信陽師範學院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS
-1975-

# 第12章 MATLAB多元统计分析

**讲授人：** 牛言涛

**日期：** 2020年5月4日

# 目录

# CONTENTS



# 1. 典型相关分析简介

- 在一元统计分析中，用相关系数来衡量两个随机变量之间的线性相关关系；用复相关系数研究一个随机变量和多个随机变量的线性相关关系。然而，这些统计方法在研究两组变量之间的相关关系时却无能为力。
- 比如要研究生理指标与训练指标的关系，居民生活环境与健康状况的关系，人口统计变量（户主年龄、家庭年收入、户主受教育程度）与消费变量（每年去餐馆就餐的频率、每年出外看电影的频率）之间是否具有相关关系？阅读能力变量（阅读速度、阅读才能）与数学运算能力变量（数学运算速度、数学运算才能）是否相关？这些多变量间的相关性如何分析？
- 1936年霍特林（Hotelling）最早就“大学表现”和“入学前成绩”的关系、政府政策变量与经济目标变量的关系等问题进行了研究，提出了典型相关分析技术。之后，Cooley和Hohnes(1971)，Tatsuoka(1971)及Mardia，Kent和Bibby(1979)等人对典型相关分析的应用进行了讨论，Kshirsagar(1972)则从理论上给出了最好的分析。
- 典型相关分析的目的是识别并量化两组变量之间的联系，将两组变量相关关系的分析，转化为一组变量的线性组合与另一组变量线性组合之间的相关关系分析。

# 1. 典型相关分析简介

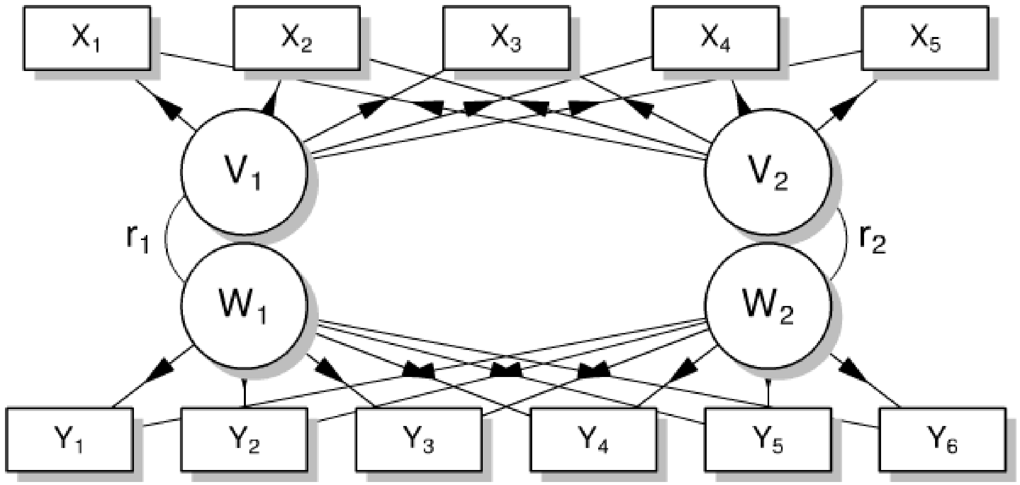


- 在对经济和管理问题的研究中，不仅经常需要考察两个变量之间的相关程度，而且还经常需要考察多个变量与多个变量之间即两组变量之间的相关性。
- 比如工厂管理人员需要了解原料的主要质量指标$X_1$，$X_2$,⋯，$X_p$与产品的主要质量指标$Y_1$，$Y_2$,⋯，$Y_q$之间的相关性，以便提高产品质量；医生要根据病人的一组体检化验指标与一些疾病之间的相关性，以便确定治疗方法等等。
- 典型相关分析CCA（canonical correlation analysis）就是测度两组变量之间相关程度的一种多元统计方法，利用综合变量对之间的相关关系来反映两组指标之间的整体相关性。它是两个随机变量之间的相关性在两组变量之下的推广。目前，典型相关分析已被应用于心理学、市场营销等领域。如用于研究个人性格与职业兴趣的关系，市场促销活动与消费者响应之间的关系等问题的分析研究。
- CCA的基本原理是：为了从总体上把握两组指标之间的相关关系，分别在两组变量中提取有代表性的综合变量$U_i$和$V_i$（分别为两个变量组中各变量的线性组合），利用综合变量之间的相关关系来反映两组指标之间的整体相关性。

# 1. 典型相关分析简介

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 按照典型相关分析，隐变量$V_i$和$W_i$分别是各自的可观察变量$X$和$Y$的线性组合，它们之间的直线箭头所代表的关系称为典型函数系数（canonical function coefficients），这些系数可以标准化或非标准化的形式出现；
- $X$的每个变量和每个$V$（或者$Y$的每个变量和每个$W$）之间的相关系数称为典型结构系数（canonical structure coefficient）或典型载荷（canonical loadinig）；
- 每个可观察变量与对立的典型变量之间，即$X$的每个变量和每个$W$（或者$Y$的每个变量和每个$V$）之间的相关系数称为典型交叉载荷（canonical cross-loading）。
- $V_i$和$W_i$的相关系数$r_i$为典型相关系数。称$(V_i, W_i)$为第$i$典型变量对（canonical variate pair），比如$(V_1, W_1)$为第一典型变量对，$(V_2, W_2)$为第二典型变量对等等。
- 典型变量对的个数和两组变量中少的变量数目一样。



# 1. 典型相关分析简介

为了更好地应用典型相关方法做实际分析，首先要弄清楚几个基础问题：

（1）严格说，一个典型相关系数描述的只是一对典型变量$(U_i, V_i)$之间关系，而不是两个变量组之间的相关，各对典型变量之间构成的多维典型相关才能共同揭示两个观测变量组之间的相关形式。

（2）典型相关模型的基本假设和数据要求：

- 首先，要求两组变量之间为线性关系，从而每对典型变量之间也为线性关系，这是很多统计分析方法都涉及的隐性假设，因为相关系数衡量的是线性相关。
- 其次，每个典型变量与本组所有观测变量的关系也是线性关系。如果不是，可先线性化。例如以先取对数（log），再做典型相关分析。

（3）所有观测变量均以数值形式进入模型。处理定性数据时，可按照一定形式设为虚拟变量后，再放入典型相关模型中进行分析。

# 2. 典型相关分析的基本原理

信阳师范学院 数学与统计学院 SCHOOL OF MATHEMATICS AND STATISTICS

- 对于两组随机变量$(X_1, X_2, \cdots, X_p)$和$(Y_1, Y_2, \cdots, Y_q)$，类似主成分分析，考虑$(X_1, X_2, \cdots, X_p)$一个线性组合$U$及$(Y_1, Y_2, \cdots, Y_q)$的一个线性组合$V$，

$$U_i = a_1^{(i)} X_1^{(1)} + a_2^{(i)} X_2^{(1)} + \cdots + a_P^{(i)} X_P^{(1)} \triangleq a^{(i)\prime} X^{(1)}$$
$$V_i = b_1^{(i)} Y_1^{(1)} + b_2^{(i)} Y_2^{(1)} + \cdots + b_q^{(i)} Y_q^{(1)} \triangleq b^{(i)\prime} Y^{(1)}$$

- 希望找到的***U*和*V*之间有最大可能的相关系数**，以充分反映两组变量间的关系。这样就把研究两组随机变量间相关关系的问题转化为研究两个随机变量间的相关关系。

- 如果一对变量$(U,V)$还不能完全刻划两组变量间的相关关系时，可以继续找第二对变量，希望这对变量在与第一对变量$(U,V)$不相关的情况下也具有尽可能大的相关系数。直到进行到找不到相关变量对时为止。这便引导出典型相关变量的概念。

# （1）总体典型相关变量

设有两组随机变量$X=(X_1,X_2,\cdots,X_p)^T,Y=(Y_1,Y_2,\cdots,Y_q)^T$，$(p\leq q)$，将$X,Y$合并

$$\left(X^T,Y^T\right)^T=\left(X_1,X_2,\cdots,X_p,Y_1,Y_2,\cdots,Y_q\right)^T$$

举个简单的例子。我们想考察一个人解题能力$X$（解题速度$x_1$，解题正确率$x_2$）与他的阅读能力$Y$（阅读速度$y_1$，理解程度$y_2$）之间的关系，那么形式化为：$U=a_1x_1+a_2x_2$，$V=b_1y_1+b_2y_2$

然后使用Pearson相关系数 $\rho_{X,Y}=\dfrac{cov(X,Y)}{\sigma_X\sigma_Y}$ 来度量$U$和$V$的关系，我们期望寻求一组最优的解$a$和$b$，使得$\rho_{X,Y}$最大，这样得到的$a$和$b$就是使得$U$和$V$就有最大关联的权重。

# （1）总体典型相关变量

设协方差矩阵为 $\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$，其中 $\Sigma_{11}=\mathrm{cov}(X)$，$\Sigma_{22}=\mathrm{cov}(Y)$，$\Sigma_{12}=\Sigma_{21}^T=\mathrm{cov}(X,Y)$

$$Cov(X,Y)=\left[\begin{array}{c:c}\underset{(p\times p)}{\Sigma_{11}} & \underset{(p\times q)}{\Sigma_{12}}\\ \hdashline \underset{(q\times p)}{\Sigma_{21}} & \underset{(q\times q)}{\Sigma_{22}}\end{array}\right]$$

根据典型相关思想是要寻找$X,Y$的线性组合：

$$U_1=a_1^TX=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p,\quad V_1=b_1^TY=b_{11}Y_1+b_{12}Y_2+\cdots+b_{1q}Y_q$$

使$U_1,V_1$的相关系数$\rho(U_1,V_1)$达到最大，这里 $a_1^T=(a_{11},a_{12},\cdots,a_{1p})$，$b_1^T=(b_{11},b_{12},\cdots,b_{1q})$

$$Var(U_1)=a_1^T\Sigma_{11}a_1,\quad Var(V_1)=b_1^T\Sigma_{22}b_1,\quad \mathrm{cov}(U_1,V_1)=a_1^T\Sigma_{12}b_1$$

所以$U_1,V_1$的相关系数为 $\rho_{U_1,V_1}=\dfrac{a_1^T\Sigma_{12}b_1}{\sqrt{a_1^T\Sigma_{11}a_1}\sqrt{b_1^T\Sigma_{22}b_1}}$

又由于相关系数与量纲无关，因此可设约束条件 $a_1^T\Sigma_{11}a_1=b_1^T\Sigma_{22}b_1=1$

满足约束条件的相关系数的最大值称为第一典型相关系数，$U_1,V_1$称为第一对典型相关变量。

# （1）总体典型相关变量

信阳师范学院 数学与统计学院 SCHOOL OF MATHEMATICS AND STATISTICS

典型相关分析在约束条件 $a_1^T\Sigma_{11}a_1=b_1^T\Sigma_{22}b_1=1$下，求$a_1,b_1$，使得 $\rho_{U_1,V_1}$ 取得最大值。

如果$(U_1,V_1)$还不足以反映$X,Y$之间的相关性，还可构造第二对线性组合：

$$U_2=a_2^TX=a_{21}X_1+a_{22}X_2+\cdots+a_{2p}X_p,\quad V_2=b_2^TY=b_{21}Y_1+b_{22}Y_2+\cdots+b_{2q}Y_q$$

使得$(U_1,V_1)$与$(U_2,V_2)$不相关，即$cov(U_1,U_2)=cov(U_1,V_2)=cov(U_2,V_1)=cov(V_1,V_2)=0$，在约束条件$Var(U_1)=Var(V_1)=Var(U_2)=Var(V_2)=1$下求$a_2,b_2$，使得$\rho_{U_2,V_2}$取得最大值。

一般地，若前$k-1$对典型变量还不足以反映$X,Y$之间的相关性，还可构造第$k$对线性组合：

$$U_k=a_k^TX=a_{k1}X_1+a_{k2}X_2+\cdots+a_{kp}X_p,\quad V_k=b_k^TY=b_{k1}Y_1+b_{k2}Y_2+\cdots+b_{kq}Y_q$$

在约束条件$cov(U_k,U_j)=cov(U_k,V_j)=cov(V_k,U_j)=cov(V_k,V_j)=0(1\le j\le k)$及$Var(U_k)=Var(V_k)=1$下求$a_k,b_k$，使得$\rho_{U_k,V_k}$ 取得最大值。

如此确定的$(U_k,V_k)$称为$X,Y$的第$k$对典型变量，相应的$\rho_{U_k,V_k}$称为第$k$个典型相关系数。

# （1）总体典型相关变量

根据条件极值的求法引入Lagrange乘数，将问题转化为求

$$\varphi(a,b)=a'\Sigma_{12}b-\frac{\lambda}{2}(a'\Sigma_{11}a-1)-\frac{\nu}{2}(b'\Sigma_{22}b-1)$$

的极大值，其中λ，ν是Lagrange乘数。根据求极值的必要条件得 $\begin{cases}\dfrac{\partial\varphi}{\partial a}=\Sigma_{12}b-\lambda\Sigma_{11}a=0\\ \dfrac{\partial\varphi}{\partial b}=\Sigma_{21}a-\nu\Sigma_{22}b=0\end{cases}$

将方程组的二式分别左乘$a'$与$b'$则得 $\begin{cases}a'\Sigma_{12}b-\lambda a'\Sigma_{11}a=0\\ b'\Sigma_{21}a-\nu b'\Sigma_{22}b=0\end{cases}$

即有 $\begin{cases}a'\Sigma_{12}b=\lambda a'\Sigma_{11}a=\lambda\\ b'\Sigma_{21}a=\nu b'\Sigma_{22}b=\nu\end{cases}$ 。因为$(b'\Sigma_{12}a)'=a'\Sigma_{12}b$，所以$\lambda=\nu=a'\Sigma_{12}b$ ，知$\lambda$为线性组合

$U,V$的相关系数。用$\lambda$代替方程组中的$\nu$，则方程组写为：$\begin{cases}\Sigma_{12}b-\lambda\Sigma_{11}a=0\\ \Sigma_{21}a-\lambda\Sigma_{22}b=0\end{cases}$

## (1) 总体典型相关变量

假定各随机变量协方差阵的逆矩阵存在，则由方程组式中的$\begin{cases}\Sigma_{12}b-\lambda\Sigma_{11}a=0\\ \Sigma_{21}a-\lambda\Sigma_{22}b=0\end{cases}$中得第二式，可得$b=\dfrac{1}{\lambda}\Sigma_{22}^{-1}\Sigma_{21}a$，代入第一式得$\dfrac{1}{\lambda}\Sigma_{21}\Sigma_{22}^{-1}\Sigma_{21}a-\lambda\Sigma_{11}a=0$，即有$\Sigma_{21}\Sigma_{22}^{-1}\Sigma_{21}a-\lambda^2\Sigma_{11}a=0$ (1)

同理，由方程组$\begin{cases}\dfrac{\partial\varphi}{\partial a}=\Sigma_{12}b-\lambda\Sigma_{11}a=0\\ \dfrac{\partial\varphi}{\partial b}=\Sigma_{21}a-\nu\Sigma_{22}b=0\end{cases}$，可得$\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}b-\lambda^2\Sigma_{22}b=0$ (2)

用$\Sigma_{11}^{-1}$和$\Sigma_{22}^{-1}$分别左乘(1)和(2)式，得$\begin{cases}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a-\lambda^2a=0\\ \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b-\lambda^2b=0\end{cases}\Rightarrow\begin{cases}\left(\overbrace{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}^{A}-\lambda^2I_p\right)a=0\\ \left(\underbrace{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}_{B}-\lambda^2I_q\right)b=0\end{cases}$

由此可见，具有相同的特征根$\lambda^2$，$a,b$则是其相应的特征向量。

# （1）总体典型相关变量



因为 $\lambda = a'\Sigma_{12}b = Corr(U,V)$ ，求$Corr(U,V)$最大值也就是求$\lambda$的最大值，而求$\lambda$的最大值又转化为求$A$和$B$的最大特征根。

可以证明，$A$和$B$的特征根和特征向量有如下性质：

1. $A$和$B$具有相同的非零特征根，且所有特征根非负。

2. $A$和$B$的特征根均在0 ~ 1之间。

3. 设$A$和$B$的非零特征根 $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2$ ，$r = rank(A) = rank(B)$， $a^{(1)}, a^{(2)}, \cdots, a^{(r)}$ 为$A$对应于$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2$ 的特征向量， $b^{(1)}, b^{(2)}, \cdots, b^{(r)}$ 为 $B$ 对应于$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2$的特征向量。

则最大特征值对应最大特征向量，第一典型相关变量可表示如下：

$$U_1 = a^{(1)\prime} X^{(1)} = a_1^{(1)} X_1^{(1)} + a_2^{(1)} X_2^{(1)} + \cdots + a_P^{(1)} X_P^{(1)}$$
$$V_1 = b^{(1)\prime} X^{(2)} = b_1^{(1)} X_1^{(2)} + b_2^{(1)} X_2^{(2)} + \cdots + b_q^{(1)} X_q^{(2)}$$

，此时注意约束条件 $a_1^T \Sigma_{11} a_1 = b_1^T \Sigma_{22} b_1 = 1$

$$a = a\left(a^T \Sigma_{11} a\right)^{-1/2}$$
$$b = b\left(b^T \Sigma_{22} b\right)^{-1/2}$$

典型相关系数为最大特征值平方根$\lambda_1$，称为第一典型相关系数。

# （1）总体典型相关变量

如果第一典型变量不足以代表两组原始变量的信息，则需要求得第二对典型变量，即

$$\begin{cases} U_2 = a^{(2)\prime} X^{(1)} = a_1^{(2)} X_1^{(1)} + a_2^{(2)} X_2^{(1)} + \cdots + a_P^{(2)} X_P^{(1)} \\ V_2 = b^{(2)\prime} X^{(2)} = b_1^{(2)} X_1^{(2)} + b_2^{(2)} X_2^{(2)} + \cdots + b_q^{(2)} X_q^{(2)} \end{cases}$$

且要求 $D(U_2)=1$，$D(V_2)=1$，$Cov(U_1,U_2)=0$，$Cov(V_1,V_2)=0$，第二对典型变量应不再包含第一对典型变量已包含的信息。典型相关系数为第二大特征值平方根$\lambda_2$，称为第二典型相关系数。

类似可求出第$r$对典型变量 $\begin{cases} U_r = a^{(r)\prime} X^{(1)} \\ V_r = b^{(r)\prime} X^{(2)} \end{cases}$，其系数向量$a^{(r)}, b^{(r)}$分别为矩阵$A$和$B$的第$r$特征根$\lambda_r^2$对应的特征向量，为第$r$典型相关系数。

# （1）总体典型相关变量

典型相关分析的步骤:

(1) 计算矩阵 $(X^T, Y^T)^T$ 的协方差矩阵或相关系数矩阵 $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$

(2) 令 $A = (\Sigma_{11})^{-1}\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}, \quad B = (\Sigma_{22})^{-1}\Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}$

或 $A = (R_{11})^{-1}R_{12}(R_{22})^{-1}R_{21}, \quad B = (R_{22})^{-1}R_{21}(R_{11})^{-1}R_{12}$

求$A, B$的特征值$\rho_1^2, \rho_2^2, \cdots, \rho_p^2$与对应的正交单位特征向量$e_k, f_k, \quad k = 1, 2, \cdots, p$

(3) $X, Y$的第$k$对典型相关变量为($k = 1, 2, \ldots, p$)

$$U_k = e_k^T\left(e_k^T \Sigma_{11} a_k\right)^{-1/2}, \quad V_k = f_k^T\left(f_k^T \Sigma_{22} f_k\right)^{-1/2}$$

(4) $X, Y$的第$k$个典型相关系数为：$\rho_k, \quad k = 1, 2, \cdots, p$

# （2）样本的典型变量与典型相关系数



在实际问题中$(X^T, Y^T)^T$的协方差矩阵Σ(或相关系数矩阵$R$)一般是未知的，我们所具有的资料通常是关于$X$和$Y$的$n$组观测数据:

$$X_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T, i = 1, 2, \cdots, p, \quad Y_j = (y_{j1}, y_{j2}, \cdots, y_{jn})^T, j = 1, 2, \cdots, q$$

同主成分分析一样，以$S$代替$\Sigma$或$R$代替$\rho$，作为Σ或ρ的估计，利用观测数据的样本协方差矩阵或者相关系数矩阵

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \quad 或 \quad R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

$$\hat{\Sigma} = \left[\begin{array}{c:c} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hdashline \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{array}\right] \xleftrightarrow{\text{标准化样本数据}} \hat{R} = \left[\begin{array}{c:c} \hat{R}_{11} & \hat{R}_{12} \\ \hdashline \hat{R}_{21} & \hat{R}_{22} \end{array}\right]$$

$$S_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T, \quad S_{22} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^T, \quad S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})^T$$

$$S_{21} = S_{12}^T, \quad \hat{A} = \hat{R}_{11}^{-1}\hat{R}_{12}\hat{R}_{22}^{-1}\hat{R}_{21}, \quad \hat{B} = \hat{R}_{22}^{-1}\hat{R}_{21}\hat{R}_{11}^{-1}\hat{R}_{12}$$

求解$\hat{A}$和$\hat{B}$的特征根和相应的特征向量，根据约束条件，即可得到样本典型变量及样本典型相关系数。

# 3. 典型相关系数的显著性检验

数学与统计学院

设总体$X,Y$的各对典型相关系数为$\rho_1 \ge \rho_2 \ge \cdots \ge \rho_p \ge 0$，首先提出检验原假设与备择假设

$$H_0^{(1)}: \rho_1 = 0 \leftrightarrow H_1^{(1)}: \rho_1 \ne 0$$

若不能拒绝原假设，则$\rho_1 = \rho_2 = \cdots = \rho_k = 0$，此时不能做典型相关分析；若拒绝$H_0^{(1)}$，继续检验

$$H_0^{(2)}: \rho_2 = 0 \leftrightarrow H_1^{(2)}: \rho_2 \ne 0$$

若不能拒绝$H_0^{(2)}$，表明只有第一对典型变量显著相关外，其余变量均不显著，实际应用只需考虑第一对典型变量；若拒绝$H_0^{(2)}$，则需检验$\rho_3$是否为零，以此类推，若假设$\rho_{k-1} = 0$被拒绝，则检验若不能拒绝$H_0^{(k)}$，则只需考虑前$k-1$对典型相关变量，否则继续检验，直至检验$\rho_p$是否为零。

$$H_0^{(k)}: \rho_k = 0 \leftrightarrow H_1^{(k)}: \rho_k \ne 0$$

# 3. 典型相关系数的显著性检验

在总体服从$p+q$维正态分布条件下，可用如下的似然比统计量进行检验

$$\Lambda_k=\prod_{j=k}^{p}(1-\hat{\rho}_j^2), T_k=-[n-(p+q+3)/2]\ln\Lambda_k \sim \chi^2(d_{1k})$$ ，其中 $d_{1k}=(p-\mathrm{k}+1)(q-\mathrm{k}+1)$

对于给定的$\alpha$，计算概率

$$p_k=P_{H_0}(T_k\ge t_k)=P(\chi^2(d_{1k})\ge t_k)$$

若$p_k<\alpha$，即认为第k对典型变量显著相关。上述检验依次对$k=1,2,\dots,p$进行，若对某个$k$检验概率首次大于$\alpha$，则检验停止，即认为只有前$k-1$对典型变量显著相关。

# 4. MATLAB程序实现

信阳师范学院 数学与统计学院

canonCorr_Demo.m

```
function cca = canonCorr_Demo(X,Y)
% ConoCorrA函数用于求解典型相关分析
% 输入参数X和Y为两组样本数据，具有相同的行数
% 输出参数cca为一结构体，包括典型相关分析求解过程中的值和结果

    %% 数据的预处理
    [n,p] = size(X); %X样本量和指标数
    q = size(Y,2); %Y样本的指标数
    data = [X,Y]; %合并样本
    dcov = corrcoef(data); %求协方差
    cov11 = dcov(1:p,1:p); %提取X协方差
    cov12 = dcov(1:p,p+1:end); %提取X和Y协方差,
    cov21 = cov12'; %对称
    cov22 = dcov(p+1:end,p+1:end); %提取Y协方差

    %% 求取A和B的特征值和特征向量，即典型相关系数和典型相关矩阵
    A = cov11\cov12*(cov22\cov21);%求A
    B = cov22\cov21*(cov11\cov12); %求B

    [VA,DA] = eig(A);%求A的特征值和特征向量
    [DA,ind] = sort(diag(DA),'descend');%从到到小排序
    VA = VA(:,ind);%对特征向量按照特征值的顺序排序
    A = VA*(VA'*cov11*VA)^(-1/2);

    [VB,DB] = eig(B);%求B的特征值和特征向量
    [DB,ind] = sort(diag(DB),'descend'); %从到到小排序
    VB = VB(:,ind);
    B = VB*(VB'*cov22*VB)^(-1/2);
    if p > q
        r = sqrt(DB)'; %典型相关系数
    else
        r = sqrt(DA)'; %典型相关系数
    end

    U = zscore(X*A); %求X的得分矩阵
    V = zscore(Y*B); %求Y的得分矩阵

    %% 典型相关分析的检验
    p = min([p,q]);
    D = r.^2;
    lambda = zeros(p,1);
    T = zeros(p,1);
    f = zeros(p,1);
    pChisq = zeros(1,p);
    for k = 1:p
        lambda(k) = prod(1-D(k:p));
        T(k) = -(n-(p+q+3)/2)*log(lambda(k));
        f(k) = (p-k+1)*(q-k+1);
        pChisq(k) = 1-chi2cdf(T(k),f(k)); %求卡方检验概率值
    end

    %% 输出变量的组合
    cca.A = A;
    cca.B = B;
    cca.r = r;
    cca.U = U;
    cca.V = V;
    cca.pChisq = pChisq;
end
```

# 4. MATLAB程序实现

```
>> ar = xlsread('air-soil.xlsx');
>> X = ar(:,1:3);
>> Y = ar(:,4:6);
>> cca = CCA_demo(X,Y)
cca =
  包含以下字段的 struct:
        A: [3×3 double]
        B: [3×3 double]
        r: [0.9279 0.5622 0.1660]
        U: [46×3 double]
        V: [46×3 double]
   pChisq: [0 0.0020 0.2816]
```

```
>> cca.A
ans =
   -0.6485    0.5550    2.0575
    0.1149    1.6993    0.2749
   -0.4600   -1.6963   -2.3422
>> cca.B
ans =
   -0.0863    0.2302   -1.6609
   -0.2016    2.8436   -0.3950
    1.2527   -2.7674    1.6293
```

<u>自编程序与样本数据标准化后的MATLAB自带函数运行结果一致。</u>

```
>> arc = zscore(ar);
>> X1 = arc(:,1:3);
>> Y1 = ar(:,4:6);
>> [A,B,r,U,V,stats] = canoncorr(X1,Y1);
```

```
>> A
A =
   -0.6485   -0.5550   -2.0575
    0.1149   -1.6993   -0.2749
   -0.4600    1.6963    2.3422
>> B
B =
    0.0116   -0.0309   -0.2226
    0.0612   -0.8627   -0.1198
   -0.0624    0.1379    0.0812
>> r
r =
    0.9279    0.5622    0.1660
```

# 4. MATLAB程序实现



ConoCorrA.m

```matlab
function cca = ConoCorrA(X,Y)
% ConoCorrA函数用于求解典型相关分析
% 输入参数X和Y为两组样本数据，具有相同的行数
% 输出参数cca为一结构体，包括典型相关分析求解过程中的值和结果

    %% 数据的预处理
    [n,p] = size(X); %X样本量和指标数
    q = size(Y,2); %Y样本的指标数
    data = [X,Y]; %合并样本
    dcov = cov(data); %求协方差
    cov11 = dcov(1:p,1:p); %提取X协方差
    cov12 = dcov(1:p,p+1:end); %提取X和Y协方差，cov21 = cov12'对称
    cov22 = dcov(p+1:end,p+1:end); %提取Y协方差

    %%求取特征值和特征向量，即典型相关系数和典型相关矩阵
    Hx = (cov11)^(-1/2);
    Hy = (cov22)^(-1/2);
    H = Hx * cov12 * Hy;
    [U,D,V] = svd(H,'econ'); %奇异值分解
    D1 = Hx * U;
    D2 = Hy * V;
    A = D1 * diag(diag((eye(size(D1, 2)) ./ sqrt(D1' * cov11 * D1))));
    B = D2 * diag(diag((eye(size(D2, 2)) ./ sqrt(D2' * cov22 * D2))));
    r = diag(D)'; %典型相关系数
    U = zscore(X*A); %求X的得分矩阵
    V = zscore(Y*B); %求Y的得分矩阵

    %% 典型相关分析的检验
    p = min([p,q]);
    D = r.^2;
    lambda = zeros(p,1);
    T = zeros(p,1);
    f = zeros(p,1);
    pChisq = zeros(1,p);
    for k = 1:p
        lambda(k) = prod(1-D(k:p));
        T(k) = -(n-(p+q+3)/2)*log(lambda(k));
        f(k) = (p-k+1)*(q-k+1);
        pChisq(k) = 1-chi2cdf(T(k),f(k)); %求卡方检验概率值
    end

    %% 输出变量的组合
    cca.A = A;
    cca.B = B;
    cca.r = r;
    cca.U = U;
    cca.V = V;
    cca.pChisq = pChisq;
end
```

根据奇异值分解求解，该函数与上一自编函数以及matlab自带函数canoncorr结果一致，自行验证。

# 5. MATLAB函数canoncorr

- 在MATLAB中，样本典型相关分析的命令为[A,B,r,U,V,stats] = canoncorr(X,Y)：
  - 其中输入X表示第一组向量的观测矩阵，Y表示第二组向量的观测矩阵；
  - 输出A,B是典型相关变量的系数矩阵；
  - r表示典型相关系数；U,V表示典型相关变量的得分；
  - 输出stats包括wilks 、 chisq及F统计量以及相应的概率。

| Field | Description |
|---|---|
| Wilks | Wilks' lambda (likelihood ratio) statistic |
| df1 | Degrees of freedom for the chi-squared statistic, and the numerator degrees of freedom for the $F$ statistic |
| df2 | Denominator degrees of freedom for the $F$ statistic |
| F | Rao's approximate $F$ statistic for |
| pF | Right-tail significance level for F |
| chisq | Bartlett's approximate chi-squared statistic for with Lawley's modification |
| pChisq | Right-tail significance level for chisq |

# 6. 案例分析——空气温度与土壤温度关系



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46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 |

# 案例求解与分析

```
>> as  = xlsread('air-soil.xlsx');
>> X = as(:,1:3);  %土壤温度数据
>> Y = as(:,4:6);  %空气温度数据
>> [A,B,r,U,V,stats] = canoncorr(X,Y);
```

```
A =
   -0.1280     0.1096     0.4061
    0.0313     0.4635     0.0750
   -0.0220    -0.0810    -0.1119
B =
    0.0116     0.0309     0.2226
    0.0612     0.8627     0.1198
   -0.0624    -0.1379    -0.0812
r =
    0.9279     0.5622     0.1660
```

第一典型相关对相关系数为0.9279，

第二典型相关对相关系数为0.5622

```
stats =
  包含以下字段的 struct:

     Wilks: [0.0925 0.6651 0.9725]
       df1: [9 4 1]
       df2: [97.5001 82 42]
         F: [17.9776 4.6366 1.1898]
        pF: [3.1880e-17 0.0020 0.2816]
     chisq: [98.7907 16.9896 1.2242]
    pChisq: [2.7628e-17 0.0019 0.2685]
       dfe: [9 4 1]
         p: [2.7628e-17 0.0019 0.2685]
```

$$\begin{cases} U_1 = -0.1280x_1 + 0.0313x_2 - 0.0220x_3 \\ V_1 = 0.0116y_1 + 0.0612y_2 - 0.0624y_3 \end{cases}$$

$$\begin{cases} U_2 = 0.1096x_1 + 0.4635x_2 - 0.0810x_3 \\ V_2 = 0.03096y_1 + 0.8627y_2 - 0.1379y_3 \end{cases}$$

Wilks：似然比统计量

df1、df2：自由度

F：统计量

pF：F统计量右边检验概率值

chisq：卡方统计量

pChisq：卡方统计量检验概率值

def：卡方检验自由度

通过前两个p值判别拒绝了原假设，即第一、第二典型相关对显著，而第三典型相关对不显著。

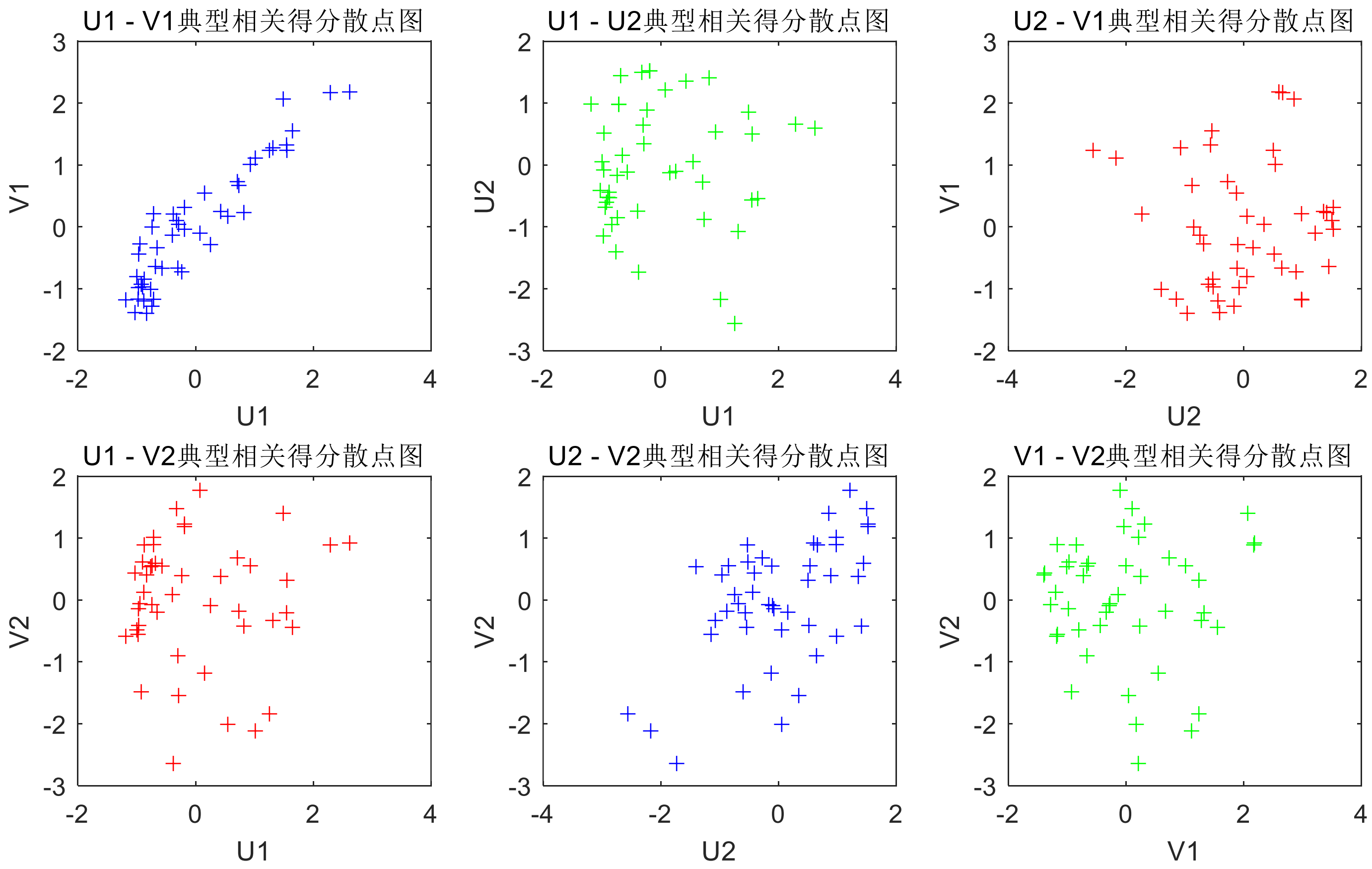
# 案例求解与分析

$$\begin{cases} U_1 = -0.1280x_1 + 0.0313x_2 - 0.0220x_3 \\ V_1 = 0.0116y_1 + 0.0612y_2 - 0.0624y_3 \end{cases}, \quad \begin{cases} U_2 = 0.1096x_1 + 0.4635x_2 - 0.0810x_3 \\ V_2 = 0.03096y_1 + 0.8627y_2 - 0.1379y_3 \end{cases}$$

- $U_1$是土壤温度各指标的线性组合，其中 X1日最高土壤温度比其他变量的载荷系数更大，说明日最高土壤温度是土壤温度的主要指标，它在土壤温度中占主导地位；而日最低土壤温度比日土壤温度曲线积分值的载荷要大，且正负相反，说明随着日最低温度的降低，日土壤温度曲线积分值在增加。
- $V_1$是空气温度各指标的线性组合，其中 Y2日最低气温和Y3日气温曲线积分值的载荷系数较大，且有相互抑制作用。日最低气温与气温曲线积分值是各空气温度中与土壤温度相关联的主要指标。
- $U_2$中 X2日最低土壤温度是土壤温度的主要指标，$V_1$中 Y2日最低气温是空气温度的主要指标。且X2与Y2符号相同，说明随着最低气温的降低，土壤最低温度也在降低。

# 典型变量之间的相关性

```
subplot(2,2,1)
plot(U(:,1),V(:,1),'b+')
xlabel('U1');ylabel('V1')
title('U1 - V1典型相关得分散点图')
subplot(2,2,2)
plot(U(:,2),V(:,1),'r+')
xlabel('U2');ylabel('V1')
title('U2 - V1典型相关得分散点图')
subplot(2,2,3)
plot(U(:,1),V(:,2),'r+')
xlabel('U1');ylabel('V2')
title('U1 - V2典型相关得分散点图')
subplot(2,2,4)
plot(U(:,2),V(:,2),'b+')
xlabel('U2');ylabel('V2')
title('U2 - V2典型相关得分散点图')
```




可见(U1,V1)和(U2,V2)之间是相关的，其余不相关。可通过cov(U(:,1),V(:,2)), cov(U(:,2),V(:,1))进行验证。

信阳师范学院 数学与统计学院

# 6. 案例分析——健康指标

例2：某康复俱乐部对20名中年人测量了三项生理指标：体重(weight)、腰围(waist)、脉搏(pulse )和三项训练指标：引体向上(chins )、起坐次数(situps )、跳跃次数(jumps )。其数据列于表。试分析这两组变量间的相关性。

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

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

# 案例求解与分析



```
club = xlsread('clubdata.xlsx');
%%三项生理指标作为第一组向量X
X = club(:,2:4);
%三项训练指标作为第二组向量Y
Y = club(:,5:7);
[A,B,r,U,V,stats] = canoncorr(X,Y)
```

```
A =
   -0.0314   -0.0763    0.0077
    0.4932    0.3687   -0.1580
   -0.0082   -0.0321   -0.1457
B =
   -0.0661   -0.0710    0.2453
   -0.0168    0.0020   -0.0198
    0.0140    0.0207    0.0082
r =
    0.7956    0.2006    0.0726
```

```
stats =
  包含以下字段的 struct:

     Wilks: [0.3504 0.9547 0.9947]
       df1: [9 4 1]
       df2: [34.2229 30 16]
         F: [2.0482 0.1758 0.0847]
        pF: [0.0635 0.9491 0.7748]
     chisq: [16.2550 0.7450 0.2109]
    pChisq: [0.0617 0.9457 0.6461]
       dfe: [9 4 1]
         p: [0.0617 0.9457 0.6461]
```

通过p值判别均拒绝了原假设。这里选取第一对典型相关变量分析。

$$\begin{cases} U_1 = -0.0314x_1 + 0.4932x_2 - 0.0082x_3 \\ V_1 = -0.0661y_1 - 0.0168y_2 + 0.0140y_3 \end{cases}$$

- $U_1$是生理指标的线性组合，其中X2腰围载荷系数最大，占主导地位，且与其他两个变量正负号相反，说明有一定的抑制作用；
- $V_1$是训练指标的线性组合，其中Y1引体向上载荷系数最大，占主导地位；且与X2正负号相反，说明腰围越大，引体向上次数就越少，起坐次数也相对较少，这也符合实际。

# 6. 案例分析——收入与支出

例3：我国31个省市自治区财政收入和支出的资料（中国统计年鉴，2008）。反映各地方财政收入的主要指标有(单位：万元)：X1:国内增值税，X2: 营业税，X3:企业所得税，X4:个人所得税，X5:专项收入，X6: 行政事业性收费收入；反映各地财政支出的主要指标有（单位：万元）：Y1: 一般公共服务，Y2: 国防，Y3: 公共安全，Y4: 教育，Y5: 科学技术，Y6: 社会保障和就业，Y7: 医疗卫生，Y8: 环境保护。

```
>> [re,text] = xlsread('Revenue_expenditure.xlsx');
>> rezs = zscore(re);
>> X = rezs(:,1:6);
>> Y = rezs(:,7:end);
>> [A,B,r,U,V,stats] = canoncorr(X,Y)
r =
    0.9953    0.9718    0.8850    0.8087    0.6320
0.3747
```

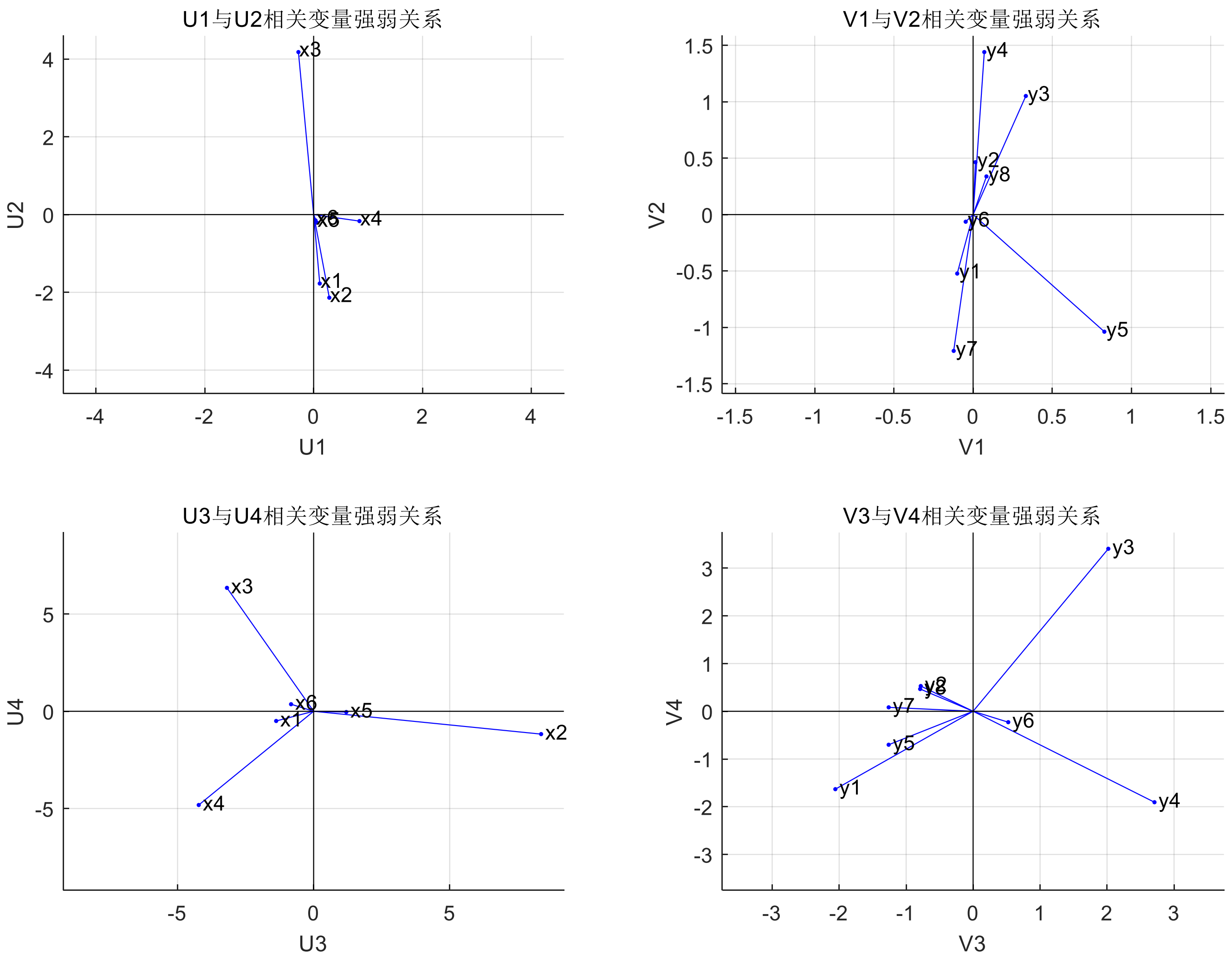
```
stats =
  包含以下字段的 struct:

    Wilks: [2.0150e-05 0.0022 0.0387 0.1787 0.5163 0.8596]
      df1: [48 35 24 15 8 3]
      df2: [87.7094 78.1490 67.4931 55.6126 42 22]
        F: [14.6217 7.3748 4.3286 3.2114 2.0567 1.1981]
       pF: [4.9986e-26 1.3925e-13 1.0573e-06 7.7408e-04 0.0625 0.3336]
    chisq: [243.2770 138.1780 73.3667 39.3467 15.4536 3.7647]
   pChisq: [6.4156e-28 3.3781e-14 6.6824e-07 5.6867e-04 0.0509 0.2880]
      dfe: [48 35 24 15 8 3]
        p: [6.4156e-28 3.3781e-14 6.6824e-07 5.6867e-04 0.0509 0.2880]
```

前四个典型相关对比较显著，故选取前四对典型相关变量。

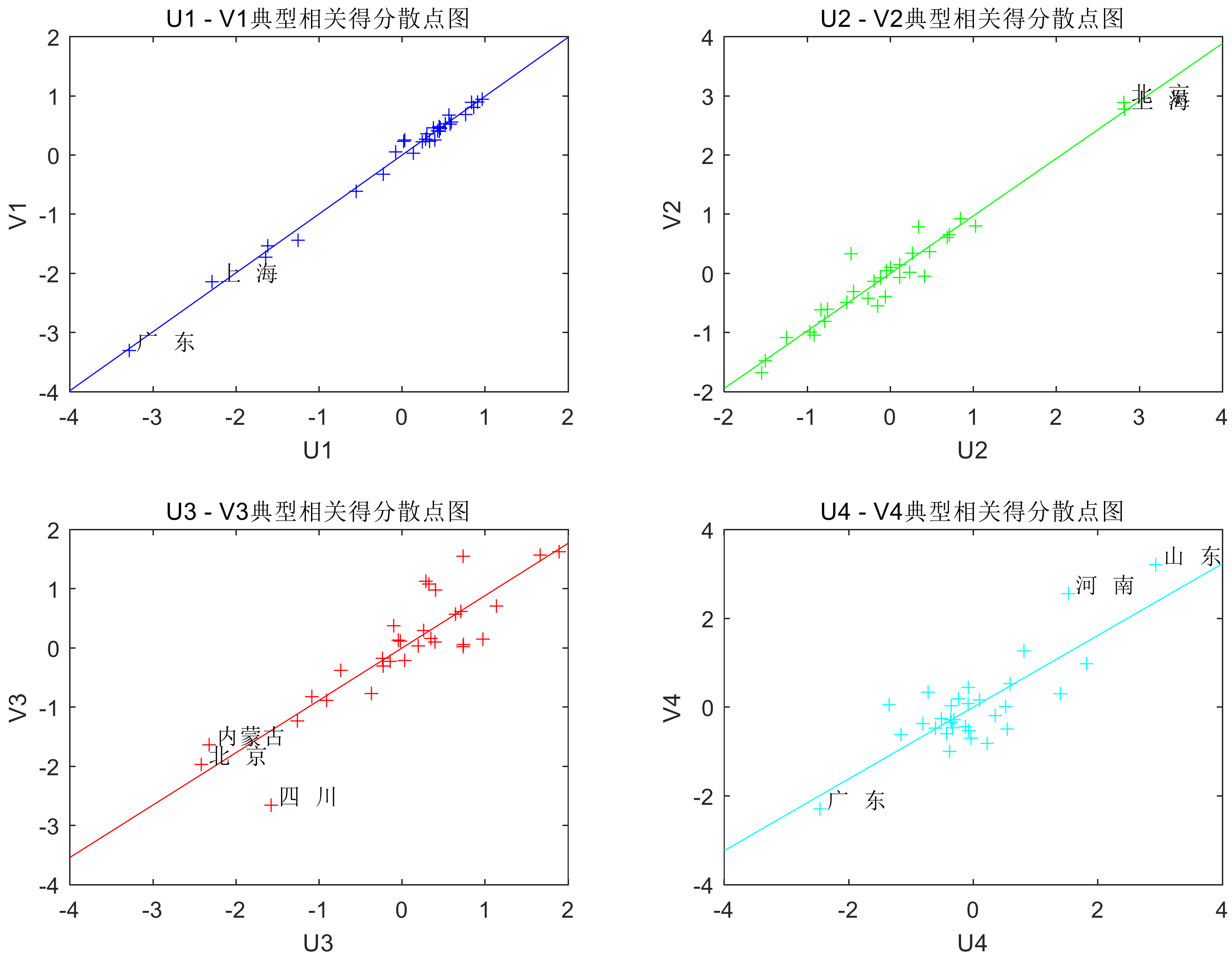
# 典型相关变量之间的关系

```
varnames = text(1,2:end);
subplot(2,2,1)
biplot(A(:,1:2),'VarLabels',varnames(1:6))
title('U1与U2相关变量强弱关系')
subplot(2,2,2)
biplot(B(:,1:2),'VarLabels',varnames(7:end))
title('V1与V2相关变量强弱关系')
subplot(2,2,3)
biplot(A(:,3:4),'VarLabels',varnames(1:6))
title('U3与U4相关变量强弱关系')
subplot(2,2,4)
biplot(B(:,3:4),'VarLabels',varnames(7:end))
title('V3与V4相关变量强弱关系')
xlabel('V3'); ylabel('V4')
```



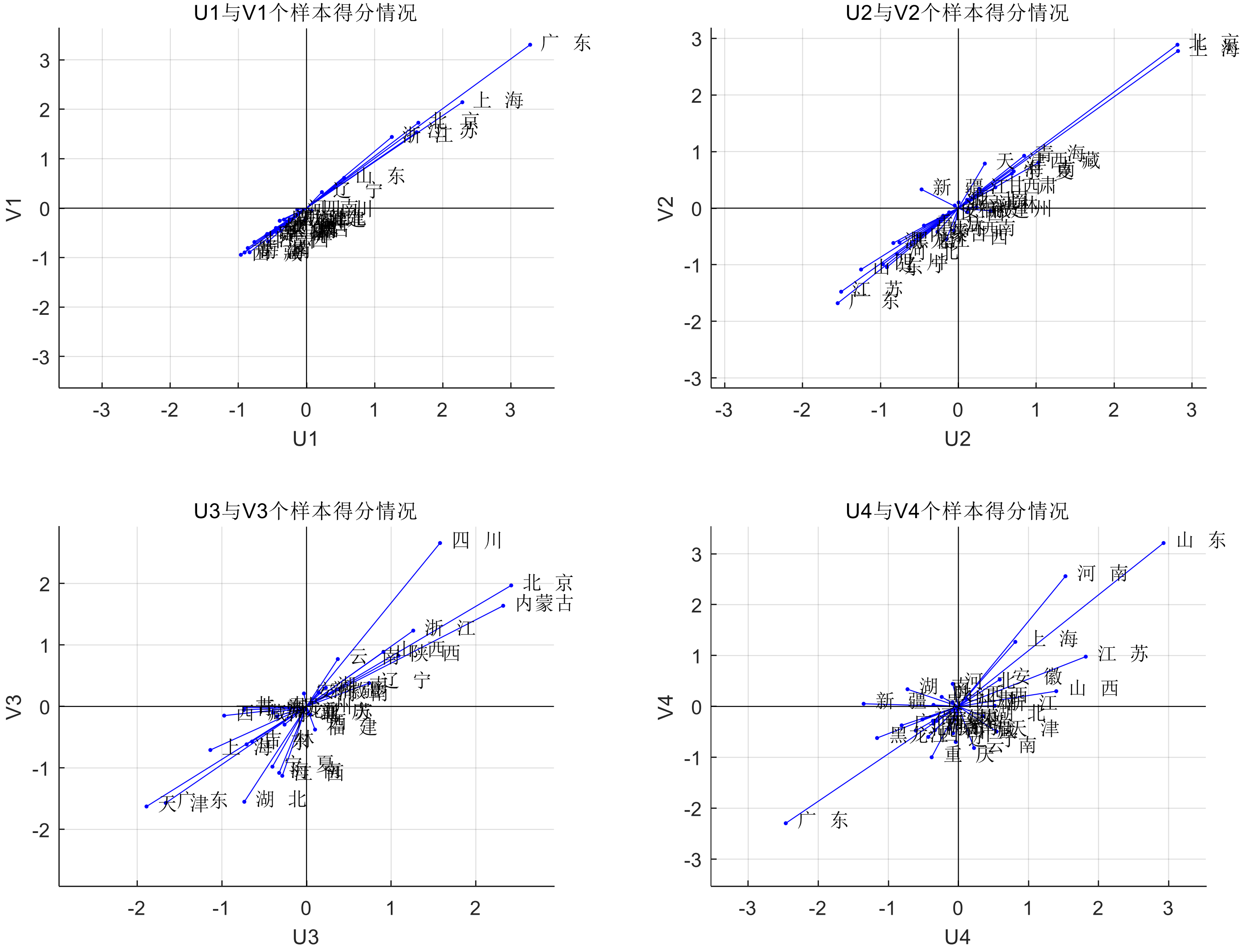
# 个体得分与典型相关变量的关系

```
rownames = text(2:end,1);
subplot(2,2,1)
plot(U(:,1),V(:,1),'b+');  lsline
title('U1 - V1典型相关得分散点图')
subplot(2,2,2)
plot(U(:,2),V(:,2),'g+');  lsline
title('U2 - V2典型相关得分散点图')
subplot(2,2,3)
plot(U(:,3),V(:,3),'r+');  lsline
title('U3 - V3典型相关得分散点图')
subplot(2,2,4)
plot(U(:,4),V(:,4),'c+');  lsline
xlabel('U4');  ylabel('V4')
title('U4 - V4典型相关得分散点图')
gname(rownames)  %标记异常值
```



# 个体得分与典型相关变量的关系

```
subplot(2,2,1)
biplot([U(:,1),V(:,1)],'VarLabels',rownames)
title('U1与V1个样本得分情况')
xlabel('U1'); ylabel('V1')
subplot(2,2,2)
biplot([U(:,2),V(:,2)],'VarLabels',rownames)
title('U2与V2个样本得分情况')
xlabel('U2'); ylabel('V2')
subplot(2,2,3)
biplot([U(:,3),V(:,3)],'VarLabels',rownames)
title('U3与V3个样本得分情况')
xlabel('U3'); ylabel('V3')
subplot(2,2,4)
biplot([U(:,4),V(:,4)],'VarLabels',rownames)
title('U4与V4个样本得分情况')
xlabel('U4'); ylabel('V4')
```



# 典型相关变量分析

-1975-
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

A =

| | | | | | |
|---|---|---|---|---|---|
| -0.1121 | -1.7732 | 1.3771 | -0.4985 | 2.6856 | 0.4096 |
| -0.2874 | -2.1374 | -8.3570 | -1.1742 | 6.4614 | -5.8815 |
| 0.2799 | 4.1789 | 3.1854 | 6.3480 | -2.2595 | 0.8466 |
| -0.8415 | -0.1674 | 4.2224 | -4.8204 | -5.3968 | 4.3433 |
| -0.0575 | -0.2042 | -1.1999 | -0.0396 | -0.7918 | 0.8756 |
| -0.0332 | -0.1481 | 0.8302 | 0.3630 | -1.5703 | -0.3630 |

B =

| | | | | | |
|---|---|---|---|---|---|
| 0.1004 | 0.5226 | -2.0569 | 1.6322 | -2.3025 | 3.2056 |
| -0.0144 | -0.4640 | -0.7817 | -0.5290 | 0.9095 | -0.8289 |
| -0.3329 | -1.0507 | 2.0187 | -3.4046 | -1.1009 | 0.5337 |
| -0.0706 | -1.4398 | 2.7079 | 1.9057 | 3.8549 | -2.6611 |
| -0.8287 | 1.0378 | -1.2608 | 0.7011 | 0.2081 | 1.5083 |
| 0.0469 | 0.0633 | 0.5247 | 0.2287 | -1.3257 | -0.5800 |
| 0.1226 | 1.2080 | -1.2606 | -0.0825 | -0.7833 | -1.5681 |
| -0.0842 | -0.3379 | -0.7907 | -0.4661 | 0.5016 | 0.6330 |

r =

| | | | | | |
|---|---|---|---|---|---|
| 0.9953 | 0.9718 | 0.8850 | 0.8087 | 0.6320 | 0.3747 |

- 第一对典型相关对中， $U_1$各地方财政收入中X4个人所得税载荷系数最大，而$V_1$各地财政支出中Y5科学技术载荷系数最大，说明个人所得税收入主要用于科学技术的支出，而且同符号发展。
- 第二对典型相关对中，$U_2$各地方财政收入中主要有X1国内增值税、X2营业税、X3企业所得税构成，$V_2$各地财政支出主要有Y3公共卫生、Y4教育、Y5科学技术、Y7医疗卫生构成。其中X1、X2与Y3、Y4同符号发展，X3与Y5、Y7同符号发展，说明国内增值税和营业税收入越高，则公共卫生和教育支出越多，企业所得税收入越高，科学技术和医疗卫生的支出也越多。

$$\begin{cases} U_1 = -0.112X_1 - 0.287X_2 + 0.280X_3 - 0.841X_4 - 0.058X_5 - 0.033X_6 \\ V_1 = 0.100Y_1 - 0.014Y_2 - 0.333Y_3 - 0.071Y_4 - 0.829Y_5 + 0.047Y_6 + 0.123Y_7 - 0.084Y_8 \end{cases}$$

$$\begin{cases} U_2 = -1.773X_1 - 2.137X_2 + 4.179X_3 - 0.167X_4 - 0.204X_5 - 0.148X_6 \\ V_2 = 0.523Y_1 - 0.464Y_2 - 1.051Y_3 - 1.440Y_4 + 1.038Y_5 + 0.063Y_6 + 1.208Y_7 - 0.338Y_8 \end{cases}$$

……

# 6. 案例分析——农业生成分析



例4：选取1980-2008年安徽省人均粮食总产量(吨/人)、人均农业总产值(亿元/万人)、人均粮食播种面积(千公顷/万人)、人均农业机械总动力(千瓦/人)、单位面积化肥施用(万吨/千公顷)、人均受灾面积(千公顷/万人)以及农业生产资料价格指数指标，分别记为：$x_1, x_2, x_3, y_1, y_2, y_3, y_4$，部分数据如下表。解决以下问题：

(1) 对安徽省粮食生产进行主成分分析，在此基础上给出适当的分类；

(2) 对安徽省粮食生产影响因素进行典型相关分析。

| 年份 | x1 | x2 | x3 | y1 | y2 | y3 | y4 |
|---|---|---|---|---|---|---|---|
| 1980 | 0.8704 | 0.0411 | 4.6332 | 0.3979 | 0.0071 | 0.2628 | 102.1000 |
| 1981 | 1.0538 | 0.0568 | 4.5664 | 0.3929 | 0.0091 | 0.6130 | 101.7000 |
| 1982 | 1.0818 | 0.0586 | 4.4808 | 0.4047 | 0.0114 | 0.2349 | 101.3000 |
| 1983 | 1.0898 | 0.0604 | 4.2600 | 0.4147 | 0.0115 | 0.1339 | 102.8000 |
| 1984 | 1.1576 | 0.0664 | 4.1872 | 0.4191 | 0.0127 | 0.4036 | 107.0000 |
| 1985 | 1.0983 | 0.0736 | 4.1470 | 0.4223 | 0.0139 | 0.2131 | 101.7000 |
| 1986 | 1.1649 | 0.0817 | 4.0089 | 0.4503 | 0.0141 | 0.3617 | 102.1000 |
| … | … | … | … | … | … | … | … |

# 主成分求解与分析



```
>> ahdata = xlsread('anhuiny.xlsx');
>> ah = ahdata(:,2:end);
>> ahzs = zscore(ah);
>> [coef,score,latent,~,explained] = pca(ahzs);
>> cse = cumsum(explained) %累积贡献率
```

```
>> d = pdist(score(:,1:4)); %按照主成分得分聚类
>> z = linkage(d,'average');
>> dendrogram(z,'ColorThreshold',0.5*max(z(:,3)),'Labels',
num2str(year),'Orientation','left'); %谱系聚类图
>> title('安徽省粮食生产主成分后的聚类图')
```

```
cse =
    57.9799
    74.7204
    88.0378
    96.8787
    99.2633
    99.8708
   100.0000
>> coef
coef =
```

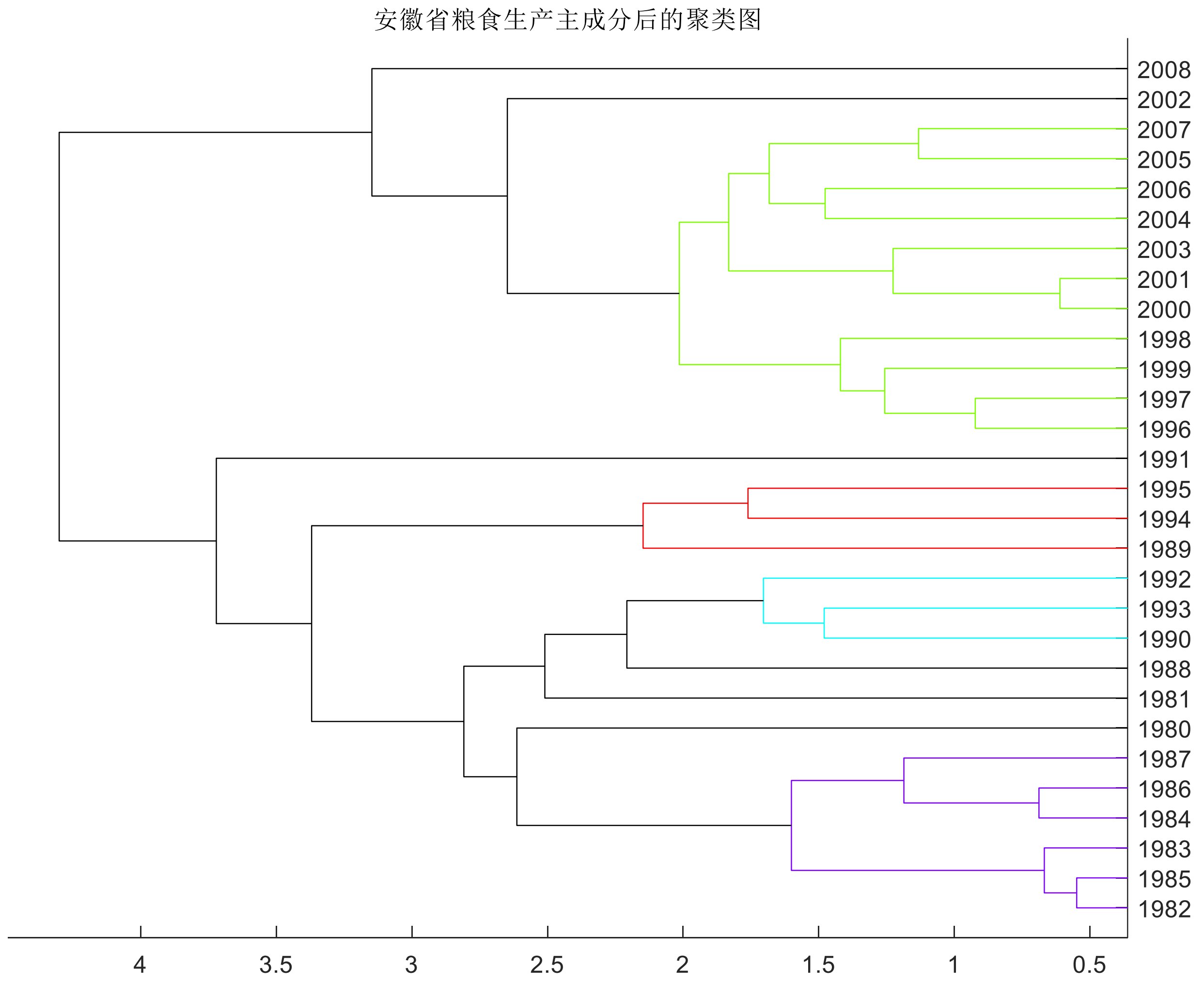
| | | | | | | |
|---|---|---|---|---|---|---|
| -0.2873 | 0.3544 | -0.3638 | 0.7966 | -0.0786 | -0.1355 | -0.0208 |
| 0.4717 | 0.1536 | -0.1271 | 0.1866 | 0.2091 | 0.7621 | -0.2801 |
| -0.4675 | -0.1189 | -0.0881 | -0.0291 | 0.7095 | 0.2684 | 0.4282 |
| 0.4721 | 0.0385 | 0.0313 | 0.1303 | 0.6413 | -0.5642 | -0.1676 |
| 0.4888 | 0.0319 | 0.0032 | 0.1693 | -0.1727 | 0.0053 | 0.8376 |
| -0.1207 | 0.4221 | 0.8724 | 0.1877 | 0.0612 | 0.0816 | 0.0251 |
| -0.0112 | 0.8100 | -0.2858 | -0.4989 | 0.0433 | -0.0618 | 0.0869 |

从累积贡献率中选取前四个主成分，信息解释度为96.8787%。

- 第一主成分主要反映：人均农业总产值$x_2$、人均粮食播种面积$x_3$、人均农业机械总动力$y_1$、单位面积化肥施用$y_2$，即农业基础设施方面。
- 第二主成分主要反映 $y_4$ 农业生产资料价格指数；
- 第三主成分主要反映人均受灾面积$y_3$；
- 第四主成分呢主要反映人均粮食总产量$x_1$。

# 层次聚类分析

考虑自然和社会因素以及年份的连续型，根据主成分得分安徽省农业生产分为三阶段：1980-1987、1988-1995、1996-2008。其中1991年和2008年最为特殊，可能存在异常值，需具体分析这两个年份情况。



# 典型相关求解与分析

```
>> X = ahzs(:,1:3);
>> Y = ahzs(:,4:7);
>> [A,B,r,U,V,stats] = canoncorr(X,Y)
```

```
A =
   -0.0038    0.6432    0.9858
    0.5102   -1.9126    0.7600
   -0.5203   -2.2055    0.3043
B =
   -0.0646   -2.4950   -0.0807
    1.0609    2.3885   -0.0516
    0.0160    0.4842   -0.2698
    0.1240    0.0694    0.9879
r =
    0.9915    0.6771    0.2671
```

```
stats =
  包含以下字段的 struct:

     Wilks: [0.0085 0.5029 0.9287]
       df1: [12 6 2]
       df2: [58.4980 46 24]
         F: [24.6485 3.1444 0.9218]
        pF: [1.6034e-18 0.0115 0.4114]
     chisq: [114.3659 16.5091 1.8649]
    pChisq: [8.1637e-19 0.0113 0.3936]
       dfe: [12 6 2]
         p: [8.1637e-19 0.0113 0.3936]
```

从相关系数及其典型相关检验，得出选取前两对典型相关组

- $U_1$主要反映了人均农业产值与人均耕地面积信息，$V_1$主要反映了单位面积化肥施用量的信息，因此第一对典型变量主要反映了人均农业产值与人均耕地面与单位面积化肥施用量的相关性；
- 第二对典型变量主要反映了人均农业产值和人均粮食播种面积与人均农业机械总动力、单位面积化肥施用量的相关性。

$$\begin{cases} U_1 = -0.0038x_1 + 0.5102x_2 - 0.5203x_3 \\ V_1 = -0.0646y_1 + 1.0609y_2 + 0.0160y_3 + 0.1240y_4 \end{cases}, \begin{cases} U_2 = 0.6432x_1 - 1.9126x_2 - 2.2055x_3 \\ V_2 = -2.4950y_1 + 2.3885y_2 + 0.4842y_3 + 0.0694y_4 \end{cases}$$


EDUCATION
BUSINESS
01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
PHOTO BOOK

# 感谢聆听