

- 1. 模型名称
- 2. 适用范围
- 3. 形式
- 4. 求解过程
  - 4.1 步骤
  - 4.2 实例
  - 4.3 代码实现
- 5. 参考资料

模型-机器学习-降维-典型相关分析

1. 模型名称

典型相关分析（Canonical Correlation Analysis, CCA）

2. 适用范围

多个随机变量和多个随机变量之间的相关系数

3. 形式

X	属性1	属性2	...
样本1	样本1-属性1得分	样本1-属性2得分	...
样本2	样本2-属性1得分	样本2-属性2得分	...
...	...	...	...

Y	属性1	属性2	...
样本1	样本1-属性1得分	样本1-属性2得分	...
样本2	样本2-属性1得分	样本2-属性2得分	...
...	...	...	...

4. 求解过程

4.1 步骤

1. 得到输入矩阵X和Y

输入两组随机变量 $x = [x_1, x_2, \dots, x_p]^T$ ， $y = [y_1, y_2, \dots, y_q]^T$ ，有n个样本，则

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$
$$Y = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q} \\ y_{21} & y_{22} & \dots & y_{2q} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{bmatrix}$$

## 2. 计算原始变量的相关系数阵 $R$

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

$R_{11}$ 和 $R_{22}$ 分别是第一组变量 $XX$ 和第二组变量 $YY$ 的相关系数阵

$R_{12} = R_{21}^T$ 是第一组与第二组变量 $XY$ 与 $YX$ 的相关系数阵

## 3. 求完整典型变量系数矩阵 $A'$ , $B'$ , 典型变量的相关系数矩阵 $r'$ 和典型变量值矩阵 $U'$ , $V'$

求 $M_1 = R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ 的特征值 $\lambda_i^2$ , 特征向量 $\rho^{(i)}$ ; 求 $M_2 = R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$ 的特征值 $\lambda_j^2$ , 特征向量 $\gamma^{(j)}$   
则典型变量为

$$u_1 = \rho^{(1)T}X, v_1 = \gamma^{(1)T}Y, \dots, u_s = \rho^{(s)T}X, v_s = \gamma^{(s)T}Y \quad s = \min(p, q)$$

记

$$A' = [\rho^{(1)}, \rho^{(2)}, \dots, \rho^{(s)}]_{p \times s} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1s} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2s} \\ \dots & \dots & \dots & \dots \\ \alpha_{p1} & \alpha_{p2} & \dots & \alpha_{ps} \end{bmatrix}$$

$$B' = [\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(s)}]_{q \times s} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1s} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2s} \\ \dots & \dots & \dots & \dots \\ \alpha_{q1} & \alpha_{q2} & \dots & \alpha_{qs} \end{bmatrix}$$

$$r' = R(U', V'), U' = [u_1, u_2, \dots, u_s]^T, V' = [v_1, v_2, \dots, v_s]^T$$

## 4. 典型相关系数的显著性检验

a) 假设 $H_0^{(k)}: \lambda_k = 0, k = 1, 2, \dots, p$

b) 给定显著性水平 $\alpha$

c) 计算显著性概率 $p_k = P(V > Q_k)$ <sup>1</sup>: 若 $p_k < \alpha$ , 则否定 $H_0^{(k)}$ , 即第 $k$ 对典型变量有效

## 5. 得到通过显著性检验的典型相关方程

## 6. 分析典型结构

a) 计算原始变量与本组典型变量之间的相关系数阵 $R(X, U)$ 和 $R(Y, V)$

b) 计算原始变量与对应组典型变量之间的相关系数阵 $R(X, V)$ 和 $R(Y, U)$

## 7. 进行典型冗余分析

a) 计算任何一个典型变量 $u_k$ (或 $v_k$ )解释本组变量 $X$ (或 $Y$ )总变差的百分比 $R_d(X; u_k)$ (或 $R_d(Y; v_k)$ )

b) 同时可求得前 $s$ 个典型变量 $u_1, u_2, \dots, u_s$ (或 $v_1, v_2, \dots, v_s$ )解释本组变量 $X$ (或 $Y$ )总变差的累计百分比 $R_d(X; u_1, u_2, \dots, u_s)$ (或 $R_d(Y; v_1, v_2, \dots, v_s)$ )

其中第一组典型变量的冗余测度: 第一组中典型变量解释的变差被第二组中典型变量重复解释的百分比, 即第二组典型变量 $v_d$ 解释第一组变量 $X$ 总变差的百分比 $R_d(X, v_k)$

其中第二组典型变量的冗余测度: 第二组中典型变量解释的变差被第一组中典型变量重复解释的百分比, 即第一组典型变量 $u_d$ 解释第二组变量 $Y$ 总变差的百分比 $R_d(Y, u_k)$

4.2 实例

城市竞争力指标体系主要包含4个指标：市场占有率、GDP增长率、劳动生产率和居民人均收入

城市基础设施指标体系主要包含6个指标：对外设施指数（由城市货运量和客运量指标综合构成），对内基本设施指数（由城市能源、交通、道路、住房等具体指标综合而成），每百人拥有电话机数，技术性设施指数（是城市现代交通、通信、信息设施的综合指数，由港口个数、机场等级、高速公路、高速铁路、地铁个数、光缆线路数等加权综合构成），文化设施指数（由公共藏书量、文化馆数量、影剧院数量等指标加权综合构成），卫生设施指数（由医院个数、万人医院床位数综合构成）

选取20个最具有代表性的城市，城市名称和竞争力、基础设施各项指标数据下表所列

试分析城市竞争力与城市基础设施之间相关联关系

	对外设施指数	对内设施指数	每百人电话数	技术设施指数	文化设施指数	卫生设施指数
上海	1.03	0.42	50	2.15	1.23	1.64
深圳	1.34	0.13	131	0.33	-0.27	-0.64
广州	1.07	0.4	48	1.31	0.49	0.09
北京	-0.43	0.19	20	0.87	3.57	1.8
厦门	-0.53	0.25	32	-0.09	-0.33	-0.84
天津	-0.11	0.07	27	0.68	-0.12	0.87
大连	0.35	0.06	31	0.28	-0.3	-0.16
杭州	-0.5	0.27	38	-0.78	-0.12	1.61
南京	0.31	0.25	43	0.49	-0.09	-0.06
珠海	-0.28	0.84	37	-0.79	-0.49	-0.98
青岛	0.01	-0.14	24	0.37	-0.4	-0.49
武汉	0.02	-0.47	28	0.03	0.15	0.26
温州	-0.47	0.03	45	-0.76	-0.46	-0.75
福州	-0.45	-0.2	34	-0.45	-0.34	-0.52
重庆	0.72	-0.83	13	0.05	-0.09	0.56
成都	0.37	-0.54	21	-0.11	-0.24	-0.02
宁波	0.01	0.38	40	-0.17	-0.4	-0.71
石家庄	-0.81	-0.49	22	-0.38	-0.21	-0.59
西安	-0.24	-0.91	18	-0.05	-0.27	0.61
哈尔滨	-0.53	-0.77	27	-0.45	-0.18	1.08

	劳动生产率	市场占有率	居民人均收入	长期经济增长率
上海	45623.05	2.5	8439	16.27
深圳	52256.67	1.3	18579	21.5
广州	46551.87	1.13	10445	11.92
北京	28146.76	1.38	7813	15
厦门	38670.43	0.12	8980	26.71
天津	26316.96	1.37	6609	11.07
大连	45330.53	0.56	6070	12.4
杭州	45853.89	0.28	7896	13.93
南京	35964.64	0.74	6497	8.97
珠海	55832.61	-0.12	13149	9.22
青岛	33334.62	0.63	6222	11.63
武汉	24633.27	0.59	5573	16.39
温州	39258.78	-0.69	9034	22.43
福州	38201.47	-0.34	7083	18.53
重庆	16524.32	0.44	5323	12.22
成都	31855.63	-0.02	6019	11.88
宁波	22528.8	-0.16	9069	15.7
石家庄	21831.94	-0.15	5497	13.56
西安	19966.36	-0.15	5344	12.43
哈尔滨	19225.71	-0.16	4233	10.16

1. 得到输入矩阵 $X$ 和 $Y$

对外设施指数	对内设施指数	每百人电话数	技术设施指数	文化设施指数	卫生设施指数
$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
劳动生产率	市场占有率	居民人均收入	长期经济增长率		
$y_1$	$y_2$	$y_3$	$y_4$		

$$X = \begin{bmatrix} 1.03 & 0.42 & 50 & 2.15 & 1.23 & 1.64 \\ 1.34 & 0.13 & 131 & 0.33 & -0.27 & -0.64 \\ 1.07 & 0.4 & 48 & 1.31 & 0.49 & 0.09 \\ -0.43 & 0.19 & 20 & 0.87 & 3.57 & 1.8 \\ -0.53 & 0.25 & 32 & -0.09 & -0.33 & -0.84 \\ -0.11 & 0.07 & 27 & 0.68 & -0.12 & 0.87 \\ 0.35 & 0.06 & 31 & 0.28 & -0.3 & -0.16 \\ -0.5 & 0.27 & 38 & -0.78 & -0.12 & 1.61 \\ 0.31 & 0.25 & 43 & 0.49 & -0.09 & -0.06 \\ -0.28 & 0.84 & 37 & -0.79 & -0.49 & -0.98 \\ 0.01 & -0.14 & 24 & 0.37 & -0.4 & -0.49 \\ 0.02 & -0.47 & 28 & 0.03 & 0.15 & 0.26 \\ -0.47 & 0.03 & 45 & -0.76 & -0.46 & -0.75 \\ -0.45 & -0.2 & 34 & -0.45 & -0.34 & -0.52 \\ 0.72 & -0.83 & 13 & 0.05 & -0.09 & 0.56 \\ 0.37 & -0.54 & 21 & -0.11 & -0.24 & -0.02 \\ 0.01 & 0.38 & 40 & -0.17 & -0.4 & -0.71 \\ -0.81 & -0.49 & 22 & -0.38 & -0.21 & -0.59 \\ -0.24 & -0.91 & 18 & -0.05 & -0.27 & 0.61 \\ -0.53 & -0.77 & 27 & -0.45 & -0.18 & 1.08 \end{bmatrix} \quad Y = \begin{bmatrix} 45623.05 & 2.5 & 8439 & 16.27 \\ 52256.67 & 1.3 & 18579 & 21.5 \\ 46551.87 & 1.13 & 10445 & 11.92 \\ 28146.76 & 1.38 & 7813 & 15 \\ 38670.43 & 0.12 & 8980 & 26.71 \\ 26316.96 & 1.37 & 6609 & 11.07 \\ 45330.53 & 0.56 & 6070 & 12.4 \\ 45853.89 & 0.28 & 7896 & 13.93 \\ 35964.64 & 0.74 & 6497 & 8.97 \\ 55832.61 & -0.12 & 13149 & 9.22 \\ 33334.62 & 0.63 & 6222 & 11.63 \\ 24633.27 & 0.59 & 5573 & 16.39 \\ 39258.78 & -0.69 & 9034 & 22.43 \\ 38201.47 & -0.34 & 7083 & 18.53 \\ 16524.32 & 0.44 & 5323 & 12.22 \\ 31855.63 & -0.02 & 6019 & 11.88 \\ 22528.8 & -0.16 & 9069 & 15.7 \\ 21831.94 & -0.15 & 5497 & 13.56 \\ 19966.36 & -0.15 & 5344 & 12.43 \\ 19225.71 & -0.16 & 4233 & 10.16 \end{bmatrix}$$

## 2. 计算相关系数阵R

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	1.0000	0.1600	0.5556	0.6502	0.0603	0.0499	0.3374	0.6350	0.4319	-0.0595
$x_2$	0.1600	1.0000	0.3899	0.2399	0.1716	-0.1531	0.7532	0.3353	0.6132	0.1407
$x_3$	0.5556	0.3899	1.0000	0.1622	-0.1032	-0.2386	0.5938	0.3161	0.8643	0.3989
$x_4$	0.6502	0.2399	0.1622	1.0000	0.5564	0.4261	0.1235	0.9066	0.0662	-0.0871
$x_5$	0.0603	0.1716	-0.1032	0.5564	1.0000	0.6368	-0.0561	0.5722	-0.0215	-0.0103
$x_6$	0.0499	-0.1531	-0.2386	0.4261	0.6368	1.0000	-0.2286	0.5265	-0.3490	-0.2956
$y_1$	0.3374	0.7532	0.5938	0.1235	-0.0561	-0.2286	1.0000	0.2509	0.7107	0.2099
$y_2$	0.6350	0.3353	0.3161	0.9066	0.5722	0.5265	0.2509	1.0000	0.2253	-0.0755
$y_3$	0.4319	0.6132	0.8643	0.0662	-0.0215	-0.3490	0.7107	0.2253	1.0000	0.3880
$y_4$	-0.0595	0.1407	0.3989	-0.0871	-0.0103	-0.2956	0.2099	-0.0755	0.3880	1.0000

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
$x_1$	1.0000	0.1600	0.5556	0.6502	0.0603	0.0499
$x_2$	0.1600	1.0000	0.3899	0.2399	0.1716	-0.1531
$x_3$	0.5556	0.3899	1.0000	0.1622	-0.1032	-0.2386
$x_4$	0.6502	0.2399	0.1622	1.0000	0.5564	0.4261
$x_5$	0.0603	0.1716	-0.1032	0.5564	1.0000	0.6368
$x_6$	0.0499	-0.1531	-0.2386	0.4261	0.6368	1.0000

	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	0.3374	0.6350	0.4319	-0.0595
$x_2$	0.7532	0.3353	0.6132	0.1407
$x_3$	0.5938	0.3161	0.8643	0.3989
$x_4$	0.1235	0.9066	0.0662	-0.0871
$x_5$	-0.0561	0.5722	-0.0215	-0.0103
$x_6$	-0.2286	0.5265	-0.3490	-0.2956

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
$y_1$	0.3374	0.7532	0.5938	0.1235	-0.0561	-0.2286
$y_2$	0.6350	0.3353	0.3161	0.9066	0.5722	0.5265
$y_3$	0.4319	0.6132	0.8643	0.0662	-0.0215	-0.3490
$y_4$	-0.0595	0.1407	0.3989	-0.0871	-0.0103	-0.2956

	$y_1$	$y_2$	$y_3$	$y_4$
$y_1$	1.0000	0.2509	0.7107	0.2099
$y_2$	0.2509	1.0000	0.2253	-0.0755
$y_3$	0.7107	0.2253	1.0000	0.3880
$y_4$	0.2099	-0.0755	0.3880	1.0000

3. 求完整典型变量系数矩阵  $A'$ ,  $B'$ , 典型变量的相关系数矩阵  $r'$  和典型变量值矩阵  $U'$ ,  $V'$   
完整典型变量系数矩阵  $A'$

	$u_1$	$u_2$	$u_3$	$u_4$
$x_1$	0.1535	-0.2134	0.6966	1.6312
$x_2$	0.3423	-0.2637	1.0577	-0.2266
$x_3$	0.4913	-0.3953	-0.9142	-0.4179
$x_4$	0.3372	0.8690	-0.4921	-1.3411
$x_5$	0.1149	-0.2429	-0.6687	0.5082
$x_6$	0.1419	0.3856	0.5884	0.2076

完整典型变量系数矩阵 $B'$

	$v_1$	$v_2$	$v_3$	$v_4$
$y_1$	0.1395	0.1322	-1.1342	0.8639
$y_2$	0.7185	-0.7361	0.2058	0.0628
$y_3$	0.4270	0.7720	0.6261	-1.0788
$y_4$	0.0285	0.0059	0.6004	0.9294

典型变量的相关系数矩阵 $r'$

序号	1	2	3	4
典型相关系数	0.9601	0.9499	0.6470	0.3571

典型变量值矩阵 $U'$



$u_1$	$u_2$	$u_3$	$u_4$
-2.1783	1.9159	-0.4551	0.5128
-2.2771	-2.0881	2.0484	-1.0352
-1.4059	0.4706	-0.6046	-0.3986
-0.7766	0.9680	1.3537	0.1197
0.3209	-0.4754	0.0403	1.6528
-0.2103	1.1600	-0.6937	1.2483
-0.0271	0.0648	-0.7630	-0.3088
0.0767	-0.3988	-1.7549	-0.2698
-0.5327	0.0171	-0.4642	0.3366
0.0985	-1.8180	-1.8628	0.1912
0.3664	0.3965	0.0158	0.6843
0.4878	0.2892	0.5611	-0.5350
0.5138	-1.3041	0.3394	0.4347
0.6996	-0.5710	0.5604	0.5308
0.8431	0.7031	-0.3599	-2.7349
0.7483	0.1338	-0.1371	-1.5973
-0.0438	-0.8847	-0.7590	0.2680
1.2013	-0.0734	1.3575	1.2241
1.1029	0.9541	0.8687	-0.2137
0.9925	0.5403	0.7090	-0.1100

典型变量值矩阵  $V'$

$v_1$	$v_2$	$v_3$	$v_4$
-2.1115	1.6866	-0.2197	1.1457
-2.4264	-1.9281	1.4324	-0.7581
-1.0817	-0.0977	-0.8527	-0.4268
-0.7720	0.9639	0.8843	-0.2773
0.0523	-0.6460	1.2668	2.3484
-0.5600	1.2636	0.3210	-0.8012
0.0306	0.4024	-1.6609	0.9733
0.0377	-0.3027	-1.2397	0.6934
-0.0605	0.5832	-1.0702	-0.5260
-0.3639	-2.0281	-1.9133	-1.2760
0.0927	0.5693	-0.5551	-0.1064
0.2886	0.7759	0.7640	0.4153
0.8186	-1.4325	0.4521	1.4530
0.7835	-0.6222	-0.2285	1.2671
0.5826	0.7882	0.9205	-0.9434
0.7403	0.0134	-0.5923	-0.1512
0.5604	-0.7386	1.3459	-1.0891
1.0380	0.1230	0.4572	-0.3910
1.0870	0.1815	0.4622	-0.7034
1.2638	0.4449	0.0258	-0.8463

4. 典型相关系数的显著性检验

序号	1	2	3	4
$pF$	0.0000	0.0005	0.3337	0.6065

由于1和2的 $pF < \alpha$ ，所以1和2通过了显著性检验，只选取1和2进行研究

5. 得到通过显著性检验的典型相关方程

典型变量系数矩阵 $A$

	$u_1$	$u_2$
$x_1$	0.1535	-0.2134
$x_2$	0.3423	-0.2637
$x_3$	0.4913	-0.3953
$x_4$	0.3372	0.8690
$x_5$	0.1149	-0.2429
$x_6$	0.1419	0.3856

典型变量系数矩阵  $B$

	$v_1$	$v_2$
$y_1$	0.1395	0.1322
$y_2$	0.7185	-0.7361
$y_3$	0.4270	0.7720
$y_4$	0.0285	0.0059

第一组典型相关方程： $u_1 = 0.1535x_1 + 0.3423x_2 + 0.4913x_3 + 0.3372x_4 + 0.1149x_5 + 0.1419x_6$   
 $v_1 = 0.1395y_1 + 0.7185y_2 + 0.4270y_3 + 0.0285y_4$   
 第二组典型相关方程： $u_2 = -0.2134x_1 - 0.2637x_2 - 0.3953x_3 + 0.8690x_4 - 0.2429x_5 + 0.3856x_6$   
 $v_2 = 0.1322y_1 - 0.7361y_2 + 0.7720y_3 + 0.0059y_4$

由第一组典型相关方程可知，基础设施方面的主要因素是  $x_2$ ， $x_3$ ， $x_4$ （典型系数分别为0.3423，0.4913，0.3372），说明基础设施中影响城市竞争力的主要因素是对内设施指数（ $x_2$ ）、每百人电话数（ $x_3$ ）和技术设施指数（ $x_4$ ）；城市竞争力的第一典型变量  $v_1$  与  $y_2$  呈高度相关，说明在城市竞争力中，市场占有率（ $y_2$ ）占有主要地位

由第二组典型相关方程可知， $x_4$ （技术设施指数）是基础设施方面的主要因素，而居民人均收入（ $y_3$ ）是反映城市竞争力的一个重要指标

由于第一组典型变量占有信息量比重较大，所以总体上基础设施方面的主要因素按重要程度依次是  $x_3$ ， $x_2$ ， $x_4$ ，反映城市竞争力的主要指标是  $y_2$ ， $y_3$

6. 分析典型结构

a) 计算原始变量与本组典型变量之间的相关系数阵  $R(X, U)$  和  $R(Y, V)$

$R(X, U)$

相关系数	$u_1$	$u_2$
$x_1$	0.7145	0.0945
$x_2$	0.6373	-0.3442
$x_3$	0.7190	-0.5426
$x_4$	0.7232	0.6320
$x_5$	0.4102	0.4688
$x_6$	0.1968	0.7252

$$R(Y, V)$$

相关系数	$v_1$	$v_2$
$y_1$	0.6292	0.4974
$y_2$	0.8475	-0.5295
$y_3$	0.6991	0.7024
$y_4$	0.1693	0.3887

b) 计算原始变量与对应组典型变量之间的相关系数阵 $R(X, V)$ 和 $R(Y, U)$

$$R(X, V)$$

相关系数	$v_1$	$v_2$
$x_1$	0.6860	-0.0897
$x_2$	0.6119	0.3270
$x_3$	0.6903	0.5154
$x_4$	0.6944	-0.6004
$x_5$	0.3938	-0.4453
$x_6$	0.1889	-0.6889

$$R(Y, U)$$

	$u_1$	$u_2$
$y_1$	0.6041	-0.4725
$y_2$	0.8137	0.5030
$y_3$	0.6712	-0.6672
$y_4$	0.1625	-0.3693

由此可知， $x_1, x_2, x_3, x_4$ 与“基础设施组”的第一典型变量 $u_1$ 呈高度相关，说明对外设施、对内设施、每百人电话数和技术设施在反映城市基础设施方面占有主导地位，其中又以技术设施居于首位。 $x_4$ 与基础设施组的第二典型变量和竞争力组的第一典型变量都呈高度相关。

“竞争力组”的第一典型变量 $v_1$ 与 $y_2$ 的相关系数均较高，体现了 $y_2$ 在反映城市竞争力中占有主导地位。 $y_3$ 与 $v_1$ 呈较高相关，与 $v_2$ 呈高相关，但 $v_2$ 凝聚的信息量有限，因而 $y_3$ 在“竞争力”中的贡献低于 $y_2$ 。由于第一对典型变量之间的高度相关，导致“基础设施组”中四个主要变量与“竞争力组”中的第一典型变量呈高度相关；而“竞争力组”中的 $y_2$ 则与“影响组”的第一典型变量也呈高度相关。这种一致性从数量上体现了“基础设施组”对“竞争力组”的本质影响作用，与指标的实际经济联系非常吻合，说明典型相关分析结果具有较高的可信度。

值得一提的是，与线性回归模型不同，相关系数与典型系数可以有不同的符号。如基础设施方面的 $u_2$ 与 $x_5$ 相关系数为正值（0.4688），而典型系数却为负值（-0.2429）。由于出现这种反号的情况，因此称 $x_5$ 为抑制变量（Suppressor）。由表还可以看出，“影响组”的第一典型变量 $u_1$ 对 $y_2$ （市场占有率）有相当高的预测能力，相关系数为0.8137，而对 $y_4$ （长期经济增长率）预测能力较差，相关系数仅为0.1625。

7. 进行典型冗余分析

a) 计算任何一个典型变量 $u_k$ (或 $v_k$ )解释本组变量 $X$ (或 $Y$ )总变差的百分比 $R_d(X; u_k)$ (或 $R_d(Y; v_k)$ )

b) 同时可求得前 $s$ 个典型变量 $u_1, u_2, \dots, u_s$ (或 $v_1, v_2, \dots, v_s$ )解释本组变量 $X$ (或 $Y$ )总变差的累计百分比 $R_d(X; u_1, u_2, \dots, u_s)$ (或 $R_d(Y; v_1, v_2, \dots, v_s)$ )

原始数据 $X$ 被典型变量解释的方差

X	比例	累计比例	X	比例	累计比例
$u_1$	0.3606	0.3636	$v_1$	0.3324	0.3324
$u_2$	0.2612	0.6218	$v_2$	0.2357	0.5681
$u_3$	0.0631	0.6849	$v_3$	0.0264	0.5945
$u_4$	0.0795	0.7644	$v_4$	0.0101	0.6046

原始数据 $Y$ 被典型变量解释的方差

Y	比例	累计比例	Y	比例	累计比例
$u_1$	0.4079	0.4079	$v_1$	0.3760	0.3760
$u_2$	0.2930	0.7009	$v_2$	0.2644	0.6404
$u_3$	0.1549	0.8558	$v_3$	0.0648	0.7053
$u_4$	0.1442	1.0000	$v_4$	0.0184	0.7237

典型相关系数平方（实际意义为一对典型变量之间的共享方差在两个典型变量各自方差中的比例）

1	2	3	4
0.9218	0.9024	0.4186	0.1275

由此可知，两对典型变量 $u_1$ 、 $u_2$ 和 $v_1$ 、 $v_2$ 均较好地预测了对应的那组变量，而且交互解释能力也较强。来自城市“竞争力组”的方差被“基础设施组”典型变量 $u_1$ 、 $u_2$ 解释的比例和为64.04%；来自“基础设施组”的方差被“竞争力组”典型变量 $v_1$ 、 $v_2$ 解释的方差比例和为56.81%。城市竞争力变量组被其自身及其对立典型变量解释的百分比、基础设施变量组被其自身及其对立典型变量解释的百分比均较高，尤其是第一对型变量具有较高的解释百分比，反映两者之间较高的相关性。

### 4.3 代码实现

#### [CCA.m](#)

代码：

```
clc, clear

% x, y行数为样本数量，列数为属性数
X = readmatrix("Book1.xlsx", "Sheet", "Sheet2");
Y = readmatrix("Book1.xlsx", "Sheet", "Sheet1");
% 设定显著性水平
a = 0.05;

% p为x的属性数，q为y的属性数，n为样本数
p = size(X,2);
q = size(Y,2);
n = size(X,1);
% 标准化矩阵x, y
X = zscore(X);
Y = zscore(Y);
% 得到相关系数矩阵R
C = [X Y];
R = corrcoef(C);
R11 = R((1:p), (1:p));
R12 = R((1:p), (p+1:end));
R21 = R12';
R22 = R((p+1:end), (p+1:end));
disp('相关系数矩阵:');
disp(R);

% A1和B1为典型变量系数，r1为典型相关系数
% U1和V1为典型变量的值，stats为假设检验的统计量的值
[A1,B1,r1,U1,V1,stats] = canoncorr(X,Y);
% 修正正负号，使每一列的系数和为正
A1 = A1 .* repmat(sign(sum(A1)),size(A1,1),1);
B1 = B1 .* repmat(sign(sum(B1)),size(B1,1),1);
U1 = U1 .* repmat(sign(sum(A1)),size(U1,1),1);
V1 = V1 .* repmat(sign(sum(B1)),size(V1,1),1);

% 输出统计数据
disp('CCA统计数据:');
disp(stats);
% 得到显著性检验后被选取的典型变量数量
count = 0;
for i=1:length(stats.pF)
```

```

        if(stats.pF(i) < a)
            count = count + 1;
        end
    end
end
% 选取经过显著性检验的典型变量
A = A1(:,1:count);
B = B1(:,1:count);
r = r1(:,1:count);
U = U1(:,1:count);
V = V1(:,1:count);

disp(' ');
disp('----- 分割线 -----');
disp(' ');
disp('完整数据');
disp(' ');
disp('X典型变量系数矩阵A': ' ');
disp(A1);
disp('Y典型变量系数矩阵B': ' ');
disp(B1);
disp('典型相关系数矩阵r': ' ');
disp(r1);
disp('X典型变量值矩阵U': ' ');
disp(U1);
disp('Y典型变量值矩阵V': ' ');
disp(V1);

% 计算R(X,U'), R(Y,V')
X_U_R1 = R11 * A1;
Y_V_R1 = R22 * B1;
disp('原始变量x与本组典型变量u之间的相关系数: ');
disp(X_U_R1);
disp('原始变量y与本组典型变量v之间的相关系数: ');
disp(Y_V_R1);
% 计算R(X,V'), R(Y,U')
X_V_R1 = R12 * B1;
Y_U_R1 = R21 * A1;
disp('原始变量x与对应组典型变量v之间的相关系数: ');
disp(X_V_R1);
disp('原始变量y与对应组典型变量u之间的相关系数: ');
disp(Y_U_R1);

% 计算方差比例与累计方差比例
%  $x \sim u$ 
ux = sum(X_U_R1.^2) / p;
disp('x组原始变量被u_i解释的方差比例: ');
disp(ux);
ux_cum = cumsum(ux);
disp('x组原始变量被u_i解释的方差累计比例: ');
disp(ux_cum);
%  $x \sim v$ 

```

```

vx = sum(X_V_R1.^2) / p;
disp('x组原始变量被v_i解释的方差比例: ');
disp(vx);
vx_cum = cumsum(vx);
disp('x组原始变量被v_i解释的方差累计比例: ');
disp(vx_cum);
% y~v
vy = sum(Y_V_R1.^2) / q;
disp('y组原始变量被v_i解释的方差比例: ');
disp(vy);
vy_cum = cumsum(vy);
disp('y组原始变量被v_i解释的方差累计比例: ');
disp(vy_cum);
% y~u
uy = sum(Y_U_R1.^2) / q;
disp('y组原始变量被u_i解释的方差比例: ');
disp(uy);
uy_cum = cumsum(uy);
disp('y组原始变量被u_i解释的方差累计比例: ');
disp(uy_cum);
disp(' ');
disp('----- 分割线 -----');
disp(' ');
disp('显著性检验后数据');
disp(' ');
disp('x典型变量系数矩阵A: ');
disp(A);
disp('y典型变量系数矩阵B: ');
disp(B);
disp('典型相关系数矩阵r: ');
disp(r);
disp('x典型变量值矩阵U: ');
disp(U);
disp('y典型变量值矩阵V: ');
disp(V);

% 计算R(X,U), R(Y,V)
X_U_R = R11 * A;
Y_V_R = R22 * B;
disp('原始变量x与本组典型变量u之间的相关系数: ');
disp(X_U_R);
disp('原始变量y与本组典型变量v之间的相关系数: ');
disp(Y_V_R);
% 计算R(X,V), R(Y,U)
X_V_R = R12 * B;
Y_U_R = R21 * A;
disp('原始变量x与对应组典型变量v之间的相关系数: ');
disp(X_V_R);
disp('原始变量y与对应组典型变量u之间的相关系数: ');
disp(Y_U_R);

```



```

% 计算方差比例与累计方差比例
%  $x \sim u$ 
ux = sum(X_U_R.^2) / p;
disp('x组原始变量被u_i解释的方差比例: ');
disp(ux);
ux_cum = cumsum(ux);
disp('x组原始变量被u_i解释的方差累计比例: ');
disp(ux_cum);
%  $x \sim v$ 
vx = sum(X_V_R.^2) / p;
disp('x组原始变量被v_i解释的方差比例: ');
disp(vx);
vx_cum = cumsum(vx);
disp('x组原始变量被v_i解释的方差累计比例: ');
disp(vx_cum);
%  $y \sim v$ 
vy = sum(Y_V_R.^2) / q;
disp('y组原始变量被v_i解释的方差比例: ');
disp(vy);
vy_cum = cumsum(vy);
disp('y组原始变量被v_i解释的方差累计比例: ');
disp(vy_cum);
%  $y \sim u$ 
uy = sum(Y_U_R.^2) / q;
disp('y组原始变量被u_i解释的方差比例: ');
disp(uy);
uy_cum = cumsum(uy);
disp('y组原始变量被u_i解释的方差累计比例: ');
disp(uy_cum);
disp('----- END -----');

```

结果：

相关系数矩阵:

Columns 1 through 6

1.0000	0.1600	0.5556	0.6502	0.0603	0.0499
0.1600	1.0000	0.3899	0.2399	0.1716	-0.1531
0.5556	0.3899	1.0000	0.1622	-0.1032	-0.2386
0.6502	0.2399	0.1622	1.0000	0.5564	0.4261
0.0603	0.1716	-0.1032	0.5564	1.0000	0.6368
0.0499	-0.1531	-0.2386	0.4261	0.6368	1.0000
0.3374	0.7532	0.5938	0.1235	-0.0561	-0.2286
0.6350	0.3353	0.3161	0.9066	0.5722	0.5265
0.4319	0.6132	0.8643	0.0662	-0.0215	-0.3490
-0.0595	0.1407	0.3989	-0.0871	-0.0103	-0.2956

Columns 7 through 10

0.3374	0.6350	0.4319	-0.0595
0.7532	0.3353	0.6132	0.1407
0.5938	0.3161	0.8643	0.3989
0.1235	0.9066	0.0662	-0.0871
-0.0561	0.5722	-0.0215	-0.0103
-0.2286	0.5265	-0.3490	-0.2956
1.0000	0.2509	0.7107	0.2099
0.2509	1.0000	0.2253	-0.0755
0.7107	0.2253	1.0000	0.3880
0.2099	-0.0755	0.3880	1.0000

CCA统计数据:

Wilks: [0.0039 0.0495 0.5072 0.8725]  
df1: [24 15 8 3]  
df2: [36.0959 30.7676 24 13]  
F: [5.8861 4.0417 1.2122 0.6335]  
pF: [1.1245e-06 5.0878e-04 0.3337 0.6065]  
chisq: [74.9775 40.8284 9.2942 2.0579]  
pChisq: [3.7608e-07 3.3963e-04 0.3181 0.5605]  
dfe: [24 15 8 3]  
p: [3.7608e-07 3.3963e-04 0.3181 0.5605]

## 完整数据

### X典型变量系数矩阵A'：

0.1535	-0.2134	0.6966	1.6312
0.3423	-0.2637	1.0577	-0.2266
0.4913	-0.3953	-0.9142	-0.4179
0.3372	0.8690	-0.4921	-1.3411
0.1149	-0.2429	-0.6687	0.5082
0.1419	0.3856	0.5884	0.2076

### Y典型变量系数矩阵B'：

0.1395	0.1322	-1.1342	0.8639
0.7185	-0.7361	0.2058	0.0628
0.4270	0.7720	0.6261	-1.0788
0.0285	0.0059	0.6004	0.9294

### 典型相关系数矩阵r'：

0.9601	0.9499	0.6470	0.3571
--------	--------	--------	--------

### X典型变量值矩阵U'：

-2.1783	1.9159	-0.4551	0.5128
-2.2771	-2.0881	2.0484	-1.0352
-1.4059	0.4706	-0.6046	-0.3986
-0.7766	0.9680	1.3537	0.1197
0.3209	-0.4754	0.0403	1.6528
-0.2103	1.1600	-0.6937	1.2483
-0.0271	0.0648	-0.7630	-0.3088
0.0767	-0.3988	-1.7549	-0.2698
-0.5327	0.0171	-0.4642	0.3366
0.0985	-1.8180	-1.8628	0.1912
0.3664	0.3965	0.0158	0.6843
0.4878	0.2892	0.5611	-0.5350
0.5138	-1.3041	0.3394	0.4347
0.6996	-0.5710	0.5604	0.5308
0.8431	0.7031	-0.3599	-2.7349
0.7483	0.1338	-0.1371	-1.5973
-0.0438	-0.8847	-0.7590	0.2680
1.2013	-0.0734	1.3575	1.2241
1.1029	0.9541	0.8687	-0.2137
0.9925	0.5403	0.7090	-0.1100

Y典型变量值矩阵V'：

-2.1115	1.6866	-0.2197	1.1457
-2.4264	-1.9281	1.4324	-0.7581
-1.0817	-0.0977	-0.8527	-0.4268
-0.7720	0.9639	0.8843	-0.2773
0.0523	-0.6460	1.2668	2.3484
-0.5600	1.2636	0.3210	-0.8012
0.0306	0.4024	-1.6609	0.9733
0.0377	-0.3027	-1.2397	0.6934
-0.0605	0.5832	-1.0702	-0.5260
-0.3639	-2.0281	-1.9133	-1.2760
0.0927	0.5693	-0.5551	-0.1064
0.2886	0.7759	0.7640	0.4153
0.8186	-1.4325	0.4521	1.4530
0.7835	-0.6222	-0.2285	1.2671
0.5826	0.7882	0.9205	-0.9434
0.7403	0.0134	-0.5923	-0.1512
0.5604	-0.7386	1.3459	-1.0891
1.0380	0.1230	0.4572	-0.3910
1.0870	0.1815	0.4622	-0.7034
1.2638	0.4449	0.0258	-0.8463

原始变量X与本组典型变量U之间的相关系数：

0.7145	0.0945	0.0269	0.5317
0.6373	-0.3442	0.4898	-0.3949
0.7190	-0.5426	-0.2660	0.0806
0.7232	0.6320	-0.0551	-0.0314
0.4102	0.4688	-0.2500	-0.0032
0.1968	0.7252	0.0438	0.1756

原始变量Y与本组典型变量V之间的相关系数：

0.6292	0.4974	-0.5116	0.3080
0.8475	-0.5295	0.0169	-0.0336
0.6991	0.7024	0.0993	-0.0900
0.1693	0.3887	0.5897	0.6874

原始变量X与对应组典型变量V之间的相关系数：

0.6860	-0.0897	-0.0174	-0.1899
0.6119	0.3270	-0.3169	0.1410
0.6903	0.5154	0.1721	-0.0288
0.6944	-0.6004	0.0356	0.0112
0.3938	-0.4453	0.1617	0.0012
0.1889	-0.6889	-0.0283	-0.0627

原始变量Y与对应组典型变量U之间的相关系数：

0.6041	-0.4725	0.3310	-0.1100
0.8137	0.5030	-0.0109	0.0120
0.6712	-0.6672	-0.0642	0.0322
0.1625	-0.3693	-0.3815	-0.2455

X组原始变量被u\_i解释的方差比例:  
0.3606      0.2612      0.0631      0.0795

X组原始变量被u\_i解释的方差累计比例:  
0.3606      0.6218      0.6849      0.7644

X组原始变量被v\_i解释的方差比例:  
0.3324      0.2357      0.0264      0.0101

X组原始变量被v\_i解释的方差累计比例:  
0.3324      0.5681      0.5945      0.6046

Y组原始变量被v\_i解释的方差比例:  
0.4079      0.2930      0.1549      0.1442

Y组原始变量被v\_i解释的方差累计比例:  
0.4079      0.7009      0.8558      1.0000

Y组原始变量被u\_i解释的方差比例:  
0.3760      0.2644      0.0648      0.0184

Y组原始变量被u\_i解释的方差累计比例:  
0.3760      0.6404      0.7053      0.7237

----- 分割线 -----

## 显著性检验后数据

### X典型变量系数矩阵A:

0.1535	-0.2134
0.3423	-0.2637
0.4913	-0.3953
0.3372	0.8690
0.1149	-0.2429
0.1419	0.3856

### Y典型变量系数矩阵B:

0.1395	0.1322
0.7185	-0.7361
0.4270	0.7720
0.0285	0.0059

### 典型相关系数矩阵r:

0.9601	0.9499
--------	--------

### X典型变量值矩阵U:

-2.1783	1.9159
-2.2771	-2.0881
-1.4059	0.4706
-0.7766	0.9680
0.3209	-0.4754
-0.2103	1.1600
-0.0271	0.0648
0.0767	-0.3988
-0.5327	0.0171
0.0985	-1.8180
0.3664	0.3965
0.4878	0.2892
0.5138	-1.3041
0.6996	-0.5710
0.8431	0.7031
0.7483	0.1338
-0.0438	-0.8847
1.2013	-0.0734
1.1029	0.9541
0.9925	0.5403

Y典型变量值矩阵V:

-2.1115	1.6866
-2.4264	-1.9281
-1.0817	-0.0977
-0.7720	0.9639
0.0523	-0.6460
-0.5600	1.2636
0.0306	0.4024
0.0377	-0.3027
-0.0605	0.5832
-0.3639	-2.0281
0.0927	0.5693
0.2886	0.7759
0.8186	-1.4325
0.7835	-0.6222
0.5826	0.7882
0.7403	0.0134
0.5604	-0.7386
1.0380	0.1230
1.0870	0.1815
1.2638	0.4449

原始变量X与本组典型变量U之间的相关系数:

0.7145	0.0945
0.6373	-0.3442
0.7190	-0.5426
0.7232	0.6320
0.4102	0.4688
0.1968	0.7252

原始变量Y与本组典型变量V之间的相关系数:

0.6292	0.4974
0.8475	-0.5295
0.6991	0.7024
0.1693	0.3887

原始变量X与对应组典型变量V之间的相关系数:

0.6860	-0.0897
0.6119	0.3270
0.6903	0.5154
0.6944	-0.6004
0.3938	-0.4453
0.1889	-0.6889

原始变量Y与对应组典型变量U之间的相关系数:

0.6041	-0.4725
0.8137	0.5030
0.6712	-0.6672
0.1625	-0.3693

X组原始变量被u\_i解释的方差比例:

0.3606      0.2612

X组原始变量被u\_i解释的方差累计比例:

0.3606      0.6218

X组原始变量被v\_i解释的方差比例:

0.3324      0.2357

X组原始变量被v\_i解释的方差累计比例:

0.3324      0.5681

Y组原始变量被v\_i解释的方差比例:

0.4079      0.2930

Y组原始变量被v\_i解释的方差累计比例:

0.4079      0.7009

Y组原始变量被u\_i解释的方差比例:

0.3760      0.2644

Y组原始变量被u\_i解释的方差累计比例:

0.3760      0.6404

----- END -----

## 5. 参考资料

1. [数模官网-CCA](#)
2. 《数学建模方法与应用》P313-P328

---

1. 其中,  $Q_k = -[n - k - \frac{1}{2}(p + q + 1)] \sum_{i=k}^p \ln(1 - \lambda_i^2)$ ,  $V \sim \chi^2(f_k)$ ,  $f_k = (p - k + 1)(q - k + 1)$  ↩