典型相关分析

（Canonical correlation analysis）

引例

对某高中一年级男生38人进行体力测试（共有七项指标）及运动能力测试（共有五项指标），如何讨论这两组指标之间的相关关系？

体力测试指标：

X_1 －反复横向跳（次）

X_2 －纵跳（cm）

X_3 －背力（kg）

X_4 －握力（kg）

X_5 －台阶试验（指数）

X_6 －立定体前曲（cm）

X_7 －俯卧上体后仰（cm）

运动能力测试指标：

X_8 －50米跑（秒）

X_9 －跳远（cm）

X_{10} －投球（m）

X_{11} －引体向上（次）

X_{12} －耐力跑（秒）

- **方法一：**

  分别研究X_i与X_j（i=1,…,7；j=8,…,12）之间的相关关系。但当两组变量较多时，这种方法不仅繁琐，也不易抓住问题的本质。

- **方法二：**

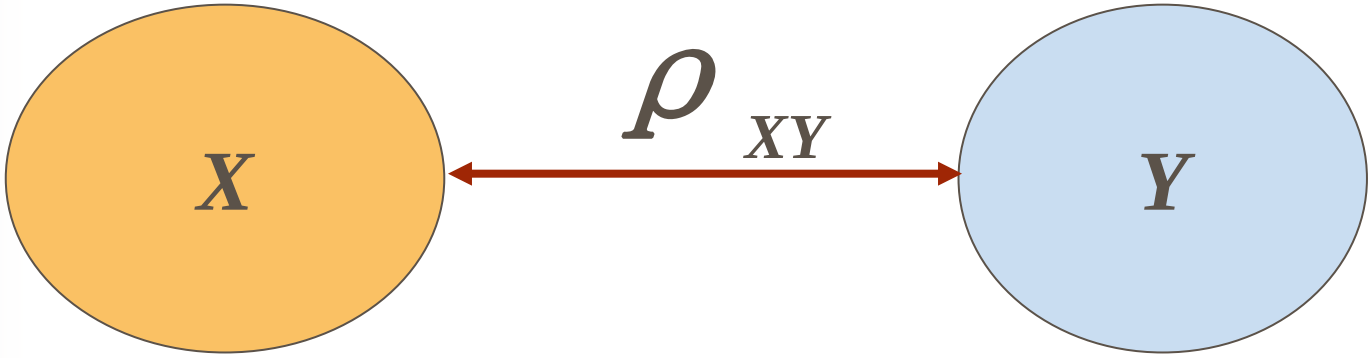
  采用类似主成分分析的方法，在每一组变量中都选择若干个有代表性的综合指标（变量的线性组合），通过研究两组的综合指标之间的关系来反映两组变量之间的相关关系，要求它们之间有最大的相关性。

一、何谓典型相关分析

- 相关系数（简单相关系数）——
  两个随机变量之间的线性相关关系：

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{VarX}\sqrt{VarY}}$$

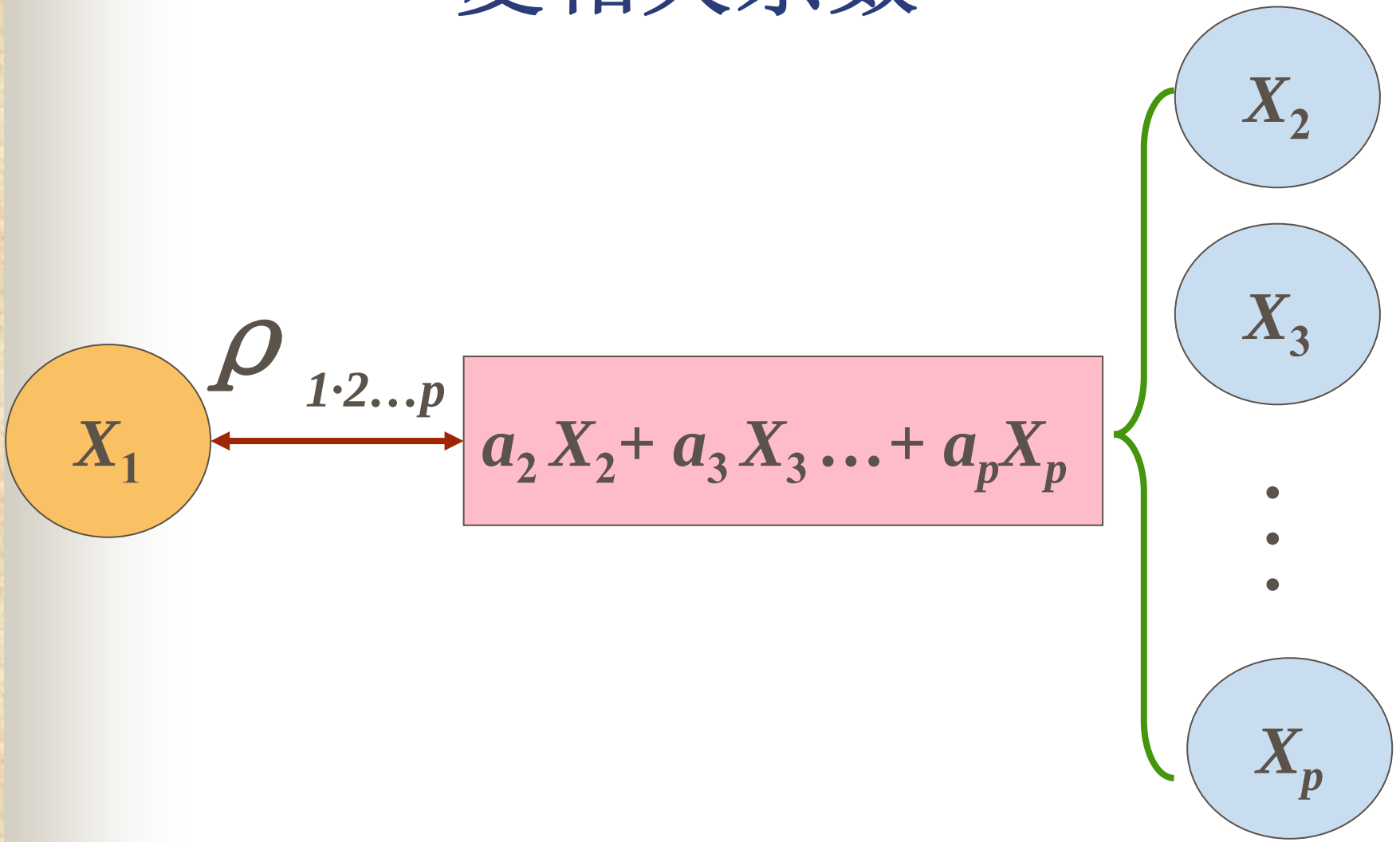
简单相关系数



- 复相关系数—— 一个随机变量与多个随机变量之间的线性相关关系:

若希望研究X_1与$X_2,\cdots,X_p$之间的相关关系，一个很自然的想法是考虑X_1与$X_2,\cdots,X_p$的线性组合之间的相关系数，当然我们应该选择X_1与$X_2,\cdots,X_p$的所有线性组合中最大的相关系数作为刻画X_1与$X_2,\cdots,X_p$的相关关系的度量，称之为复相关系数（或全相关系数）。

复相关系数



对p维随机向量$\mathbf{X}=\left(X_1,X_2,\cdots,X_p\right)'$及其协方差阵$\mathbf{\Sigma}$作剖分：

$$\mathbf{X}=\begin{pmatrix}X_1\\ \mathbf{X}_2\end{pmatrix},\quad \mathbf{\Sigma}=\begin{pmatrix}\sigma_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{12}' & \mathbf{\Sigma}_{22}\end{pmatrix}$$

则 $$\rho_{1\bullet 23\cdots p}=\left(\frac{\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}'}{\sigma_{11}}\right)^{1/2}$$

即为X_1与$X_2,\cdots,X_p$之间的复相关系数。

Hotelling(1936)首先将复相关系数推广到研究多个随机变量与多个随机变量的相关关系的讨论中，提出了典型相关分析。

- 实际背景：

投资性变量（如劳动者人数、货物周转量、生产建设投资）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）具有相关关系。

运动员的体力测试指标（如反复横向跳、纵跳、背力、握力等）与运动体能测试指标（如耐力跑、跳远、投球等）之间具有相关关系。

……

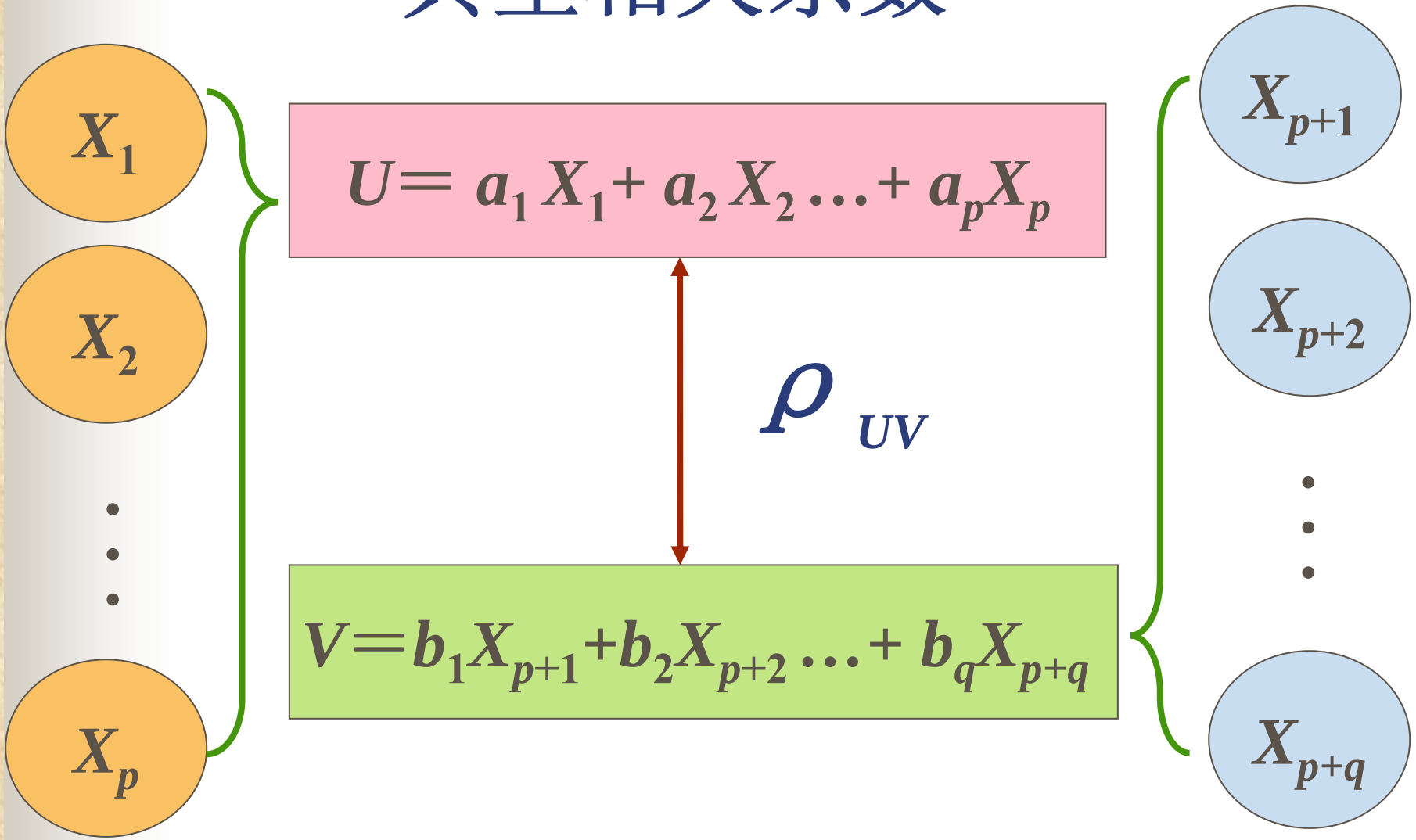
- 基本思想：

为研究两组随机变量$X_1, X_2, \cdots, X_p$及$X_{p+1}, X_{p+2}, \cdots, X_{p+q}$之间的相关关系，可采用类似主成分分析的思想，在每一组变量中都选择若干个有代表性的综合指标(变量的线性组合)，通过研究两组变量的综合指标之间的关系来反映两组变量之间的相关关系。

即构造$U = a_1X_1 + a_2X_2 + \cdots + a_pX_p$及

$V = b_1X_{p+1} + b_2X_{p+2} + \cdots + b_qX_{p+q}$，则两组变量之间的相关问题转化为两个变量$U$与$V$之间的相关问题，希望使$U$与$V$的相关达到最大。 称这种相关为典型相关。

典型相关系数


X_1
X_2
X_p
$U = a_1X_1 + a_2X_2 \ldots + a_pX_p$
ρ_{UV}
$V = b_1X_{p+1} + b_2X_{p+2} \ldots + b_qX_{p+q}$
X_{p+1}
X_{p+2}
X_{p+q}

例：F. V. Waugh(1942) 研究了美国1921年至1940年每年牛肉、猪肉的价格与按人口平均的牛肉和猪肉的消费量之间的相关关系。猪肉价格和牛肉价格用X_1，X_2表示，它们的消费量用X_3，X_4表示，研究这两组变量之间的相关关系。

构造 $U_1 = a_1 X_1 + a_2 X_2$ ——价格指标，

构造 $V_1 = b_1 X_3 + b_2 X_4$ ——消费量指标，

要求它们之间具有最大相关性，这就是一个典型相关分析问题。

典型相关分析的数学描述

设有两组变量$\mathbf{X}^{(1)}=(X_1,\cdots,X_p)',\mathbf{X}^{(2)}=(X_{p+1},\cdots,X_{p+q})'$,

不妨设$p\le q$,假定$\mathbf{X}=\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}$的协方差阵$\mathbf{\Sigma}>0$,均值向量

$\boldsymbol{\mu}=\mathbf{0}$,相应地将$\mathbf{\Sigma}$剖分为

$$\mathbf{\Sigma}=\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix}$$

其中,$\mathbf{\Sigma}_{11}$是$\mathbf{X}^{(1)}$的协方差阵,$\mathbf{\Sigma}_{22}$是$\mathbf{X}^{(2)}$的协方差阵,$\mathbf{\Sigma}_{12}$是$\mathbf{X}^{(1)}$与$\mathbf{X}^{(2)}$的协方差阵。

要研究$\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$两组变量之间的相关关系，作两组变量的线性组合：

$$U = a_1X_1 + a_2X_2 + \cdots + a_pX_p \triangleq \boldsymbol{\alpha}'\mathbf{X}^{(1)}$$

$$V = b_1X_{p+1} + b_2X_{p+2} + \cdots + b_qX_{p+q} \triangleq \boldsymbol{\beta}'\mathbf{X}^{(2)}$$

其中$\boldsymbol{\alpha}=(a_1, a_2, \cdots, a_p)'$，$\boldsymbol{\beta}=(b_1, b_2, \cdots, b_q)'$为任意非零常数向量，易见：

$$Var(U) = Var(\boldsymbol{\alpha}'\mathbf{X}^{(1)}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}$$

$$Var(V) = Var(\boldsymbol{\beta}'\mathbf{X}^{(2)}) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}$$

$$Cov(U, V) = \boldsymbol{\alpha}'Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})\boldsymbol{\beta} = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}$$

$$\rho_{UV} = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}}}$$

由于对任意非零同号常数k_1和k_2，有

$$\rho_{k_1U,\ k_2V} = \rho_{UV}$$

因此，为避免不必要的结果重复出现，通常限制

$Var(U) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = 1,\ \ Var(V) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = 1$。

于是，问题归结为在约束条件：

$$Var(U) = 1,\ \ Var(V) = 1$$

之下，寻求$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$使 $\rho_{UV} = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}$ 达到最大。

二、总体的典型相关系数和典型变量的导出

在约束条件：

$$Var(U)=1,\quad Var(V)=1$$

之下，如何寻求$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$使 $\rho_{UV}=\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}$ 达到最大？

由求条件极值的方法，引入*Lagrange*乘数，可将问题化为求

$$f(\boldsymbol{\alpha},\ \boldsymbol{\beta})=\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}-\frac{\lambda}{2}(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}-1)-\frac{\nu}{2}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}-1)$$

的极大值，其中λ，ν 是*Lagrange*乘数。

由极值的必要条件

$$\begin{cases} \dfrac{\partial f}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}_{12}\boldsymbol{\beta} - \lambda \boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \mathbf{0} \\ \dfrac{\partial f}{\partial \boldsymbol{\beta}} = \boldsymbol{\Sigma}_{21}\boldsymbol{\alpha} - \nu \boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = \mathbf{0} \end{cases} \tag{1}$$

将上式分别左乘$\boldsymbol{\alpha}'$与$\boldsymbol{\beta}'$，则得

$$\begin{cases} \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} = \lambda \;\; \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = \lambda \\ \boldsymbol{\beta}'\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha} = \nu \;\; \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = \nu \end{cases} \Rightarrow \lambda = \nu$$

因此可得$\rho_{UV} = \lambda$。

于是，方程组(1)化为

$$\begin{cases} \boldsymbol{\Sigma}_{12}\boldsymbol{\beta}-\lambda\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}=\mathbf{0} \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}-\lambda\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}=\mathbf{0} \end{cases} \qquad (2)$$

解方程组(2)可得

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}-\lambda^2\boldsymbol{\alpha}=\mathbf{0}$$

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}-\lambda^2\boldsymbol{\beta}=\mathbf{0}$$

记 $\mathbf{A}=\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

$\mathbf{B}=\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$

则得 $\mathbf{A}\boldsymbol{\alpha}=\lambda^2\boldsymbol{\alpha}$ $\mathbf{B}\boldsymbol{\beta}=\lambda^2\boldsymbol{\beta}$

即λ^2既是**A**又是**B**的特征根，$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$分别是其相应于**A**和**B**的特征向量。

A和**B**的特征根有如下性质：

(1) **A**和**B**有相同的非零特征根，且相等的非零特征根数目为p。

(2) **A**和**B**的特征根非负。

(3) **A**和**B**的特征根在0和1之间。

记**A**和**B** 的p 个非零特征值为$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_p^2 > 0$，称$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$为典型相关系数，相应的单位特征向量为$\boldsymbol{\alpha}^{(1)}$，$\boldsymbol{\alpha}^{(2)}$,…，$\boldsymbol{\alpha}^{(p)}$和$\boldsymbol{\beta}^{(1)}$，$\boldsymbol{\beta}^{(2)}$,…，$\boldsymbol{\beta}^{(p)}$，从而可得$p$对线性组合：

$U_1 = \boldsymbol{\alpha}^{(1)\prime}\mathbf{X}^{(1)}$，$V_1 = \boldsymbol{\beta}^{(1)\prime}\mathbf{X}^{(2)}$；$U_2 = \boldsymbol{\alpha}^{(2)\prime}\mathbf{X}^{(1)}$，$V_2 = \boldsymbol{\beta}^{(2)\prime}\mathbf{X}^{(2)}$；…；$U_p = \boldsymbol{\alpha}^{(p)\prime}\mathbf{X}^{(1)}$，$V_p = \boldsymbol{\beta}^{(p)\prime}\mathbf{X}^{(2)}$。称每一对变量为典型变量，因此，求典型相关系数和典型变量归结为求**A**和**B**的特征值及特征向量。

称$U_1=\boldsymbol{\alpha}^{(1)\prime}\mathbf{X}^{(1)}$，$V_1=\boldsymbol{\beta}^{(1)\prime}\mathbf{X}^{(2)}$为第一对典型变量，

λ_1为第一典型相关系数。

若第一对典型变量不足以代表原来两组变量的信息，

需要第二对典型变量：

$$U_2=\boldsymbol{\alpha}^{(2)\prime}\mathbf{X}^{(1)},\quad V_2=\boldsymbol{\beta}^{(2)\prime}\mathbf{X}^{(2)}$$

其中$\boldsymbol{\alpha}^{(2)}$、$\boldsymbol{\beta}^{(2)}$满足条件：$\boldsymbol{\alpha}^{(2)\prime}\Sigma_{11}\boldsymbol{\alpha}^{(2)}=1,\ \boldsymbol{\beta}^{(2)\prime}\Sigma_{22}\boldsymbol{\beta}^{(2)}=1,$

且使$\rho_{U_2V_2}=\boldsymbol{\alpha}^{(2)\prime}\Sigma_{12}\boldsymbol{\beta}^{(2)}$达到最大，而且

$$Cov(U_1,\ U_2)=\boldsymbol{\alpha}^{(1)\prime}\Sigma_{11}\boldsymbol{\alpha}^{(2)}=\mathbf{0}$$

$$Cov(V_1,\ V_2)=\boldsymbol{\beta}^{(1)\prime}\Sigma_{22}\boldsymbol{\beta}^{(2)}=\mathbf{0}$$

即第二对典型变量不应包含第一对典型变量的相关信息。

依此类推……

典型变量的性质

1. 由$\mathbf{X}^{(1)}$中出现的一切典型变量均不相关，并且方差为1，对于$\mathbf{X}^{(2)}$中出现的一切典型变量也是如此，即

$$Cov(U_i,\ U_j)=\delta_{ij}$$

$$Cov(V_i,\ V_j)=\delta_{ij}$$

2. $\mathbf{X}^{(1)}$与$\mathbf{X}^{(2)}$的同一对典型变量U_i与V_i之间的相关系数为λ_i，不同对的典型变量U_i与$V_j(i\neq j)$之间不相关，即

$$Cov(U_i,\ V_j)=\begin{cases}\lambda_i, & i=j=1,\cdots,\ p\\ 0, & i\neq j\end{cases}$$

三、样本的典型相关系数和典型变量

当总体的均值向量和协方差阵未知时，无法求总体的典型相关系数和典型变量，因而需要给出样本的典型相关系数和典型变量。

设$\mathbf{X}_{(1)}$，$\mathbf{X}_{(2)}$,…，$\mathbf{X}_{(n)}$为来自总体容量为n的样本，此时$\boldsymbol{\Sigma}$的极大似然估计为

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_{(j)}-\bar{\mathbf{X}})(\mathbf{X}_{(j)}-\bar{\mathbf{X}})'$$

其中$\bar{\mathbf{X}}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_{(j)}$。

用$\hat{\boldsymbol{\Sigma}}$替代$\boldsymbol{\Sigma}$即可求出样本典型相关系数 $\tilde{\lambda}_i$ 和 $\hat{\boldsymbol{\alpha}}^{(i)}$，$\hat{\boldsymbol{\beta}}^{(i)}$，

称 $\tilde{U}_i=\hat{\boldsymbol{\alpha}}^{(i)}{}'\mathbf{x}^{(1)}$，$\tilde{V}_i=\hat{\boldsymbol{\beta}}^{(i)}{}'\mathbf{x}^{(2)}(i=1,\cdots,p)$为样本的典型变量。

可以证明$\tilde{\lambda}_i$，$\hat{\boldsymbol{\alpha}}^{(i)}$，$\hat{\boldsymbol{\beta}}^{(i)}$分别为总体典型相关系数和典型相关系数向量$\lambda_i$，$\boldsymbol{\alpha}^{(i)}$，$\boldsymbol{\beta}^{(i)}$的极大似然估计量。

注：当指标间数据量纲不同或大小差异非常大时，建议先对数据实施标准化处理或直接从样本相关系数矩阵出发求样本的典型相关系数和典型变量。

四、典型相关系数的显著性检验

在作两组变量$\mathbf{X}^{(1)}$，$\mathbf{X}^{(2)}$的典型相关分析之前，首先应检验两组变量是否相关，若不相关，即$Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \mathbf{0}$，则讨论两组变量的典型相关就毫无意义。

设总体$\mathbf{X}$的两组变量$\mathbf{X}^{(1)} = (X_1, \cdots, X_p)', \mathbf{X}^{(2)} = (X_{p+1}, \cdots, X_{p+q})'$，且$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。

H_0：$Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \boldsymbol{\Sigma}_{12} = \mathbf{0}$

由似然比方法构造检验统计量为：

$$\Lambda=\prod_{i=1}^{p}(1-\hat{\lambda}_i^2)$$

其中$\hat{\lambda}_i^2$是$\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$的特征根，按大小次序排列为$\hat{\lambda}_1^2 \geq \hat{\lambda}_2^2 \geq \cdots \hat{\lambda}_p^2 > 0$，当$n >> 1$时，在$H_0$成立下$Q_0 = -m\ln\Lambda$近似服从$\chi^2(pq)$分布，其中$m = n-1-\frac{1}{2}(p+q+1)$。

在给定检验水平α之下，若$Q_0 > \chi_\alpha^2$，则拒绝H_0，认为第一对典型变量$\hat{U}_1$，$\hat{V}_1$具有相关性，其相关系数为$\hat{\lambda}_1$，即至少可以认为第一个典型相关系数是显著的。

将$\hat{\lambda}_1$除去后，再检验其余$p-1$个典型相关系数的显著性。此时检验统计量为：

$$\Lambda_1=\prod_{i=2}^{p}(1-\hat{\lambda}_i^2)$$

则$Q_1=-m_1\ln\Lambda$近似服从$\chi^2((p-1)(q-1))$分布，其中$m_1=n-2-\frac{1}{2}(p+q+1)$，若$Q_1>\chi_\alpha^2$，则认为$\hat{\lambda}_2$显著，即第二对典型变量$\hat{U}_2$，$\hat{V}_2$相关。依此类推，直到某一个相关系数$\hat{\lambda}_k$检验为不显著时截止。这时，我们找出了反映两组变量相互关系的$k-1$对典型变量。

一般，检验第r个$(r<k)$典型相关系数的显著性时，作统计量：

$$Q_{r-1}=-[n-r-\tfrac{1}{2}(p+q+1)]\ln\Lambda_{r-1}$$

它近似服从自由度为$(p-r+1)(q-r+1)$的χ^2分布，其中$\Lambda_{r-1}=\prod_{i=r}^{p}(1-\hat{\lambda}_i^2)$。该统计量*Hotelling*（1936），*Girshik*（1939）和*Anderson*（1958）均给出过精确分布，但形式非常复杂。*Bartlett*给出了在大样本情形下，它近似服从χ^2分布。

五、简单相关，复相关和典型相关之间的关系

- 当$p=q=1$时，$\mathbf{X}^{(1)}$与$\mathbf{X}^{(2)}$之间的（唯一）典型相关就是它们之间的简单相关；当$p=1$或$q=1$时，$\mathbf{X}^{(1)}$与$\mathbf{X}^{(2)}$之间的（唯一）典型相关就是它们之间的复相关(由此可得复相关系数的公式)。因此复相关是典型相关的一个特例，而简单相关是复相关的一个特例。

- 从第一个典型相关的定义可以看出，第一个典型相关系数至少同$\mathbf{X}^{(1)}$（或$\mathbf{X}^{(2)}$）的任一分量与$\mathbf{X}^{(2)}$（或$\mathbf{X}^{(1)}$）的复相关系数一样大，即使所有这些复相关系数都很小，第一典型相关系数仍可能很大；同样，从复相关系数的定义也可看出，当$p=1$（或$q=1$时），$\mathbf{X}^{(1)}$（或$\mathbf{X}^{(2)}$）与$\mathbf{X}^{(2)}$（或$\mathbf{X}^{(1)}$）之间的复相关系数也不会小于$\mathbf{X}^{(1)}$（或$\mathbf{X}^{(2)}$）与$\mathbf{X}^{(2)}$（或$\mathbf{X}^{(1)}$）的任一分量之间的相关系数，即使所有这些相关系数都很小，复相关系数仍可能很大。

六、典型载荷分析

- 典型载荷分析是指原始变量与典型变量之间的相关性分析。进行典型载荷分析有助于更好分析已提取的典型变量。

令 $\mathbf{G}=(\boldsymbol{\alpha}^{(1)},\ \boldsymbol{\alpha}^{(2)},\cdots,\ \boldsymbol{\alpha}^{(p)})'$, $\mathbf{H}=(\boldsymbol{\beta}^{(1)},\ \boldsymbol{\beta}^{(2)},\cdots,\ \boldsymbol{\beta}^{(q)})'$

$\mathbf{U}=(U_1,\ U_2,\cdots,\ U_p)'$, $\mathbf{V}=(V_1,\ V_2,\cdots,\ V_q)'$

其中，$\mathbf{G}$、$\mathbf{H}$为典型变量系数组成的矩阵；

$\mathbf{U}$、$\mathbf{V}$为典型变量组成的向量；

则 $\mathbf{U}=\mathbf{G}\mathbf{X}^{(1)}$、$\mathbf{V}=\mathbf{H}\mathbf{X}^{(2)}$

注：$\boldsymbol{\beta}^{(p+1)},\cdots,\ \boldsymbol{\beta}^{(q)}$为$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$的零特征根所对应的正交特征向量。

考查$\mathbf{U}$与$\mathbf{X}^{(1)}$，$\mathbf{V}$与$\mathbf{X}^{(2)}$，$\mathbf{U}$与$\mathbf{X}^{(2)}$，$\mathbf{V}$与$\mathbf{X}^{(1)}$之间的相关系数矩阵。

记 $\mathbf{V}_{11}^{-1/2}=\mathrm{diag}(\sigma_{11}^{-1/2},\sigma_{22}^{-1/2},...,\sigma_{pp}^{-1/2})$

则 $\mathbf{R}_{\mathbf{U},\mathbf{X}^{(1)}}=\mathrm{corr}(\mathbf{U},\mathbf{X}^{(1)})=\mathrm{cov}(\mathbf{U},\mathbf{V}_{11}^{-1/2}\mathbf{X}^{(1)})$

$$=\mathrm{cov}(\mathbf{G}\mathbf{X}^{(1)},\mathbf{V}_{11}^{-1/2}\mathbf{X}^{(1)})=\mathbf{G}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2}$$

类似可得 $\mathbf{R}_{\mathbf{V},\mathbf{X}^{(2)}}=\mathbf{H}\boldsymbol{\Sigma}_{22}\mathbf{V}_{22}^{-1/2}$

$$\mathbf{R}_{\mathbf{U},\mathbf{X}^{(2)}}=\mathbf{H}\boldsymbol{\Sigma}_{22}\mathbf{V}_{22}^{-1/2}$$

$$\mathbf{R}_{\mathbf{V},\mathbf{X}^{(1)}}=\mathbf{G}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2}$$

- 对于经过标准化处理的典型变量有

$$\mathbf{R}_{\mathbf{U},\mathbf{Z}^{(1)}} = \mathbf{H}_{\mathbf{Z}}\mathbf{R}_{11}$$

$$\mathbf{R}_{\mathbf{V},\mathbf{Z}^{(2)}} = \mathbf{G}_{\mathbf{Z}}\mathbf{R}_{22}$$

$$\mathbf{R}_{\mathbf{U},\mathbf{Z}^{(2)}} = \mathbf{H}_{\mathbf{Z}}\mathbf{R}_{22}$$

$$\mathbf{R}_{\mathbf{V},\mathbf{Z}^{(1)}} = \mathbf{G}_{\mathbf{Z}}\mathbf{R}_{11}$$

七、典型冗余分析

- 典型冗余分析用于讨论原始变量被典型变量解释的百分比，从而度量典型变量所包含的原始信息量的大小。
- 考虑标准化情形：由于 $\mathbf{U}=\mathbf{G}_Z\mathbf{Z}^{(1)}$、$\mathbf{V}=\mathbf{H}_Z\mathbf{Z}^{(2)}$，则

$$\operatorname{cov}(\mathbf{Z}^{(1)},\mathbf{U})=\operatorname{cov}(\mathbf{G}_Z^{-1}\mathbf{U},\mathbf{U})=\mathbf{G}_Z^{-1}=\begin{pmatrix} r_{z_1^{(1)},U_1} & \cdots & r_{z_1^{(1)},U_p} \\ \vdots & & \vdots \\ r_{z_p^{(1)},U_1} & \cdots & r_{z_p^{(1)},U_p} \end{pmatrix}$$

$$\operatorname{cov}(\mathbf{Z}^{(2)},\mathbf{V})=\operatorname{cov}(\mathbf{H}_Z^{-1}\mathbf{V},\mathbf{V})=\mathbf{H}_Z^{-1}=\begin{pmatrix} r_{z_1^{(2)},V_1} & \cdots & r_{z_1^{(2)},V_q} \\ \vdots & & \vdots \\ r_{z_q^{(2)},V_1} & \cdots & r_{z_q^{(2)},V_q} \end{pmatrix}$$

- 定义前r对典型变量对样本总方差的贡献为

$$tr(\boldsymbol{\alpha}_{\mathbf{Z}}^{(1)}\boldsymbol{\alpha}_{\mathbf{Z}}^{(1)\prime}+\boldsymbol{\alpha}_{\mathbf{Z}}^{(2)}\boldsymbol{\alpha}_{\mathbf{Z}}^{(2)\prime}+\ldots+\boldsymbol{\alpha}_{\mathbf{Z}}^{(r)}\boldsymbol{\alpha}_{\mathbf{Z}}^{(r)\prime})=\sum_{i=1}^{r}\sum_{k=1}^{p}r_{z_k^{(1)},U_i}^2$$

$$tr(\boldsymbol{\beta}_{\mathbf{Z}}^{(1)}\boldsymbol{\beta}_{\mathbf{Z}}^{(1)\prime}+\boldsymbol{\beta}_{\mathbf{Z}}^{(2)}\boldsymbol{\beta}_{\mathbf{Z}}^{(2)\prime}+\ldots+\boldsymbol{\beta}_{\mathbf{Z}}^{(r)}\boldsymbol{\beta}_{\mathbf{Z}}^{(r)\prime})=\sum_{i=1}^{r}\sum_{k=1}^{q}r_{z_k^{(2)},V_i}^2$$

则第一组变量方差由前r个典型变量解释的比例为

$$\mathbf{R}d_{\mathbf{Z}^{(1)}|\mathbf{U}}=\sum_{i=1}^{r}\sum_{k=1}^{p}r_{z_k^{(1)},U_i}^2\Big/p$$

则第二组变量方差由前r个典型变量解释的比例为

$$\mathbf{R}d_{\mathbf{Z}^{(2)}|\mathbf{V}}=\sum_{i=1}^{r}\sum_{k=1}^{q}r_{z_k^{(2)},V_i}^2\Big/q$$

例1. 计算标准化随机变量的典型变量和典型相关系数

假设$\mathbf{X}^{(1)}=(X_1^{(1)},\ X_2^{(1)})$和$\mathbf{X}^{(2)}=(X_1^{(2)},\ X_2^{(2)})$均为标准化随机变量。记$\mathbf{X}=(\mathbf{X}^{(1)},\ \mathbf{X}^{(2)})$，并且已知

$$Cov(\mathbf{X})=\left(\begin{array}{cc:cc} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ \hdashline 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{array}\right)$$

求$\mathbf{X}^{(1)}$和$\mathbf{X}^{(2)}$的典型相关系数及典型变量。

记$\mathbf{R}_{11}=\begin{pmatrix}1 & 0.4\\ 0.4 & 1\end{pmatrix}$, $\mathbf{R}_{12}=\begin{pmatrix}0.5 & 0.6\\ 0.3 & 0.4\end{pmatrix}$, $\mathbf{R}_{22}=\begin{pmatrix}1 & 0.2\\ 0.2 & 1\end{pmatrix}$,

则$\mathbf{A}=\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{12}^{'}=\begin{pmatrix}0.4519 & 0.2892\\ 0.1463 & 0.0947\end{pmatrix}$,

$\mathbf{B}=\mathbf{R}_{22}^{-1}\mathbf{R}_{12}^{'}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}=\begin{pmatrix}0.2063 & 0.251\\ 0.2778 & 0.3402\end{pmatrix}$,

A和**B**的特征值为$\lambda_1^2=0.5457$，$\lambda_2^2=0.0009$。

A的特征向量为(0.951，0.309)'，(-0.54，0.842)';

B的特征向量为(0.595，0.804)'，(-0.774，0.633)'。

因此，$U_1=0.951X_1^{(1)}+0.309X_2^{(1)}$，$V_1=0.595X_1^{(2)}+0.804X_2^{(2)}$

为第一对典型变量，且$\rho_{UV}=\lambda_1=\sqrt{0.5457}=0.7387$。

- SAS程序如下：

```
data a1(type=corr);
_type_='corr';
input name $ x1-x4;
cards;
x1 1   0.4 0.5 0.6
x2 0.4 1   0.3 0.4
x3 0.5 0.3 1   0.2
x4 0.6 0.4 0.2 1
;
proc cancorr data=a1 vprefix=v wprefix=w;
with x3 x4;
var x1 x2;
run;
```

例2. 为了研究长子的头型与次子的头型之间的关系，研究者随机抽查了25个家庭的两兄弟的头长和头宽（数据见can1.xls），希望求得两组变量的典型变量及典型相关系数。

解： 以long1与width1分别表示长子的头长和头宽，以long2与width2分别表示次子的头长和头宽。

求得两组变量的相关系数矩阵如下：

$$\begin{pmatrix} 1 & 0.7346 & 0.7108 & 0.704 \\ & 1 & 0.6932 & 0.7086 \\ & & 1 & 0.8393 \\ & & & 1 \end{pmatrix}$$

SAS程序如下：

```
data sasuser.can1;
input long1 width1 long2 width2;
n=_n_;/*n用于记录观测序号*/
cards;
…… /*输入数据*/
;
run;
proc cancorr data=sasuser.can1 simple
   corr redundancy out=ocan1 vprefix=v
   wprefix=w
     vname='长子' wname='次子';
   with long2 width2;
   var long1 width1;
run;
```

```
proc print data=ocan1;
var v1 w1 long1 width1 long2 width2;
run;
proc plot data=ocan1;
plot w1*v1 $ n='*'/vref=0 href=0（图中显示
v1=0和w1=0两条直线）;
run;
```

由SAS运行结果可得第一对典型变量：

V1=0.0566*long1+0.0707*width1

W1=0.0502*long2+0.0802*width2

第一典型相关系数，即V1与w1的相关系数为0.7885，检验结果为显著；第二典型相关系数为0.0537，检验结果为不显著。注意到第一典型相关系数比两组变量间最大的相关系数（long1与long2的相关系数为0.7108）都大。

SPSS程序如下：

include ‘SPSS 所在路径\Canonical correlation.sps'.
cancorr set1=第一组变量的列表
/set2=第二组变量的列表.

注： **Canonical correlation.sps**为**SPSS**提供的典型相关分析的宏程序

此例中，在语法窗口中键入的程序如下：

```
include 'C:\Program Files\Spss\Canonical
    correlation.sps'.
cancorr set1=long1 width1
   /set2=long2 width2.
```

SPSS运行结果解释：

```
Run MATRIX procedure:

Correlations for Set-1
          LONG1   WIDTH1
 LONG1   1.0000    .7346
WIDTH1    .7346   1.0000

Correlations for Set-2
          LONG2   WIDTH2
 LONG2   1.0000    .8393
WIDTH2    .8393   1.0000
```

上表为两组变量内部各自的相关矩阵。

Correlations Between Set-1 and Set-2

	LONG2	WIDTH2
LONG1	.7108	.7040
WIDTH1	.6932	.7086

上表为两组变量间各变量的两两相关矩阵。

Canonical Correlations

1	.789
2	.054

上表为两个典型相关系数的值。

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	.377	20.964	4.000	.000
2	.997	.062	1.000	.803

上表为检验各典型相关系数的显著性，可见第一典型相关系数显著，而第二典型相关系数不显著。可用来判断需提取多少对典型变量。

Standardized Canonical Coefficients for Set-1

	1	2
LONG1	-.552	-1.366
WIDTH1	-.522	1.378

Raw Canonical Coefficients for Set-1

	1	2
LONG1	-.057	-.140
WIDTH1	-.071	.187

以上为各典型变量与第一组变量中各变量间标准化与未标准化的系数列表。由此可得，典型变量的转换公式（原始的）为

U_1 ＝ 0.057* long1 +0.071*width1

U_2 ＝ 0.14* long1－0.187*width1

Standardized Canonical Coefficients for Set-2

	1	2
LONG2	-.504	-1.769
WIDTH2	-.538	1.759

Raw Canonical Coefficients for Set-2

	1	2
LONG2	-.050	-.176
WIDTH2	-.080	.262

以上为各典型变量与第二组变量中各变量间标准化与未标准化的系数列表。由此可得，典型变量的转换公式（原始的）为

$V_1 = 0.05*\text{long2} + 0.08*\text{width2}$

$V_2 = 0.176*\text{long2} - 0.262*\text{width2}$

Canonical Loadings for Set-1

	1	2
LONG1	-.935	-.354
WIDTH1	-.927	.375

Cross Loadings for Set-1

	1	2
LONG1	-.737	-.019
WIDTH1	-.731	.020

以上为各第一组变量中各变量分别与自身、相对的典型变量的相关系数列表，可见它们主要与第一对典型变量的关系较为密切。

Canonical Loadings for Set-2

	1	2
LONG2	-.956	-.293
WIDTH2	-.962	.274

Cross Loadings for Set-2

	1	2
LONG2	-.754	-.016
WIDTH2	-.758	.015

以上为各第二组变量中各变量分别与自身、相对的典型变量的相关系数列表，可见它们主要与第一对典型变量的关系较为密切。

Redundancy Analysis:

Proportion of Variance of Set-1 Explained by Its Own Can. Var.

	Prop Var
CV1-1	.867
CV1-2	.133

Proportion of Variance of Set-1 Explained by Opposite Can.Var.

	Prop Var
CV2-1	.539
CV2-2	.000

以上为冗余度（Redundancy）分析结果，列出了各典型相关系数所能解释原变量变异的比例，可以用来辅助判断需要保留多少个典型相关系数。可见第一组变量的变异被自身的第一典型变量解释了**86.7%**，第二典型变量解释了**13 .3%**；被相对的第一典型变量解释了**53.9%**，第二典型变量解释了**0%**。

```
 Proportion of Variance of Set-2 Explained by Its Own Can. Var.
                Prop Var
CV2-1               .920
CV2-2               .080

Proportion of Variance of Set-2 Explained by Opposite Can. Var.
                Prop Var
CV1-1               .572
CV1-2               .000

------ END MATRIX -----
```

上表为第二组变量的变异被自身、相对典型变量所解释的比例，结果与上面类似。因此，综合上述冗余度分析的结果，我们只需保留第一对典型变量即可。

The canonical scores have been written to the active file.
Also, a file containing an SPSS Scoring program has been written
To use this file GET a system file with the SAME variables
Which were used in the present analysis. Then use an INCLUDE command to run the scoring program.
For example :

```
GET FILE anotherfilename
INCLUDE FILE "CC__.INC".
EXECUTE.
```

最后系统给出说明：标准化变量已被写入当前文件，同时相应的程序也以文件形式被储存在当前目录中，可以使用**GET**命令进入数据文件，再使用**INCLUDE**命令来调用相应程序。

例3. 工作满意感的典型相关分析

为了研究工作满意感与工作特征之间相关程度如何，**Dunham**通过问卷调查的形式，对一家大型商业集团分支机构的**784**位经理的**5**个工作特征指标与**7**个工作满意感指标进行调查分析。研究结果将为工作设计提供参考。

工作特征变量：

X_1 = **feedback**（反馈）

X_2 = **task significance**（任务重要性）

X_3 = **task variety**（任务多样性）

X_4 = **task identity**（任务完整性）

X_5 = **autonomy**（自主性）

工作满意感变量：

Y_1 = supervisor satisfaction（上级满意感）

Y_2 = career-future satisfaction（职业前景满意感）

Y_3 = financial satisfaction（报酬满意感）

Y_4 = workload satisfaction（工作负荷满意感）

Y_5 = company identification（公司认同感）

Y_6 = kind-of-work-satisfaction（工作类别满意感）

Y_7 = general satisfaction（总体满意感）

这12个变量的样本相关系数矩阵见下表。

1.00 . . . . . .

0.49 1.00 . . . . . .

0.53 0.57 1.00 . . . . . .

0.49 0.46 0.48 1.00 . . . . . .

0.51 0.53 0.57 0.57 1.00 . . . . . .

0.33 0.30 0.31 0.24 0.38 1.00 . . . . . .

0.32 0.21 0.23 0.22 0.32 0.43 1.00 . . . . .

0.20 0.16 0.14 0.12 0.17 0.27 0.33 1.00 . . . .

0.19 0.08 0.07 0.19 0.23 0.24 0.26 0.25 1.00 . . .

0.30 0.27 0.24 0.21 0.32 0.34 0.54 0.46 0.28 1.00 . .

0.37 0.35 0.37 0.29 0.36 0.37 0.32 0.29 0.30 0.35 1.00 .

0.21 0.20 0.18 0.16 0.27 0.40 0.58 0.45 0.27 0.59 0.31 1.00

- 试利用上述相关系数矩阵对工作满意感与工作特征之间进行典型相关分析。要求：
- （1）给出典型变量的个数；
- （2）给出典型相关系数；
- （3）写出典型变量的表达式；
- （4）由典型相关分析的结果可以得到怎样的信息？

案例分析作业：

试对我国城市竞争力与基础设施状况作典型相关分析（要求利用2007年或2006年的数据）

案例分析实施的步骤：

- （1）查阅问题的相关背景知识（这一步至关重要，涉及到指标的选取）；
- （2）数据的搜集；
- （3）数据的初步整理；
- （4）统计分析；
- （5）结合问题的背景知识和统计分析结果给出结论并进行讨论；
- （6）撰写完整的案例分析报告

案例分析报告的要求：

- 学生在做完案例分析并进行小组讨论后，需提交一份分析报告。报告内容应包括：
- （1）案例分析的步骤，每一步骤的方法、结果，最终的结论；
- （2）小组讨论的内容、结果；
- （3）案例分析及小组讨论中未解决的问题；
- （4）对本案例中提供的方法进行评价，并提出改进的意见或采用其它方法的建议。