

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第八章 典型相关分析

主讲教师 周仕君

- 在对经济问题的研究和管理中，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。
- 例如，为了研究扩张性财政政策实施以后对宏观经济发展的影响，就需要考察有关财政政策的一系列指标如财政支出总额的增长率、财政赤字增长率、国债发行额的增长率、税率降低率等与经济发展的一系列指标如国内生产总值增长率、就业增长率、物价上涨率等两组变量之间的相关程度。



第八章 典型相关分析

典型相关分析 (Canonical Correlation) 是研究两组变量之间内在联系的基本统计方法。它能够揭示出两组变量之间的内在联系。

✓ 第二节 典型相关系数的显著性检验

一、典型相关分析的基本思想

□ 一般情况，设 $\mathbf{X}^{(1)} = (X_1^{(1)}, X_2^{(1)}, \dots, X_p^{(1)})$ $\mathbf{X}^{(2)} = (X_1^{(2)}, X_2^{(2)}, \dots, X_q^{(2)})$

是两个相互关联的随机向量，分别在两组变量中选取若干有代表性的综合变量 U_i 、 V_i ，使得每一个综合变量是原变量的线性组合，即

$$U_i = a_1^{(i)} X_1^{(1)} + a_2^{(i)} X_2^{(1)} + \dots + a_p^{(i)} X_p^{(1)} \square \mathbf{a}^{(i)'} \mathbf{X}^{(1)}$$

$$V_i = b_1^{(i)} X_1^{(2)} + b_2^{(i)} X_2^{(2)} + \dots + b_q^{(i)} X_q^{(2)} \square \mathbf{b}^{(i)'} \mathbf{X}^{(2)}$$

□ 典型相关分析的基本思想即是找 $\mathbf{X}^{(1)}$ 、 $\mathbf{X}^{(2)}$ 的一组线性函数 $\mathbf{a}^{(i)'} \mathbf{X}^{(1)}$ 与 $\mathbf{b}^{(i)'} \mathbf{X}^{(2)}$ ，以使得它们的相关系数达到最大。



□ 若存在常向量 $\mathbf{a}^{(1)}, \mathbf{b}^{(1)}$, 在 $D(\mathbf{a}^{(1)'} \mathbf{X}^{(1)}) = D(\mathbf{b}^{(1)'} \mathbf{X}^{(2)}) = 1$ 的条件下, 使得 $\rho(\mathbf{a}^{(1)'} \mathbf{X}^{(1)}, \mathbf{b}^{(1)'} \mathbf{X}^{(2)})$ 达到最大, 则称 $\mathbf{a}^{(1)'} \mathbf{X}^{(1)}, \mathbf{b}^{(1)'} \mathbf{X}^{(2)}$ 是 $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ 的第一对典型相关变量。

□ 求出第一对典型相关变量之后, 可以类似的求出各对之间互不相关的第二对、第三对等典型相关变量。这些典型相关变量就反映了 $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ 之间的线性相关情况。



二、典型相关分析原理及方法

□ 设有两组随机向量, $X^{(1)}$ 代表第一组的 p 个变量, $X^{(2)}$ 代表第二组的 q 个变量, 假设 $p \leq q$ 。令

$$\text{Cov}(X^{(1)}) = \Sigma_{11}, \quad \textcircled{1}$$

$$\text{Cov}(X^{(2)}) = \Sigma_{22},$$

$$\text{Cov}(X^{(1)}, X^{(2)}) = \Sigma_{12} = \Sigma'_{21}$$

$$\text{Cov}(X, X) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$\begin{matrix} (p \times p) & (p \times q) \\ (q \times p) & (q \times q) \end{matrix}$

$\textcircled{3}$

$$X_{(p+q) \times 1} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_p^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_q^{(2)} \end{bmatrix}$$

$\textcircled{2}$



- 首先要计算出各组变量的线性组合——典型变量，并使其相关系数达到最大。因此，设两组变量的线性组合分别为：

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

$$V = \mathbf{b}'\mathbf{X}^{(2)} = b_1X_1^{(2)} + b_2X_2^{(2)} + \cdots + b_qX_q^{(2)}$$

- 易见：

$$D(U) = D(\mathbf{a}'\mathbf{X}^{(1)}) = \mathbf{a}'\text{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(1)})\mathbf{a} = \mathbf{a}'\Sigma_{11}\mathbf{a}$$

$$D(V) = D(\mathbf{b}'\mathbf{X}^{(2)}) = \mathbf{b}'\text{Cov}(\mathbf{X}^{(2)}, \mathbf{X}^{(2)})\mathbf{b} = \mathbf{b}'\Sigma_{22}\mathbf{b}$$

$$\text{Cov}(U, V) = \mathbf{a}'\text{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})\mathbf{b} = \mathbf{a}'\Sigma_{12}\mathbf{b}$$

$$\text{Corr}(U, V) = \frac{\text{Cov}(U, V)}{\sqrt{D(U)}\sqrt{D(V)}} = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}}$$



$$\text{Corr}(c\mathbf{a}'\mathbf{X}^{(1)}, c\mathbf{b}'\mathbf{X}^{(2)}) = \frac{c\mathbf{a}'\Sigma_{12}\mathbf{b}c'}{c\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}c\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}} = \text{Corr}(U, V)$$

- 寻找使相关系数达到最大的向量 $\mathbf{a}$ 与 $\mathbf{b}$ ，由于随机向量乘以常数时并不改变它们的相关系数，为防止不必要的结果重复出现（有唯一解），最好的限制是

$$D(U) = \mathbf{a}'\Sigma_{11}\mathbf{a} = 1, D(V) = \mathbf{b}'\Sigma_{22}\mathbf{b} = 1 \quad (9.1)$$

令

$$\text{Corr}(U, V) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}} = \mathbf{a}'\Sigma_{12}\mathbf{b}$$

- 那么，问题就成为在 (9.1) 式的约束条件下，求使

$\text{Corr}(U, V) = \mathbf{a}'\Sigma_{12}\mathbf{b}$ ，达到最大的系数向量 $\mathbf{a}$ 与 $\mathbf{b}$ 。



□ 根据条件极值的求法引入Lagrange乘数，将问题转化为求

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

的极大值，其中 λ ， ν 是Lagrange乘数。

📖 根据求极值的必要条件得

$$\begin{cases} \frac{\partial \varphi}{\partial \mathbf{a}} = \Sigma_{12}\mathbf{b} - \lambda\Sigma_{11}\mathbf{a} = 0 \\ \frac{\partial \varphi}{\partial \mathbf{b}} = \Sigma_{21}\mathbf{a} - \nu\Sigma_{22}\mathbf{b} = 0 \end{cases}$$



□ 将方程组的二式分别左乘 $\mathbf{a}'$ 与 $\mathbf{b}'$ 则得

$$\begin{cases} \mathbf{a}'\Sigma_{12}\mathbf{b} - \lambda\mathbf{a}'\Sigma_{11}\mathbf{a} = 0 \\ \mathbf{b}'\Sigma_{21}\mathbf{a} - \nu\mathbf{b}'\Sigma_{22}\mathbf{b} = 0 \end{cases}$$

即有

$$\begin{cases} \mathbf{a}'\Sigma_{12}\mathbf{b} = \lambda\mathbf{a}'\Sigma_{11}\mathbf{a} = \lambda \\ \mathbf{b}'\Sigma_{21}\mathbf{a} = \nu\mathbf{b}'\Sigma_{22}\mathbf{b} = \nu \end{cases}$$

$$\begin{cases} \frac{\partial \phi}{\partial \mathbf{a}} = \Sigma_{12}\mathbf{b} - \lambda\Sigma_{11}\mathbf{a} = 0 \\ \frac{\partial \phi}{\partial \mathbf{b}} = \Sigma_{21}\mathbf{a} - \nu\Sigma_{22}\mathbf{b} = 0 \end{cases}$$

□ 所以 $\lambda = \nu = \mathbf{a}'\Sigma_{12}\mathbf{b}$ ，知 λ 为线性组合 U, V 的相关系数。用 λ 代替方程组中的 ν ，则方程组写为：

$$\begin{cases} \Sigma_{12}\mathbf{b} - \lambda\Sigma_{11}\mathbf{a} = 0 \\ \Sigma_{21}\mathbf{a} - \lambda\Sigma_{22}\mathbf{b} = 0 \end{cases}$$



$$\mathbf{b} = \frac{1}{\lambda} \Sigma_{22}^{-1} \Sigma_{21} \mathbf{a}$$

$$\begin{cases} \Sigma_{21} \mathbf{a} - \lambda \Sigma_{22} \mathbf{b} = 0 \\ \Sigma_{12} \mathbf{b} - \lambda \Sigma_{11} \mathbf{a} = 0 \end{cases}$$

$$\frac{1}{\lambda} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \mathbf{a} - \lambda \Sigma_{11} \mathbf{a} = 0$$

$$\Sigma_{11}^{-1} \times \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \mathbf{a} - \lambda^2 \Sigma_{11} \mathbf{a} = 0$$

$$\Sigma_{22}^{-1} \times \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \mathbf{b} - \lambda^2 \Sigma_{22} \mathbf{b} = 0$$

$\text{Corr}(U, V) = \lambda$

由此可见, $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$ 和 $\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ 具有相同的特征根 λ^2 , $\mathbf{a}, \mathbf{b}$ 则是其相应的特征向量。

A

$(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \lambda^2 \mathbf{I}_p) \mathbf{a} = 0$

B

$(\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 \mathbf{I}_q) \mathbf{b} = 0$



□ 令 $C \times D$ $D \times C$

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

其中 $\mathbf{A}$ 为 $p \times p$ 阶矩阵, $\mathbf{B}$ 为 $q \times q$ 阶矩阵。

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



□ 因此，最大特征根 λ_1^2 对应的特征向量

$\mathbf{a}^{(1)} = (a_1^{(1)}, a_2^{(1)}, \dots, a_p^{(1)})'$ 和 $\mathbf{b}^{(1)} = (b_1^{(1)}, b_2^{(1)}, \dots, b_q^{(1)})'$ 就是所求的典型变量的系数向量。

□ 即可得

$$U_1 = \mathbf{a}^{(1)'} \mathbf{X}^{(1)} = a_1^{(1)} X_1^{(1)} + a_2^{(1)} X_2^{(1)} + \dots + a_p^{(1)} X_p^{(1)}$$

$$V_1 = \mathbf{b}^{(1)'} \mathbf{X}^{(2)} = b_1^{(1)} X_1^{(2)} + b_2^{(1)} X_2^{(2)} + \dots + b_q^{(1)} X_q^{(2)}$$

□ 称 U_1 、 V_1 为第一对典型变量，最大特征根的平方根 λ_1 即为两典型变量的相关系数，称为第一典型相关系数。



□ **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，所有特征根非负。

2. **A** 和 **B** 的特征根均在 $0 \sim 1$ 之间。

3. 设 **A** 和 **B** 的非零特征根为 $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_r^2$ ， $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(r)}$

为 **A** 对应于 $\lambda_1^2, \lambda_2^2, \dots, \lambda_r^2$ 的特征向量， $\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \dots, \mathbf{b}^{(r)}$ 为 **B**

对应于 $\lambda_1^2, \lambda_2^2, \dots, \lambda_r^2$ 的特征向量。



- 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

$$U_2 = \mathbf{a}^{(2)'} \mathbf{X}^{(1)}$$

$$V_2 = \mathbf{b}^{(2)'} \mathbf{X}^{(2)}$$

- 要求第二对典型变量也要满足如下约束条件：

$$D(U_2) = \mathbf{a}^{(2)'} \boldsymbol{\Sigma}_{11} \mathbf{a}^{(2)} = 1$$

$$D(V_2) = \mathbf{b}^{(2)'} \boldsymbol{\Sigma}_{22} \mathbf{b}^{(2)} = 1$$

- 除此之外，为了有效测度两组变量的相关信息，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，因而，需增加约束条件：

$$\text{Cov}(U_1, U_2) = \text{Cov}(\mathbf{a}^{(1)'} \mathbf{X}^{(1)}, \mathbf{a}^{(2)'} \mathbf{X}^{(1)}) = \mathbf{a}^{(1)'} \boldsymbol{\Sigma}_{11} \mathbf{a}^{(2)} = 0$$

$$\text{Cov}(V_1, V_2) = \text{Cov}(\mathbf{b}^{(1)'} \mathbf{X}^{(2)}, \mathbf{b}^{(2)'} \mathbf{X}^{(2)}) = \mathbf{b}^{(1)'} \boldsymbol{\Sigma}_{22} \mathbf{b}^{(2)} = 0$$



□ 在式的约束条件下，可求得其相关系数：

$$\text{Corr}(U_2, V_2) = \mathbf{a}^{(2)'} \boldsymbol{\Sigma}_{12} \mathbf{b}^{(2)}$$

的最大值为上述矩阵 **A** 和 **B** 的第二大特征根 λ_2 的平方根 λ_2^2 ，其对应的单位特征向量 $\mathbf{a}^{(2)}$ 、 $\mathbf{b}^{(2)}$ 就是第二对典型变量的系数向量，称 $U_2 = \mathbf{a}^{(2)'} \mathbf{X}^{(1)}$ 和 $V_2 = \mathbf{b}^{(2)'} \mathbf{X}^{(2)}$ 为第二对典型变量， λ_2 为第二典型相关系数。

□ 类似地，依次可求出第 r 对典型变量： $U_r = \mathbf{a}^{(r)'} \mathbf{X}^{(1)}$ 和 $V_r = \mathbf{b}^{(r)'} \mathbf{X}^{(2)}$ ，其系数向量 $\mathbf{a}^{(r)}$ 和 $\mathbf{b}^{(r)}$ 分别为矩阵 **A** 和 **B** 的第 r 特征根 λ_r^2 对应的特征向量。即 λ_r 为第 r 典型相关系数。



□ 综上所述，典型变量和典型相关系数的计算可归结为矩阵A和B特征根及相应特征向量的求解。如果矩阵A和B的秩为 r ，则共有 r 对典型变量，第 k 对($1 \leq k \leq r$)典型变量的系数向量分别是矩阵A和B第 k 特征根 λ_k^2 相应的特征向量，典型相关系数为 λ_k 。



□ 典型变量具有如下性质：

$$D(U_k) = 1, D(V_k) = 1 \quad (k = 1, 2, \dots, r)$$

$$\text{Cov}(U_i, U_j) = 0, \quad \text{Cov}(V_i, V_j) = 0 \quad (i \neq j)$$

$$\text{Cov}(U_i, V_j) = \begin{cases} \lambda_i \neq 0 & (i = j, i = 1, 2, \dots, r) \\ 0 & (i \neq j) \\ 0 & (j > r) \end{cases}$$



第三节 样本典型相关分析

- 样本典型相关变量及典型相关系数的计算
- 典型相关系数的显著性检验



一、样本典型相关变量及典型相关系数的计算

□ 设 $\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}$ 服从正态分布 $N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 从该总体中

抽取样本容量为 n 的样本, 得到下列数据矩阵:

$$\mathbf{X}^{(1)} = \begin{bmatrix} X_{11}^{(1)} & X_{12}^{(1)} & \cdots & X_{1p}^{(1)} \\ X_{21}^{(1)} & X_{22}^{(1)} & \cdots & X_{2p}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}^{(1)} & X_{n2}^{(1)} & \cdots & X_{np}^{(1)} \end{bmatrix} \quad \mathbf{X}^{(2)} = \begin{bmatrix} X_{11}^{(2)} & X_{12}^{(2)} & \cdots & X_{1q}^{(2)} \\ X_{21}^{(2)} & X_{22}^{(2)} & \cdots & X_{2q}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}^{(2)} & X_{n2}^{(2)} & \cdots & X_{nq}^{(2)} \end{bmatrix}$$



□ 样本均值向量: $\bar{\mathbf{X}} = \begin{bmatrix} \bar{\mathbf{X}}^{(1)} \\ \bar{\mathbf{X}}^{(2)} \end{bmatrix}$

$$\text{其中, } \bar{\mathbf{X}}^{(1)} = \frac{1}{n} \sum_{\alpha=1}^n \mathbf{X}_{\alpha}^{(1)} \quad \bar{\mathbf{X}}^{(2)} = \frac{1}{n} \sum_{\alpha=1}^n \mathbf{X}_{\alpha}^{(2)}$$

□ 样本协差阵: $\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}$

$$\text{其中, } \hat{\Sigma}_{kl} = \frac{1}{n-1} \sum_{j=1}^n (\mathbf{X}_j^{(k)} - \bar{\mathbf{X}}^{(k)})(\mathbf{X}_j^{(l)} - \bar{\mathbf{X}}^{(l)})', \quad k, l = 1, 2$$

□ 由此可得 **A** 和 **B** 的样本估计:

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



- 另外，若样本数据矩阵已经标准化处理，此时样本的协差阵就等于样本的相关系数矩阵。

$$\hat{\mathbf{R}} = \begin{bmatrix} \hat{\mathbf{R}}_{11} & \hat{\mathbf{R}}_{12} \\ \hat{\mathbf{R}}_{21} & \hat{\mathbf{R}}_{22} \end{bmatrix}$$

- 由此可得矩阵 $\mathbf{A}$ 和 $\mathbf{B}$ 的样本估计：

$$\hat{\mathbf{A}}^* = \hat{\mathbf{R}}_{11}^{-1} \hat{\mathbf{R}}_{12} \hat{\mathbf{R}}_{22}^{-1} \hat{\mathbf{R}}_{21} \quad \hat{\mathbf{B}}^* = \hat{\mathbf{R}}_{22}^{-1} \hat{\mathbf{R}}_{21} \hat{\mathbf{R}}_{11}^{-1} \hat{\mathbf{R}}_{12}$$

- 求解 $\hat{\mathbf{A}}^*$ 和 $\hat{\mathbf{B}}^*$ 的特征根及相应的特征向量，即可得到典型变量及典型相关系数。此时相当于从相关矩阵出发计算典型变量。



二、典型相关系数的显著性检验

- 在利用样本进行两组变量的典型相关分析时，应就两组变量的相关性进行检验。
- 这是因为，如果两个随机向量 $X^{(1)}$ 、 $X^{(2)}$ 互不相关， $Cov(X^{(1)}, X^{(2)}) = 0$ ，那么即使其估计 $\hat{\Sigma}_{12} \neq 0$ ，进行典型相关分析也是无意义的。
- 因此，在用样本数据进行典型相关分析时应就两组变量的协差阵是否为零进行检验。



□ 即检验假设 $H_0 : \Sigma_{12} = 0$, $H_1 : \Sigma_{12} \neq 0$

□ 根据随机向量的检验理论，构造似然比统计量（用于进行随机向量的独立性检验）为：

$$\Lambda_0 = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_{11}| |\hat{\Sigma}_{22}|} = \prod_{i=1}^r (1 - \hat{\lambda}_i^2)$$

📖 在上述的式子中 $\hat{\lambda}_i^2$ 是矩阵 A 的第 i 特征根的估计值

, $r = \min(p, q) = p$ 。巴特莱特 (Bartlett) 证明，当

H_0 成立时, $Q_0 = -m \ln \Lambda_0$ 近似服从 $\chi^2(f)$ 分布，其中

$$m = (n-1) - \frac{1}{2}(p+q+1), \text{ 自由度 } f = pq。$$



$$Q_0 = -m \ln \Lambda_0 \approx \chi^2(f)$$

 在给定的显著性水平 α 下，当由样本计算的临界值 $Q_0 > \chi_\alpha^2$ 时，拒绝原假设，认为两组变量间存在相关性。



□ 在进行典型相关分析时，对于两随机向量 $X^{(1)}$ 、 $X^{(2)}$ ，可以提取出 p 对典型变量，问题是进行典型相关分析的目的就是要减少分析变量，简化两组变量间关系分析，提取 p 对变量是否必要？如何确定保留多少对典型变量？



- 巴特莱特 (Bartlett) 提出了一个根据样本数据检验总体典型相关系数 $\lambda_1, \lambda_2, \dots, \lambda_r$ 是否等于零的方法。检验假设为

$$H_0: \lambda_{k+1} = \lambda_{k+2} = \dots = \lambda_r = 0 \quad H_1: \lambda_{k+1} \neq 0$$

- 用于检验的似然比统计量为:

$$\Lambda_k = \prod_{i=k+1}^r (1 - \hat{\lambda}_i^2)$$

- 可以证明, $Q_k = -m_k \ln \Lambda_k$ 近似服从 $\chi^2(f_k)$ 分布, 其中自由度 $f_k = (p-k)(q-k)$

$$m_k = (n-k-1) - \frac{1}{2}(p+q+1)$$



- 首先检验 $H_0: \lambda_1 = \lambda_2 = \dots = \lambda_r = 0$ 。此时 $k=0$ ，则似然比统计量式为

$$\Lambda_0 = \prod_{i=1}^r (1 - \hat{\lambda}_i^2) = (1 - \hat{\lambda}_1)(1 - \hat{\lambda}_2) \cdots (1 - \hat{\lambda}_r)$$

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

- 若 $Q_0 > \chi_\alpha^2(f_0)$ ，则拒绝原假设，也就是说至少有一个典型相关系数大于零，自然应是最大的典型相关系数 $\lambda_1 > 0$ 。



- 若已判定 $\lambda_1 > 0$ ，则再检验 $H_0: \lambda_2 = \lambda_3 = \cdots = \lambda_r = 0$
此时 $k=1$ ，则似然比统计量为：

$$\Lambda_1 = \prod_{i=2}^r (1 - \hat{\lambda}_i^2) = (1 - \hat{\lambda}_2)(1 - \hat{\lambda}_3) \cdots (1 - \hat{\lambda}_r)$$

$$Q_1 = -m_1 \ln \Lambda_1 = -[(n-1-1) - \frac{1}{2}(p+q+1)] \ln \Lambda_1$$

- Q_1 近似服从 $\chi^2(f_1)$ 分布，其中 $f_1 = (p-1)(q-1)$ ，
如果 $Q_1 > \chi_{\alpha}^2(f_1)$ ，则拒绝原假设，也即认为
 $\lambda_2, \lambda_3, \cdots, \lambda_r$ 至少有一个大于零，自然是 $\lambda_2 > 0$



□ 若已判断 λ_1 和 λ_2 大于零，重复以上步骤直至

$H_0 : \lambda_j = \lambda_{j+1} = \dots = \lambda_r = 0$ ，此时令

$$Q_{j-1} = -m_{j-1} \ln \Lambda_{j-1} = -[(n-j) - \frac{1}{2}(p+q+1)] \ln \Lambda_{j-1}$$

$$Q_{j-1} \square \chi^2(f_{j-1}) \quad f_{j-1} = (p-j+1)(q-j+1)$$

$$\Lambda_{j-1} = \prod_{i=j}^r (1 - \hat{\lambda}_i^2) = (1 - \hat{\lambda}_j)(1 - \hat{\lambda}_{j+1}) \dots (1 - \hat{\lambda}_r)$$

□ 如果 $Q_{j-1} < \chi^2_\alpha(f_{j-1})$ ，则 $\lambda_j = \lambda_{j+1} = \dots = \lambda_r = 0$ ，

于是总体只有 $j-1$ 个典型相关系数不为零，提取 $j-1$ 对典型变量进行分析。



- 【例9.1】 康复俱乐部对20名中年人测量了三个生理指标：体重(x_1),腰围(x_2), 脉搏(x_3); 三个训练指标：引体向上次数(y_1), 起坐次数(y_2), 跳跃次数(y_3)。分析生理指标与训练指标的相关性。数据详见表9.1。

变量 样本	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



变量 样本	X1	X2	X3	Y1	Y2	Y3
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



□ 根据表9.1数据可得：

$$\hat{\Sigma}_{11} = \begin{bmatrix} 579.14 & 65.36 & -61.86 \\ 65.36 & 9.74 & -7.74 \\ -61.86 & -7.74 & 49.39 \end{bmatrix}$$

$$\hat{\Sigma}_{12} = \begin{bmatrix} -48.32 & -723.63 & -272.18 \\ -8.88 & -122.87 & -29.87 \\ 5.46 & 96.45 & 12.27 \end{bmatrix}$$

$$\hat{\Sigma}_{22} = \begin{bmatrix} 26.55 & 218.60 & 127.67 \\ 218.60 & 3718.85 & 2039.64 \\ 127.67 & 2039.64 & 2497.91 \end{bmatrix}$$

$$\hat{\Sigma}_{21} = \begin{bmatrix} -48.32 & -8.88 & 5.46 \\ -723.63 & -122.87 & 96.45 \\ -272.18 & -29.87 & 12.27 \end{bmatrix}$$



$$\hat{\Sigma}_{11}^{-1} = \begin{bmatrix} 0.00723237 & -0.047214 & 0.00165941 \\ -0.047214 & 0.42549329 & 0.00754531 \\ 0.00165941 & 0.00754531 & 0.02350784 \end{bmatrix}$$

$$\hat{\Sigma}_{22}^{-1} = \begin{bmatrix} 0.0732399 & -0.0040789 & -0.00041 \\ -0.0040789 & 0.00071416 & -0.00037 \\ -0.0004126 & -0.0003747 & 0.000727 \end{bmatrix}$$



计算得

$$\hat{\mathbf{A}} = \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{21} = \begin{bmatrix} -0.2459454 & -0.0551887 & 0.04651367 \\ 4.498811 & 0.90714323 & -0.7392212 \\ -0.0575041 & -0.0138964 & 0.01728371 \end{bmatrix}$$
$$\hat{\mathbf{B}} = \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} = \begin{bmatrix} 0.16178831 & 2.03428439 & 0.223085 \\ 0.04076171 & 0.54877371 & 0.091339 \\ -0.0328274 & -0.4227509 & -0.03208 \end{bmatrix}$$

求得特征值为: $\lambda_1^2 = 0.632994993$, $\lambda_2^2 = 0.040214862$,

$\lambda_3^2 = 0.005267145$ 。典型相关系数分别为: $\lambda_1 = 0.796$,

$\lambda_2 = 0.201$, $\lambda_3 = 0.073$ 。



□ $\hat{A}$ 和 $\hat{B}$ 相应的特征向量分别为：

$$\mathbf{a}^{(1)} = (0.031, -0.493, 0.008)' \quad \mathbf{b}^{(1)} = (0.066, 0.017, -0.014)'$$

$$\mathbf{a}^{(2)} = (-0.076, 0.369, -0.032)' \quad \mathbf{b}^{(2)} = (-0.071, 0.002, 0.021)'$$

$$\mathbf{a}^{(3)} = (-0.008, 0.158, 0.146)' \quad \mathbf{b}^{(3)} = (-0.245, 0.020, -0.008)'$$



□ 根据前述的典型相关系数显著性检验方法，对于
 $H_0: \lambda_1 = \lambda_2 = \lambda_3 = 0$, H_1 : 至少有一个不为零。

$$\Lambda_0 = \prod_{i=1}^3 (1 - \hat{\lambda}_i^2)$$

$$= (1 - 0.632994993) \times (1 - 0.040214862) \\ \times (1 - 0.005267145) = 0.350390621$$

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

$$= -[(20-1) - \frac{1}{2}(3+3+1)] \ln \Lambda_0$$

$$= -15.5 \ln \Lambda_0 = 16.255$$

$Q_0 < \chi_{0.05}^2(9) = 16.91896016$ ，故在 $\alpha=0.05$ 下，生理指标与训练指标之间不存在相关性；

而在 $\alpha=0.10$ 下， $Q_0 > \chi_{0.10}^2(9) = 14.68366$ ，生理指标与训练指标之间存在相关性，且第一对典型变量相关性显著。



$$\Lambda_1 = \prod_{i=2}^3 (1 - \hat{\lambda}_i^2) = (1 - 0.040214862)(1 - 0.005267145) = 0.954729811$$

$$Q_1 = -m \ln \Lambda_1 = -[(n-1) - \frac{1}{2}(p+q+1)] \ln \Lambda_1$$

$$= -[(20-1) - \frac{1}{2}(3+3+1)] \ln \Lambda_0$$

$$= -15.5 \ln \Lambda_1 = 0.718$$

- $Q_1 < \chi_{0.10}^2(4) = 7.779434$, 故在 $\alpha = 0.10$ 下, 第二对典型变量间相关性不显著。
- 说明生理指标和训练指标之间只有一对典型变量, 即:

$$U_1 = 0.031X_1 - 0.493X_2 + 0.008X_3$$

$$V_1 = 0.066Y_1 + 0.017Y_2 - 0.014Y_3$$



从相关矩阵出发计算典型相关

- ❑ 为了消除量纲和数量级别的影响，必须对数据先做标准化变换处理，然后再做典型相关分析。
- ❑ 显然，经标准化变换之后的协差阵就是相关系数矩阵，也即通常应从相关矩阵出发进行典型相关分析。



□ 【例9.2】 对于例9.1从相关系数矩阵出发进行典型相关分析。

$$\hat{\mathbf{R}}_{11} = \begin{bmatrix} 1 & & \\ 0.87024349 & 1 & \\ -0.365762 & -0.3528921 & 1 \end{bmatrix}$$

$$\hat{\mathbf{R}}_{22} = \begin{bmatrix} 1 & & \\ 0.69572742 & 1 & \\ 0.49576018 & 0.66920608 & 1 \end{bmatrix}$$

$$\hat{\mathbf{R}}_{12} = \begin{bmatrix} -0.3896937 & -0.4930836 & -0.2262956 \\ -0.5522321 & -0.645598 & -0.1914994 \\ 0.15064802 & 0.22503808 & 0.03493306 \end{bmatrix} = \hat{\mathbf{R}}'_{21}$$



$$\hat{\mathbf{A}}_z = \hat{\mathbf{R}}_{11}^{-1} \hat{\mathbf{R}}_{12} \hat{\mathbf{R}}_{22}^{-1} \hat{\mathbf{R}}_{21} = \begin{bmatrix} -0.2459455 & -0.4255619 & 0.15927699 \\ 0.58342554 & 0.9071432 & -0.3282724 \\ -0.0167929 & -0.0312927 & 0.01728372 \end{bmatrix}$$

$$\hat{\mathbf{B}}_z = \hat{\mathbf{R}}_{22}^{-1} \hat{\mathbf{R}}_{21} \hat{\mathbf{R}}_{11}^{-1} \hat{\mathbf{R}}_{12} = \begin{bmatrix} 0.16178827 & 0.1718776 & 0.022998202 \\ 0.48244159 & 0.5487737 & 0.111448282 \\ -0.3184294 & -0.3464725 & -0.032080511 \end{bmatrix}$$

计算得 $\hat{\mathbf{A}}_z, \hat{\mathbf{B}}_z$ 的特征值为: $\lambda_1^2 = \mathbf{0.632994993}$,

$\lambda_2^2 = \mathbf{0.040214862}$, $\lambda_3^2 = \mathbf{0.005267145}$ 。其结果同从协差阵出发计算的特征值相同, 因此检验结果也相同, 提取第一典型变量, 按照类似的方法可求得典型变量系数向量:



$$\mathbf{a}^{(1)*} = (0.775, -1.579, 0.059)'$$

$$\mathbf{a}^{(2)*} = (-1.884, 1.181, -0.231)'$$

$$\mathbf{a}^{(3)*} = (-0.191, 0.506, 1.051)'$$

$$\mathbf{b}^{(1)*} = (0.349, 1.054, -0.716)'$$

$$\mathbf{b}^{(2)*} = (-0.376, 0.123, 1.062)'$$

$$\mathbf{b}^{(3)*} = (-1.297, 1.237, -0.419)'$$

可得到标准化的第一对典型变量：

$$U_1^* = 0.7751Z_1^{(1)} - 1.579Z_2^{(1)} + 0.059Z_3^{(1)}$$

$$V_1^* = 0.349Z_1^{(2)} + 1.054Z_2^{(2)} - 0.716Z_3^{(2)}$$

其中， $Z_i^{(1)}$ 和 $Z_j^{(2)}$ 分别为原始变量 X_i 和 Y_j 标准化后的结果。



实例1

- 股份公司在发行股票(或债券)时往往作连带发行认股证书(stock warrant)、且有时赠予本公司职工购股证书(stock option)。
- 某月末随机抽取了100家股份公司，调查了与认股证书和职工购股证书有关的5个指标，它们分别是上月普通股的最低价格 X_1 ，职工购股证书价格 X_2 ，认股证书价格 X_3 ，公司季度盈利率($\%$) X_4 和公司季度红利分配额 X_5 。这5个指标的样本均值和样本协差矩阵列在了下表中。



公司股票有关指标的样本均值和协差阵

指 标	样本均值 $\bar{x}$	样本协差阵 S					
x_1	20.98	244.67	103.67	144.89	5.37	1.19	
x_2	25.45	103.67	442.83	49.48	3.69	0.62	
x_3	13.02	144.89	49.48	186.63	1.59	1.03	
x_4	0.02	5.37	3.69	1.59	4.62	0.24	
x_5	0.13	1.19	0.62	1.03	0.24	0.04	

这 5 个变量可分为两组, x_1, x_2 和 x_3 属于价格指数组, x_4 和 x_5 属于股息指标组。现对这两组指标进行典型相关分析。



记随机向量 $x = (x_1, x_2, x_3)'$, $y = (y_1, y_2)' = (x_4, x_5)'$, 则由表中给出的样本协差矩阵资料可知:

$$S_{xx} = \begin{bmatrix} 244.67 & 103.67 & 144.89 \\ 103.67 & 442.83 & 49.48 \\ 144.89 & 49.48 & 186.63 \end{bmatrix}, \quad S_{yy} = \begin{bmatrix} 4.62 & 0.24 \\ 0.24 & 0.04 \end{bmatrix}$$

$$S_{xy} = \begin{bmatrix} 5.37 & 1.19 \\ 3.69 & 0.62 \\ 1.59 & 1.03 \end{bmatrix}, \quad S_{yx} = \begin{bmatrix} 5.37 & 3.69 & 1.59 \\ 1.19 & 0.62 & 1.03 \end{bmatrix}$$



S_{xx} 与 S_{yy} 的逆矩阵分别为：

$$S_{xx}^{-1} = \begin{bmatrix} 0.008188289782 & -0.001241959618 & 0.00601993379 \\ -0.001241959618 & 0.002515746397 & 0.000297210508 \\ -0.00601993379 & 0.000297210508 & 0.009952967002 \end{bmatrix}$$

$$S_{yy}^{-1} = \begin{bmatrix} 0.314465408 & -1.886792453 \\ -1.886792453 & 36.32075472 \end{bmatrix}$$

于是，可计算得到：

$$S_{xx}^{-1}S_{xy} = \begin{bmatrix} 0.02976289 & 0.002761618074 \\ 0.00308635 & 0.000387957644 \\ -0.01540512 & 0.003272105317 \end{bmatrix}$$

$$S_{yy}^{-1}S_{yx} = \begin{bmatrix} -0.55660377 & -0.009433965 & -1.44339622 \\ 33.08962264 & 15.55660377 & 34.41037736 \end{bmatrix}$$



由此得矩阵 $\hat{M}$ 和 $\hat{N}$ 分别为：

$$\hat{M} = S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} = \begin{bmatrix} 0.047814763 & 0.042680616 & 0.052068677 \\ 0.01111195 & 0.006006187 & 0.008894949 \\ 0.116847278 & 0.051048177 & 0.13483007 \end{bmatrix}$$

$$\hat{N} = S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} = \begin{bmatrix} 0.005640439 & -0.006263731 \\ 0.502759864 & 0.210010582 \end{bmatrix}$$

因为 $\hat{M}$ 与 $\hat{N}$ 有相同的特征根，故为了简使用 $\hat{N}$ 求特征根。 $\hat{N}$ 的特征多项式为：

$$\begin{aligned} |\hat{\lambda}^2 I - \hat{N}| &= (\lambda^2 - 0.005640438718)(\lambda^2 - 0.210010582) \\ &\quad + 0.003149152775 \\ &= \lambda^4 - 0.21565102\lambda^2 + 0.004333704593 \end{aligned}$$



$$\hat{\lambda}_1^2 = 0.193222442, \quad \hat{\lambda}_2^2 = 0.022428577$$

用巴特莱特检验,对于假设:

$$H_0: \lambda_1 = \lambda_2 = 0, \quad H_1: \text{至少 } \lambda_i \neq 0$$

其似然比统计量的值为:

$$\begin{aligned} \Lambda_0 &= \prod_{i=2}^p (1 - \hat{\lambda}_i^2) \\ &= (1 - 0.193222442)(1 - 0.022428577) \\ &= 0.788682706 \end{aligned}$$

相应的 χ^2 统计量为:

$$\begin{aligned} -w \ln \Lambda_0 &= -[(n-1) - \frac{1}{2}(p+q+1)] \ln \Lambda_0 \\ &= -96 \ln \Lambda_0 = 22.7895537 \end{aligned}$$

对于显著性水平 0.05, 按自由度 $pq = 3 \times 2 = 6$, 查 χ^2 分布表得 $\chi_{0.05}^2(6) = 12.59$, 由于 $-w \ln \Lambda_0 > \chi_{0.05}^2(6)$, 故拒绝假设 H_0 , 这说明至少第一对典型变量显著相关。



对于假设：

$$H_0: \lambda_2 = 0, H_1: \lambda_2 \neq 0$$

其似然比统计量的值为：

$$\Lambda_1 = \prod_{i=1}^p (1 - \hat{\lambda}_i^2) = 1 - 0.022428577 = 0.977571423$$

相应的 χ^2 统计量为：

$$\begin{aligned} -w \ln \Lambda_1 &= -[(n-2) - \frac{1}{2}(p+q+1)] \ln \Lambda_1 \\ &= -95 \ln \Lambda_1 = 2.15497266 \end{aligned}$$

对于显著性水平 0.05, 按自由度 $(p-1)(q-1) = (3-1)(2-1) = 2$, 查 χ^2 分布表得 $\chi_{0.05}^2(2) = 5.99$, 由于 $-w \ln \Lambda_1 < \chi_{0.05}^2(2)$, 故接受原假设 H_0 , 说明第二对典型变量不相关。

这说明价格指标和股息指标两组间只有第一对典型变量相关, 即只用第一对典型变量就足以说明两组指标间的相关关系。现计算此对典型变量。



价格指标与股息指标的第一对典型变量为：

$$\begin{cases} z_1 = 0.02269172x_1 + 0.003657044396x_2 + 0.048604836x_3 \\ w_1 = -0.195668933y_1 + 5.859761848y_2 \end{cases}$$

两个典型变量之间的典型相关系数为 $\rho_{z_1w_1} = \hat{\lambda}_1 = 0.439570747$ 。



疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

