典型相关分析的应用

摘 要

本文主要研究典型相关分析的应用场景和应用步骤。

首先，我们简单介绍了典型相关分析的历史研究及应用场景。其次，我们就典型相关分析的原理进行了比较详细的推导，指出典型相关分析的问题可以向求矩阵的**特征值**和**特征向量**进行转化，并提出了**显著性检验**的必要性。紧接着，我们引入了**典型载荷分析**和**典型冗余分析**的主要思想。最后，我们以身体指标变量及训练指标变量间的关系的问题研究为例，具体地呈现了典型相关分析的步骤。

关键字： 典型相关分析　线性组合　典型载荷分析　整体相关性　典型冗余分析

目录

一、 应用场景

相关系数分析能够用于衡量两个随机变量之间的线性相关关系。然而在某些情况下，我们需要衡量两组随机变量之间的**整体相关程度**。比如，衡量身体指标变量（体重、胸围、脉搏）和训练指标变量（引体向上次数、起坐次数、跳跃次数）之间的整体相关程度。这种时候，计算两组变量两两变量之间的相关系数不失为一种方法，但它不仅繁琐，而且不能够从整体上更好地体现和说明两组数据的相关程度。

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 (Canonical Correlation Analysis, CCA) 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。

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。

二、 原理简述

典型相关分析由 Hotelling 提出，其基本思想和主成分分析非常相似。首先在每组变量中构造出变量的**线性组合**，使得两组变量的线性组合之间具有**最大的相关系数**。然后选取和最初挑选的这对线性组合**不相关**的线性组合，使其配对，并选取相关系数最大的一对，如此继续下去，直到两组变量之间的相关性被提取完毕为此。被选出的线性组合配对称为**典型变量**，它们的相关系数称为**典型相关系数**。典型相关系数度量了这两组变量之间联系的强度。

2.1 基本原理

设两组随机变量分别为

$$\mathbf{X}=\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p\end{bmatrix} \qquad \mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_q\end{bmatrix}$$

我们可以分别构造若干个它们的线性组合，形成所谓的**综合变量 $\mathbf{U_i}$、$\mathbf{V_i}$**

$$\mathbf{U_i}=a_1^{(i)}\mathbf{X}_1+a_2^{(i)}\mathbf{X}_2+\cdots+a_p^{(i)}\mathbf{X}_p=\mathbf{a^{(i)}X} \tag{1}$$

$$\mathbf{V_i}=b_1^{(i)}\mathbf{Y}_1+b_2^{(i)}\mathbf{Y}_2+\cdots+b_q^{(i)}\mathbf{Y}_q=\mathbf{b^{(i)}Y} \tag{2}$$

由于 $\rho(k\mathbf{U_i},m\mathbf{V_i})=\rho(\mathbf{U_i},\mathbf{V_i})$，为了确保典型变量的唯一性，我们添加一个限制条件，即只考虑方差为 1 的 $\mathbf{U_i},\mathbf{V_i}$ 对。

若存在 $\mathbf{a^{(1)}}$、$\mathbf{b^{(1)}}$, 在 $D(\mathbf{a^{(1)}X}) = D(\mathbf{b^{(1)}Y}) = 1$ 的前提下，能够使得 $\rho(\mathbf{a^{(1)}X}, \mathbf{b^{(1)}Y})$ 达到最大值，则称 $\mathbf{U_1} = \mathbf{a^{(1)}X}, \mathbf{V_1} = \mathbf{b^{(1)}Y}$ 是 $\mathbf{X}, \mathbf{Y}$ 的第一对**典型变量**。

求出第一对典型变量后，可以添加一个新的条件——与第一对典型变量不相关，即 $Cov(U_1, U_2) = Cov(V_1, V_2) = 0$，紧接着以相同的方法求出第二对典型变量、第三对典型变量等。

这些对典型变量 $\mathbf{U_i}$、$\mathbf{V_i}$ 能够分别代表 $\mathbf{X}, \mathbf{Y}$ 的综合。求出它们之间的相关系数，研究它们的相关性，就可以分析两组变量的整体相关性。

除此之外，我们还可以通过检验各对典型变量相关系数的显著性，来反映每一对综合变量的代表性，如果某一对的相关程度不显著，那么这对变量就不具有代表性，不具有代表性的变量就可以忽略。这样就可以通过对少数典型相关变量的研究，代替原来两组变量之间的相关关系的研究，从而容易抓住问题的本质。

2.2 数学推导

已知 $\mathbf{X}_{n\times p}, \mathbf{Y}_{n\times q}$. 设

$$Cov(\mathbf{X}) = \mathbf{\Sigma}_{11} \qquad Cov(\mathbf{Y}) = \mathbf{\Sigma}_{22} \qquad Cov(\mathbf{X}, \mathbf{Y}) = \mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}^T$$

又设 $\mathbf{M} = \begin{bmatrix} X^T \\ Y^T \end{bmatrix}$，则有

$$Cov(\mathbf{M}) = \begin{bmatrix} \underset{p\times p}{\mathbf{\Sigma}_{11}} & \underset{p\times q}{\mathbf{\Sigma}_{12}} \\ \underset{q\times p}{\mathbf{\Sigma}_{21}} & \underset{q\times q}{\mathbf{\Sigma}_{22}} \end{bmatrix} \tag{3}$$

为了表示出 $\rho(\mathbf{U}, \mathbf{V})$，需要计算 $Cov(\mathbf{U}, \mathbf{V}), D(\mathbf{U}), D(\mathbf{V})$

$$D(\mathbf{U}) = D(\mathbf{aX}) = \mathbf{a}^T Cov(\mathbf{X})\mathbf{a} = \mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a}$$

$$D(\mathbf{V}) = D(\mathbf{bY}) = \mathbf{b}^T Cov(\mathbf{Y})\mathbf{b} = \mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b}$$

$$Cov(\mathbf{U}, \mathbf{V}) = Cov(\mathbf{aX}, \mathbf{bY}) = \mathbf{a}^T Cov(\mathbf{X}, \mathbf{Y})\mathbf{b} = \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}$$

加上我们限制的条件：$D(\mathbf{U}) = \mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a} = D(\mathbf{V}) = \mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b} = 1$，可得

$$\rho(\mathbf{U}, \mathbf{V}) = \frac{Cov(\mathbf{U}, \mathbf{V})}{\sqrt{D(\mathbf{U})D(\mathbf{V})}} = \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b} \tag{4}$$

引入拉格朗日函数以求使得式 (4) 达到条件极值的 $\mathbf{a}, \mathbf{b}$

$$L(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b} - \frac{\lambda_1}{2}(\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a} - 1) - \frac{\lambda_2}{2}(\mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b} - 1) \tag{5}$$

将 $L(\mathbf{a}, \mathbf{b})$ 分别对 $\mathbf{a}, \mathbf{b}$ 求偏导，得到取条件极值的必要条件

$$\begin{cases} \frac{\partial L}{\partial \mathbf{a}} = \mathbf{\Sigma}_{12}\mathbf{b} - \lambda_1\mathbf{\Sigma}_{11}\mathbf{a} = 0 \\ \frac{\partial L}{\partial \mathbf{b}} = \mathbf{\Sigma}_{21}\mathbf{a} - \lambda_2\mathbf{\Sigma}_{22}\mathbf{b} = 0 \end{cases} \tag{6}$$

对式 (6) 分别左乘 $\mathbf{a}^T, \mathbf{b}^T$, 得

$$\begin{cases} \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b} - \mathbf{a}^T\lambda_1\mathbf{\Sigma}_{11}\mathbf{a} = 0 \\ \mathbf{b}^T\mathbf{\Sigma}_{21}\mathbf{a} - \mathbf{b}^T\lambda_2\mathbf{\Sigma}_{22}\mathbf{b} = 0 \end{cases} \tag{7}$$

即

$$\begin{cases} \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b} = \lambda_1 \\ \mathbf{b}^T\mathbf{\Sigma}_{21}\mathbf{a} = \lambda_2 \end{cases} \tag{8}$$

又 $(\mathbf{b}^T\mathbf{\Sigma}_{21}\mathbf{a})^T = \mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}$，所以可设 $\lambda = \lambda_1 = \lambda_2$

于是式 (6) 变为

$$\begin{cases} \mathbf{\Sigma}_{12}\mathbf{b} - \lambda\mathbf{\Sigma}_{11}\mathbf{a} = 0 \\ \mathbf{\Sigma}_{21}\mathbf{a} - \lambda\mathbf{\Sigma}_{22}\mathbf{b} = 0 \end{cases} \tag{9}$$

由式 (9) 中第二式得

$$\mathbf{b} = \frac{1}{\lambda}{\mathbf{\Sigma}_{22}}^{-1}\mathbf{\Sigma}_{21}\mathbf{a} \tag{10}$$

将式 (10) 代回式 (9) 中第一式得

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

同理得

$$\mathbf{\Sigma}_{21}{\mathbf{\Sigma}_{11}}^{-1}\mathbf{\Sigma}_{12}\mathbf{a} - \lambda^2\mathbf{\Sigma}_{22}\mathbf{b} = 0 \tag{12}$$

式 (11) 左乘 ${\mathbf{\Sigma}_{11}}^{-1}$，式 (12) 左乘 ${\mathbf{\Sigma}_{22}}^{-1}$ 得

$$\begin{cases} ({\mathbf{\Sigma}_{11}}^{-1}\mathbf{\Sigma}_{12}{\mathbf{\Sigma}_{22}}^{-1}\mathbf{\Sigma}_{21} - \lambda^2\mathbf{E})\mathbf{a} = 0 \\ ({\mathbf{\Sigma}_{22}}^{-1}\mathbf{\Sigma}_{21}{\mathbf{\Sigma}_{11}}^{-1}\mathbf{\Sigma}_{12} - \lambda^2\mathbf{E})\mathbf{b} = 0 \end{cases} \tag{13}$$

可见，$\mathbf{A} = {\mathbf{\Sigma}_{11}}^{-1}\mathbf{\Sigma}_{12}{\mathbf{\Sigma}_{22}}^{-1}\mathbf{\Sigma}_{21}$ 与 $\mathbf{B} = {\mathbf{\Sigma}_{22}}^{-1}\mathbf{\Sigma}_{21}{\mathbf{\Sigma}_{11}}^{-1}\mathbf{\Sigma}_{12}$ 具有相同的特征根 λ^2，而 $\mathbf{a}, \mathbf{b}$ 则是该特征值对应的特征向量。可以证明，λ 即为 $\rho(\mathbf{U}, \mathbf{V})$。

于是，求解 r 对典型向量 $\mathbf{U_i}, \mathbf{V_i}$ 的问题就转化为求解 $\mathbf{A}, \mathbf{B}$ 的 r 个共同特征根 $\lambda_1^2, \lambda_2^2, \cdots, \lambda_r^2$，以及 r 组对应特征向量 $\mathbf{a}_i, \mathbf{b}_i$ 的问题

2.3 显著性检验

如果求得的特征根 $\lambda^2 = 0$，即典型向量对的相关系数 $\lambda = 0$，那么这对典型向量反映的信息将不起作用，我们可以忽略。对此，我们可以进行 k 次假设检验（k=1、2、⋯、r），即所谓**显著性检验**，以挑选出相关系数显著不等于 0 的有价值的典型向量对。

若相关系数 λ 从大到小排序依次为 $\lambda_1, \lambda_2, \cdots, \lambda_r$，则 k 次显著性检验的具体步骤如下：

1. **作出假设**。原假设 H_0 为 “$\lambda_{k+1}=\lambda_{k+2}=\cdots=\lambda_r=0$”，备选假设 H_1 为 “$\lambda_k\neq 0$”。
2. **计算似然比统计量** Λ_k^*。根据 $\Lambda_k^*=\prod_{i=k}^{r}(1-\lambda_i^2)$ 计算似然比统计量 Λ_k^*。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 $f_k=(p-k+1)(q-k+1), m_k=(n-k)-\frac{1}{2}(p+q+1)+\sum_{i=1}^{k-1}\lambda_i^{-2}$。
3. **计算 p 值**。根据概率密度函数或其他手段计算出 $\Lambda_k\geq\Lambda_k^*$ 的概率，即 p 值（此处为单侧检验）。
4. **选择显著性水平** α。$p<\alpha$ 代表在 $(1-\alpha)\times 100\%$ 的置信水平上拒绝原假设。常见的是 $\alpha=0.05$。
5. **得出结论**。若拒绝原假设，说明 λ_k 显著地不等于 0。若不能拒绝原假设，则可判断 $\lambda_k=\lambda_{k+1}=\cdots=\lambda_r=0$，不必再进行假设检验。

经过假设检验，我们可以提取出 p 对有价值的典型向量。

2.4 从相关系数矩阵出发

进行典型相关分析，为了去除量纲的影响，我们一般要对 $\mathbf{\Sigma}_{11},\mathbf{\Sigma}_{12},\mathbf{\Sigma}_{21},\mathbf{\Sigma}_{22}$ 进行标准化。由此，式 (3) 变为

$$Cov(\mathbf{M})=\mathbf{R}(\mathbf{M})=\begin{bmatrix}\underset{p\times p}{\boldsymbol{R}_{11}} & \underset{p\times q}{\boldsymbol{R}_{12}}\\ \underset{q\times p}{\boldsymbol{R}_{21}} & \underset{q\times q}{\boldsymbol{R}_{22}}\end{bmatrix} \tag{14}$$

其中，$\mathbf{R}$ 代表相应的相关系数矩阵。

2.5 典型载荷分析

进行典型载荷分析有助于更好解释分析已提取的 p 对典型变量。所谓的**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。即计算以下四个相关系数矩阵：

$$\mathbf{R}(\mathbf{U},\mathbf{X})\qquad\mathbf{R}(\mathbf{U},\mathbf{Y})\qquad\mathbf{R}(\mathbf{V},\mathbf{X})\qquad\mathbf{R}(\mathbf{V},\mathbf{Y}) \tag{15}$$

此处的 $\mathbf{U},\mathbf{V}$ 是由各组典型变量的转置组成的矩阵。

2.6 典型冗余分析

在进行样本典型相关分析时，我们也想了解每组变量提取出的典型变量所能解释的该组样本总方差的比例，从而定量测度典型变量所包含的原始信息量的大小。

这个比例可以通过特征值之比进行展现。第 i 对典型变量对 $\mathbf{X},\mathbf{Y}$ 的解释度为

$$\mathbf{R}d_{\mathbf{X}|\mathbf{U_i}}=\frac{\sum_{k=1}^{p}r^2(\mathbf{U}_i,\mathbf{X}_k)}{p}\qquad\mathbf{R}d_{\mathbf{Y}|\mathbf{V_i}}=\frac{\sum_{k=1}^{q}r^2(\mathbf{V}_i,\mathbf{Y}_k)}{q}$$

三、 实例展示

测量 20 名受试者的身体指标数据以及训练指标数据如表 1。第一组是身体指标变量，有体重、腰围、脉搏；第二组是训练指标变量，有引体向上次数、起坐次数和跳跃次数。试分析身体指标数据以及训练指标数据这两组变量之间的关系。

表 1　身体形态及健康情况指标表

体重 $\mathbf{X}_1$	腰围 $\mathbf{X}_2$	脉搏 $\mathbf{X}_3$	引体向上次数 $\mathbf{Y}_1$	起坐次数 $\mathbf{Y}_2$	跳跃次数 $\mathbf{Y}_3$
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

3.1 正态性检验

典型相关分析建立在变量服从联合正态分布的前提上。可以进行 Jarque-Bera 检验（$n \geq 30$）或 Shapiro-wilk 检验（$3 \leq n \leq 50$）。这里我们选择 Shapiro-wilk 检验。检验结果如表 2。

表 2　Shapiro-wilk 检验结果

变量	体重	胸围	脉搏	引体向上次数	起坐次数	跳跃次数
p 值	0.1255	0.0025	0.2250	0.0955	0.0959	0.0001

可见，除体重和脉搏外，其它变量都能以 90% 以上的置信水平近似服从正态分布。

3.2 典型相关分析

对 $\mathbf{X}=\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3\end{bmatrix}$ 及 $\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2 & \mathbf{Y}_3\end{bmatrix}$ 进行典型相关分析。

第一对典型变量为

$$\mathbf{U}_1 = -0.031\mathbf{X_1} + 0.493\mathbf{X_2} - 0.008\mathbf{X_3}$$

$$\mathbf{V}_1 = -0.066\mathbf{Y_1} - 0.017\mathbf{Y_2} + 0.014\mathbf{Y_3}$$

第一对标准化典型变量为

$$\mathbf{U}_1^* = 0.7754\mathbf{X}_1^* - 1.5793\mathbf{X}_2^* + 0.0591\mathbf{X}_3^*$$

$$\mathbf{V}_1^* = -0.3495\mathbf{Y}_1^* - 1.0540\mathbf{Y}_2^* + 0.7164\mathbf{Y}_3^*$$

它们的相关系数为 $\rho_1 = 0.7956$，p 值为 0.0617 ，我们能以 90% 的置信水平判断这对典型变量有较大价值。

第二对，第三对典型变量的相关系数分别为 $\rho_2 = 0.2006, \rho_3 = 0.0726$，p 值分别为 0.9457、0.6461。由此可见，这两对典型变量对我们的研究几乎毫无价值。

$\rho_1 = 0.7956$，这说明变量组 $\mathbf{X}, \mathbf{Y}$ 在 $\mathbf{U_1}, \mathbf{V_1}$ 的意义上具有较强的相关性。

3.3 典型载荷分析

$\mathbf{U_1}$ 与 $\mathbf{X}$ 各列的相关系数为 $\mathbf{R}(\mathbf{U_1}, \mathbf{X}) = \begin{bmatrix}0.6206 & 0.9254 & -0.3328\end{bmatrix}$

$\mathbf{U_1^*}$ 与 $\mathbf{X}^*$ 各列的相关系数为 $\mathbf{R}(\mathbf{U_1^*}, \mathbf{X}^*) = \begin{bmatrix}-0.6206 & -0.9254 & 0.3328\end{bmatrix}$

$\mathbf{V_1}$ 与 $\mathbf{Y}$ 各列的相关系数为 $\mathbf{R}(\mathbf{V_1}, \mathbf{Y}) = \begin{bmatrix}-0.7276 & -0.8177 & -0.1622\end{bmatrix}$

$\mathbf{V_1^*}$ 与 $\mathbf{Y}^*$ 各列的相关系数为 $\mathbf{R}(\mathbf{V_1^*}, \mathbf{Y}^*) = \begin{bmatrix}-0.7276 & -0.8177 & -0.1622\end{bmatrix}$

这说明第一对典型变量 $\mathbf{U}_1, \mathbf{V}_1$ 主要由体重 $\mathbf{X}_1$、胸围 $\mathbf{X}_2$ 和引体向上次数 $\mathbf{Y}_1$、起坐次数 $\mathbf{Y}_2$ 决定。

3.4 典型冗余分析

第一对典型变量 $\mathbf{U}_1, \mathbf{V}_1$ 对 $\mathbf{X}, \mathbf{Y}$ 的解释度分别为 0.4508、0.4081，解释性较强。

附录 D　Matlab 代码

```
clc,clear

%% 导入数据
data = readmatrix("data\健康指标.csv");
[m,n] = size(data);
X = data(:,1:3);
Y = data(:,4:6);
p = size(X,2);
q = size(Y,2);

%% 进行正态性检验
% Shapiro-wilk 检验
sw_H = ones([n,1]);
sw_P = ones([n,1]);
for i=1:n
    [sw_h,sw_p] = swtest(data(:,i),0.1);
    sw_H(i,1)=sw_h;
    sw_P(i,1)=sw_p;
end
disp("  是否通过     p")
disp([sw_H,sw_P]);

%% 未进行标准化的典型相关分析
[A,B,r,U,V,stats] = canoncorr(X,Y);
disp("r=")
disp(r);
disp("p=");
disp(stats.p);
disp("A=")
disp(A);
disp("B=")
disp(B);

%% 进行标准化的典型相关分析
R11 = corr(X,X);
R12 = corr(X,Y);
R21 = corr(Y,X);
R22 = corr(Y,Y);
M = R11\R12/(R22)*R21;
N = R22\R21/(R11)*R12;
[A_normalized,lambda1] = eig(M);
[B_normalized,lambda2] = eig(N);

lambda1 = sum(sqrt(lambda1));
```

```
[lambda1,I1] = sort(lambda1,"descend");
A_normalized = A_normalized(:,I1);
lambda2 = sum(sqrt(lambda2));
[lambda2,I2] = sort(lambda2,"descend");
B_normalized = B_normalized(:,I2);

for i=1:p
   d = A_normalized(:,i)' * R11 *A_normalized(:,i);
   A_normalized(:,i)=A_normalized(:,i)/(d^(1/2));
end

for i=1:q
   d = B_normalized(:,i)' * R22 *B_normalized(:,i);
   B_normalized(:,i)=B_normalized(:,i)/(d^(1/2));
end

disp("lambda_1=");
disp(lambda1);
disp("A_normalzed=");
disp(A_normalized);
disp("lambda_2=");
disp(lambda2);
disp("B_normalzed=");
disp(B_normalized);

%% 典型载荷分析
% 未标准化
RUX = corr(U,X);
RUY = corr(U,V);
RVX = corr(V,X);
RVY = corr(V,Y);
% 标准化
nRUX = A_normalized'*R11;
nRUY = A_normalized'*R12;
nRVX = B_normalized'*R21;
nRVY = B_normalized'*R22;

%% 典型冗余分析
RdU = sum(nRUX.^2,2)./p;
RdV = sum(nRVY.^2,2)./q;

clear d I1 I2 i
```

附录 E　Python 代码

```
## 导库
import numpy as np
import pandas as pd
from scipy import stats

np.set_printoptions(suppress=True)

## 数据初始化
data = pd.read_csv("../data/健康指标.csv", header=None)
data.columns = ["体重", "腰围", "脉搏", "引体向上次数", "起坐次数", "跳跃次数"]
display(data)

m, n = data.shape
D = np.array(data)
X = np.array(data[["体重", "腰围", "脉搏"]])
Y = np.array(data[["引体向上次数", "起坐次数", "跳跃次数"]])
p = X.shape[1]
q = Y.shape[1]

## 正态性检验
swtest_result = []
for i in range(n):
    W, sw_p = stats.shapiro(D[:, i])
    swtest_result.append([W, sw_p])
swtest_result = pd.DataFrame(swtest_result, columns=["W值", "p值"],
     index=data.columns)
display(swtest_result)

def cov(X, Y,normalized=False):
"""计算两个矩阵的协方差矩阵"""
    p = X.shape[1]
    q = Y.shape[1]
    result = np.ones((p, q))
    for i in range(p):
        for j in range(q):
            if normalized:
                result[i, j] = np.corrcoef(X[:, i], Y[:, j])[0, 1]
            else:
                result[i,j] = np.cov(X[:, i], Y[:, j])[0, 1]
    return result

def sortVbyL(L,V):
"""按特征值大小排序特征向量"""
    vectors = V.T.tolist()
    tmp = zip(L,vectors)
    tmp = sorted(tmp,key=lambda x:x[0],reverse=True)
    L,V = zip(*tmp)
```

```
    L = np.array(L)
    V = np.array(V).T
    return L,V

def canoncorr(X, Y,normalized=False):
"""典型相关分析"""
    # 计算协方差矩阵
    S11 = cov(X, X,normalized)
    S12 = cov(X, Y,normalized)
    S21 = cov(Y, X,normalized)
    S22 = cov(Y, Y,normalized)
    # 求特征值和特征向量并排序
    M = np.linalg.inv(S11) @ S12 @ np.linalg.inv(S22) @ S21
    N = np.linalg.inv(S22) @ S21 @ np.linalg.inv(S11) @ S12

    lambda_1, A = np.linalg.eig(M)
    lambda_1 = np.sqrt(lambda_1)
    lambda_1, A = sortVbyL(lambda_1,A)

    lambda_2, B = np.linalg.eig(N)
    lambda_2 = np.sqrt(lambda_2)
    lambda_2, B = sortVbyL(lambda_2,B)

    if np.linalg.norm(lambda_1 - lambda_2) > 0.0001:
        return None
    r = lambda_1

    # 限制方差为1
    for i in range(p):
        d1 = A[:, i].T @ S11 @ A[:, i]
        A[:, i] = A[:, i] * (d1 ** (-1 / 2))
    for j in range(q):
        d2 = B[:, j].T @ S22 @ B[:, j]
        B[:, j] = B[:, j] * (d2 ** (-1 / 2))

    # 计算典型向量
    U = X @ A
    V = Y @ B

    # 假设检验
    P = []
    for k in range(r.shape[0]):
        Lambda_k = np.prod([1 - x ** 2 for x in r[k:]])
        m_k = (m - k - 1) - 1 / 2 * (p + q + 1) + np.sum([x ** (-2) for x in
            r[:k]])
        f_k = (p - k) * (q - k)
        P.append(stats.chi2.sf(-m_k * np.log(Lambda_k), f_k))
```

```
    return A, B, r, U, V, P

## 未进行标准化的典型相关分析
A, B, r, U, V, P = canoncorr(X, Y)
display(r, P, A, B)

## 进行标准化的典型相关分析
nA, nB, nr, nU, nV, nP = canoncorr(X, Y,normalized=True)
display(nr, nP, nA, nB)

## 典型载荷分析
# 未标准化
RUX = cov(U,X,normalized=True)
RUY = cov(U,V,normalized=True)
RVX = cov(V,X,normalized=True)
RVY = cov(V,Y,normalized=True)
# 标准化
nRUX = nA.T @ cov(X,X,normalized=True)
nRUY = nA.T @ cov(X,Y,normalized=True)
nRVX = nB.T @ cov(Y,X,normalized=True)
nRVY = nB.T @ cov(Y,Y,normalized=True)
display(RUX)
display(RVY)
display(nRUX)
display(nRVY)

## 典型冗余分析
RdU = np.sum(nRUX**2,1)/p
RdV = np.sum(nRVY**2,1)/q
display(RdU)
display(RdV)
```