典则相关分析

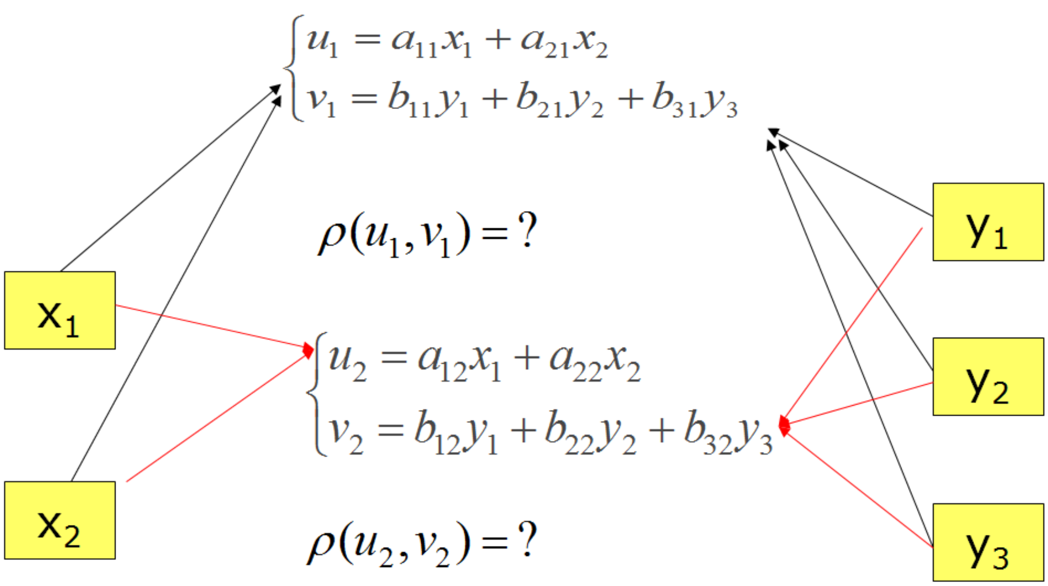
王宁宁

典型相关分析的基本思想

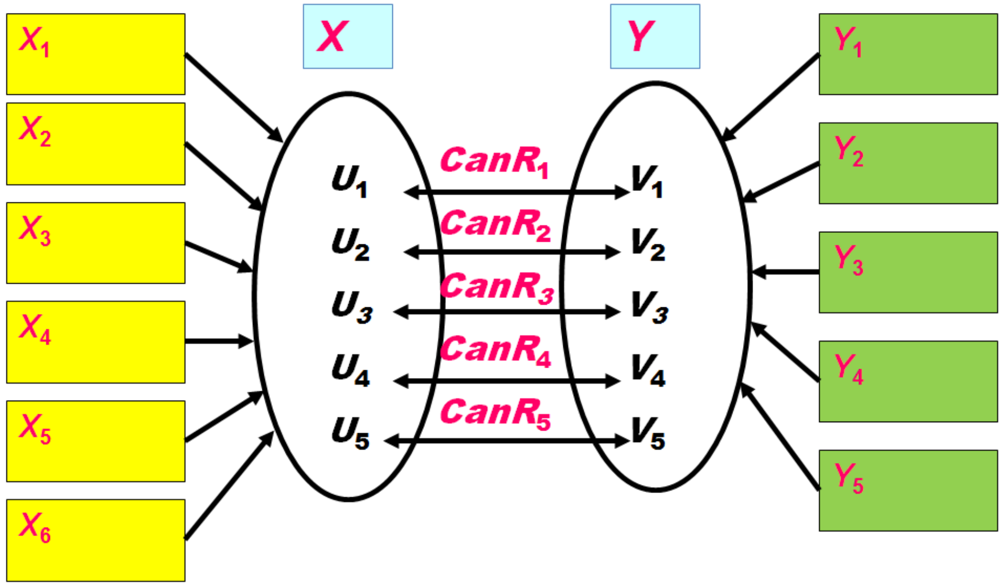
例如为了了解家庭的特征与其消费模式之间的关系。调查了若干个家庭的下面两组变量：

$$\begin{cases} x_1\text{：每年去餐馆就餐的频率} \\ x_2\text{：每年外出看电影频率} \end{cases} \qquad \begin{cases} y_1\text{：户主的年龄} \\ y_2\text{：家庭的年收入} \\ y_3\text{：户主受教育程度} \end{cases}$$

现在要研究两组变量的关系，如果是两个变量可以计算相关系数，但是两组变量如何研究呢？



典型相关分析示意图



- 典型相关分析，又称为典则相关分析，（canonical correlation analysis），是分析两组变量间线性相关关系的一种统计分析方法。
- 典型相关分析的基本思想类似主成分分析，它根据变量间的相关关系，寻找几个简单的综合变量（可看作主成分）替代关系复杂的实际观测变量，将两组变量间的多重线性相关关系转化为少数几对综合变量（主成分）间的简单线性相关。此时，少数几对综合变量（主成分）所包含的相关性信息覆盖了原变量组间所包含的大部分信息。

如果我们记两组变量的第一对线性组合为：

$$\begin{aligned} X &= (X_1, \cdots, X_p) \\ Y &= (Y_1, \cdots, Y_q) \\ u_1 &= \alpha_1' X \\ v_1 &= \beta_1' Y \end{aligned}$$

其中：$\alpha_1 = (\alpha_{11}, \alpha_{21}, \cdots, \alpha_{p1})'$ $\beta_1 = (\beta_{11}, \beta_{21}, \cdots, \beta_{q1})'$ Σ是协方差矩阵 考虑：

$$\begin{aligned} Var(u_1) &= \alpha_1' Var(X) \alpha_1 = \alpha_1' \Sigma_{11} \alpha = 1 \\ Var(v_1) &= \beta_1' Var(Y) \beta_1 = \beta_1' \Sigma_{22} \beta = 1 \\ \rho_{u_1, v_1} &= Cov(u_1, v_1) = \alpha_1' Cov(X, Y) \beta_1 = \alpha_1' \Sigma_{12} \beta_1 \end{aligned}$$

典型相关分析就是求α_1, β_1，使两者的相关系数ρ 达到最大。

计算过程

1. 对变量X 和Y 标准化
2. 求X 和Y 的相关系数矩阵 R：

$$R=\begin{bmatrix}R_{XX} & R_{XY}\\ R_{YX} & R_{YY}\end{bmatrix}=\begin{bmatrix}\sum XX & \sum XY\\ \sum YX & \sum YY\end{bmatrix}_{(p+q)\times(p+q)}$$

3. 求A 和B：

$$A=\quad (R_{XX})^{-1}R_{XY}(R_{YY})^{-1}R_{YX}$$
$$B=\quad (R_{YY})^{-1}R_{YX}(R_{XX})^{-1}R_{XY}$$

4. 分布求A 和B 的特征值和特征根，设A 的第 i 个特征值λ_i对应的特征向量为：$(a_{i1},\cdots,a_{ip})$，设B 的第 i 个特征值λ_i对应的特征向量为：$(b_{i1},\cdots,b_{iq})$
5. 计算V_i 和W_i：

$$V_i=\quad a_{i1}X_1+\cdots+a_{ip}X_p$$
$$W_i=\quad b_{i1}Y_1+\cdots+b_{iq}Y_q$$

6. V_i 和W_i 的相关系数为$r_i=\sqrt{\lambda_i}$

典型相关系数的检验

典型相关分析是否恰当，应该取决于两组原变量之间是否相关，如果两组变量之间毫无相关性而言，则不应该作典型相关分析。用样本来估计总体的典型相关系数是否有误，需要进行检验。

H_0: $\rho_1=\cdots=\rho_r=0$ H_1: ρ_i 中，至少ρ_1 不为 0

检验统计量： 不放设：$r=\min(p,q)=p$

$$\Lambda_0=\frac{|S|}{|S_{XX}||S_{YY}|}=\prod_{i=1}^{p}(1-\lambda_i^2)$$

Λ_0 越小，越支持备则假设。

$$Q_0=-[(n-1)-\frac{1}{2}(p+q+1)]\ln\Lambda_0$$

在原假设成立的条件下，Q_0 服从自由度为$p\times q$ 的卡方分布。

实例

康复俱乐部对 20 名中年人测量了三个生理指标：体重(x1),腰围(x2)，脉搏(x3)；三个训练指标：引体向上次数(y1)，起坐次数(y2)，跳跃次数(y3)。分析生理指标与训练指标的相关性。

```
ex1=read.table("http://statstudy.github.io/data/9-1.txt",head=T)
ex1
```

```
##     x1 x2 x3 y1  y2  y3
## 1  191 36 50  5 162  60
## 2  189 37 52  2 110  60
## 3  193 38 58 12 101 101
## 4  162 35 62 12 105  37
## 5  189 35 46 13 155  58
## 6  182 36 56  4 101  42
## 7  211 38 56  8 101  38
## 8  167 34 60  6 125  40
## 9  176 31 74 15 200  40
## 10 154 33 56 17 251 250
## 11 169 34 50 17 120  38
## 12 166 33 52 13 210 115
## 13 154 34 64 14 215 105
## 14 247 46 50  1  50  50
## 15 193 36 46  6  70  31
## 16 202 37 62 12 210 120
## 17 176 37 54  4  60  25
## 18 157 32 52 11 230  80
## 19 156 33 54 15 225  73
## 20 138 33 68  2 110  43
```

```
x=ex1[,1:3]
y=ex1[,4:6]

s11=cor(x);s11
```

```
##            x1         x2         x3
## x1  1.0000000  0.8702435 -0.3657620
## x2  0.8702435  1.0000000 -0.3528921
## x3 -0.3657620 -0.3528921  1.0000000
```

```
s22=cor(y);s22
```

```
##           y1        y2        y3
## y1 1.0000000 0.6957274 0.4957602
## y2 0.6957274 1.0000000 0.6692061
## y3 0.4957602 0.6692061 1.0000000
```

```
s12=cor(ex1)[1:3,4:6];s12
```

```
##            y1         y2          y3
## x1 -0.3896937 -0.4930836 -0.22629556
## x2 -0.5522321 -0.6455980 -0.19149937
## x3  0.1506480  0.2250381  0.03493306
```

```
s21=cor(ex1)[4:6,1:3];s21
```

```
##            x1         x2         x3
## y1 -0.3896937 -0.5522321 0.15064802
## y2 -0.4930836 -0.6455980 0.22503808
## y3 -0.2262956 -0.1914994 0.03493306
```

```
A=solve(s11)%*%s12%*%solve(s22)%*%s21
A

##              x1          x2          x3
## x1 -0.24594544 -0.42556193  0.15927694
## x2  0.58342555  0.90714323 -0.32827241
## x3 -0.01679293 -0.03129267  0.01728371

eigen(A)$vectors[,1]

## [1]  0.4404622 -0.8971428  0.0335830

sqrt(eigen(A)$values)

## [1] 0.79560815 0.20055604 0.07257029

B=solve(s22)%*%s21%*%solve(s11)%*%s12
B

##             y1         y2          y3
## y1  0.1617883  0.1718776  0.02299820
## y2  0.4824417  0.5487737  0.11144827
## y3 -0.3184295 -0.3464725 -0.03208051

eigen(B)

## $values
## [1] 0.632992335 0.040222726 0.005266446
##
## $vectors
##            [,1]       [,2]       [,3]
## [1,] -0.2644705 -0.3313572 -0.7046067
## [2,] -0.7975886  0.1089606  0.6721095
## [3,]  0.5421327  0.9371926 -0.2275921

sqrt(eigen(B)$values)

## [1] 0.79560815 0.20055604 0.07257029

A0=prod(1-eigen(A)$values)
A0

## [1] 0.3503905

Q0=-(20-1-0.5*(3+3+1))*log(A0)
pr=1-pchisq(Q0,9)
pr

## [1] 0.06174456

m1=cancor(x,y)
m1

## $cor
## [1] 0.79560815 0.20055604 0.07257029
```

```
##
## $xcoef
##              [,1]         [,2]         [,3]
## x1 -0.007204730 -0.017508896  0.001774541
## x2  0.113157401  0.084590855 -0.036255405
## x3 -0.001881052 -0.007353232 -0.033433269
##
## $ycoef
##              [,1]          [,2]         [,3]
## y1 -0.015167589 -0.0162979716  0.056270024
## y2 -0.003864790  0.0004528082 -0.004535007
## y3  0.003205298  0.0047521419  0.001873747
##
## $xcenter
##    x1    x2    x3
## 178.6  35.4  56.1
##
## $ycenter
##     y1     y2     y3
##   9.45 145.55  70.30
```

```
corcoef.test<-function(r, n, p, q, alpha=0.1){
  #r为相关系数 n为样本个数 且n>p+q
  m<-length(r);  Q<-rep(0, m);  lambda <- 1
  for (k in m:1){
    lambda<-lambda*(1-r[k]^2);   #检验统计量
    Q[k]<- -log(lambda)   #检验统计量取对数
  }
  s<-0;  i<-m
  for (k in 1:m){
    Q[k]<- (n-k+1-1/2*(p+q+3)+s)*Q[k]  #统计量
    chi<-1-pchisq(Q[k], (p-k+1)*(q-k+1))
        if (chi>alpha){
      i<-k-1; break
    }
    s<-s+1/r[k]^2
  }
  i  #显示输出结果 选用第几对典型变量
}

corcoef.test(cancor(x,y)$cor,n=20,p=3,q=3,alpha=0.1)
```

```
## [1] 1
```