第五章 典型相关分析

对于两个变量，用它们的相关系数来衡量它们之间的线性相关关系。当考虑一个变量与一组变量的线性相关关系时，用它们的多重相关系数来衡量。但是，在许多实际问题中，常常会碰到两组变量之间的线性相关性问题研究。

例如，考虑几种主要产品的价格（作为第一组变量）和相应这些产品的销售量（作为第二组变量）之间的相关关系；考虑投资性变量（如劳动者人数、货物周转量、生产建设投资等）与国民收入变量（如工农业国民收入、运输业国民收入、建筑业国民收入等）之间的相关关系等等；再如研究患者的各种临床症状 $(x_1, x_2, \cdots, x_p)$ 与所患各种疾病 $(y_1, y_2, \cdots, y_q)$ 之间的线性相关性。

5.1 典型相关分析简介

对于这类问题的研究，可以用最原始的方法，分别计算两组变量之间的全部相关系数，一共有 pq个简单相关系数（多重线性相关性），这样又繁琐又不能抓住问题的本质。

Hotelling在主成分分析和因子分析的基础上引进了典型相关系数的概念，从而找到了揭示两组变量之间线性相关关系的一种统计分析方法—典型相关分析。

第五章 典型相关分析

典型相关分析是分析两组变量之间相关性的一种统计分析方法，它包含了简单的Pearson相关分析（两个组均含一个变量）和复相关分析（一个组含有一个变量，而另一个组含有多个变量）这两种特殊情况。典型相关分析的基本思想和主成分分析的基本思想相似，它将一组变量与另一组变量之间单变量的多重线性相关性研究，

转化为少数几对综合变量之间的简单线性相关性的研究，并且这少数几对变量所包含的线性相关性的信息几乎覆盖了原变量组所包含的全部相应信息。

5.2 典型相关分析的基本思想

典型相关分析方法的基本原理是：所有研究的两组变量为x组和y组，x组有p个变量 $(x_1, x_2, \cdots, x_p)$，y组有q个变量 $(y_1, y_2, \cdots, y_q)$ ，则分别对这两组变量各做线性组合后，再计算此两加权和的简单相关系数，然后以这个简单相关系数当做这两组变量之间相关性的衡量指标，即

5.2 典型相关分析的基本思想

$$s = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

$$t = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q$$

这样的 s 和 t 称为**典型变量**，典型变量 s 和 t 之间的相关系数称为**典型相关系数**。

注：1. 典型变量成对出现，有很多，因为改变权值系数，就得到不同的典型变量。由于典型变量s 和 t线性变换，保持典型相关系数不变，所以仅考虑方差为1的典型变量s 和 t 。

5.2 典型相关分析的基本思想

注： 2. 典型相关系数最大的那对典型变量称为第1（对）典型变量；典型相关系数第2大的那对典型变量称为第2（对）典型变量；以此类推。

注： 3. 如何求典型相关系数及其典型变量呢?

5.2 典型相关分析的基本思想

通常情况下，为了研究两组变量

$$[x_1,x_2,\cdots,x_p],\ [y_1,y_2,\cdots,y_q]$$

首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，

$$\begin{cases} u_1=\alpha_{11}x_1+\alpha_{21}x_2+\cdots+\alpha_{p1}x_p, \\ v_1=\beta_{11}y_1+\beta_{21}y_2+\cdots+\beta_{q1}y_q. \end{cases}$$

然后再在每组变量中找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性。

$$\begin{cases} u_2=\alpha_{12}x_1+\alpha_{22}x_2+\cdots+\alpha_{p2}x_p, \\ v_2=\beta_{12}y_1+\beta_{22}y_2+\cdots+\beta_{q2}y_q. \end{cases}$$

5.2 典型相关分析的基本思想

u_2与u_1、v_2与v_1不相关，但u_2和v_2相关。如此继续下去，直至进行到r步，两组变量的相关性被提取完为止，可以得到r组变量，这里$r \le \min(p,q)$。

5.3 典型相关分析的具体计算

令 $$x_1^* = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p$$

$$= (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p})(x_1, x_2, \cdots, x_p)^T \hat{=} \alpha_1^T x$$

$$y_1^* = \beta_{11}y_1 + \beta_{12}y_2 + \cdots + \beta_{1q}y_q$$

$$= (\beta_{11}, \beta_{12}, \cdots, \beta_{1q})(y_1, y_2, \cdots, y_q)^T \hat{=} \beta_1^T y$$

5.3 典型相关分析的具体计算

$$[x_1,x_2,\cdots,x_p],\ [y_1,y_2,\cdots,y_q]$$

设x组的协方差阵为Σ_{11}，y组的协方差阵为 Σ_{22} ，x与y的协方差阵为 Σ_{12} ，则x1*和y1*之间的简单相关系数为

$$\rho(x_1^*,y_1^*)=\frac{Cov(x_1^*,y_1^*)}{\sqrt{Var(x_1^*)\times Var(y_1^*)}}$$

$$=\frac{Cov(\alpha_1^T x,\beta_1^T y)}{\sqrt{Var(\alpha_1^T x)\times Var(\beta_1^T y)}}=\frac{\alpha_1^T\Sigma_{12}\beta_1}{\sqrt{\alpha_1^T\Sigma_{11}\alpha_1\times\beta_1^T\Sigma_{22}\beta_1}}$$

$$x_1^*=\alpha_{11}x_1+\alpha_{12}x_2+\cdots+\alpha_{1p}x_p$$
$$=(\alpha_{11},\alpha_{12},\cdots,\alpha_{1p})(x_1,x_2,\cdots,x_p)^T\overset{\wedge}{=}\alpha_1^T x$$

$$y_1^*=\beta_{11}y_1+\beta_{12}y_2+\cdots+\beta_{1q}y_q$$
$$=(\beta_{11},\beta_{12},\cdots,\beta_{1q})(y_1,y_2,\cdots,y_q)^T\overset{\wedge}{=}\beta_1^T y$$

5.3 典型相关分析的具体计算

x_1*和y_1*之间的简单相关系数 $\rho(x_1^*, y_1^*)$ 最大等价于

在 $\alpha_1^T \Sigma_{11} \alpha_1 = 1$ 和 $\beta_1^T \Sigma_{22} \beta_1 = 1$ 条件下，

$\rho(x_1^*, y_1^*) = \alpha_1^T \Sigma_{12} \beta_1$ —>max, 达到最大，

等价于求使 $\rho(x_1^*, y_1^*) = \alpha_1^T \Sigma_{12} \beta_1$

达到最大的 α_1, β_1 。

5.3 典型相关分析的具体计算

构造拉格朗日函数

$$Q=\alpha_1^T\Sigma_{12}\beta_1-\frac{1}{2}\lambda_1(\alpha_1^T\Sigma_{11}\alpha_1-1)-\frac{1}{2}\lambda_2(\beta_1^T\Sigma_{22}\beta_1-1)$$

求Q的偏导数，并令其为零，有

5.3 典型相关分析的具体计算

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_1} = \Sigma_{12}\beta_1 - \lambda_1\Sigma_{11}\alpha_1 = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = \Sigma_{21}\alpha_1 - \lambda_2\Sigma_{22}\beta_1 = 0 \\ \dfrac{\partial Q}{\partial \lambda_1} = -\dfrac{1}{2}(\alpha_1^T\Sigma_{11}\alpha_1 - 1) = 0 \\ \dfrac{\partial Q}{\partial \lambda_2} = -\dfrac{1}{2}(\beta_1^T\Sigma_{22}\beta_1 - 1) = 0 \end{cases}$$

5.3 典型相关分析的具体计算

前两个方程分别乘以 α_1^T, β_1^T ，有

$$\begin{cases} \alpha_1^T \Sigma_{12} \beta_1 = \lambda_1 \alpha_1^T \Sigma_{11} \alpha_1 = \lambda_1 \\ \beta_1^T \Sigma_{21} \alpha_1 = \lambda_2 \beta_1^T \Sigma_{22} \beta_1 = \lambda_2 \end{cases}$$

故 $\lambda_1 = \lambda_2 = \lambda$

$$\rho(x_1^*, y_1^*) = \alpha_1^T \Sigma_{12} \beta_1 = \lambda$$

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_1} = \Sigma_{12}\beta_1 - \lambda_1 \Sigma_{11}\alpha_1 = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = \Sigma_{21}\alpha_1 - \lambda_2 \Sigma_{22}\beta_1 = 0 \\ \dfrac{\partial Q}{\partial \lambda_1} = -\dfrac{1}{2}(\alpha_1^T \Sigma_{11}\alpha_1 - 1) = 0 \\ \dfrac{\partial Q}{\partial \lambda_2} = -\dfrac{1}{2}(\beta_1^T \Sigma_{22}\beta_1 - 1) = 0 \end{cases}$$

5.3 典型相关分析的具体计算

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_1} = \Sigma_{12}\beta_1 - \lambda_1\Sigma_{11}\alpha_1 = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = \Sigma_{21}\alpha_1 - \lambda_2\Sigma_{22}\beta_1 = 0 \\ \dfrac{\partial Q}{\partial \lambda_1} = -\dfrac{1}{2}(\alpha_1^T\Sigma_{11}\alpha_1 - 1) = 0 \\ \dfrac{\partial Q}{\partial \lambda_2} = -\dfrac{1}{2}(\beta_1^T\Sigma_{22}\beta_1 - 1) = 0 \end{cases}$$

从而前两个方程变为

$$\Sigma_{12}\beta_1 = \lambda\Sigma_{11}\alpha_1$$

$$\Sigma_{21}\alpha_1 = \lambda\Sigma_{22}\beta_1$$

故 $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\beta_1 = \lambda^2\beta_1$

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha_1 = \lambda^2\alpha_1$$

5.3 典型相关分析的具体计算

令

$$A=\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$B=\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

则

$$B\beta_1=\lambda^2\beta_1$$

$$A\alpha_1=\lambda^2\alpha_1$$

5.3 典型相关分析的具体计算

定理1：A和B有相同的非零特征值，且特征值非负。

定理2：第1典型相关系数就是A和B最大的特征值的平方根。

5.3 典型相关分析的具体计算

定理3：

第1对典型变量x1*和y1*中的 α_1, β_1

分别满足

$$A\alpha_1 = \lambda^2\alpha_1$$

$$B\beta_1 = \lambda^2\beta_1$$

5.3 典型相关分析的具体计算

问题：

第2对典型变量怎么计算，α_1, β_1 需要满足什么条件？

第五章 典型相关分析

对于

$$A\alpha_1 = \lambda^2 \alpha_1$$

$$B\beta_1 = \lambda^2 \beta_1$$

A为p阶矩阵，B为q阶矩阵，非零特征值的个数等于$\min(p,q)$，不妨设为q。

将q个非零特征值排序为$\lambda_1^2 \ge \lambda_2^2 \ge \cdots \ge \lambda_q^2$，其余$p-q$个特征值为0，我们称$\lambda_1,\lambda_2,\cdots,\lambda_q$为典型相关系数。

相应地，从$A\alpha_i = \lambda_i^2 \alpha_i$和$B\beta_i = \lambda_i^2 \beta_i$解出的特征向量为线性组合系数

$$\boxed{A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}$$

典型相关分析的步骤： $(x_1, x_2, \cdots, x_p)$ $(y_1, y_2, \cdots, y_q)$

Step1： 计算x组的协方差阵为Σ_{11}，y组的协方差阵为Σ_{22} ，x与y的协方差阵为Σ_{12}

Step2： 计算 $A=\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ $B=\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Step3： 确定非零特征根的数量 $\min(p,q)$

Step4： 由A或B计算，q个非零特征值排序为$\lambda_1^2 \ge \lambda_2^2 \ge \cdots \ge \lambda_q^2$

$\lambda_1, \lambda_2, \cdots, \lambda_q$为典型相关系数

Step5： 由$A\alpha_i = \lambda_i^2 \alpha_i$和$B\beta_i = \lambda_i^2 \beta_i$，计算特征向量 α_i 和 β_i，为典型变量的线性组合系数

第五章 典型相关分析

例1：为研究空气温度与土壤温度的关系，考虑如下六个变量：X1日最高土壤温度，X2日最低土壤温度，X3日土壤温度曲线积分值，Y1日最高气温，Y2日最低气温，Y3日气温曲线积分值，共观测了20天，部分数据如表所示，试做土壤温度与空气温度的典型相关分析。

Obs	X1	X2	X3	Y1	Y2	Y3	Obs	X1	X2	X3	Y1	Y2	Y3
1	85	59	151	84	65	147	11	91	76	206	88	73	176
2	86	61	159	84	65	149	12	94	76	211	90	74	187
3	83	64	152	79	66	142	13	94	75	211	88	72	171
4	83	65	158	81	67	147	14	92	70	201	58	72	171
5	88	69	180	84	68	167	15	87	68	167	81	69	154
6	77	67	147	74	66	131	16	83	68	162	79	68	149
7	78	69	159	73	66	131	17	87	66	173	84	69	160
8	84	68	159	75	67	134	18	87	68	177	84	70	160
9	89	71	195	84	68	161	19	88	70	169	84	70	168
10	91	76	206	86	72	169	20	83	66	170	77	67	147

第五章 典型相关分析

计算x组的协方差阵为Σ_{11}，y组的协方差阵为Σ_{22}，x与y的协方差阵为Σ_{12}

假设有X组和Y组变量，样本容量为20，观测值矩阵为

$$\begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & y_{1,1} & y_{1,2} & y_{1,3} \\ x_{2,1} & x_{2,1} & x_{2,3} & y_{2,1} & y_{2,2} & y_{2,3} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{20,1} & x_{20,2} & x_{20,3} & y_{20,1} & y_{20,2} & y_{20,3} \end{bmatrix}_{20\times 6},$$

Data=xlsread(“data_chapter5.xls”)

Covz=cov(data)

$$\begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}$$

第五章 典型相关分析

计算x组的协方差阵为Σ_{11}，y组的协方差阵为Σ_{22} ，x与y

的协方差阵为Σ_{12}

$$\begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}$$

```
cgm11=covz(1:3,1:3)
cgm12=covz(1:3,4:6)
cgm21=cgm12'
cgm22=covz(4:6,4:6)
```

$$A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

```
A=inv(cgm11)*cgm12*inv(cgm22)*cgm21
B=inv(cgm22)*cgm21*inv(cgm11)*cgm12
```

[a,lbta]=eig(A)

[b,lbtb]=eig(B)

```
A = 3x3

    0.7396    0.4091    2.7630
    0.1253    0.8293    1.7860
    0.0128   -0.0705   -0.0189

B = 3x3

    0.0830    0.0365    0.0122
    0.3192    0.7781    1.2377
    0.0948    0.0178    0.6888
```

```
a = 3x3

   -0.8259   -0.2915   -0.8174
   -0.5630    0.9488   -0.5078
    0.0312   -0.1217    0.2720

lbta = 3x3

    0.9140         0         0
         0    0.5617         0
         0         0    0.0742
```

```
b = 3x3

    0.9717   -0.0451   -0.0699
   -0.1867   -0.9942   -0.9794
   -0.1444   -0.0976    0.1893

lbtb = 3x3

    0.0742         0         0
         0    0.9140         0
         0         0    0.5617
```

第五章 典型相关分析

lbt=sqrt(diag(lbta))

3个典型相关系数： 0.9561　　0.7495　　0.2724

```
a = 3×3
   -0.8259   -0.2915   -0.8174
   -0.5630    0.9488   -0.5078
    0.0312   -0.1217    0.2720

b = 3×3
    0.9717   -0.0451   -0.0699
   -0.1867   -0.9942   -0.9794
   -0.1444   -0.0976    0.1893

lbta = 3×3
    0.9140         0         0
         0    0.5617         0
         0         0    0.0742

lbtb = 3×3
    0.0742         0         0
         0    0.9140         0
         0         0    0.5617
```

3对典型变量：

S1=-0.8259 x1 -0.5630 x2 +0.0312 x3

T1=-0.0451 y1 -0.9942 y2 -0.0976 y3

S2=-0.2915 x1 +0.9488 x2 -0.1217 x3

T2=-0.0699 y1 -0.9794 y2 +0.1893 y3

S3= -0.8174 x1 -0.5078 x2 +0.2720 x3

T3= 0.9717 y1 -0.1867 y2 – 0.1444y3

第五章 典型相关分析

例1：为研究空气温度与土壤温度的关系，考虑如下六个变量：X1日最高土壤温度，X2日最低土壤温度，X3日土壤温度曲线积分值，Y1日最高气温，Y2日最低气温，Y3日气温曲线积分值，共观测了20天，部分数据如表所示，试做土壤温度与空气温度的典型相关分析。

Obs	X1	X2	X3	Y1	Y2	Y3	Obs	X1	X2	X3	Y1	Y2	Y3
1	85	59	151	84	65	147	11	91	76	206	88	73	176
2	86	61	159	84	65	149	12	94	76	211	90	74	187
3	83	64	152	79	66	142	13	94	75	211	88	72	171
4	83	65	158	81	67	147	14	92	70	201	58	72	171
5	88	69	180	84	68	167	15	87	68	167	81	69	154
6	77	67	147	74	66	131	16	83	68	162	79	68	149
7	78	69	159	73	66	131	17	87	66	173	84	69	160
8	84	68	159	75	67	134	18	87	68	177	84	70	160
9	89	71	195	84	68	161	19	88	70	169	84	70	168
10	91	76	206	86	72	169	20	83	66	170	77	67	147

Matlb中进行典型相关分析：

Matlab在其统计与机器学习工具箱中，提供了进行典型相关分析的函数：

[A,B,r] = canoncorr(X,Y)

其中，X, Y分别为x组和y组的数据观测矩阵；

A, B矩阵中每一列，分别为x组和y组的线性组合系数

r 向量给出，每对典型变量的典型相关系数

Matlb中进行典型相关分析：

```
data=xlsread("data_chapter5.xls")

X=data(:,1:3)

Y=data(:,4:6)

[A,B,r] = canoncorr(X,Y)
```

```
A = 3×3
     -0.1583    0.1059   -0.5136
     -0.1079   -0.3448   -0.3191
      0.0060    0.0442    0.1709

B = 3×3
     -0.0107   -0.0543   -0.1525
     -0.2354   -0.7619    0.0293
     -0.0231    0.1472    0.0227

r = 1×3
      0.9561    0.7495    0.2724
```