE, sep="", na.strings="NA",
dec=".", strip.white=TRUE)
init.values<-ifa.init.random(Dataset,2)
ifafit<-
ifa.em(Dataset,c(2,2),it=50,eps=0.0001,in
it.values)
ifafit$H # 估计因子负荷矩阵
ifa.predict(Dataset, ifafit, method =
"bartlett") #预测因子得分
```

*:必须不是正态分布

贝叶斯主成分 /因子分析

Bayesian PCA and FA Estimator

贝叶斯主成分看bpca(pcaMethods)

贝叶斯因子分析：

MCMCfactanal(MCMCpack)　　Markov Chain Monte Carlo for Normal Theory Factor Analysis Model
MCMCmixfactanal(MCMCpack)　Markov Chain Monte Carlo for Mixed Data Factor Analysis Model
MCMCordfactanal(MCMCpack)　Markov Chain Monte Carlo for Ordinal Data Factor Analysis Model

关键输入参数：

burnin=MCMC（抽样器）开始时的不作数迭代的抽样样本，称之为“燃烧”，其作用是为降低起始值的影响，选取迭代较稳定的数据

mcmc=“燃烧”后的MCMC抽样迭代的数值，总共产生的抽样样本是：burnin+mcmc

thin=使用模拟的间隔。迭代次数必须整除这个值。表示每隔 thin= 个单位保留一笔资料

verbose=是否输出抽样器迭代的信息，如果大于 0，每隔输出 verbose= 蒙特卡罗采样（Metropolis-Hastings acceptance rate Metropolis-Hastings接受比）

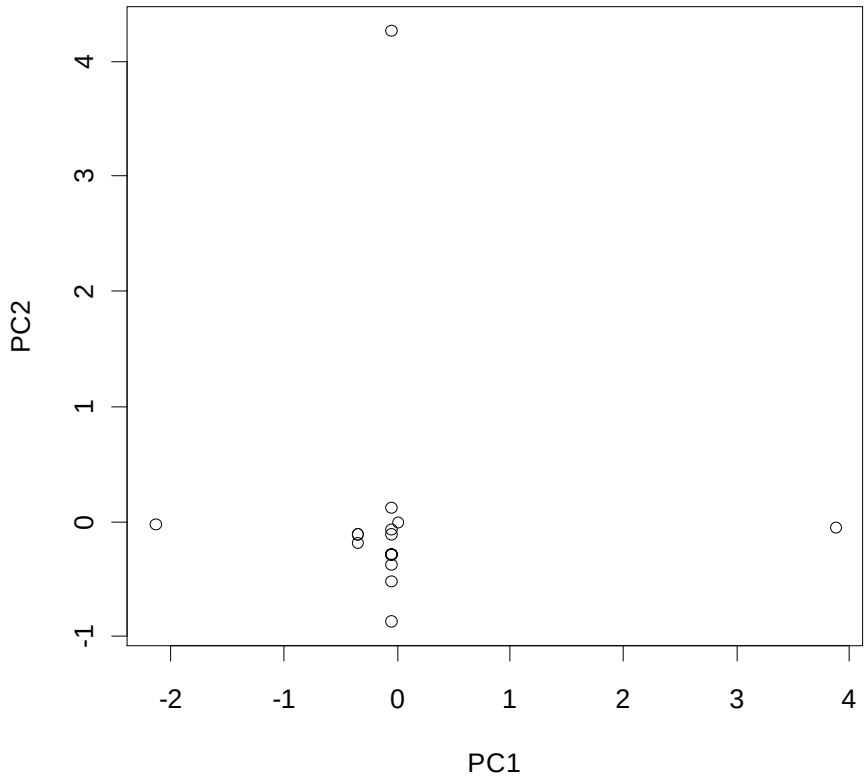
小提示：

注意区分有约束和无约束以及如何设置约束

如果是单核CPU不推荐使用：MCMCmixfactanal(混合数据的贝叶斯因子分析)和MCMCordfactanal（定序变量的贝叶斯因子分析），多核请使用并行运算

核估计主成分 /因子分析

如果（PC1,PC2）之间存在着非线性的关系，并且根据先验的知识可知旋转角是最优的主成分。则在这种情况下，PCA就会失效。但是，如果加入先验的知识，对数据进行某种划归，就可以将数据转化为以为线性的空间中。这类根据先验知识对数据预先进行非线性转换的方法就成为kernel-PCA，它扩展了PCA能够处理的问题的范围，又可以结合一些先验约束，是比较流行的方法。《独立主成分分析》一书




```
library(kernlab)
Dataset <- read.table("C:/Documents and
Settings/Administrator/桌面/Dataset.txt",
  header=TRUE, sep="", na.strings="NA", dec=".",
strip.white=TRUE)
kpc <-
kpca(~.,data=Dataset,kernel="rbfdot",kpar=list(sigm
a=0.2),features=2)
plot(rotated(kpc),xlab="PC1",ylab="PC2")
pcv(kpc)
```

投影追踪因子分析

Gradient Projection Algorithm Rotation for Factor Analysis

GPArotation包一共22种旋转方法

投影寻踪 (简称 PP)是国际统计界于 70年代中期发展起来的一种新的、有价值的高新技术，是统计学、应用数学和计算机技术的交叉学科，属当今前沿领域，它是用来分析和处理高维观测数据，尤其是非正态非线性高维数据的一种新兴统计方法，它通过把高维数据投影到低维子空间上，寻找出能反映原高维数据的结构或特征的投形，选到研究分析高维数据的目的，它具有稳健性、抗干扰性和准确度高等优点，因而在许多领域得到广泛应用。

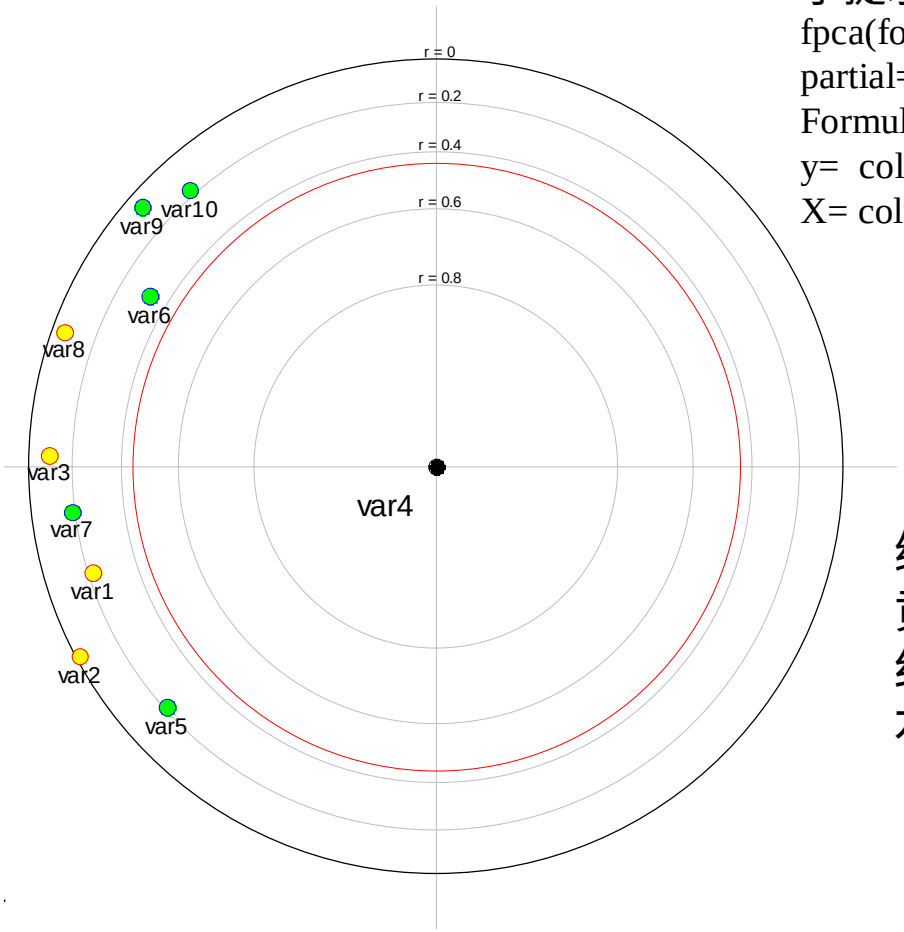
R中投影追踪的算法在是应用在解主成分分析的非线性和稳健估计上。而在应用在因子分析时，主要为了因子旋转估计负荷用（优化）

唯一可以和R抗衡的是STATA，其他主流统计软件无法相比

小提示：
GPForth是做正交旋转算法。GPFoblq是做斜交旋转算法。
正交旋转："varimax","quartimax", "tandemI", "tandemII", "entropy", "mccammon"
斜交旋转："quartimin", "oblimin", "simplimax", "oblimax"
无论正斜均可："bentler","geomin", "target", "pst", "cf", "infomax"

非独立变量主成分分析

Focused Principal Components Analysis “有监督”因子分析 **Psy包**



小提示：研究变量之间关系

fpca(formula=NULL,y=NULL, x=NULL, data, cx=0.75, pvalues="No", partial="Yes", input="data", contraction="No", sample.size=1)

Formula="model" formula, of the form y ~ x

y= column number of the dependent variable var4

X= column numbers of the independent (explanatory) variables 其他变量

```
fpca(y="var4",x=c(1:3,5:10),data=mydata)
fpca(y=4,x=c(1:3,5:10),data=mydata)
```

绿色表示与因变量变量呈正相关

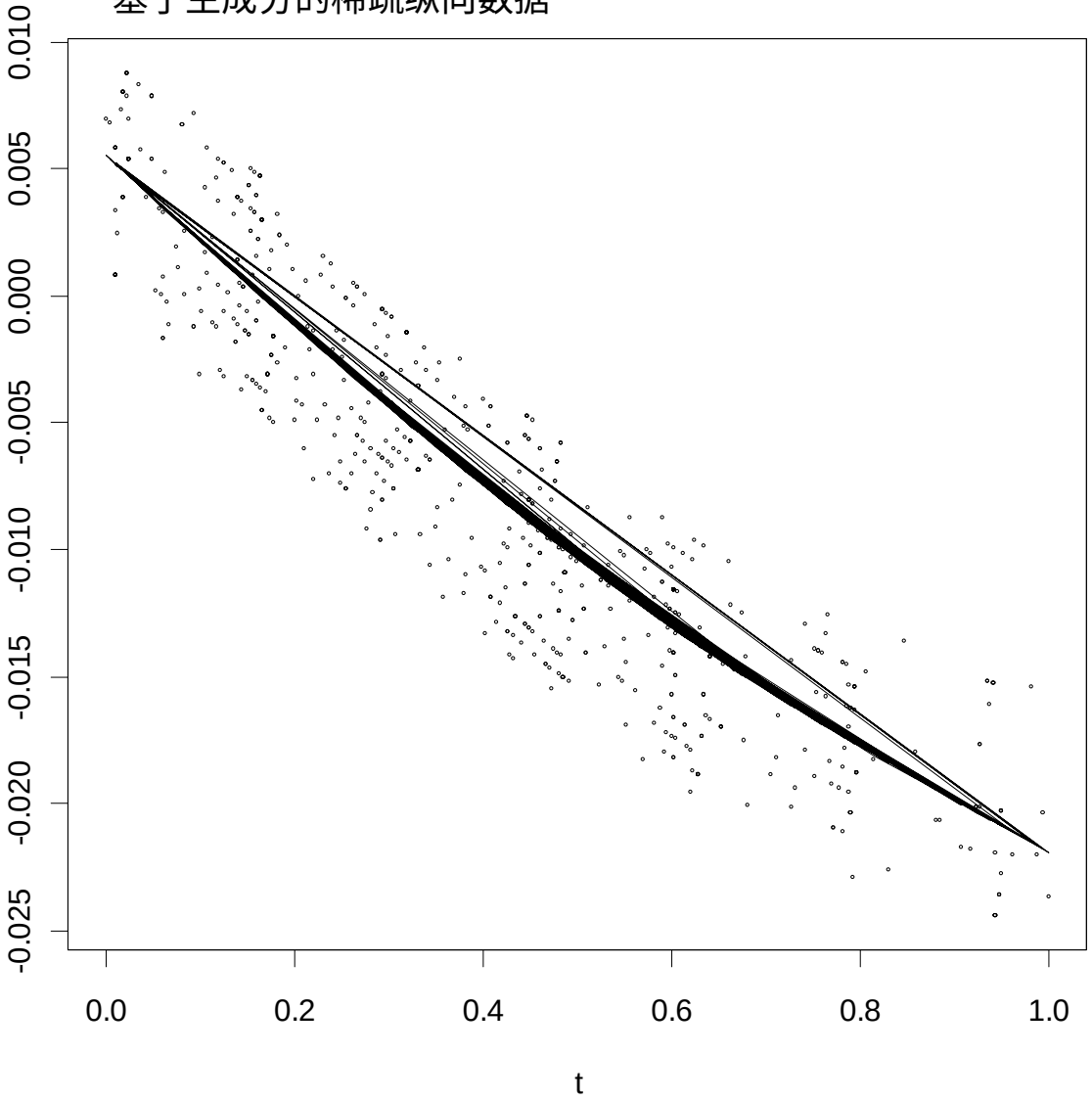
黄色表示与因变量是负相关。

红色圈子内的变量与与因变量的关系显著性水平p <0.05

时间序列主成份分析

Restricted MLE for Functional Principal Components Analysis

基于主成分的稀疏纵向数据



fpca包

```
fpca.mle(data.m, M.set,r.set,
ini.method="EM", basis.method="bs",
sl.v=rep(0.5,10), max.step=50,
grid.l=seq(0,1,0.01),
grids=seq(0,1,0.002))
```

小提示：
不能有缺失值
data.m数据格式： id个案号
measurement测量变量 measurement
time 测量时间
ini.method="EM" 默认 or "loc"
basis.method="bs" 默认 or "ns"

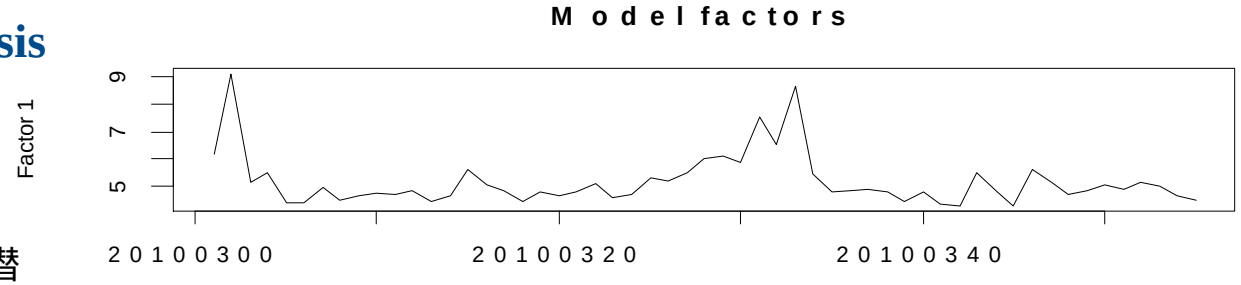
结果
特征根： result$eigenvalues
特征根函数： result$eigenfunctions
误差： result$error_var
得分： result$fitted_mean

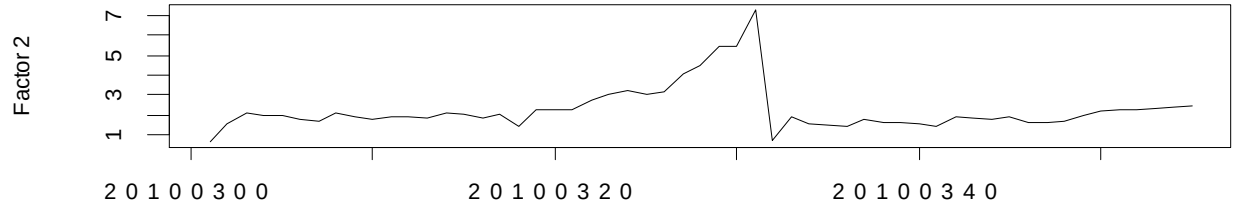
时间序列因子分析

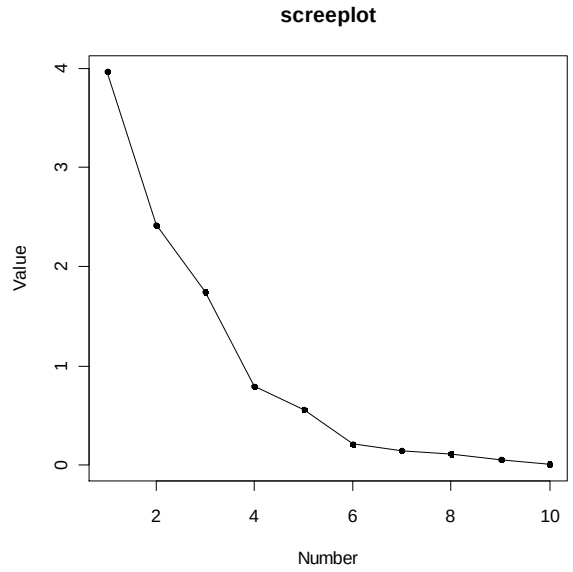
Time Series Factor Analysis

tsfa包

用于研究多元变量在时间序列上存在的潜在结构。
本数据是10个企业经营指标在一个月（单位：天）内的数据，数据本身有缺失。







	Factor 1	Factor 2
M1	-0.32965816	-0.41024079
M2	-0.11046719	0.91286781
M3	0.40049239	0.30379800
M4	0.36053399	0.83967576
M5	0.67312276	0.40092086
M6	0.15099952	0.04579166
M7	-0.06875299	0.31530028
M8	0.22681799	-0.01081203
M9	1.03336988	-0.22096505
M10	1.03372249	-0.22669041

时间序列的因子得分

小提示：
Stata 中可以用Dynamic Factor Analysis做类似的分析

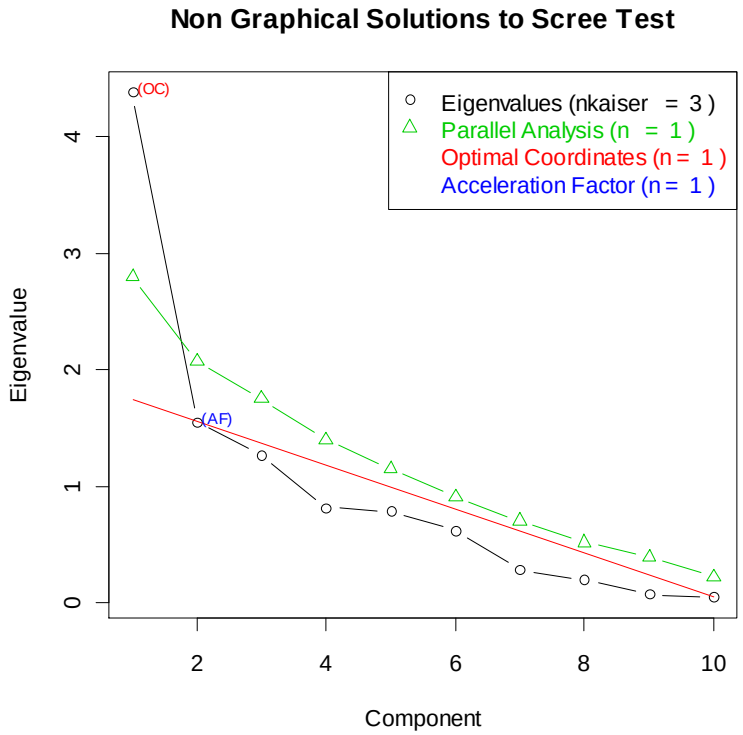
R中主成分/因子分析作图

主成分 /因子分析作图

1. 碎石图(因子数目选择)
2. 平行图(因子数目选择)
3. 负荷图(因子旋转前后的结构)
4. 得分图(样本结构分布)
5. 双标图(总体因子结构分布)
6. 树状图(分类结构分布)

平行图

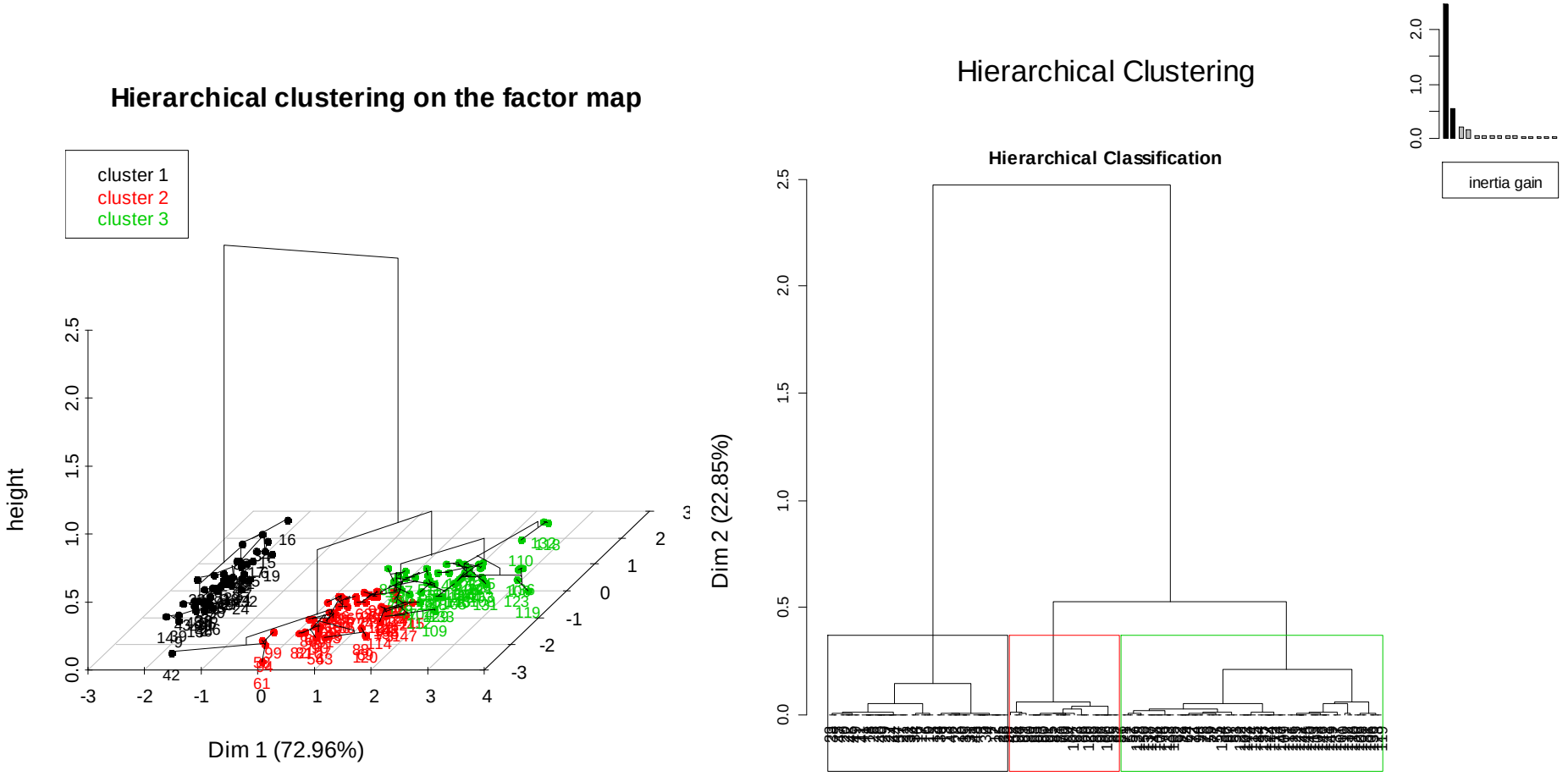
R中可以使用的包：
Paran
nFactors(推荐，并且在R.10以上安装最新版本)
Psych
random.polychor.pa(理论上可以做混合变量，但运算时间慢)




```
library(nFactors)
Dataset <- read.table("C:/Documents and Settings/Administrator/桌面
/Dataset.txt",   header=TRUE, sep="", na.strings="NA", dec=".",
strip.white=TRUE)
## PARALLEL ANALYSIS (qevpea for the centile criterion, mevpea for the
mean criterion)
aparallel <-
parallel(subject=nrow(Dataset),var=ncol(Dataset),rep=100,cent=.95)$eigen$q
evpea  # The 95 centile
## NOMBER OF FACTORS RETAINED ACCORDING TO DIFFERENT
RULES
 results <- nScree(eig=eigen(cor(Dataset))$values,aparallel=aparallel)
 results
## PLOT ACCORDING TO THE nScree CLASS
 plotnScree(results)
```

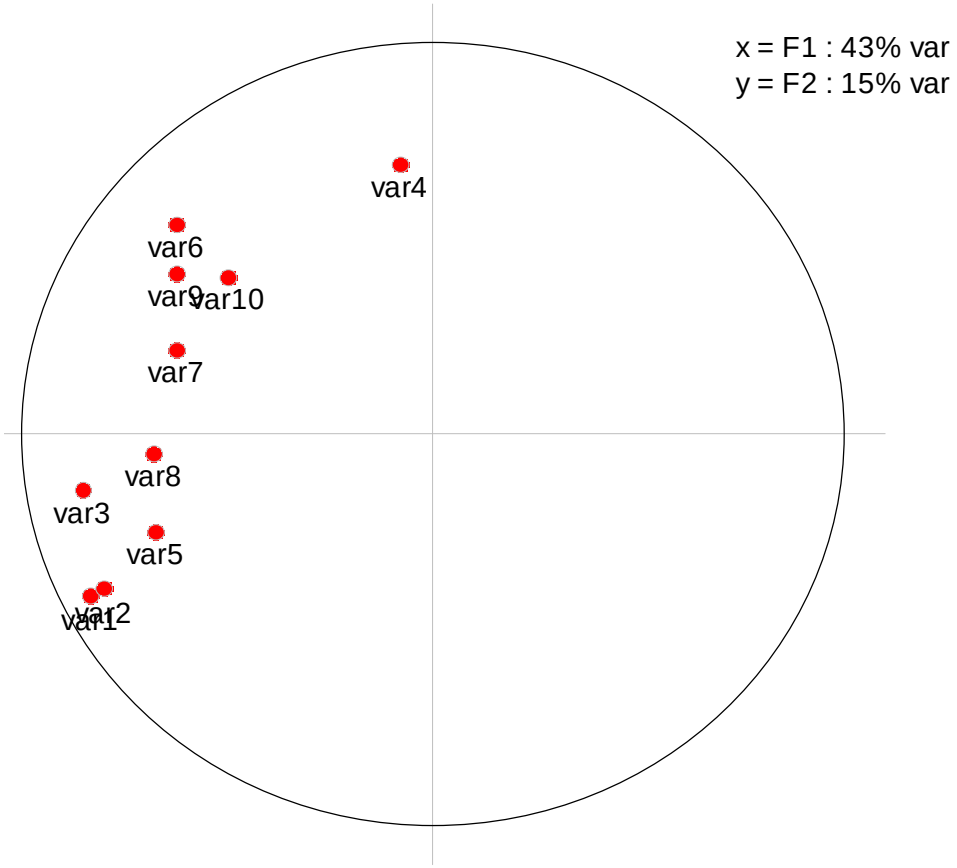
树状图

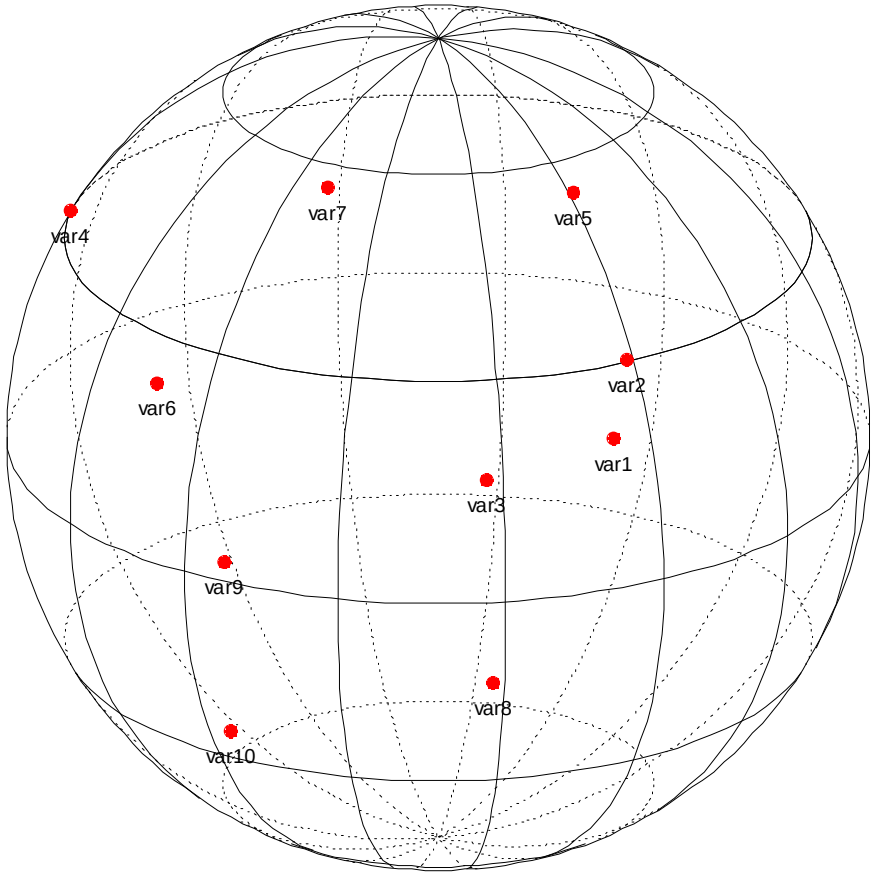
Hierarchical Classification on Principle Components (HCPC)



小提示：运算会相当慢

负荷图

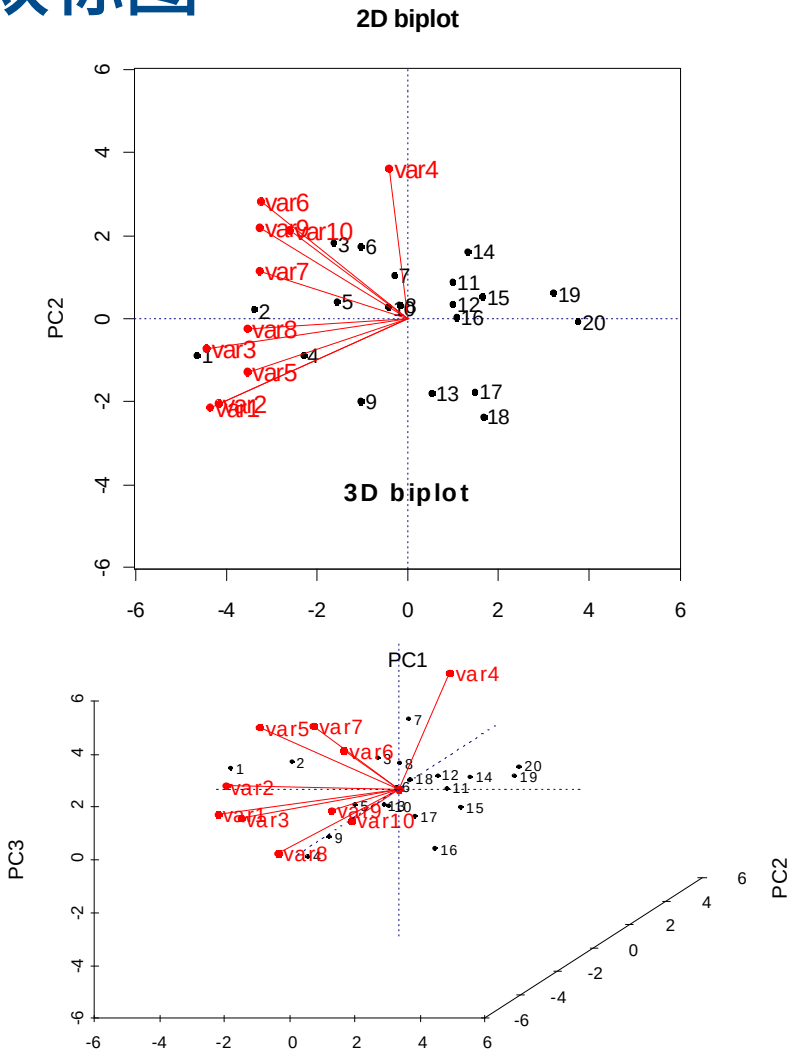




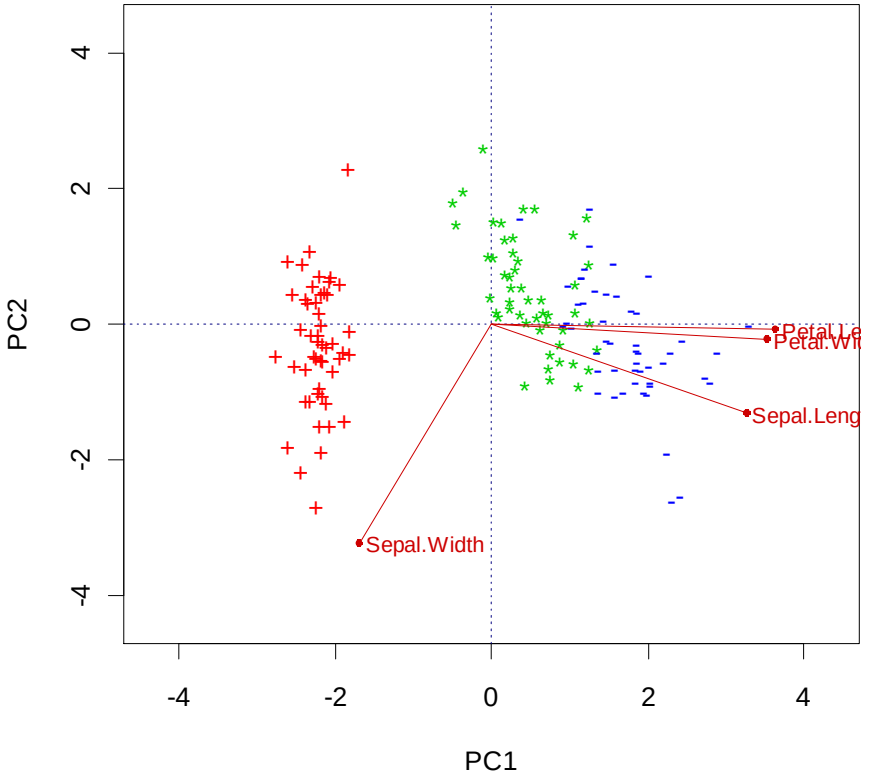

```
library(psy)
mydata <-
read.table("D:/data/r2010/Dataset1.txt",
  header=TRUE, sep="", na.strings="NA",
dec=".", strip.white=TRUE)
mdspca(mydata[,c(2:11)],cx=1)
```

```
cordata <-
as.data.frame(cor(mydata[,c(2:11)],use="p
airwise.complete.obs"))
sphpca(cordata,method="rscal",input="Cor"
,h=180,f=180,nbsphere=1,back=TRUE)
```

双标图




```
library(bpca)
plot(bpca(Dataset), var.factor=1.2,var.cex=1.2, var.col='red', obj.cex=1, main="2D biplot")
plot(bpca(Dataset, lambda.end=3),var.factor=1.2,var.cex=1, var.col='red', obj.cex=0.7,
main="3D biplot")
```




```
library(bpca)
plot(bpca(iris[-5]),var.factor=.3,
var.cex=.7,obj.names=FALSE,
obj.cex=1.5,obj.col=c('red', 'green3',
'blue')[unclass(iris$Species)], obj.pch=c(‘1’, ‘2’,
‘3')[unclass(iris$Species)])
```

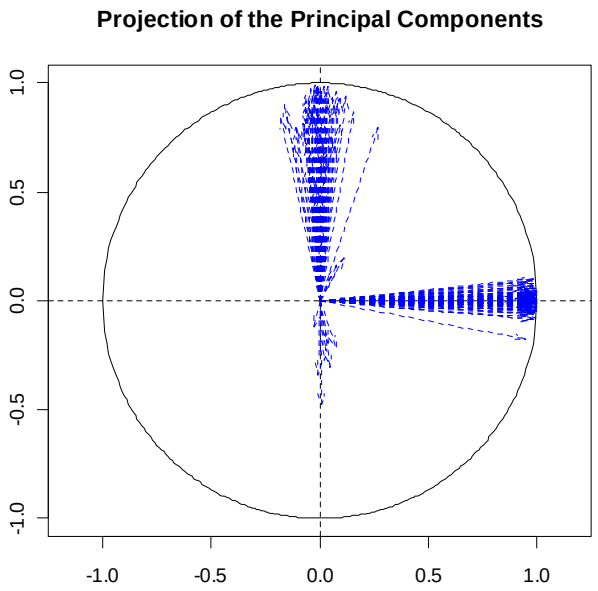
R中因子分析的应用

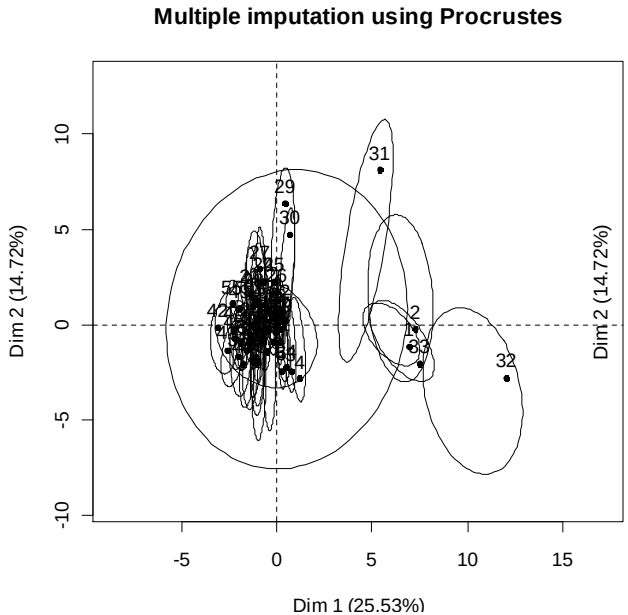
主成分 /因子分析用于缺失值处理

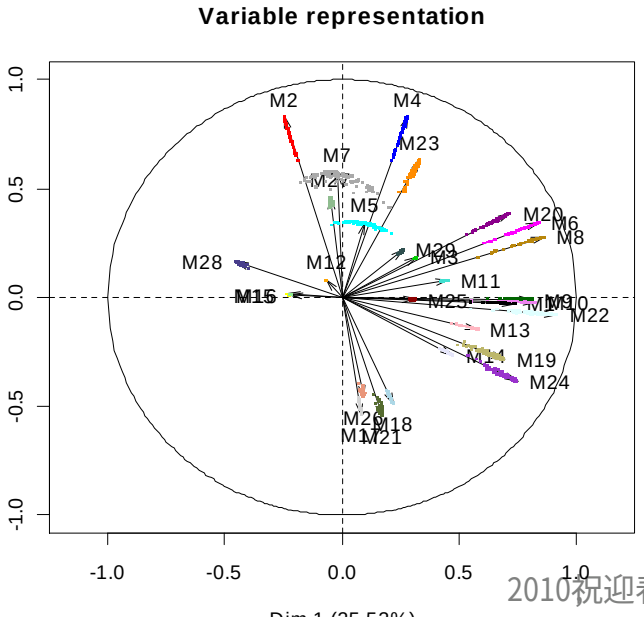
PCA and FA Missing Value Estimator

```
library(missMDA)
tdata <- read.table("D:/data/r2010/tdata.txt", header=TRUE, sep="", na.strings="NA", dec=".",
strip.white=TRUE)
npc <- estim_ncpPCA(tdata,ncp.min=1,ncp.max=4) #估计成分数目npc=1

resMI <- MIPCA(tdata,ncp=2) #多重插补Multiple Imputation with PCA
resMI$res.imputePCA
plot(resMI)
```







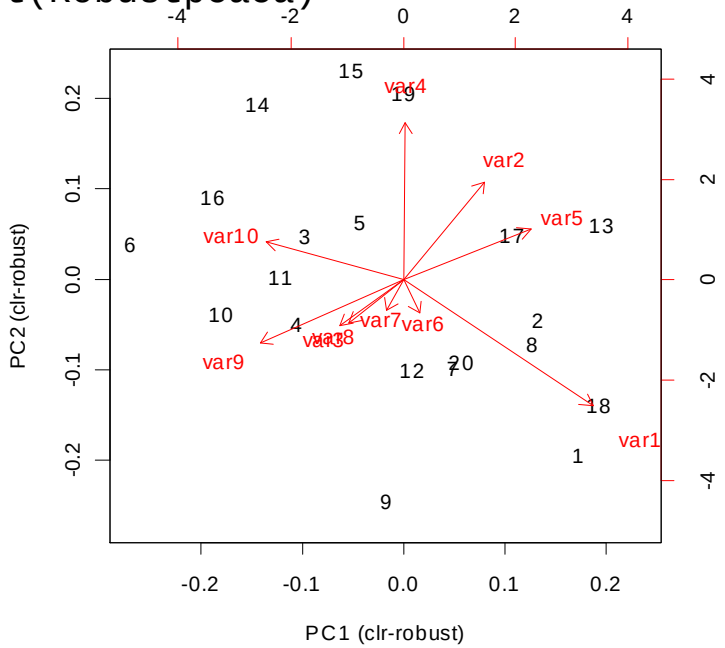
成分数据主成分 /因子分析

PCA and FA for compositional data

成分数据：$\Sigma X_j=1$，$0 \le X_j \le 1$，绝对共线性

PCA

```
x <- read.table("C:/Documents and
Settings/Administrator/桌面/Dataset.txt",
  header=TRUE, sep="", na.strings="NA",
dec=".", strip.white=TRUE)
Robustpcaca <- pcaCoDa(x)
p1$loadings  #负荷
p1$eigenvalues  #特征根
p1$scores  #得分
plot(Robustpcaca)
```



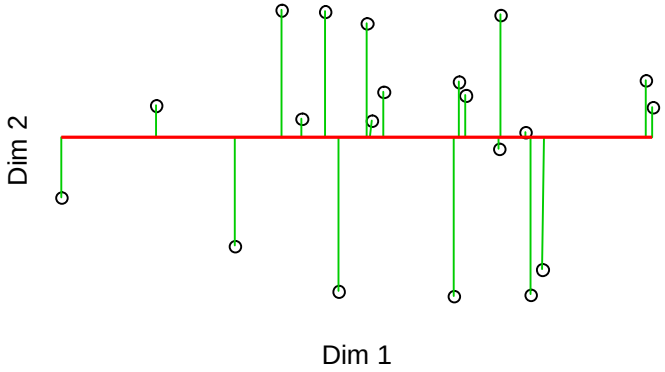
FA（主因子+varimax旋转）

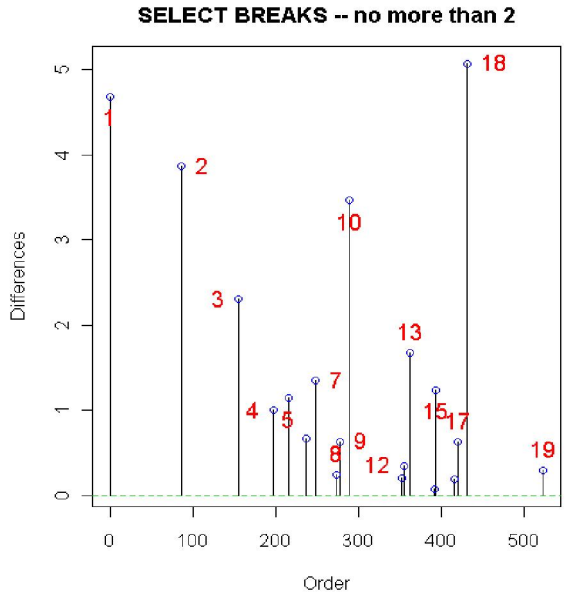
```
library(robCompositions)
x <- read.table("C:/Documents and
Settings/Administrator/桌面/Dataset.txt",
  header=TRUE, sep="", na.strings="NA",
dec=".", strip.white=TRUE)
# construct orthonormal basis:
V <- matrix(0,nrow=ncol(x),ncol=ncol(x)-1)
for (i in 1:ncol(V)){
  V[1:i,i] <- 1/i
  V[i+1,i] <- (-1)
  V[,i] <- V[,i]*sqrt(i/(i+1))
}
z <- ilr(x) #ilr transformed data
y <- z
set.seed(1245)
require(robustbase) 注意此包帮助少了 “结果$scores”
z.mcd <- covMcd(z)
mean_z <- z.mcd$center
mean_y <- V
var_z <- z.mcd$cov
var_y <- V
#classical scaling
y.sc <- scale(y,scale=FALSE) #only centering
#robust scaling
#y.rsc <- scale(y,mean_y,scale=FALSE) #only
centering
res1logcentr <- pfa(y.sc, factors=1,
scores="Bartlett", rotation="varimax")
```

主曲线分析

Principal Curve analysis

主曲线(principal curves)是第一主成分的非线性推广,第一主成分是对数据集的一维线性最优描述.主曲线强调寻找通过数据分布的“中间”(middle)并满足“自相合”的光滑一维曲线,其理论基础是寻找嵌入高维空间的非欧氏低维流形.《主曲线研究综述》（张军平）





小提示：pcdiagsplt <- pcdiags.plt(zz=pc2afit,xx=Dataset,pch=1,graphics=TRUE)

1.请用鼠标选择：5.Response-residuals plot

2.使用图形查看各个主成分是否有非线性关系

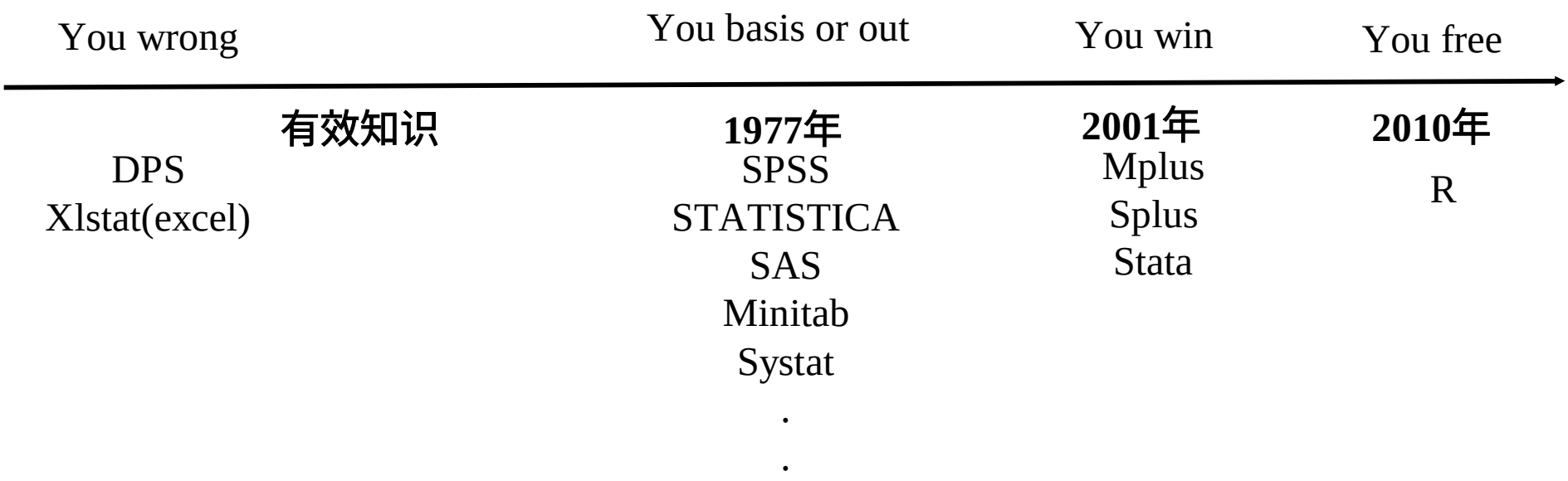
3.不推荐princurve包（不完整），推荐用pcurve

4.可以做诊断图

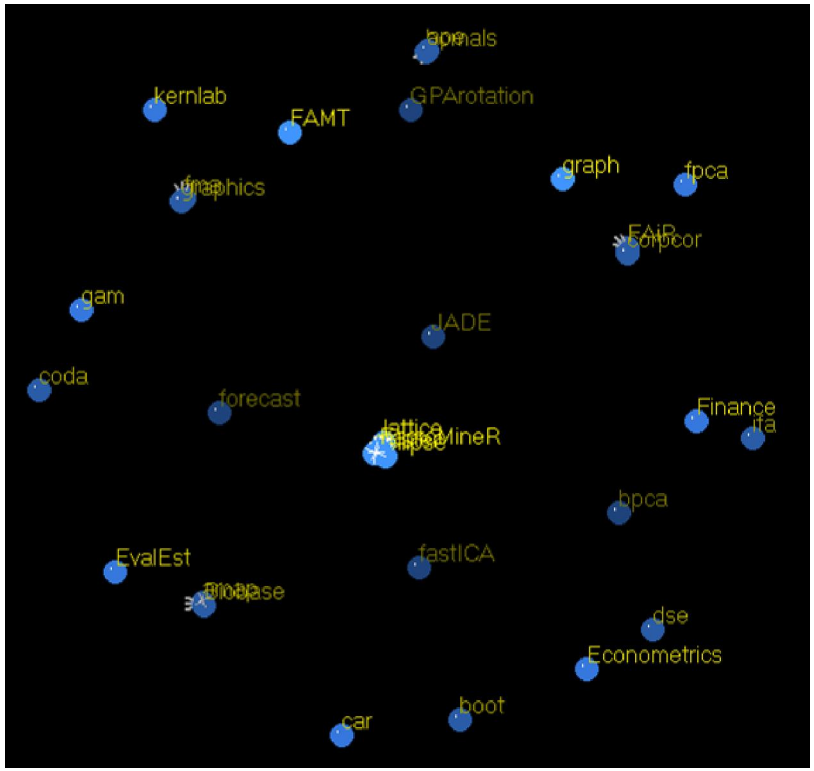
总结与展望

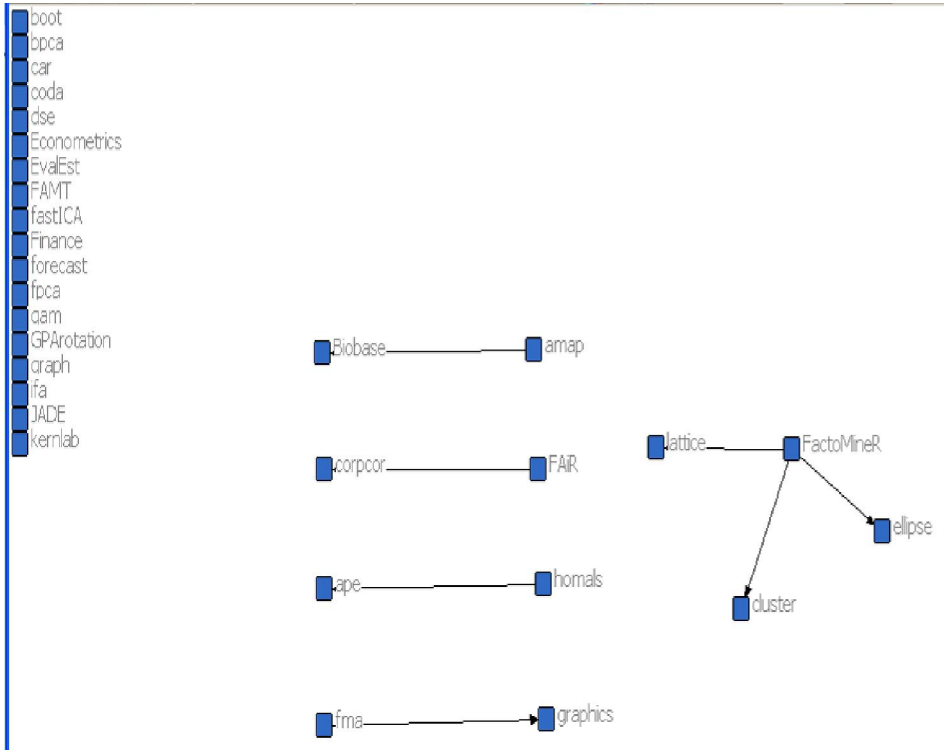
不同软件的因子分析技术

所谓时尚，就是享受当代各行各业顶尖的人的创作成果。”《越读者》（郝明义）



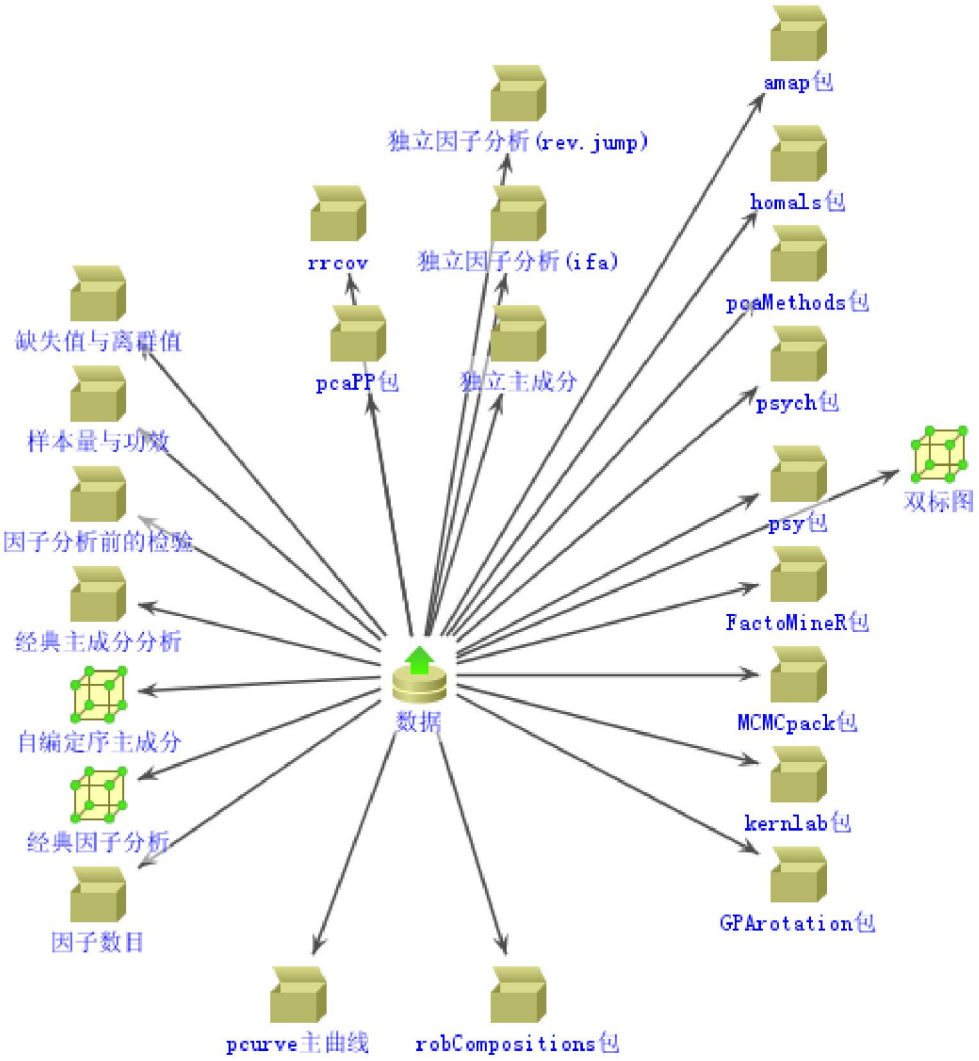
30个包的社会网络关系

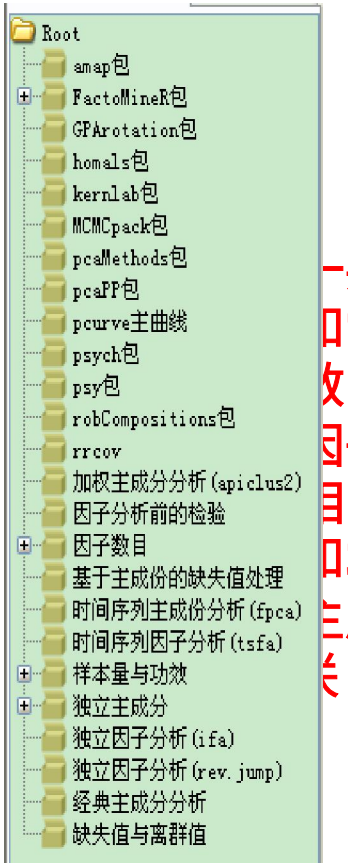




R是统计学的“九阴真经”：必须要用“梵文”？
解释第二届R会议的发言：代码意义上的算法与统计学算法

本讲义中所涉及的包和代码





一共30包
和75个函
数：36种
因子分析
相关技术
和35种与
主成分相
关的技术

主成分 /因子分析的变量选择

http://mo161.soci.ous.ac.jp/vasmm/indexE.html

日本人在因子分析技术上有很多贡献

逐步因子分析Stepwise variable selection in factor analysis

VASMM Since Feb., 2002

VASMM Project
Y.Mori
M.Iizuka
T.Tarumi
Y.Tanaka

Our URL
http://mo161.soci.ous.ac.jp/vasmm/

Select one among the selection modules below.

VASpca
Variable Selection in Principal Component Analysis

VASfa
Variable Selection in Factor Analysis

VAScorres
Variable Selection in Correspondence Analysis

VASMM

VARIABLE SELECTION IN MULTIVARIATE METHODS

- Introduction
- Variable selection in multivariate methods without external variables
- Statistical system VASMM
  - Implemented modules
  - Selection procedures
  - Flow of VASMM
  - Architecture of VASMM
- Policy to provide computational environment in this site
- References

感谢你的耐心

转载请保留知识产权

Thank You !

问题？