

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



第七章 相应分析

主讲教师 周仕君

第七章 相应分析

第一节 引言

第二节 列联表

第三节 相应分析的基本理论

第四节 相应分析中应注意的问题

第五节 实例分析与计算机实现

第一节 引言

- 相应分析(correspondence analysis)也叫对应分析，其特点是它所研究的变量可以是定性的。通常意义下的相应分析，是指对两个定性变量（因素）的多种水平进行相应性研究，因而它的应用越来越广泛，现在这种方法已经成为常用的多元分析方法之一。



- 在社会、经济以及其他领域中，进行数据分析时经常要处理因素与因素之间的关系，及因素内部各个水平之间的相互关系。
- 例如，评价某一个行业所属企业的经济效益，我们不仅要研究因素A，即企业按照经济效益好坏的分类情况，以及要研究因素B，即经济效益指标之间的关系，还要研究哪些企业与哪些经济效益指标更密切一些。这就需要相应分析的方法，将经济效益指标和企业状况放在一起进行分类、作图，以便更好的描述两者之间的关系，在经济意义上做出切合实际的解释。



-
- 相应分析也可以称为R-Q型分析，它是在R型和Q型因子分析基础上发展起来的一种多元统计分析方法。相应分析把R型和Q型因子分析统一起来处理，不仅能反映出变量之间的结构和样品之间的结构，而且更重要的是能反映变量结构和样品结构之间的关系，给分析结构带来很大的便利，是一种十分有用的多元统计分析方法。



第二节 列联表

□ 列联表的概念

□ 有关记号



一、列联表的概念

- 设有两组因素 A 和 B ，其中因素 A 包含 r 个水平，即 $A_1, A_2, \dots, A_r$ ；因素 B 包含 c 个水平，即 $B_1, B_2, \dots, B_c$ 。又设有受制于这两个因素的载体(或客体)的集合总体 G 。我们希望通过总体 G 关于这两组因素的有关资料(或抽样资料)，来分析这两组因素的关系。



例如，要考查在某个人群中关于吸烟或不吸烟(因素A)与得肺癌或不得肺癌(因素B)两组因素之间的关系。通常的作法是，随机地从该人群中抽样，对这两种因素进行调查，设调查了 k 个人，得到一个二维列联表，见表8.1。

因素 A	因素 B		
	得肺癌 (B_1)	不得肺癌 (B_2)	
吸烟 (A_1)	k_{11}	k_{12}	$k_{1.}$
不吸烟 (A_2)	k_{21}	k_{22}	$k_{2.}$
	$k_{.1}$	$k_{.2}$	$k = k_{..} = \sum k_{ij}$

表8.1 2×2 二维列联表



一般地，设受制于某个载体总体的两个因素为 A 和 B ，其中因素 A 包含 r 个水平，即 $A_1, A_2, \dots, A_r$ ；因素 B 包含 c 个水平，即 $B_1, B_2, \dots, B_c$ 。对这两组因素作随机抽样调查，得到一个 $r \times c$ 的二维列联表，记为 $K = (k_{ij})_{r \times c}$ ，见表 8.2。

		因素 B				
		B_1	B_2	$\dots$	B_c	
因素 A	A_1	k_{11}	k_{12}	$\dots$	k_{1c}	$k_{1.}$
	A_2	k_{21}	k_{22}	$\dots$	k_{2c}	$k_{2.}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	A_r	k_{r1}	k_{r2}	$\dots$	k_{rc}	$k_{r.}$
		$k_{.1}$	$k_{.2}$	$\dots$	$k_{.c}$	$k = k_{..} = \sum k_{ij}$



二、有关记号

- 为了叙述方便，先引进一些基本概念和记号。
- 将列联表 K (表8.2)转化为频率矩阵，记为 $F=(f_{ij})_{r \times c}$ ，见表8.3。

		因素 B				
		B_1	B_2	$\cdots$	B_c	
因 素 A	A_1	f_{11}	f_{12}	$\cdots$	f_{1c}	$f_{1.}$
	A_2	f_{21}	f_{22}	$\cdots$	f_{2c}	$f_{2.}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	A_r	f_{r1}	f_{r2}	$\cdots$	f_{rc}	$f_{r.}$
		$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.c}$	$1 = f_{..} = \sum f_{ij}$

表8.3 一般的二维频率表



表 8.3 中 $f_{ij} = k_{ij} / k$ 是样本中属于因素 A 的第 i 个水平和

因素 B 的第 j 个水平的百分比； $f_{i.} = \sum_{j=1}^c f_{ij}$ ， $f_{.j} = \sum_{i=1}^r f_{ij}$ ，

$i = 1, 2, \dots, r$ ， $j = 1, 2, \dots, c$ 。这里我们记

$$\mathbf{f}_r = (f_{1.}, f_{2.}, \dots, f_{r.})', \quad \mathbf{f}_c = (f_{.1}, f_{.2}, \dots, f_{.c})',$$

$$\mathbf{D}_r = \text{diag}(f_{1.}, f_{2.}, \dots, f_{r.}) = \text{diag}(\mathbf{f}_r),$$

$$\mathbf{D}_c = \text{diag}(f_{.1}, f_{.2}, \dots, f_{.c}) = \text{diag}(\mathbf{f}_c)$$

那么有，

$$\mathbf{f}_r = \mathbf{F}\mathbf{I}_c, \quad \mathbf{f}_c = \mathbf{F}'\mathbf{I}_r \quad (8.1)$$

$$\mathbf{I}_r' \mathbf{f}_r = \mathbf{I}_c' \mathbf{f}_c = \mathbf{I}_r' \mathbf{F} \mathbf{I}_c = 1 \quad (8.2)$$

其中 $\mathbf{I}_r = (1, 1, \dots, 1)'_{r \times 1}$ ， $\mathbf{I}_c = (1, 1, \dots, 1)'_{c \times 1}$ 。



从数理统计的角度， K 可视为对两个随机变量(记为 ξ 和 η) 调查得到的二维列联表，频率矩阵 F 则表示它们相应的经验联合抽样分布为

$$P\{\xi = i, \eta = j\} = f_{ij}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

其中 ξ 与 η 分别表示因素 A 和因素 B 的随机变量。

$(f_{1.}, f_{2.}, \dots, f_{r.})$ 和 $(f_{.1}, f_{.2}, \dots, f_{.c})$ 分别为二维随机变量 (ξ, η) 的抽样边际分布。在此，我们称 D_r 和 D_c 分别为 ξ 和 η 的边际阵。那么，有条件概率为

$$P\{\eta = j \mid \xi = i\} = \frac{P\{\xi = i, \eta = j\}}{P\{\xi = i\}} = \frac{f_{ij}}{f_{i.}}, \quad j = 1, 2, \dots, c$$



在此称

$$\mathbf{f}_c^i = \left(\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, \dots, \frac{f_{ic}}{f_{i.}} \right)' \in \mathbf{R}^c \quad (8.3)$$

为因素 A 的第 i 个水平分布轮廓。称 $\mathbf{D}_r^{-1} \mathbf{F}$ 为因素 A 的轮廓矩阵。这里应该注意到， $\mathbf{f}_c^i$ ， $i=1,2,\dots,r$ 是超平面 $x_1 + x_2 + \dots + x_{xc} = 1$ 的一点集。

同理，因素 B 的第 j 个水平的分布轮廓为

$$\mathbf{f}_r^j = \left(\frac{f_{1j}}{f_{.j}}, \frac{f_{2j}}{f_{.j}}, \dots, \frac{f_{cj}}{f_{.j}} \right)' \in \mathbf{R}^r \quad (8.4)$$



并称 $D_c^{-1}F'$ 为因素 B 的轮廓矩阵，同样 f_r^j ， $j=1,2,\dots,c$ 是超平面 $y_1 + y_2 + \dots + y_{yr} = 1$ 的一点集。这里有

$$P\{\xi = i \mid \eta = j\} = \frac{P\{\xi = i, \eta = j\}}{P\{\eta = j\}} = \frac{f_{ij}}{f_{.j}}, \quad i = 1, 2, \dots, r$$

最后，由 (8.1) 式和 (8.2) 式我们应该明确

$$D_r I_r = F I_c, \quad I_r' D_r I_r = I_r' F I_c = 1, \quad (8.5)$$

$$D_c I_c = F' I_r, \quad I_c' D_c I_c = I_c' F' I_r = 1, \quad (8.6)$$

从 (8.5) 式和 (8.6) 式我们清楚地看到， D_r 和 D_c 中的元素起到了权重的作用，称其为权重矩阵。



第三节 相应分析的基本理论

- 原始资料的变换
- 基于矩阵的分析过程



- 相应分析的主要目的是寻求列联表行因素A和列因素B的基本分析特征和它们的最优联立表示。
- 为了实现行因素A与列因素B最优联立表示，进一步剖析行因素A内部之间，列因素B内部之间，以及因素A和列因素B之间的关系，
- 这里将介绍原始的列联资料 $K=(k_{ij}) r \times c$ 变换成矩阵 $Z=(z_{ij}) r \times c$ 的具体过程，这样使得 z_{ij} 对因素A和列因素B具有对等性，在此基础上进行相应分析。



为什么要把列联资料换成矩阵Z?

- 若要进行R型因子分析，需要计算所考察变量的协差阵。对于所考察的p个变量，其变量间的叉积矩阵是 $A = (a_{ij})_{p \times p}$ ，其中

$$a_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_{.i})(x_{kj} - \bar{x}_{.j})$$

- 若要进行Q型因子分析，需要计算所考察变量的协差阵。对于所考察的n个样品，其样品间的叉积矩阵是 $A^* = (a_{ij}^*)_{n \times n}$ ，其中

$$a_{ij}^* = \sum_{k=1}^p (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.})$$



- 显然，变量间的叉积矩阵和样品的叉积矩阵阶数不同，一般来说，二者的非零特征根也不一样。
- 如果能将列联表做一变换，将其变换为 Z ，使得 $Z'Z$ 和 ZZ' 能起到叉积矩阵 A 和 A^* 同样作用，那么由矩阵理论可知，矩阵 $Z'Z$ 和 ZZ' 有相同的非零特征根，他们相应的特征向量也有密切的关系，这样不仅在计算时带来了方便，而且由于变量的各因子和样品的各相应因子的尺度相同，从而可以将变量的主因子和样品的主因子在同一个坐标系内给出，使对变量的因子分析和样品的因子分析相对照，从而对观测数据的分析就更清晰明了了。



一、原始资料的变换

设 $K = (k_{ij})_{r \times c}$ 为一个 $r \times c$ 的列联资料，其转化后的频率矩阵为 $F = (f_{ij})_{r \times c}$ 。我们针对因素 A 而言，由 (8.3) 式知，第 i 个水平分布轮廓 $f_c^i \in \mathbf{R}^c$ $i = 1, 2, \dots, r$ 为超平面 $x_1 + x_2 + \dots + x_r = 1$ 的一点集。如果我考虑因素 A 中各水平之间的远近，引入欧氏距离，那么第 i 个水平和第 i' 个水平之间的欧氏距离为

$$D^2(i, i') = \sum_{j=1}^c \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \quad (8.7)$$



这样定义的距离没有考虑到因素 B 的各水平边际概率的影响，为了消除因素 B 各个水平数量级的影响，应该对每一项加一个权数 $1/f_{.j}$ ，即有：

$$\begin{aligned} D_w^2(i, i') &= \sum_{j=1}^c \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \frac{1}{f_{.j}} \\ &= \sum_{j=1}^c \left(\frac{f_{ij}}{f_{i.} \sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.} \sqrt{f_{.j}}} \right)^2 \quad (8.8) \end{aligned}$$

我们称 $D_w^2(i, i')$ 为因素 A 中第 i 个水平和第 i' 个水平之间 χ^2 距离。



这里应该注意到，(8.8) 式所定义的距离 $D_w^2(i, i')$ ，也可以

看作是点集 $\left(\frac{f_{i1}}{f_{i.}\sqrt{f_{.1}}}, \frac{f_{i2}}{f_{i.}\sqrt{f_{.2}}}, \dots, \frac{f_{ic}}{f_{i.}\sqrt{f_{.c}}} \right)'$ 中两点 i 和 i' 之间的欧氏距离 ($i = 1, 2, \dots, r$)。那么，我们从加权的角度考察这 r 个点的平均水平，其第 j 个分量的平均水平为

$$\sum_{i=1}^r \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} \cdot f_{i.} = \frac{1}{\sqrt{f_{.j}}} \sum_{i=1}^r f_{ij} = \sqrt{f_{.j}}, \quad j = 1, 2, \dots, c \quad (8.9)$$

从而，计算出关于因素 B 各水平构成的协差阵为

$$\Sigma_c = (a_{ij})_{c \times c} \quad (8.10)$$



其中,

$$\begin{aligned} a_{ij} &= \sum_{\alpha=1}^r \left(\frac{f_{\alpha i}}{f_{\alpha.} \sqrt{f_{.i}}} - \sqrt{f_{.i}} \right) \left(\frac{f_{\alpha j}}{f_{\alpha.} \sqrt{f_{.j}}} - \sqrt{f_{.j}} \right) \cdot f_{\alpha.} \\ &= \sum_{\alpha=1}^r \left(\frac{f_{\alpha i} - f_{\alpha.} f_{.i}}{\sqrt{f_{\alpha.} f_{.i}}} \right) \left(\frac{f_{\alpha j} - f_{\alpha.} f_{.j}}{\sqrt{f_{\alpha.} f_{.j}}} \right) \\ &= \sum_{\alpha=1}^r z_{\alpha i} \cdot z_{\alpha j} \end{aligned}$$



这里

$$\begin{aligned} z_{\alpha i} &= \frac{f_{\alpha i} - f_{\alpha.} f_{.i}}{\sqrt{f_{\alpha.} f_{.i}}} = \frac{k_{\alpha i} / x_{..} - (k_{\alpha.} / x_{..})(k_{.i} / x_{..})}{\sqrt{(k_{\alpha.} / x_{..})(k_{.i} / x_{..})}} \\ &= \frac{k_{\alpha i} - (k_{\alpha.} k_{.i} / k_{..})}{\sqrt{k_{\alpha.} k_{.i}}} \end{aligned}$$

$$\alpha = 1, 2, \dots, r \quad i = 1, 2, \dots, c$$

令 $\mathbf{Z} = (z_{ij})_{r \times c}$ ，则 (8.10) 式可表示为

$$\Sigma_c = \mathbf{Z}'\mathbf{Z} \quad (8.11)$$



类似地，由 (8.4) 式知，针对因素 B 的第 y_r 个水平的分布轮廓

$f_r^j \in \mathbf{R}^r$ ，它是超平面 $y_1 + y_2 + \cdots + y_c = 1$ 的一点集， $j = 1, 2, \cdots, c$ 。同样，变换以后所得到的关于因素 A 各水平构成的协差阵为

$$\Sigma_r = \mathbf{Z}\mathbf{Z}' \quad (8.12)$$

这里我们需要说明的是，将原始列联表设 $\mathbf{K} = (k_{ij})_{r \times c}$ 中的数据变换成矩阵 $\mathbf{Z} = (z_{ij})_{r \times c}$ 时，则因素 A 和因素 B 各个水平构成的协差阵分别为 $\Sigma_r = \mathbf{Z}\mathbf{Z}'$ 和 $\Sigma_c = \mathbf{Z}'\mathbf{Z}$ ，矩阵 Σ_r 和 Σ_c 存在简单的对等关系，这样如果把原始列联表中的数据 k_{ij} 变换成 z_{ij} 以后， z_{ij} 对于两个因素具有对等性。



二、基于矩阵的分析过程

由矩阵的知识我们知道， $\Sigma_r = \mathbf{Z}\mathbf{Z}'$ 和 $\Sigma_c = \mathbf{Z}'\mathbf{Z}$ 有完全相同的非零特征根，记为 $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ ， $0 < m \leq \min\{r, c\}$ ，设 $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ 为相对于特征根 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 的关于因素 B 各水平构成的协差阵 Σ_c 的特征向量，则有

$$\Sigma_c \mathbf{u}_j = \mathbf{Z}'\mathbf{Z}\mathbf{u}_j = \lambda_j \mathbf{u}_j \quad (8.13)$$

用矩阵 $\mathbf{Z}$ 左乘 (8.13) 式两端得

$$\mathbf{Z}\mathbf{Z}'(\mathbf{Z}\mathbf{u}_j) = \lambda_j(\mathbf{Z}\mathbf{u}_j)$$

即有

$$\Sigma_r(\mathbf{Z}\mathbf{u}_j) = \lambda_j(\mathbf{Z}\mathbf{u}_j) \quad (8.14)$$



-
- (8.14) 式表明 Zu_j 为相对于特征值 λ_j 的关于因素 A 各水平构成的协差阵 $\sum_r$ 的特征向量。
 - 这样我们就建立了相应分析中 R 型因子分析和 Q 型因子分析的关系。即 Q 型因子分析的载荷矩阵可由 R 型因子分析的载荷矩阵经过一个线性变换而得到。也就是说, 我们可以从 R 型因子分析出发而直接得到 Q 型因子分析的结果。



- 这里需要强调的是，由于 $\sum_r$ 和 $\sum_c$ 有相同的特征根，而这些特征根又表示各个公共因子所提供的方差。那么，在因素 B 的 c 维空间 R^c 中的第一公共因子，第二公共因子直到第 m 个公共因子与因素 A 的 r 维空间 R^r 中相对于的各个主因子在总方差中所占的百分比就完全相同。这样就可以用相同的因子轴同时描述两个因素各个水平的情况，把两个因素的各个水平的状况同时反映到具有相同坐标轴的因子平面上。
- 一般情形，我们取两个公共因子，这样就可以在一张二维平面图上绘出两个因素各个水平的情况，即可以直观地描述两个因素 A 和因素 B 以及各个水平之间的相关关系。



-
- 实践中，由于变量间协差阵的阶数通常小于样品间协差阵的阶数，所以从R型因子分析出发进行计算比较简单。



第四节 相应分析中应注意的问题

- 相应分析是分析两组或多组变量之间关系的有效方法，在离散情况下，它是从资料出发通过建立因素间的二维或多维列联表来对数据进行分析。在此我们要问，这种分析是否有意义，或者说对于所给的数据是否值得做这种相应分析。



设二维列联资料为 $K = (k_{ij})_{r \times c}$ （见表 8.2），其频率阵为 $F = (f_{ij})_{r \times c}$ （见表 8.3）。用 $p_{i.}$ 表示因素 A 中第 i 水平发生时的概率； $p_{.j}$ 表示因素 B 中第 j 水平发生时的概率，那么其估计值分别为

$$f_{i.} = \frac{k_{i.}}{k} \quad \text{和} \quad f_{.j} = \frac{k_{.j}}{k}$$

这里我们关心的是因素 A 和因素 B 是否独立，由此提出要检验的问题是

H_0 ：因素 A 和因素 B 是独立的

H_1 ：因素 A 和因素 B 不独立



统计量为：

$$\begin{aligned}\chi^2 &= \sum_{i=1}^r \sum_{j=1}^c \frac{[k_{ij} - \hat{E}(k_{ij})]^2}{\hat{E}(k_{ij})} \\ &= \sum_{i=1}^r \sum_{j=1}^c \frac{[k_{ij} - k_{i.}k_{.j} / k]^2}{k_{i.}k_{.j} / k} \\ &= k \sum_{i=1}^r \sum_{j=1}^c (z_{ij})^2\end{aligned}\tag{8.15}$$

其中 $z_{ij} = (k_{ij} - k_{i.}k_{.j} / k) / \sqrt{k_{i.}k_{.j}}$ ，当假设 H_0 ：因素 A 和因素 B 是独立成立时，在 n 足够大的条件下， χ^2 服从自由度为 $(r-1)(c-1)$ 的 χ^2 分布。拒绝区域为：

$$\chi^2 > \chi_{1-\alpha}^2 [(r-1)(c-1)]$$



通过上面的分析，我们应该注意几个问题。

第一，这里的 z_{ij} 是原始列联资料 $K = (k_{ij})_{r \times c}$ 通过相应变换以后得到的资料阵 $Z = (z_{ij})_{r \times c}$ 的元素。说明 z_{ij} 与 χ^2 统计量有着内在的联系。

第二，关于因素 B 和因素 A 各水平构成的协差阵 Σ_c 和 Σ_r ，由 (8.15) 式知， $tr(\Sigma_c) = tr(\Sigma_r) = \chi^2 / k$ ，这里 $tr(.)$ 表示矩阵的迹。

第三，独立性检验只能判断因素 A 和因素 B 是否独立。如果因素 A 和因素 B 独立，则没有必要进行相应分析；如果因素 A 和因素 B 不独立，可以进一步通过相应分析考察两因素各个水平之间的相关关系。



例1:

各区域每种有害气体的平均浓度

区域	氯	硫化氢	二氧化硫	碳 4	环氧氯丙烷	环己烷
A	0.056	0.084	0.031	0.038	0.0081	0.0220
B	0.049	0.055	0.100	0.110	0.0220	0.0073
C	0.038	0.130	0.079	0.170	0.0580	0.0430
D	0.034	0.095	0.058	0.160	0.2000	0.0290
E	0.084	0.066	0.029	0.320	0.0120	0.0410
F	0.064	0.072	0.100	0.210	0.0280	1.3800
G	0.048	0.089	0.062	0.260	0.0380	0.0360
H	0.069	0.087	0.027	0.050	0.0890	0.0210



由于变量间协差阵与样品间协差阵具有相同的非零特征根,而此例中变量的个数少于样品的个数,故变量间协差阵阶数较低,便于计算特征根和特征向量,故可计算变量间协差阵。该协差阵为:

$$A = \begin{bmatrix} 0.0423 & 0.0370 & 0.0136 & 0.0242 & 0.0104 & -0.0818 \\ 0.0370 & 0.0612 & 0.0251 & 0.0236 & 0.0416 & -0.1176 \\ 0.0136 & 0.0251 & 0.0404 & 0.0152 & 0.0115 & -0.0661 \\ 0.0242 & 0.0236 & 0.0152 & 0.1021 & 0.0046 & -0.1325 \\ 0.0104 & 0.0416 & 0.0115 & 0.0046 & 0.1370 & -0.1169 \\ -0.0818 & -0.1176 & -0.0661 & -0.1325 & -0.1169 & 0.3407 \end{bmatrix}$$



由此协差阵可计算出其各特征根分别为：

特征根序号	特征根	各特征根占总和的比例	累计比例
1	0.5067	0.6999	0.6999
2	0.1221	0.1687	0.8687
3	0.0566	0.0782	0.9469
4	0.0290	0.0401	0.9870
5	0.0095	0.0130	1.0000
6	0.0000	0	1.0000

由于前两个特征根的累计贡献率已达 86.87%，故可只取前两个特征根，计算相应于变量间协差阵的特征向量和相应于样品间协差阵的特征向量，可得 R 型因子分析的载荷矩阵和 Q 型因子分析的载荷矩阵分别是：

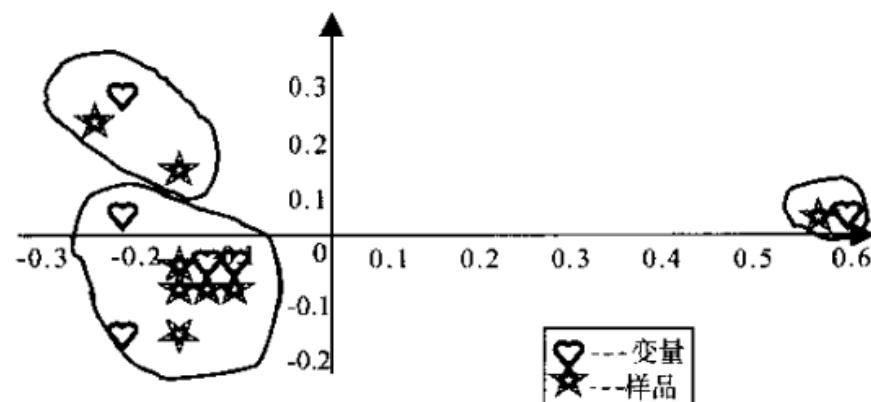


(1) R 型因子分析各变量的载荷为:

变量	因子 f_1	因子 f_2
x_1	-0.13832	-0.04385
x_2	-0.20333	0.02650
x_3	-0.11003	-0.01985
x_4	-0.21754	-0.18687
x_5	-0.21720	0.28831
x_6	0.58275	0.03279

(2) Q 型因子分析各样品的载荷为:

样品	因子 g_1	因子 g_2
A	-0.10599	-0.02354
B	-0.15369	-0.06164
C	-0.16248	-0.00928
D	-0.22388	-0.22377
E	-0.15853	-0.19307
F	0.56153	0.01900
G	-0.16656	-0.10664
H	-0.16429	-0.13644



全部变量和样品可分为三类,每一类聚合了一些变量和样品。第1类包含变量环氧氯丙烷 x_5 和区域 D 与 H,表明这两个区域的主要污染物是环氧氯丙烷;第2类包含变量 x_1 、 x_2 、 x_3 与 x_4 和区域 A、B、C、E 与 G,这5个区域的污染物主要是氯、硫化氢、二氧化硫和碳4;第3类包含变量 x_6 和区域 F,该区域的主要污染物是环己烷。



例2:

- 数据为1992年美国大选的部分数据。要求对选民的最高学历水平（degree）和所支持的总统候选人（pres92）进行相应分析。

pres92	年龄	年龄类别	教育	学位	性别
3	53	3.00	12	1	1
1	27	1.00	17	3	1
3	42	2.00	16	3	2
3	27	1.00	16	3	2
3	40	2.00	20	4	2
3	74	4.00	3	0	1
2	41	2.00	15	2	1
1	26	1.00	16	3	1
3	49	3.00	14	2	2
1	49	3.00	17	4	2
3	37	2.00	16	3	2
3	55	3.00	18	4	2
3	26	1.00	16	3	2
3	25	1.00	12	1	1
1	47	3.00	20	4	2
3	53	3.00	16	3	1
1	42	2.00	13	1	1



Correspondence Table（相应分析表），如表8.4，即列联表。

Correspondence Table						
VOTE FOR	RS HIGHEST DEGREE					
CLINTON, BUSH, PEROT	lt high school	high school	junior college	bachelor	graduate degree	Active Margin
Bush	55	349	48	146	63	661
Perot	12	159	26	62	19	278
Clinton	122	436	58	178	111	905
Active Margin	189	944	132	386	193	1844

Active Margin为边际频数。大致可以看出Clinton在各个学历层次都有最高的票数。



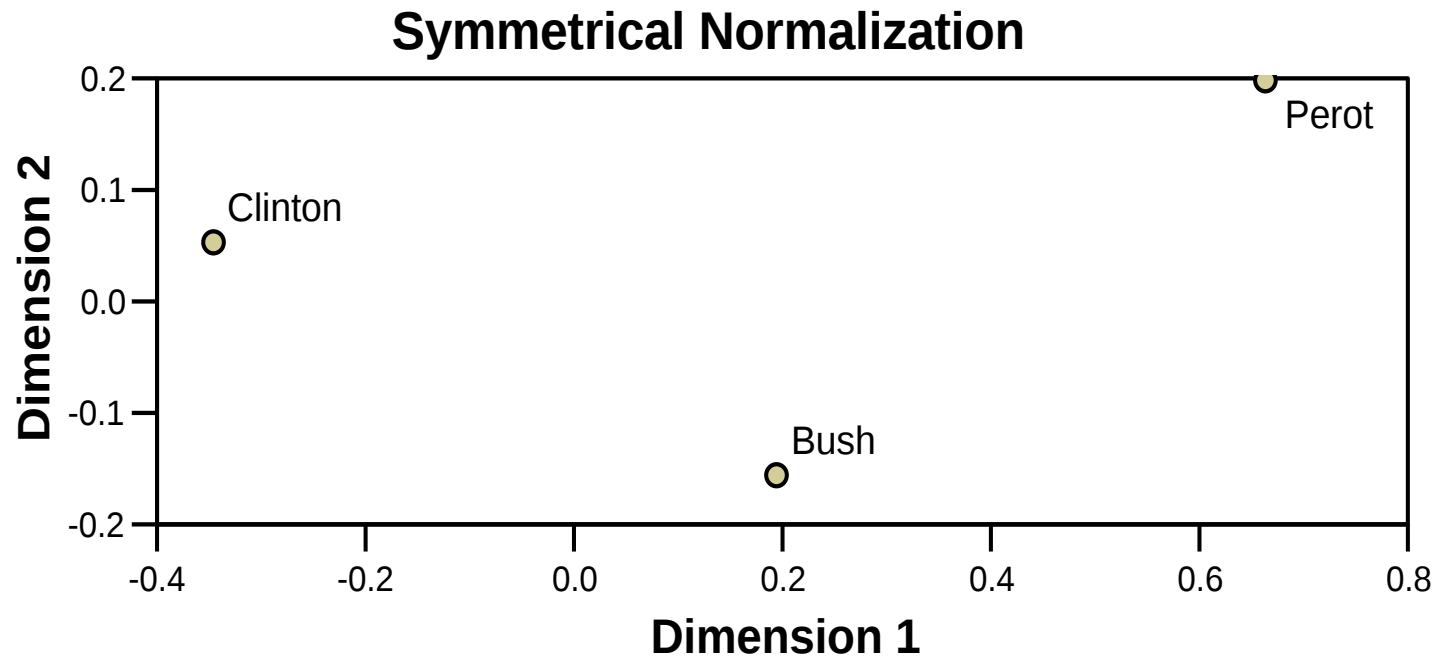
Overview Row Points(a)

VOTE FOR CLINTON, BUSH, PEROT	Mass	Score in		Inertia	Contribution				
		Dimension			Of Point to		Of Dimension to		
					Inertia of	Dimension	Inertia of Point	Inertia of Point	Total
1	2	1	2	1	2	Total			
Bush	.358	.194	-.156	.002	.098	.544	.931	.069	1.000
Perot	.151	.663	.198	.009	.479	.370	.990	.010	1.000
Clinton	.491	-.346	.053	.008	.423	.086	.997	.003	1.000
Active Total	1.000			.019	1.000	1.000			

* Symmetrical normalization



Row Points for VOTE FOR CLINTON, BUSH, PEROT



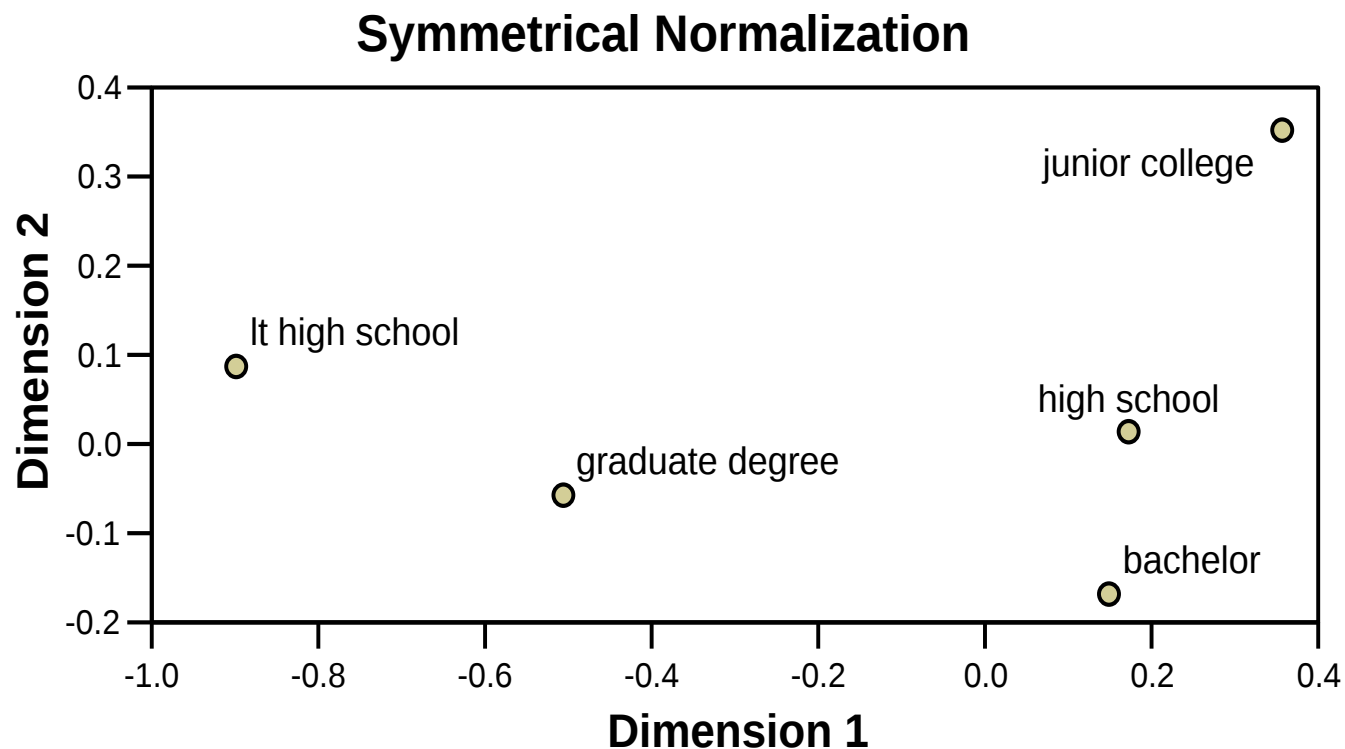
Overview Column Points(a)

RS HIGHEST DEGREE	Mass	Score in		Inertia	Contribution				
		Dimension			Of Point to		Of Dimension to		
					Inertia of	Dimension	Inertia of Point		
		1	2		1	2	1	2	Total
lt high school	.102	-.899	.087	.011	.597	.048	.999	.001	1.000
high school	.512	.173	.014	.002	.110	.006	.999	.001	1.000
junior college	.072	.357	.352	.001	.066	.554	.899	.101	1.000
bachelor	.209	.149	-.168	.001	.034	.370	.871	.129	1.000
graduate degree	.105	-.506	-.057	.004	.193	.022	.999	.001	1.000
Active Total	1.000			.019	1.000	1.000			

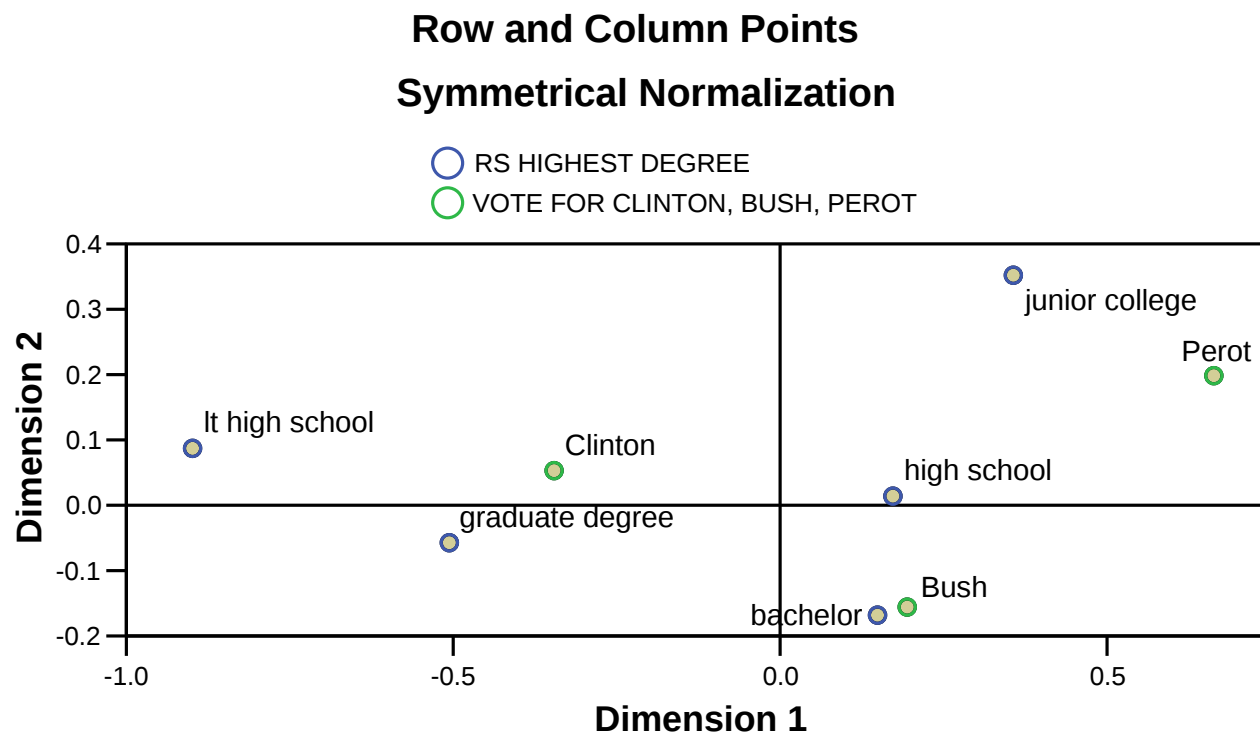
* Symmetrical normalization



Column Points for RS HIGHEST DEGREE



- ❑ 可以发现研究生层次的选民（Graduate degree）倾向于具有实干精神的Clinton，而较Clinton更为激进的Bush更受high school和Bachelor层次的选民欢迎，Perot仅和junior college层次的选民较近。



疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回复问题

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

