数据分析与R语言 第11周

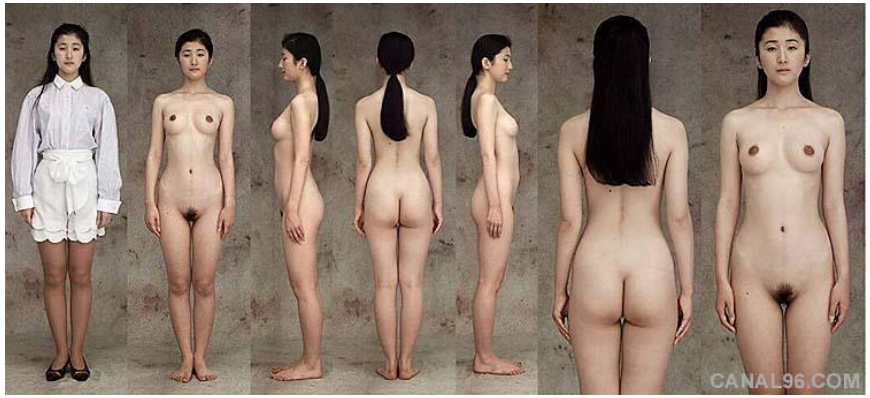
2012.7.22

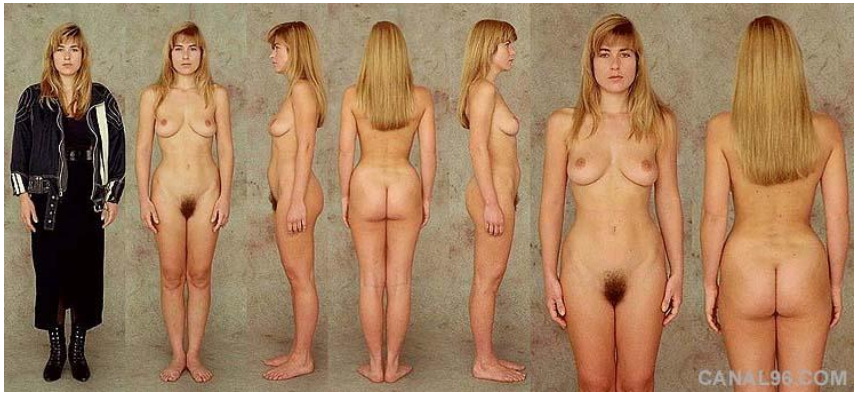
线粒体揭示人类起源的奥秘

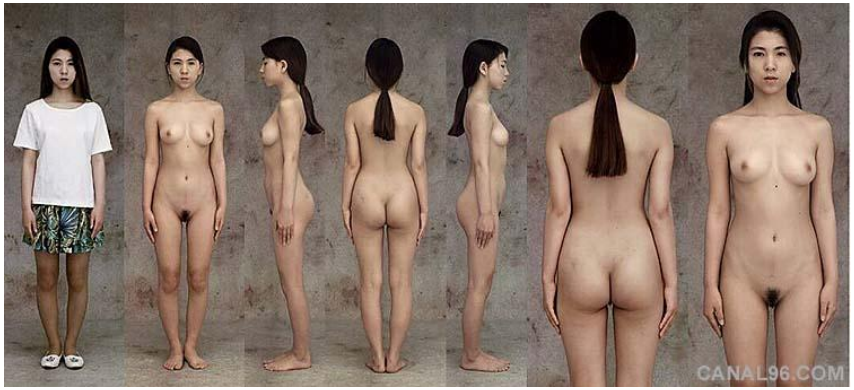
- 东非的夏娃
- Super king

2012.7.22

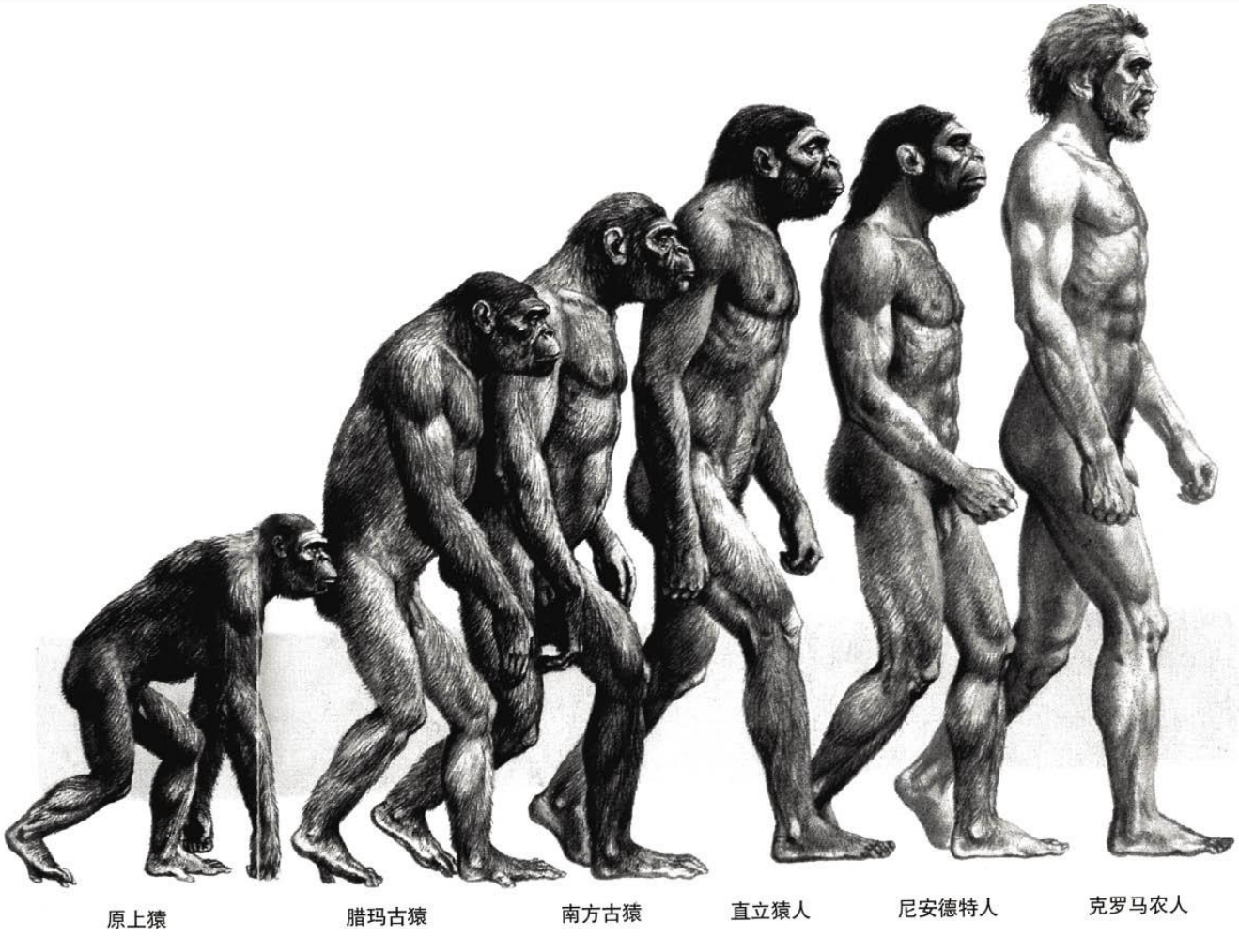
人种差异





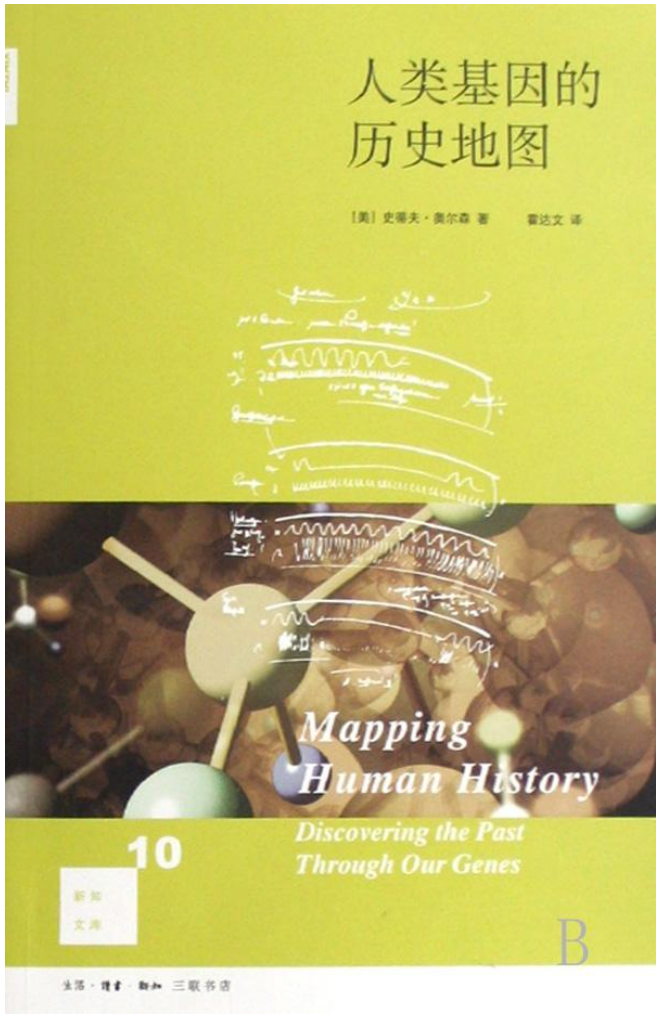


寻找原始人类的遗传物质



2012.7.22

人类基因的历史地图



2012.7.22

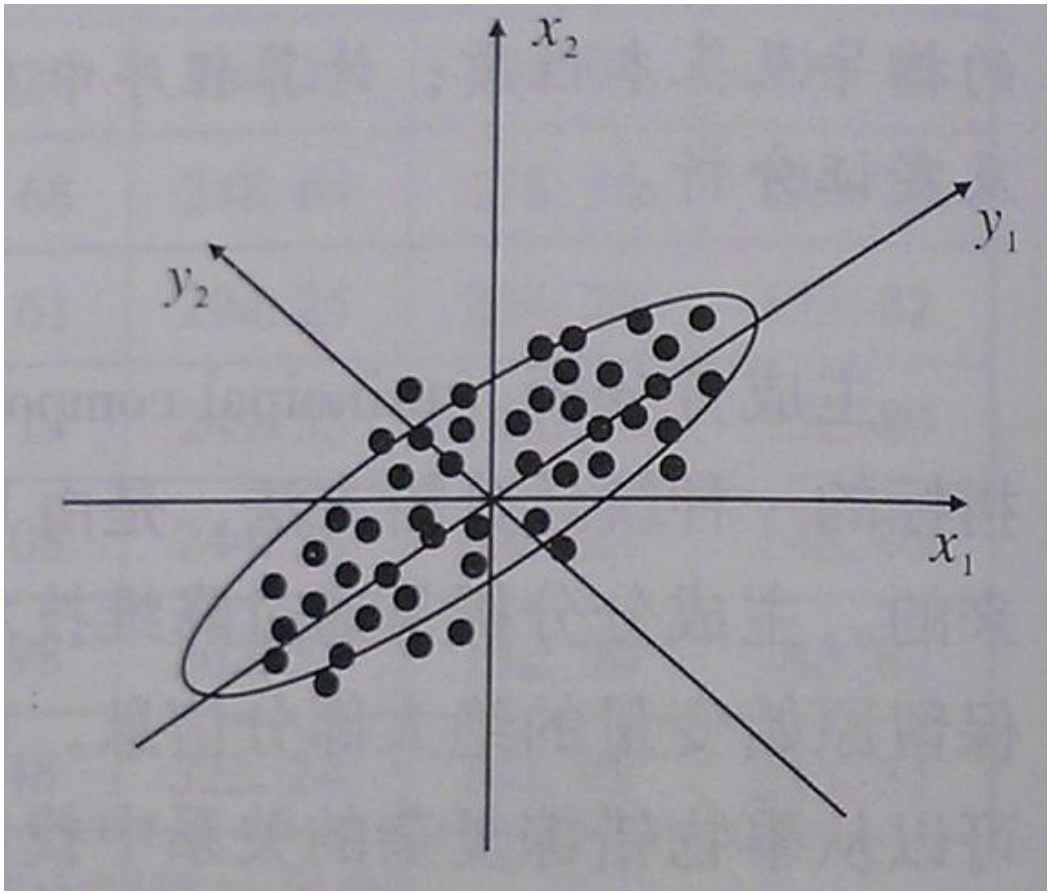
主成分分析

- 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

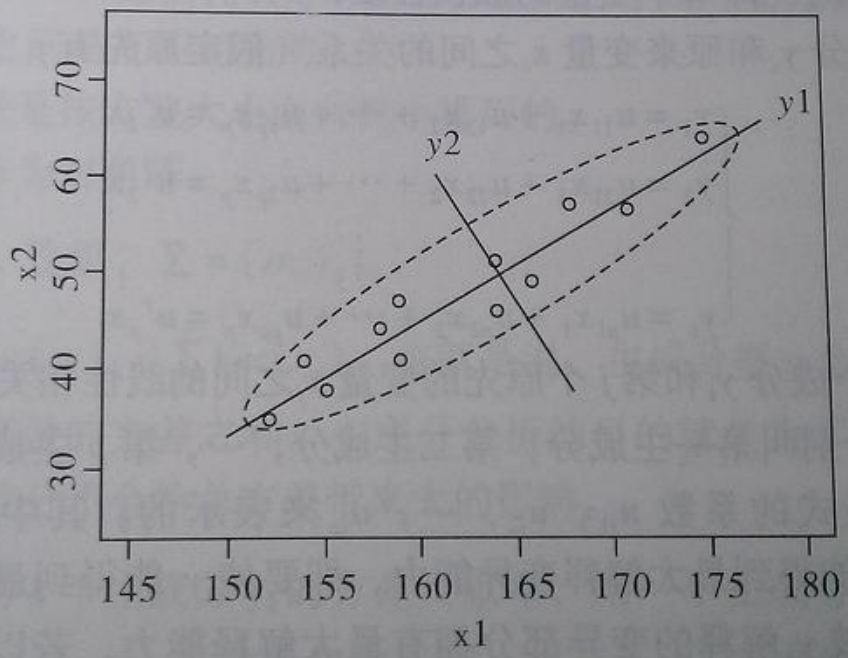
- 通过对原始变量进行线性组合，得到优化的指标
- 把原先多个指标的计算降维为少量几个经过优化指标的计算（占去绝大部分份额）
- 基本思想：**设法将原先众多具有一定相关性的指标，重新组合为一组新的互相独立的综合指标，并代替原先的指标**

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

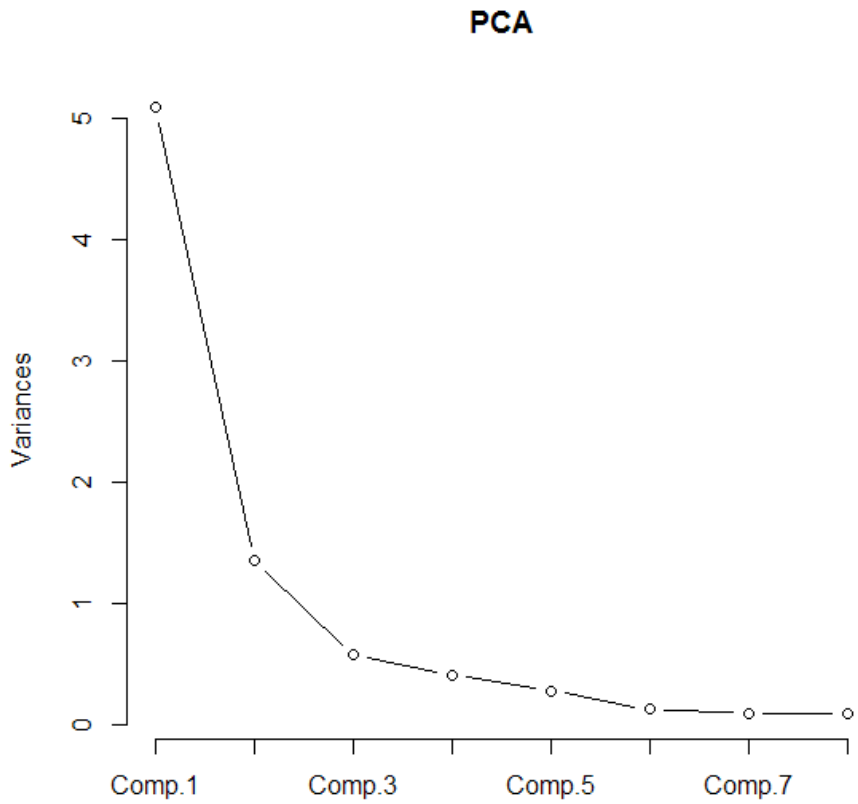
```
> PCA$loadings

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")
```



例子：主成分得分-相当于predict( )

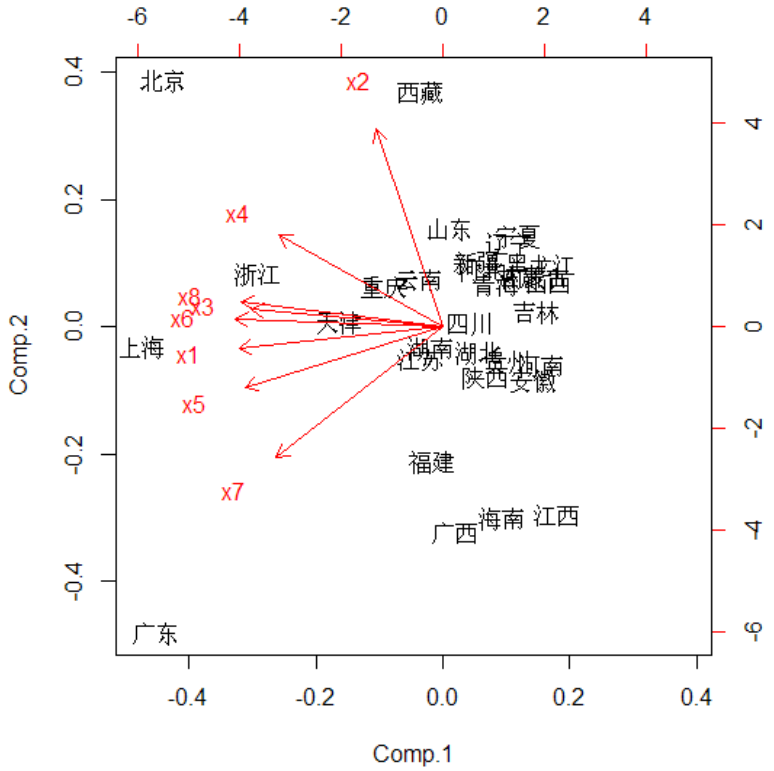
```
> PCA$score
          Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6       Comp
北京    -5.5068881  2.51368747 -0.77052784 -0.34499076 -0.48456544  0.73526042  0.142820
天津    -2.0391525  0.04696816 -0.83866069  0.84294280 -0.23905123 -0.36965072  0.438523
河北     0.7647412  0.58939950 -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

Thanks

FAQ时间