厦门大学
XIAMEN UNIVERSITY

对应分析与典型相关分析建模

谭　忠

Part 1

对应分析

对应分析与典型相关分析建模

对应分析引入

我们来看这样一个例子：

中美纯水有限公司欲为其新推出的一种纯水产品起一个合适的名字，为此专门委托了当地的策划咨询公司，取了一个名字“波澜”。一个好的名字至少应该满足两个条件：

（1）会使消费者联想到正确的产品“纯水”；

（2）会使消费者产生与正确产品密切相关的联想，如“纯净”、“清爽”等.

后来中美纯水有限公司委托调查统计研究所进行了一次全面的市场研究，调查中还包括简单的名 称测试：

问卷问题如下：

首先给出一些名词：玉泉、雪源、春溪、期望、波澜、天山绿、中美纯、雪浪花。

请您判断一下每个词最像什么商品名称?

（1）雪糕 （2）纯水 （3）碳酸饮料 （4）果汁饮料 （5）保健食品 （6）空调 （7）洗衣机 （8）毛毯

给出的每个名词最能让您产生什么感觉呢?

（1）清爽 （2）甘甜 （3）欢快 （4）纯净 （5）安闲 （6）个性 （7）兴奋 （8）高档

对这样的一个调查，最后得出的数据大致如下：

	雪糕	纯水	碳酸饮料	果汁饮料	保健食品	空调	洗衣机	毛毯
玉泉	67	454	98	34	122	101	102	22
雪源	322	207	167	34	90	65	51	64
春溪	68	421	97	67	121	90	80	56
期望	234	207	123	30	167	177	42	20
波澜	89	234	40	23	86	189	321	18
天上绿	123	67	109	89	178	190	178	66
中美纯	45	356	107	233	21	16	107	115
雪浪花	380	145	207	78	13	8	140	29
	清爽	甘甜	欢快	纯净	安闲	个性	兴奋	高档
玉泉	239	256	78	287	28	23	30	59
雪源	247	256	54	208	92	23	7	113
春溪	201	231	266	204	27	15	6	50
期望	56	23	189	66	101	208	243	114
波澜	110	81	165	109	45	78	329	83
天上绿	251	54	108	61	121	94	104	207
中美纯	69	165	177	198	34	62	137	158
雪浪花	127	25	243	67	18	165	288	67

对于这样的数据，我们应该用什么样的方法分析这些名词与商品类型和感觉之间的相互关系呢?

思考本次数据处理的目的，相比因子分析中我们去寻找样本（行变量）或者变量（列变量）之间的 关系，本例题中我们更希望找到样本（行变量）与变量（列变量）之间的相互关系，对应分析便是我们首选的方法。我们知道，在因子分析中，根据研究对象的不同，分为了R型和Q型，研究变量间的相互关系时采用R型因子分析，研究样本间相互关系时采用Q型因子分析，而无论是R 型还是Q型都不能很好地揭示变量与样品之间的相互关系，此时我们采用对应分析。

对应分析方法是在R型和Q型因子分析基础上发展起来的多元统计分析方法，又称R-Q型因子分析。我们可以看到，样品可以用 p 维空间上的 n 个点来表示，而变量可以用 n 维空间上的 p 个点来表示.而对应分析就是利用降维的思想，把样品和变量都表示在同一张二维图上，且通过后边的分析可以看到，这张二维图的坐标轴有相同的含义，也就是说可以把样品和变量的各个取值同时在一张二维图上表示出来.

对应分析的数据变换方法：

设有n个样品，每个样品观测p个指标，原始数据为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

数据变换的方法的具体步骤如下：

1、对数据阵先分别按行和列求和，再求总和：

$$X=\begin{bmatrix} & & & x_{1p} & \sum_{k=1}^{p} x_{1k}=X_1 \\ x_{11} & x_{12} & \cdots & & \\ x_{21} & x_{22} & \cdots & x_{2p} & \sum_{k=1}^{p} x_{2k}=X_2 \\ & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & x_{np} & \sum_{k=1}^{p} x_{nk}=X_n \\ x_{n1} & x_{n2} & \cdots & & \\ X_{.1} & X_{.2} & \cdots & X_{.p} & \sum_{l=1}^{n}\sum_{k=1}^{p} x_{lk}=X\cdots \overset{\text{def}}{=} T \end{bmatrix}$$

2、化数据阵X为规格化的“概率矩阵” P ,令

$$\mathrm{P} = \frac{1}{T} X \overset{\text{def}}{=} (p_{ij})_{n\times p}$$

其中$p_{ij} = \frac{1}{T} x_{ij} (i = 1, \cdots, n; j = 1, \cdots, p)$

不难看出 $0 \le p_{ij} \le 1$ 且$\sum_{i=1}^{n} \sum_{j=1}^{p} p_{ij} = 1$

p_{ij}可理解为数据 x_{ij}出现的“概率”并称P为对应阵。

类似地可以写出对应阵P的行和列和，并把P表示成如下一张列联表：

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} & P_{1.} \\ p_{21} & p_{22} & \cdots & p_{2p} & P_{2.} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{np} & P_{n.} \\ P_{.1} & P_{.2} & \cdots & P_{.p} & 1 \end{bmatrix}$$

其中$P_{.j}=\sum_{i=1}^{n}p_{ij}\ (i=1,\cdots,n)$;

可理解为第j个变量的边缘概率

$P_{i.}=\sum_{j=1}^{p}p_{ij}\ (j=1,\cdots,p)$;

可理解为第i个样品的边缘概率

记$r=\begin{bmatrix}P_{1.}\\ \vdots\\ P_{n.}\end{bmatrix}c=\begin{bmatrix}P_{.1}\\ \vdots\\ P_{.p}\end{bmatrix}$

则$r=P1_p\ \ c=P'1_n$

其中 $1_p=(1,1,\cdots,1)'$为元素全为 1 的 p 维常向量.

3、从对应阵 P 出发计算变量的协方差阵(考虑 R 型因子分析)，我们把 P 矩阵中的 n 个行作为 p 维空间中 n 个样品点.具体步骤如下

(i) 消除各样品点出现的概率大小的影响，称

$$R'_i = \left(\frac{p_{i1}}{P_{i.}}, \frac{p_{i2}}{P_{i.}}, \cdots, \frac{p_{ip}}{P_{i.}}\right), (i = 1, \cdots, n)$$

为样品 i 的形象，或 p 个变量在第 i 个样品上的分布轮廓，显然有

$$R'_i = \left(\frac{p_{i1}}{P_{i.}}, \frac{p_{i2}}{P_{i.}}, \cdots, \frac{p_{ip}}{P_{i.}}\right) = \left(\frac{x_{i1}}{X_{i.}}, \frac{x_{i2}}{X_{i.}}, \cdots, \frac{x_{ip}}{X_{i.}}\right)$$

研究样品点的相互关系一般用两个样品点的欧式距离来表示，为了消除各变量量纲不同的影响，引入第 k 个和第 l 个样品点的加权平方距离公式(或称卡方距离)：

$$D^2(k,l)=\sum_{j=1}^{p}\frac{\left(\frac{p_{kj}}{P_{k.}}-\frac{p_{lj}}{P_{l.}}\right)^2}{P_{.j}}=\sum_{j=1}^{p}\left(\frac{p_{kj}}{P_{k.}\sqrt{P_{.j}}}-\frac{p_{lj}}{P_{l.}\sqrt{P_{.j}}}\right)^2$$

(ii) 消除各变量量纲不同的影响，把第 i 个样品点的坐标化为

$$\left(\frac{p_{i1}}{P_{i.}\sqrt{P_{.1}}},\frac{p_{i2}}{P_{i.}\sqrt{P_{.2}}},\cdots,\frac{p_{ip}}{P_{i.}\sqrt{P_{.p}}}\right)(\mathrm{i}=1,\cdots,n)$$

(iii) 计算第 j 个变量(即第 j 列)的加权平均值:以第 i 个样品点的概率 $P_{i.}$ 作为权重来计算第 j 个变量的加权平均值,公式为

$$\sum_{i=1}^{n} \frac{p_{ij}}{P_{i.}\sqrt{P_{.j}}} \cdot P_{i.} = \sqrt{P_{.j}} (j = 1,2,\cdots,p)$$

(iv) 用加权方法计算第 i 个变量与第 j 个变量的 协方差:

$$\alpha_{ij} = \sum_{\alpha=1}^{n} \left(\frac{p_{\alpha i}}{P_{\alpha .} \sqrt{P_{.i}}} - \sqrt{P_{.i}} \right) \left(\frac{p_{\alpha j}}{P_{\alpha .} \sqrt{P_{.j}}} - \sqrt{P_{.j}} \right) \cdot P_{\alpha}$$

$$= \sum_{\alpha=1}^{n} \left(\frac{p_{\alpha i}}{\sqrt{P_{\alpha .} P_{.i}}} - \sqrt{P_{\alpha .} P_{.i}} \right) \left(\frac{p_{\alpha j}}{\sqrt{P_{\alpha .} P_{.j}}} - \sqrt{P_{\alpha .} P_{.j}} \right)$$

$$= \sum_{\alpha=1}^{n} \frac{p_{\alpha i} - P_{\alpha .} P_{.i}}{\sqrt{P_{\alpha .} P_{.i}}} \cdot \frac{p_{\alpha j} - P_{\alpha .} P_{.j}}{\sqrt{P_{\alpha .} P_{.j}}} \stackrel{\text{def}}{=\!=} \sum_{\alpha=1}^{n} z_{\alpha i} z_{\alpha j}$$

其中

$$z_{\alpha i}=\frac{p_{\alpha i}-P_{\alpha .}P_{.i}}{\sqrt{P_{\alpha .}P_{.i}}}=\frac{x_{\alpha i}-X_{\alpha .}X_{.i}/T}{\sqrt{X_{\alpha .}X_{.i}}}$$

令 $Z=(z_{ij})$ 为 $n\times p$ 矩阵，则变量间的协方差阵为

$$S_R=Z'Z=(a_{ij})_{p\times p}$$

4、从 P 出发计算样品间的协方差阵(考虑 Q 型因子分析). 用类似地方法可以得出 n 个样品间的协方差S_Q为

$$S_Q = ZZ' = (b_{ij})_{n\times n}$$

5、进行数据的对应变换，令

$$Z = (z_{ij})_{n\times p}$$

其中

$$z_{ij} = \frac{p_{ij} - P_{i.}P_{.j}}{\sqrt{P_{i.}P_{.j}}} = \frac{x_{ij} - \frac{X_{i.}X_{.j}}{T}}{\sqrt{X_{i.}X_{.j}}} \qquad (*)$$

$$(i = 1, \cdots n; j = 2, \cdots, p)$$

公式 $(*)$ 即是我们同时研究 R 型和 Q 型因子分析的角度导出的数据对应变换公式.

对应分析的计算步骤

1、由原始数据 X 出发计算对应阵 P 和变换后的新数据阵 Z .

2、计算行轮廓分布(或行形象分布)，记

$$R = \left(\frac{x_{ij}}{X_{i.}}\right) = \left(\frac{p_{ij}}{P_{i.}}\right)_{n\times p} \overset{\text{def}}{=} \begin{pmatrix} R'_1 \\ \vdots \\ R'_n \end{pmatrix}$$

R 矩阵由 X 矩阵(或对应阵 P)的每一行除以行和得到，其目的在于消除行点(即样品点)出现的“概率”不同的影响.

记 $N(R) = R_i, i = 1, \cdots, n, N(R)$表示 n 行向量组成的 p 维空间的点集，则点集 N(R) 的重心(每个样品点以 $P_{i.}$为权重)为

$$\sum_{i=1}^{n} P_{i.} R_i = \sum_{i=1}^{n} P_{i.} \begin{bmatrix} \frac{p_{i1}}{P_{i.}} \\ \vdots \\ \frac{p_{ip}}{P_{i.}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} p_{i1} \\ \vdots \\ \sum_{i=1}^{n} p_{ip} \end{bmatrix} = \begin{bmatrix} P_{.1} \\ \vdots \\ P_{.p} \end{bmatrix} = c$$

注:

（1）记 D_r, D_c 分别为:

$$D_r = diag(P_{1.}, \cdots, P_{n.})$$
$$D_c = diag(P_{.1}, \cdots, P_{.p})$$

（2）c 是 p 个列变量的边缘分布.

3、计算列轮廓分布(或列形象分布)，记

$$C=\left(\frac{x_{ij}}{X_{.j}}\right)=\left(\frac{p_{ij}}{P_{.j}}\right)_{n\times p}\overset{\text{def}}{=}(C_1,\cdots,C_p)$$

C 矩阵由 X 矩阵(或对应阵 P)的每一列除以列和得到，

其目的在于消除列点(即变量点)出现“概率”不同的影响.

4、χ^2统计量和总惯量

χ^2统计量：设用于检验行和列两个属性变量是否互不相关的统计量:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{ij} - m_{ij})^2}{m_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2$$

其中 $m_{ij}=\frac{x_{i.}x_{.j}}{T}$, x_{ij}^2表示第(i , j)个单位元在检验行与列两个属性变量是否不相关时对总 χ^2 统计量的贡献:

故$x_{ij}^2=\frac{(x_{ij}-m_{ij})^2}{m_{ij}}=Tz_{ij}^2$

$$\chi^2=T\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=Ttr(Z'Z)=Ttr(S_R)\overset{或}{=}Ttr(S_Q)$$

总惯量：

首先由加权平方距离公式可知，第k个与第l个样品点的χ^2距离为

$$D^2(k,l)=\sum_{j=1}^{p}(\frac{p_{kj}}{P_{k.}}-\frac{p_{lj}}{P_{l.}})^2/P_{.j}=(R_k-R_l)'D_c^{-1}(R_k-R_l)$$

我们把n个样品点（即行点）到重心c的加权平方距离的总和定义为行形象点集N(R)的总惯量Q：

$$Q=\sum_{i=1}^{n}P_{i.}D^2(i,c)=\sum_{i=1}^{n}P_{i.}\sum_{j=1}^{p}\frac{1}{P_{.j}}\left(\frac{p_{ij}}{P_{i.}}-P_{.j}\right)^2=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{P_{i.}}{P_{.j}}\frac{(P_{ij}-P_{i.}P_{.j})^2}{P_{i.}^2}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-P_{i.}P_{.j})^2}{P_{i.}P_{.j}}=\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=\frac{\chi^2}{T}$$

其中χ^2统计量是检验行点与列点是否互不相关的检验统计量。

5、对标准化后的新矩阵 Z 做奇异值分解。求 Z 的奇异值分解式其实是通过求$S_R = Z'Z$矩阵的特征值和标准化特征向量来得到.

设特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ 相应标准化特征向量为$\nu_1, \nu_2, \cdots, \nu_m$,在实际应用用常按累计贡献率

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_l}{\lambda_1+\lambda_2+\cdots+\lambda_l+\cdots+\lambda_m} \geq 0.80$$

确定所取公共因子个数 l(l ≤m)，Z 的奇异值$d_j = \sqrt{\lambda_j}$,以下我们仍用 m 表示选定的因子个数.

6、计算行轮廓的坐标G和列轮廓的坐标 F . 令

$$a_i = D_c^{-1/2} v_i$$

则 $a'_i D_c a_i = 1 (i = 1, \cdots, m)$ R 型因子分析的“因子载荷矩阵”(或列轮廓坐标)为

$$F = (d_1 a_1, d_2 a_2, \cdots, d_m a_m)$$

令 $b_i = D_c^{-1/2} u_i$

则 $b'_i D_r b_i = 1 (i = 1, \cdots, m)$ Q 型因子分子的“因子载荷矩阵”(或行轮廓坐标)为$G = (d_1 b_1, d_2 b_2, \cdots, d_m b_m)$

我们常常把 a_i或$\mathrm{b}_i (i = 1, \cdots, m)$称为加权意义下有单位长度的特征向量.

7、在相同二维平面上用行轮廓的坐标 G 和列轮廓的坐标 F (取 m=2) 绘制出点的平面图，也就是把 n 个行点(样品点)和 p 个列点和(变量点)在同一平面坐标系中点图.

8、求总惯量Q与 x^2统计量的分解式。

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij}^2 = tr(Z'Z) = \sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} d_i^2$$

其中 $\lambda_i = (i = 1, \cdots, m)$ 是Z'Z的特征值，

$d_i = \sqrt{\lambda_i}(i = 1, \cdots, m)$是Z的奇异值。

第i个因子轴末端的惯量$Q_i = d_i^2$

相应的$\chi^2 = \mathrm{TQ} = \mathrm{T}\sum_{i=1}^{m} d_i^2$

给出总χ^2统计量的分解式。

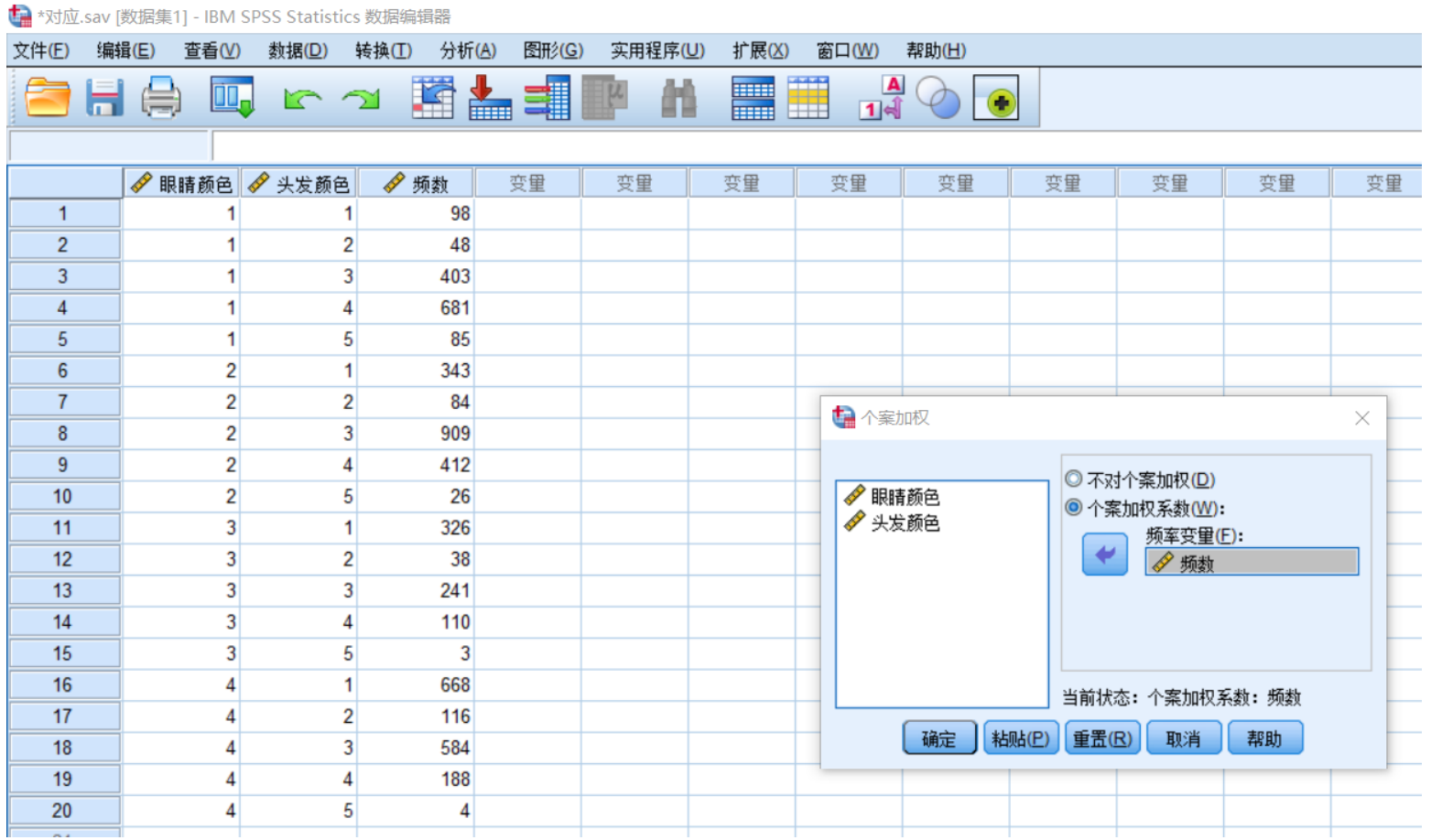
9、对样品点和变量点进行分类，并结合专业知识进行成因解释

• 对应分析案例

费希尔在1940年首次介绍列联表资料时使用的是一份关于眼睛颜色与头发颜色的调查研究数据。该研究数据包含了5387名苏格兰北部的凯斯纳斯郡的小学生的眼睛颜色与头发颜色，如下表所示。试用对应分析方法研究眼睛颜色与头发颜色之间的对应关系。

眼睛颜色	头发颜色	频数
1	1	98
1	2	48
1	3	403
1	4	681
1	5	85
2	1	343
2	2	84
2	3	909
2	4	412
2	5	26
3	1	326
3	2	38
3	3	241
3	4	110
3	5	3
4	1	668
4	2	116
4	3	584
4	4	188
4	5	4

选择数据里面的个案加权



点击分析里的降维，选择对应分析

摘要给出了奇异值，总惯量，卡方统计量，以及显著性，这里显著性为0说明两个变量有很强的相关性。

摘要

维	奇异值	惯量	卡方	显著性	惯量比例		置信度奇异值	
					占	累积	标准差	相关性 2
1	.444	.197			.866	.866	.012	.270
2	.172	.030			.130	.996	.013	
3	.031	.001			.004	1.000		
总计		.228	1221.688	.000[a]	1.000	1.000		

a. 12 自由度

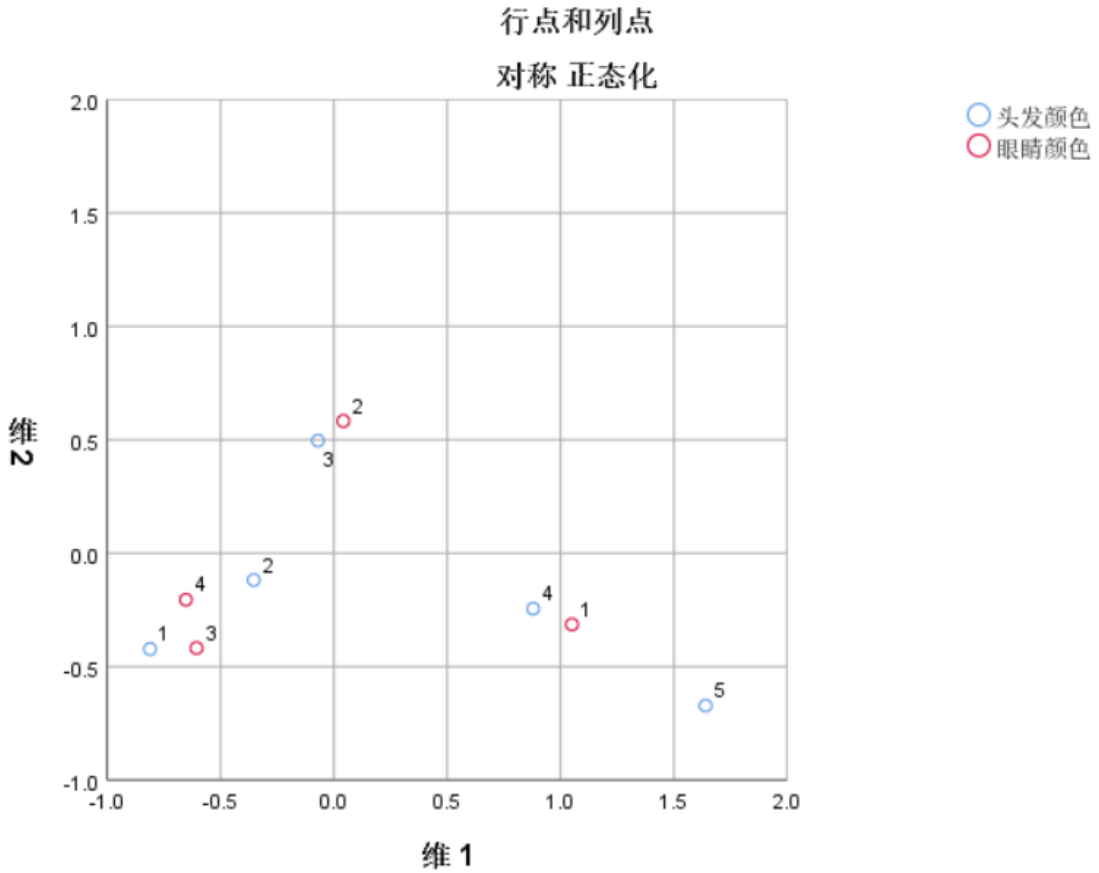
- 其中数量表示样本中各类占比，维得分就是在二维图上的坐标。

行点总览[a]

眼睛颜色	数量	维得分 1	维得分 2	惯量	贡献 点对维的惯量 1	贡献 点对维的惯量 2	贡献 维对点的惯量 1	贡献 维对点的惯量 2	贡献 总计
1	.245	1.050	-.314	.124	.608	.141	.966	.034	1.000
2	.331	.042	.583	.020	.001	.652	.013	.986	.999
3	.134	-.607	-.418	.026	.111	.136	.825	.152	.977
4	.291	-.653	-.205	.058	.279	.071	.957	.037	.994
活动总计	1.000			.228	1.000	1.000			

a. 对称正态化

- 绘制出的二维图


行点和列点
对称 正态化
头发颜色
眼睛颜色
维 2
维 1
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
1
2
3
4

Part 2

典型相关分析

典型相关分析

我们来看这样的一个例子，测量15名受试者的身体形态以及健康情况指标，如下表。第一组是身体形态变量，有年龄、体重、胸围和日抽烟量；第二组是健康状况变量，有脉搏、收缩压和舒张压，要 求分析身体形态和健康状况这两组变量之间的关系。

年龄X1	体重X2	抽烟量X3	胸围X4	脉搏Y1	收缩压Y2	舒张压Y3
25	125	30	83.5	70	130	85
26	131	25	82.9	72	135	80
28	128	35	88.1	75	140	90
29	126	40	88.4	78	140	92
27	126	45	80.6	73	138	85
32	118	20	88.4	70	130	80
31	120	18	87.8	68	135	75
34	124	23	84.6	70	135	75
36	128	25	88	75	140	80
38	124	23	85.6	72	145	86
41	135	40	86.3	76	148	88
46	143	45	84.8	80	145	90
47	141	48	87.9	82	148	92
48	139	50	81.6	85	150	95
45	140	55	88	88	160	95

观察本题给出的数据，将7个变量分成了2组，一组是身体形态变量，另一组是健康状况变量，根据题目要求，我们需要分析得出两组变量之间的相互关系，而不是单独分析单个变量之间的关系。

或者某题需要分析饲料与荤菜价格的关系，统计出了若干年玉米、大豆、稻子、麦子、鱼粉以及猪肉、牛肉、羊肉、鸡肉、鸡蛋、鸭肉、鸭蛋的价格，分析饲料与荤菜价格关系时，我们发现单独几种饲料和单独几种肉蛋禽价格关系不密切，但饲料的某种综合价格与肉蛋禽的综合价格关系密切时，我们将研究变量分成两组，分析两组变量之间的关系。

因此当遇到这样类型的题目，在进行数据处理时我们应该采取典型相关分析，其基本思想和主成分分析非常相似，首先在每组变量中找出变量的线性组合，使得两组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止。被选出的线性组合配对称为典型相关变量，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。

典型相关分析原理

设两组随机变量，X代表第一组p个变量，Y代表第二组的q个变量，假设$p \leq q$

$$cov(X,X) = \sum_{11}, cov(Y,Y) = \sum_{22}$$
$$cov(X,Y) = \sum_{12} = \sum'_{21}$$

- 根据典型相关分析的基本思想，进行两组随机向量的相关分析，首先要计算出各组变量的线性组合，即典型变量，并使其相关系数达到最大。

因此，我们设两组变量的线性组合分别为：

$$U = a'X = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

$$V = b'Y = b_1Y_1 + b_2Y_2 + \cdots + b_pY_p$$

我们希望寻找使相关系数达到最大的向量a与b，为防止结果的重复出现，令

$$Var(U) = a'\sum_{11}a = 1, Var(V) = b'\sum_{22}b = 1$$

那么 $\rho(U,V)=\frac{a\prime\sum_{12}b}{\sqrt{a\prime\sum_{11}a}\sqrt{b\prime\sum_{22}b}}=a'\sum_{12}b$

问题就转化为，在方差的约束下，求向量a和b使得

$$\rho(U,V)=a'\sum_{12}b$$

达到最大。

根据条件极值的求法引入拉格朗日乘数，问题转化为求

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

的极大值，其中λ,ν是拉格朗日乘数。

根据求极值的必要条件得

$$\begin{cases} \dfrac{\partial \varphi}{\partial a} = \sum_{12} b - \lambda \sum_{11} a = 0 \\ \dfrac{\partial \varphi}{\partial b} = \sum_{21} a - \nu \sum_{22} b = 0 \\ a' \sum_{11} a - 1 = 0 \\ b' \sum_{22} b - 1 = 0 \end{cases}$$

化简得

$$\begin{cases}\sum_{11}^{-1}\sum_{12}\sum_{22}^{-1}\sum_{21}a-\lambda^2 a=0\\ \sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}\sum_{12}b-\lambda^2 b=0\end{cases}$$

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

令$\mathrm{A}=\sum_{11}^{-1}\sum_{12}\sum_{22}^{-1}\sum_{21}, B=\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}\sum_{12}$

由$\frac{\partial \varphi}{\partial a} = \sum_{12} b - \lambda \sum_{11} a = 0$及$a' \sum_{11} a - 1 = 0$

可得$\lambda = a' \sum_{12} b$,而

$$\rho(U,V) = \frac{a' \sum_{12} b}{\sqrt{a' \sum_{11} a} \sqrt{b' \sum_{22} b}} = a' \sum_{12} b$$

所以求$\rho(U,V)$最大值也就是求λ 的最大值，而求λ的最大值又转化为求A和B的最大特征根。

设A和B的特征根为$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2$

$$\boldsymbol{r} = rank(\boldsymbol{A}) = rank(\boldsymbol{B})$$

$\mathrm{a}^{(1)}, \mathrm{a}^{(2)}, \cdots, \mathrm{a}^{(r)}$为A对于 $\lambda_1^2, \lambda_2^2, \cdots, \lambda_r^2$的特征向量，

$b^{(1)}, b^{(2)}, \cdots, b^{(r)}$为B对于 $\lambda_1^2, \lambda_2^2, \cdots, \lambda_r^2$的特征向量。

最大特征根 λ_1^2对应的特征向量 $\mathrm{a}^{(1)} = (a_1^{(1)}, a_2^{(1)}, \cdots, a_p^{(1)})'$

和 $b^{(1)} = (b_1^{(1)}, b_2^{(1)}, \cdots, b_p^{(1)})'$就是所求的典型变量的系数向量，即可得

$$U_1 = a^{(1)\prime}X = a_1^{(1)}X_1 + a_2^{(1)}X_2 + \cdots + a_p^{(1)}X_p$$

$$V_1 = b^{(1)\prime}Y = b_1^{(1)}Y_1 + b_2^{(1)}Y_2 + \cdots + b_p^{(1)}Y_p$$

我们称其为第一对典型变量，最大特征根的平方根，即为两典型变量的相关系数，我们称其为第一典型相关系数。如果第一对典型相关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，类似可求出第r对典型变量$U_r = a^{(r)\prime}X, V_r = b^{(r)\prime}Y$

其系数向量 $a^{(r)}$和$b^{(r)}$分别为A 和B 的第r个特征根λ_r^2对应的特征向量，λ_r为第r典型相关系数。

典型相关系数的显著性检验：

巴特莱特提出了一个根据样本数据检验总体典型相关系数

$\lambda_1, \lambda_2, \cdots, \lambda_r$是否等于零的方法。检验假设为

$$H_0: \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_r = 0$$

$$H_1: \lambda_{k+1} \neq 0$$

似然比统计量为：

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

可以证明 $Q_k = -mln\Lambda_k$ 近似服从 $\chi^2(f_k)$分布，

其中$f_k = (p-k)(q-k)$

自由度$m = (n-k-1) - 0.5(p+q+1)$

检验过程：

(i)首先检验

H_0： $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 0$此时k=0。

在检验水平α下，若$Q_0 > \chi_\alpha^2(f_0)$ 拒绝原假设，也就是说至少有一个典型相关系数大于零，即最大典型相关系数为$\lambda_1 > 0$

(ii)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。近似服从$\chi^2(f_1)$分布

如果 $Q_1 > \chi_\alpha^2(f_1)$拒绝原假设，即认为 $\lambda_2, \lambda_3, \cdots, \lambda_r$

至少有一个大于零，即$\lambda_2 > 0$

(iii)若已判定$\lambda_1 > 0$和$\lambda_2 > 0$重复以上步骤直至

$$H_0: \lambda_j = \lambda_{j+1} = \cdots = \lambda_r = 0$$

无法拒绝原假设时，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。

典型相关分析案例分析

- 问题背景：运动能力与身体状况的典型相关性分析。
- 下面是有关体重(x1)、腰围(x2)、脉搏(x3)、引体向上次数(y1)、起坐次数(y2)、跳跃次数(y3)的20组数据。

体重(x1)	腰围(x2)	脉搏(x3)	引体向上次数(y1)	起坐次数(y2)	跳跃次数(y3)
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

- 使用SPSS软件进行典型相关分析，具体步骤为：
- 打开SPSS软件。
- 点击“分析”中的“相关”。
- 点击典型相关分析，分别导入两组变量。

- 确定典型相关变量的个数（直接看典型相关系数对应的显著性水平）
- sig.的值越小说明相关系数越显著

Canonical Correlations

	Correlation	Eigenvalue	Wilks Statistic	F	Num D.F	Denom D.F.	Sig.
1	.796	1.725	.350	2.048	9.000	34.223	.064
2	.201	.042	.955	.176	4.000	30.000	.949
3	.073	.005	.995	.085	1.000	16.000	.775

H0 for Wilks test is that the correlations in the current and following rows are zero

这里显示的是典型相关变量的系数

Set 1 Standardized Canonical Correlation Coefficients

Variable	1	2	3
体重x1	.775	-1.884	-.191
腰围x2	-1.579	1.181	.506
脉搏x3	.059	-.231	1.051

Set 2 Standardized Canonical Correlation Coefficients

Variable	1	2	3
引体向上次数y1	.349	-.376	-1.297
起坐次数y2	1.054	.123	1.237
跳跃次数y3	-.716	1.062	-.419

- 标准化的系数由于经过标准化，因此系数相互之间是可比的，用处是用于比较不同自变量对应变量的影响程度。
- 而未标准化的系数因为每个变量没有标准化，量纲不一样，因此不能直接用系数大小比较自变量贡献程度，它的用处是可以用于计算CCA得分，（直接用系数乘以原始数据）

- 载荷分析说明生理指标的第一典型变量与体重的相关系数为-0.621，与腰围的相关系数为-0.925，与脉搏的相关系数为0.333. 从另一方面说明生理指标的第一对典型变量与体重、腰围负相关，而与脉搏正相关。其中与腰围的相关性最强。

Set 1 Canonical Loadings

Variable	1	2	3
体重x1	-.621	-.772	-.135
腰围x2	-.925	-.378	-.031
脉搏x3	.333	.041	.942

Set 2 Canonical Loadings

Variable	1	2	3
引体向上次数y1	.728	.237	-.644
起坐次数y2	.818	.573	.054
跳跃次数y3	.162	.959	-.234

已解释的方差比例

Proportion of Variance Explained

Canonical Variable	Set 1 by Self	Set 1 by Set 2	Set 2 by Self	Set 2 by Set 1
1	.451	.285	.408	.258
2	.247	.010	.434	.017
3	.302	.002	.157	.001