Correspondence Analysis

对 应 分 析

肖 磊，2026 年 5 月 26 日

Outline

- 对应分析 (Correspondence Analysis)
  - 动机 (Motivations)
  - χ^2 分解 (Chi-Square Decomposition)
  - 对应分析的应用 (Correspondence Analysis in Practice)

Preface 引言

- 对应分析是用于分析列联表的行和列之间关联关系的一种工具.
  - 列联表：两个定性变量的联合频数表.

		是否吸烟：是	是否吸烟：否
性别	男	n_{11}	n_{12}
	女	n_{21}	n_{22}

  - 列联表：更一般地，可以是一个 $n \times p$ 的频数表.

条目二 \ 条目一	A_1	A_2	$\cdots$	A_k
B_1	n_{11}	n_{12}	$\cdots$	n_{1k}
B_2	n_{21}	n_{22}	$\cdots$	n_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
B_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell k}$

Preface 引言

- 对应分析的主要思想: 建立简单的索引 (或指数) 以展示行和列的类别之间的关系.
  - 这些指数将同时告诉我们，哪些列在行类别中具有更大的权重，哪些行在列类别中具有更大的权重.
  - 类似于第 11 章中介绍的主成分分析，以及第 10 章中讨论的数据矩阵的因子分解，对应分析也涉及到降低表格维度的问题.
  - 其想法是按重要性递减的顺序提取指标，以便可以在较小维度的空间中汇总表格的主要信息.
  - 例如，如果只使用两个因子 (指标)，就可以在二维图中绘制其结果，以显示表的行和列之间的关系.

Motivation 动机

- 对应分析的目的：建立简单的指标以展示行和列的类别之间的关系.
  - 许多情形中，描述两个变量之间的关联关系，列联表是一种非常有用的工具.
  - 两个变量可以是定性 (qualitative) 变量或名义 (nominal) 变量

变量二的状态 (modality): $B_1 \quad B_2 \quad \cdots \quad B_p$

表格: $\mathcal{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{matrix}$ 变量一的状态 (modality)

x_{ij} : 观测结果同时位于状态 $A_i \bigcap B_j$ 中的数量

$i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, p.$

  - 两个变量也可以是离散型变量.
  - 两个变量还可以是连续型变量. 例如通过落在某个区间与否来定义不同状态.
  - 对应分析的应用领域十分广泛.

Motivation 动机

- 对应分析得到的关于表 $\mathcal{X}$ 的行和列之间的图形关系，是基于表示所有的行和列的状态，并根据对应于列和行的权重来解释数据点的相对位置的这样一种思想.
  - 通过推导能够给出每一行以及每一列坐标的简单指标系统来实现.
  - 将这些行与列的坐标同时表现在一张图中. 就可以清楚地看到表的行当中哪些列更重要，也可以清楚地看到表的列当中哪些行更重要.

- 指标的构建是基于与主成分分析类似的思想.
  - 使用主成分分析，总的方差被划分为来自相互独立的主成分的贡献.
  - 对应分析则分解的是一种相关性的度量，常用的是独立性检验的总 χ^2 值，而非对总的方差进行分解.

Motivation 动机

- Example: 数据包含 1976 年法国中学会考的 202,100 个观测结果，按照不同的状态或变量 (报考专业、所在地区) 统计出了相应的频数.

报考专业:

$$\begin{cases} X_1 : \mathrm{A} & \text{Philosophy-Letters} \\ X_2 : \mathrm{B} & \text{Economics and Social Sciences} \\ X_3 : \mathrm{C} & \text{Mathematics and Physics} \\ X_4 : \mathrm{D} & \text{Mathematics and Natural Sciences} \\ X_5 : \mathrm{E} & \text{Mathematics and Techniques} \\ X_6 : \mathrm{F} & \text{Industrial Techniques} \\ X_7 : \mathrm{G} & \text{Economic Techniques} \\ X_8 : \mathrm{H} & \text{Computer Techniques} \end{cases}$$

Motivation 动机

- Example: 数据包含 1976 年法国中学会考的 202,100 个观测结果，按照不同的状态或变量 (报考专业、所在地区) 统计出了相应的频数.

22 个大区:

NOPC	:	Nord-Pas-de-Calais 北部加莱海峡
PICA	:	Picardie 皮卡第
HNOR	:	Normandie 诺曼底
ILDF	:	Ile de France 巴黎大区
CHAM	:	Champagne Ardenne 香槟-阿登
LORR	:	Lorraine 洛林
ALSA	:	Alsace 阿尔萨斯
BRET	:	Bretagne 布列塔尼
PAYL	:	Pays de loire 卢瓦尔河地区
CENT	:	Centre Val de loire 中央-卢瓦河山谷
BOUR	:	Bourgogne 勃艮第
FRAC	:	Franche comte 佛朗什孔泰
PCHA	:	Poitou Charentes 普瓦图-夏郎德
LIMO	:	Limousin 利木赞
AUVE	:	Auvergne 奥佛涅
RHOA	:	Rhone Alpes 罗讷-阿尔卑斯
BNOR	:	Basse-Normandie 下诺曼底
AQUI	:	Aquitaine 阿基坦
MIDI	:	Midi pyrenees 南部-比利牛斯
PROV	:	Provence Alpes cote d'azur 普罗旺斯-阿尔卑斯南部，蓝色海岸
LARO	:	Languedoc Roussilon 朗格多克-鲁西永
CORS	:	corse 科西嘉

Motivation 动机

- Example: 数据包含 1976 年法国中学会考的 202,100 个观测结果，按照不同的状态或变量 (报考专业、所在地区) 统计出了相应的频数.

```
library(readr)
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with data/Data")
bac = read_csv("bac.csv", col_names = FALSE)
x = as.data.frame(bac)
colnames(x) = c("Region", "A", "B", "C", "D", "E", "F", "G", "H")
x
```

列联表 (contingency table)

```
> x
   Region    A    B    C    D   E    F    G  H
1    ILDF 9724 5650 8679 9432 839 3353 5355 83
2    CHAM  924  464  567  984 132  423  736 12
3    PICA 1081  490  830 1222 118  410  743 13
4    HNOR 1135  587  686  904  83  629  813 13
5    CENT 1482  667 1020 1535 173  629  989 26
6    BNOR 1033  509  553 1063 100  433  742 13
7    BOUR 1272  527  861 1116 219  769 1232 13
8    NOPC 2549 1141 2164 2752 587 1660 1951 41
9    LORR 1828  681 1364 1741 302 1289 1683 15
10   ALSA 1076  443  880 1121 145  917 1091 15
11   FRAC  827  333  481  892 137  451  618 18
12   PAYL 2213  809 1439 2623 269  990 1783 14
13   BRET 2158 1271 1633 2352 350  950 1509 22
14   PCHA 1358  503  639 1377 164  495  959 10
15   AQUI 2757  873 1466 2296 215  789 1459 17
16   MIDI 2493 1120 1494 2329 254  885 1565 28
17   LIMO  551  297  386  663  67  334  378 12
18   RHOA 3951 2127 3218 4743 545 2072 3018 36
19   AUVE 1066  579  724 1239 126  476  649 12
20   LARO 1844  816 1154 1839 156  469  993 16
21   PROV 3944 1645 2415 3616 343 1236 2404 22
22   CORS  327   31   85  178   9   27   79  0
```

- 问题：是否某些地区更倾向于报考某种类型的学士学位?

Motivation 动机

- Example: 数据包含 1976 年法国中学会考的 202,100 个观测结果，按照不同的状态或变量 (报考专业、所在地区) 统计出了相应的频数.
  - 比如洛林 (LORR) 大区，报考专业的百分比为

```
x.LORR = x[9, 2:9] # LORR 大区的数据
x.LORR
ratio.LORR = x.LORR / (sum(x.LORR)) * 100 # 计算 LORR 大区报考方向的百分比
round(ratio.LORR, digits = 1) # 显示结果
```

```
> x.LORR
     A   B    C    D   E    F    G  H
9 1828 681 1364 1741 302 1289 1683 15
```

```
> round(ratio.LORR, digits = 1) # 显示结果
    A   B    C    D   E    F    G   H
9 20.5 7.6 15.3 19.6 3.4 14.5 18.9 0.2
```

  - 全部 22 个大区报考专业的百分比为

```
ratio = 100 * apply(x[, 2:9], 2, sum) / sum(x[, 2:9]) # 计算各列的百分比
round(ratio, digits = 1) # 显示结果
```

```
> round(ratio, digits = 1) # 显示结果
   A    B    C    D   E   F    G   H
22.6 10.7 16.2 22.8 2.6 9.7 15.2 0.2
```

  - 相对于学位类型的总频数而言，我们可能会认为洛林 (LORR) 大区更喜欢 E, F, G 类型，而不喜欢 A, B, C, D 类型.

Motivation 动机

- 在对应分析中，我们尝试为这些地区建立一个指标，以便只用一个数来衡量这种代表性过高或过低的情形.
  - 同时，我们还要尝试对地区赋以权重，这样我们就可以看到在哪个地区，某些类型的学位是首选的.

Motivation 动机

- Example: 考虑位于 p 个地方的 n 种类型的公司. 是否有某种类型的公司更喜欢位于某个地点呢? 或者换句话说, 是否能建立与某一类型的公司相对应的地点指数呢?

$$\mathcal{X} = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{matrix} \longleftarrow \text{Finance} \\ \longleftarrow \text{Energy} \\ \longleftarrow \text{HiTech} \end{matrix} = (x_{ij})_{n\times p}$$

$$\begin{matrix} \uparrow \text{Frankfurt} \\ \uparrow \text{Berlin} \\ \uparrow \text{Munich} \end{matrix}$$

$$x_{\bullet j} = \sum_{i=1}^{n} x_{ij}$$

  - 假设存在一个 (关于公司的) 权重向量 $\boldsymbol{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$, 我们可以定义地点的指数 s_j 为 $s_j = c \sum_{i=1}^{n} r_i \frac{x_{ij}}{x_{\bullet j}}$. (常数)
  - 例如 $s_1 = c\left(r_1 \frac{x_{11}}{x_{\bullet 1}} + r_2 \frac{x_{21}}{x_{\bullet 1}} + \cdots + r_n \frac{x_{n1}}{x_{\bullet 1}}\right)$ 给出的是各类公司在地点 1 (法兰克福) 的 (按照 $\boldsymbol{r}$) 加权平均频数.

Motivation 动机

- Example: 考虑位于 p 个地方的 n 种类型的公司. 是否有某种类型的公司更喜欢位于某个地点呢? 或者换句话说，是否能建立与某一类型的公司相对应的地点指数呢?

$$\mathcal{X} = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{matrix} \longleftarrow \text{Finance} \\ \longleftarrow \text{Energy} \\ \longleftarrow \text{HiTech} \end{matrix} = \left(x_{ij}\right)_{n\times p}$$

$$\begin{matrix} \uparrow \text{Frankfurt} \\ \uparrow \text{Berlin} \\ \uparrow \text{Munich} \end{matrix}$$

$$x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$$

  - 如果给定地点的权重向量 $\boldsymbol{s}^* = \begin{pmatrix} s_1^* \\ s_2^* \\ \vdots \\ s_p^* \end{pmatrix}$，同样方式我们可以定义公司的指数 $r_i^* = c^* \sum_{j=1}^{p} s_j^* \frac{x_{ij}}{x_{i\bullet}}$. （常数）
  - 例如 $r_2^* = c^* \left(s_1^* \frac{x_{21}}{x_{2\bullet}} + s_2^* \frac{x_{22}}{x_{2\bullet}} + \cdots + s_p^* \frac{x_{2p}}{x_{2\bullet}} \right)$ 给出的是能源公司在不同地点的 (按照 $\boldsymbol{s}^*$) 加权平均频数.

Motivation 动机

- 对于 $\begin{cases} r_i = c^* \sum_{j=1}^{p} s_j \dfrac{x_{ij}}{x_{i\bullet}}, & i = 1, 2, \ldots, n \\ s_j = c \sum_{i=1}^{n} r_i \dfrac{x_{ij}}{x_{\bullet j}}, & j = 1, 2, \ldots, p \end{cases} \Longrightarrow \boldsymbol{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}, \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_p \end{pmatrix}$

  - 如果能同时求得 “行权重” 向量 $\boldsymbol{r}$ 和 “列权重” 向量 $\boldsymbol{s}$，我们可以将每一行的状态 r_i, $i = 1, 2, \ldots, n$ 和每一列的状态 s_j, $j = 1, 2, \ldots, p$ 表示在一维图中.
  - 如果图中 r_i 和 s_j 非常接近 (且远离原点)，这表明在 $s_j = c \sum_{i=1}^{n} r_i \dfrac{x_{ij}}{x_{\bullet j}}$ 中第 i 行的状态具有较为重要的条件频率 $\dfrac{x_{ij}}{x_{\bullet j}}$，同时在 $r_i = c^* \sum_{j=1}^{p} s_j \dfrac{x_{ij}}{x_{i\bullet}}$ 中第 j 列的状态具有较为重要的条件频率 $\dfrac{x_{ij}}{x_{i\bullet}}$.
  - 这说明第 i 行和第 j 列之间具有正的关联性.

Motivation 动机

- 对于 $\begin{cases} r_i = c^* \sum_{j=1}^{p} s_j \dfrac{x_{ij}}{x_{i\bullet}}, & i = 1, 2, \ldots, n \\ s_j = c \sum_{i=1}^{n} r_i \dfrac{x_{ij}}{x_{\bullet j}}, & j = 1, 2, \ldots, p \end{cases} \Longrightarrow \boldsymbol{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}, \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_p \end{pmatrix}$

  - 如果能同时求得 “行权重” 向量 $\boldsymbol{r}$ 和 “列权重” 向量 $\boldsymbol{s}$，我们可以将每一行的状态 r_i, $i = 1, 2, \ldots, n$ 和每一列的状态 s_j, $j = 1, 2, \ldots, p$ 表示在一维图中.

  - 如果图中 r_i 和 s_j 非常遥远 (且远离原点)，我们也可以进行类似的分析.

  - 这说明条件频率的贡献较小，或者第 i 行和第 j 列之间具有负的关联性.

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
- 二维列联表的独立性的 χ^2 检验包含两步：
  - 第一步，在独立性假设下，估计每一个单元的期望值.
  - 第二步，利用下述统计量对相应的观测值与期望值进行比较

x_{ij} 是 $(i,\ j)$ 单元的观测频数

$$t=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(x_{ij}-E_{ij}\right)^2}{E_{ij}}$$

E_{ij} 是在独立假设下 $(i,\ j)$ 单元相应期望频数的估计值

$$E_{ij}=\frac{x_{i\bullet}x_{\bullet j}}{x_{\bullet\bullet}}$$

$$x_{\bullet\bullet}=\sum_{i=1}^{n}x_{i\bullet}$$

- 在 H_0：独立性假设为真的前提下：

$$t\sim\chi^2_{(n-1)(p-1)}$$

- H_0 的拒绝域：

$$t>\chi^2_{(n-1)(p-1)}(\alpha)$$

上 α 分位数

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
- 二维列联表的独立性的 χ^2 检验包含两步：

$$x = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{matrix} \longleftarrow \text{Finance} \\ \longleftarrow \text{Energy} \\ \longleftarrow \text{HiTech} \end{matrix}$$

↑ Frankfurt ↑ Berlin ↑ Munich

```
x = data.frame(Frankfurt = c(4, 0, 1), Berlin = c(0, 1, 1), Munich = c(2, 1, 4))
row.names(x) = c("Finance", "Energy", "HiTech")
x1 = cbind(x, RowSum = rowSums(x))
X = rbind(x1, RowSum = colSums(x1))
X
```

```
> X
          Frankfurt Berlin Munich RowSum
Finance           4      0      2      6
Energy            0      1      1      2
HiTech            1      1      4      6
RowSum            5      2      7     14
```

```
t = 0
for (i in 1:3){
  for (j in 1:3) t = t + (X[i, j] - X[i, 4] * X[4, j] / X[4, 4])^2 / (X[i, 4] * X[4, j] / X[4, 4])
}
t # 检验统计量的值
```

```
> t
[1] 6.266667
```

<

```
> qchisq(0.05, df = 4, lower.tail = FALSE) # 检验的临界值
[1] 9.487729
```

```
qchisq(0.05, df = 4, lower.tail = FALSE) # 检验的临界值
pchisq(t, 4, lower.tail = FALSE) # 检验的 p 值
```

```
> pchisq(t, 4, lower.tail = FALSE) # 检验的 p 值
[1] 0.1800989
```

> 0.05

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
- 二维列联表的独立性的 χ^2 检验包含两步：

```
x = data.frame(Frankfurt = c(4, 0, 1), Berlin = c(0, 1, 1), Munich = c(2, 1, 4))
row.names(x) = c("Finance", "Energy", "HiTech")
x1 = cbind(x, RowSum = rowSums(x))
X = rbind(x1, RowSum = colSums(x1))
X
```

```
> X
         Frankfurt Berlin Munich RowSum
Finance          4      0      2      6
Energy           0      1      1      2
HiTech           1      1      4      6
RowSum           5      2      7     14
```

$$x = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{matrix} \longleftarrow \text{Finance} \\ \longleftarrow \text{Energy} \\ \longleftarrow \text{HiTech} \end{matrix}$$

↑ Frankfurt ↑ Berlin ↑ Munich

- 虽然不拒绝 H_0：独立性假设为真，但 p 值接近 0.05，因此有必要考察偏离独立性的特殊原因.

```
> t
[1] 6.266667
```

<

```
> qchisq(0.05, df = 4, lower.tail = FALSE) # 检验的临界值
[1] 9.487729
```

```
qchisq(0.05, df = 4, lower.tail = FALSE) # 检验的临界值
pchisq(t, 4, lower.tail = FALSE) # 检验的 p 值
```

```
> pchisq(t, 4, lower.tail = FALSE) # 检验的 p 值
[1] 0.1800989
```

> 0.05

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - χ^2 分解方法包括求解矩阵 $\mathscr{C} = \left(c_{ij}\right)_{n\times p}$ 的奇异值分解 (SVD)，其中

$$c_{ij} = \frac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, p$$

  - c_{ij} 可以看作是在独立性假设下，观测值 x_{ij} 与理论值 E_{ij} (加权) 偏差的度量.
  - 我们可以使用第 10 章中的因子分解方法来描述矩阵 $\mathscr{C}$ 的行与列.

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - 为简单起见，定义矩阵 $\mathscr{A}_{n\times n}$ 和 $\mathscr{B}_{p\times p}$ 如下

$$\mathscr{A} = \operatorname{diag}\left(x_{1\bullet}, x_{2\bullet}, \ldots, x_{n\bullet}\right) = \begin{pmatrix} x_{1\bullet} & & & \\ & x_{2\bullet} & & \\ & & \ddots & \\ & & & x_{n\bullet} \end{pmatrix}$$

$$\mathscr{B} = \operatorname{diag}\left(x_{\bullet 1}, x_{\bullet 2}, \ldots, x_{\bullet p}\right) = \begin{pmatrix} x_{\bullet 1} & & & \\ & x_{\bullet 2} & & \\ & & \ddots & \\ & & & x_{\bullet p} \end{pmatrix}$$

  - 这两个矩阵可以给出边缘行频数 $\boldsymbol{a}_{n\times 1}$ 和边缘列频数 $\boldsymbol{b}_{p\times 1}$：

$$\boldsymbol{a} = \mathscr{A}\mathbf{1}_n\,, \quad \boldsymbol{b} = \mathscr{B}\mathbf{1}_p$$

  - 易证：

$$\mathscr{C}\sqrt{\boldsymbol{b}} = \mathscr{C}\begin{pmatrix} \sqrt{x_{\bullet 1}} \\ \sqrt{x_{\bullet 2}} \\ \vdots \\ \sqrt{x_{\bullet p}} \end{pmatrix} = \mathbf{0}\,, \quad \mathscr{C}^{\mathrm{T}}\sqrt{\boldsymbol{a}} = \mathscr{C}^{\mathrm{T}}\begin{pmatrix} \sqrt{x_{1\bullet}} \\ \sqrt{x_{2\bullet}} \\ \vdots \\ \sqrt{x_{n\bullet}} \end{pmatrix} = \mathbf{0}$$

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - 对矩阵 $\mathscr{C} = \left(c_{ij}\right)_{n\times p}$ 做奇异值分解 (SVD)，$c_{ij} = \dfrac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}}$，$i = 1, 2, \ldots, n$；$j = 1, 2, \ldots, p$.
  - 用 R 表示矩阵 $\mathscr{C}$ 的秩：$R = \text{rank}(\mathscr{C}) \le \min\left\{(n-1), (p-1)\right\}$.

$$\mathscr{C} = \Gamma\Lambda\Delta^{\mathrm{T}}$$

- **Theorem 2.2** (Singular Value Decomposition)

  Each matrix $\mathscr{A}_{n\times p}$ with rank r can be decomposed as

  $$\mathscr{A} = \Gamma\ \Lambda\ \Delta^{\mathrm{T}}$$

  where $\Gamma_{n\times r}$ and $\Delta_{p\times r}$. Both Γ and Δ are column orthogonal, i.e.,

  $$\Gamma^{\mathrm{T}}\Gamma = \Delta^{\mathrm{T}}\Delta = \boldsymbol{I}_r$$

  and

  $\Lambda = \text{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_r\right)$

  the values $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the none zero eigenvalues of the matrices $\mathscr{A}\mathscr{A}^{\mathrm{T}}$ and $\mathscr{A}^{\mathrm{T}}\mathscr{A}$. Γ and Δ consist of the corresponding r eigenvectors of these matrices.

Γ 是 $\mathscr{C}\mathscr{C}^{\mathrm{T}}$ 的特征向量构成的矩阵

$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_R$ 是 $\mathscr{C}\mathscr{C}^{\mathrm{T}}$ 的特征值

Δ 是 $\mathscr{C}^{\mathrm{T}}\mathscr{C}$ 的特征向量构成的矩阵

$$\Lambda = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_R}\right)$$

$$c_{ij} = \sum_{k=1}^{R} \sqrt{\lambda_k}\ \gamma_{ik}\ \delta_{jk}$$

独立性的 χ^2 检验统计量:

$$t = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(x_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}^2 = \text{tr}\left(\mathscr{C}\mathscr{C}^{\mathrm{T}}\right) = \sum_{k=1}^{R}\lambda_k$$

$$= \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_R} \end{pmatrix}$$

- 上式说明 $\mathscr{C}$ 的奇异值分解，分解的是总的 χ^2 值，而非第 10 章中的总方差.

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - 利用第 10 章介绍的行与列之间的双向关系，我们有

$$\begin{cases} \boldsymbol{\delta}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{C}^{\mathrm{T}} \boldsymbol{\gamma}_k \\ \boldsymbol{\gamma}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{C} \boldsymbol{\delta}_k \end{cases}, \qquad k = 1, 2, \ldots, R$$

  - 矩阵 $\mathscr{C}$ 的行与列的投影则分别为

$$\begin{cases} \mathscr{C} \boldsymbol{\delta}_k = \sqrt{\lambda_k} \boldsymbol{\gamma}_k \\ \mathscr{C}^{\mathrm{T}} \boldsymbol{\gamma}_k = \sqrt{\lambda_k} \boldsymbol{\delta}_k \end{cases}, \qquad k = 1, 2, \ldots, R$$

  - 注意到，特征向量满足

$$\boldsymbol{\delta}_k^{\mathrm{T}} \sqrt{\boldsymbol{b}} = \mathbf{0}, \qquad \boldsymbol{\gamma}_k^{\mathrm{T}} \sqrt{\boldsymbol{a}} = \mathbf{0}$$

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - 由于

$$c_{ij} = \sum_{k=1}^{R} \sqrt{\lambda_k}\ \gamma_{ik}\ \delta_{jk}$$

  - 在行与列的对应分析中，特征向量 $\boldsymbol{\delta}_k$ 与 $\boldsymbol{\gamma}_k$ 是我们需要的量.
  - 如果假设首个特征值占有统治地位，则有

$$c_{ij} \approx \sqrt{\lambda_1}\ \gamma_{i1}\ \delta_{j1}$$

    - 此时如果坐标 γ_{i1} 与 δ_{j1} 相比其他坐标都较大 (且同号)，那么 c_{ij} 也会较大，表明列联表当中第 i 行的状态与第 j 列的状态之间呈现正的关联性.
    - 如果坐标 γ_{i1} 与 δ_{j1} 都较大且符号相反，则表明列联表当中第 i 行的状态与第 j 列的状态之间呈现负的关联性.

Chi-Square Decomposition χ^2 分解

- 行与列各状态间关联性的另一种度量方法是对 χ^2 检验统计量的值进行分解.
  - 在很多应用中，前两个特征值 λ_1 与 λ_2 会占据统治地位，由特征向量 $\boldsymbol{\gamma}_1$ 与 $\boldsymbol{\gamma}_2$ 以及 $\boldsymbol{\delta}_1$ 与 $\boldsymbol{\delta}_2$ 解释的总 χ^2 的比例会较大.
    - 此时，$\begin{cases} \mathscr{C}\boldsymbol{\delta}_k = \sqrt{\lambda_k}\boldsymbol{\gamma}_k \\ \mathscr{C}^{\mathrm{T}}\boldsymbol{\gamma}_k = \sqrt{\lambda_k}\boldsymbol{\delta}_k \end{cases}$，$(k = 1, 2)$ 与 $\left(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\right)$ 可以用来得到列联表的 n 行的一个图形表示，类似地，与 $\left(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2\right)$ 也可以用来得到列联表的 p 列的一个图形表示. 行点和列点之间的接近度的解释将被解释为如上关于10所述
    - 行与列的点之间接近程度的解释，与上述关于 $c_{ij} = \sum_{k=1}^{R} \sqrt{\lambda_k}\ \gamma_{ik}\ \delta_{jk}$ 的解释相同. 后面的例子会详细说明.

Chi-Square Decomposition χ^2 分解

- 在对应分析中，我们使用矩阵 $\mathscr{C}$ 的行加权之后的投影，以及 $\mathscr{C}$ 的列加权之后的投影进行图形表示.

  - 用 $\boldsymbol{r}_k$ $(n \times 1)$ 表示 $\mathscr{A}^{-1/2}\mathscr{C}$ 在 $\boldsymbol{\delta}_k$ 上的投影

    用 $\boldsymbol{s}_k$ $(p \times 1)$ 表示 $\mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}$ 在 $\boldsymbol{\gamma}_k$ 上的投影

$$\mathscr{A} = \operatorname{diag}\left(x_{1\bullet}, x_{2\bullet}, \ldots, x_{n\bullet}\right)$$

$$\left(k = 1, 2, \ldots, R\right)$$

$$\mathscr{B} = \operatorname{diag}\left(x_{\bullet 1}, x_{\bullet 2}, \ldots, x_{\bullet p}\right)$$

$$\boldsymbol{r}_k = \mathscr{A}^{-1/2}\mathscr{C}\boldsymbol{\delta}_k = \sqrt{\lambda_k}\mathscr{A}^{-1/2}\boldsymbol{\gamma}_k$$

$$\boldsymbol{s}_k = \mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}\boldsymbol{\gamma}_k = \sqrt{\lambda_k}\mathscr{B}^{-1/2}\boldsymbol{\delta}_k$$

$$\boldsymbol{a} = \mathscr{A}\mathbf{1}_n = \begin{pmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \vdots \\ x_{n\bullet} \end{pmatrix}$$

满足

$$\begin{cases} \boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{a} = 0 \\ \boldsymbol{s}_k^{\mathrm{T}}\boldsymbol{b} = 0 \end{cases}$$

$$\boldsymbol{b} = \mathscr{B}\mathbf{1}_p = \begin{pmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ \vdots \\ x_{\bullet p} \end{pmatrix}$$

  - 在每一个轴 $k = 1, 2, \ldots, R$ 上得到的投影值都以零为中心，行坐标 $\boldsymbol{r}_k$ 的权重由 $\boldsymbol{a}$ ($\mathscr{X}$的行的边缘频数)给出，列坐标 $\boldsymbol{s}_k$ 的权重由 $\boldsymbol{b}$ ($\mathscr{X}$的列的边缘频数)给出.

而特征向量满足的是 $\boldsymbol{\delta}_k^{\mathrm{T}}\sqrt{\boldsymbol{b}} = \mathbf{0}$, $\boldsymbol{\gamma}_k^{\mathrm{T}}\sqrt{\boldsymbol{a}} = \mathbf{0}$.

Chi-Square Decomposition χ^2 分解

- 在对应分析中，我们使用矩阵 $\mathscr{C}$ 的行加权之后的投影，以及 $\mathscr{C}$ 的列加权之后的投影进行图形表示.

  - 用 $\boldsymbol{r}_k$ $(n \times 1)$ 表示 $\mathscr{A}^{-1/2}\mathscr{C}$ 在 $\boldsymbol{\delta}_k$ 上的投影

    用 $\boldsymbol{s}_k$ $(p \times 1)$ 表示 $\mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}$ 在 $\boldsymbol{\gamma}_k$ 上的投影

$$\mathscr{A} = \operatorname{diag}\left(x_{1\bullet}, x_{2\bullet}, \ldots, x_{n\bullet}\right)$$

$$\mathscr{B} = \operatorname{diag}\left(x_{\bullet 1}, x_{\bullet 2}, \ldots, x_{\bullet p}\right)$$

$$\left(k = 1, 2, \ldots, R\right)$$

$$\boldsymbol{r}_k = \mathscr{A}^{-1/2}\mathscr{C}\boldsymbol{\delta}_k = \sqrt{\lambda_k}\mathscr{A}^{-1/2}\boldsymbol{\gamma}_k$$

$$\boldsymbol{s}_k = \mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}\boldsymbol{\gamma}_k = \sqrt{\lambda_k}\mathscr{B}^{-1/2}\boldsymbol{\delta}_k$$

满足 $\begin{cases} \boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{a} = 0 \\ \boldsymbol{s}_k^{\mathrm{T}}\boldsymbol{b} = 0 \end{cases}$

$$\boldsymbol{a} = \mathscr{A}\mathbf{1}_n = \begin{pmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \vdots \\ x_{n\bullet} \end{pmatrix}$$

$$\boldsymbol{b} = \mathscr{B}\mathbf{1}_p = \begin{pmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ \vdots \\ x_{\bullet p} \end{pmatrix}$$

  - 因此，原点是所有表示的重心.

  - 以及 $\mathscr{C}$ 的奇异值分解，我们有

$$\boldsymbol{r}_k^{\mathrm{T}}\mathscr{A}\boldsymbol{r}_k = \lambda_k$$

$$\boldsymbol{s}_k^{\mathrm{T}}\mathscr{B}\boldsymbol{s}_k = \lambda_k$$

Chi-Square Decomposition χ^2 分解

- 由 $\boldsymbol{\delta}_k$ 与 $\boldsymbol{\gamma}_k$ 的双向关系

$$\begin{cases} \boldsymbol{\delta}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{C}^{\mathrm{T}} \boldsymbol{\gamma}_k \\ \boldsymbol{\gamma}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{C} \boldsymbol{\delta}_k \end{cases}, \quad (k = 1, 2, \ldots, R)$$

我们来推导 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 的双向关系:

$$\boldsymbol{r}_k = \mathscr{A}^{-1/2}\mathscr{C}\boldsymbol{\delta}_k = \frac{1}{\sqrt{\lambda_k}}\mathscr{A}^{-1/2}\mathscr{C}\ \mathscr{C}^{\mathrm{T}}\boldsymbol{\gamma}_k = \frac{1}{\lambda_k}\mathscr{A}^{-1/2}\mathscr{C}\ \mathscr{C}^{\mathrm{T}}\ \mathscr{C}\ \boldsymbol{\delta}_k$$

$$= \frac{1}{\lambda_k}\mathscr{A}^{-1/2}\mathscr{C}\left(\lambda_k\boldsymbol{\delta}_k\right) = \mathscr{A}^{-1/2}\mathscr{C}\boldsymbol{\delta}_k = \mathscr{A}^{-1/2}\mathscr{C}\left(\mathscr{B}^{1/2}\mathscr{B}^{-1/2}\right)\boldsymbol{\delta}_k$$

$$= \frac{1}{\sqrt{\lambda_k}}\mathscr{A}^{-1/2}\mathscr{C}\mathscr{B}^{1/2}\boldsymbol{s}_k$$

$\boldsymbol{\delta}_k$ 是 $\mathscr{C}^{\mathrm{T}}\mathscr{C}$ 的特征向量

$$\boldsymbol{r}_k = \mathscr{A}^{-1/2}\mathscr{C}\boldsymbol{\delta}_k = \sqrt{\lambda_k}\mathscr{A}^{-1/2}\boldsymbol{\gamma}_k$$

$$\boldsymbol{s}_k = \mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}\boldsymbol{\gamma}_k = \sqrt{\lambda_k}\mathscr{B}^{-1/2}\boldsymbol{\delta}_k$$

- 同理可得: $\boldsymbol{s}_k = \dfrac{1}{\sqrt{\lambda_k}}\mathscr{B}^{-1/2}\mathscr{C}^{\mathrm{T}}\mathscr{A}^{1/2}\boldsymbol{r}_k$

Chi-Square Decomposition χ^2 分解

- 进一步化简可得

$$\begin{cases} \boldsymbol{r}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{A}^{-1/2} \mathscr{C} \mathscr{B}^{1/2} \boldsymbol{s}_k \\ \boldsymbol{s}_k = \dfrac{1}{\sqrt{\lambda_k}} \mathscr{B}^{-1/2} \mathscr{C}^{\mathrm{T}} \mathscr{A}^{1/2} \boldsymbol{r}_k \end{cases} \Longrightarrow \begin{cases} \boldsymbol{r}_k = \sqrt{\dfrac{x_{\bullet\bullet}}{\lambda_k}}\ \mathscr{A}^{-1} \mathscr{X} \boldsymbol{s}_k \\ \boldsymbol{s}_k = \sqrt{\dfrac{x_{\bullet\bullet}}{\lambda_k}}\ \mathscr{B}^{-1} \mathscr{X}^{\mathrm{T}} \boldsymbol{r}_k \end{cases}$$

- 这些向量还同时满足我们一开始给出的:

$$\begin{cases} \boldsymbol{r}_k = c^* \sum_{j=1}^{p} \left(\boldsymbol{s}_j \dfrac{x_{kj}}{x_{k\bullet}} \right) \\ \boldsymbol{s}_k = c \sum_{i=1}^{n} \left(\boldsymbol{r}_i \dfrac{x_{ik}}{x_{\bullet k}} \right) \end{cases}, \qquad k = 1, 2, \ \ldots, R$$

Chi-Square Decomposition χ^2 分解

- 如同第 10 章中一样，向量 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 也可以看作是因子 (分别为行因子和列因子).
  - 它们的均值为:

$$\begin{cases} \bar{\boldsymbol{r}}_k = \dfrac{1}{x_{\bullet\bullet}} \boldsymbol{r}_k^{\mathrm{T}} \boldsymbol{a} = 0 \\ \bar{\boldsymbol{s}}_k = \dfrac{1}{x_{\bullet\bullet}} \boldsymbol{s}_k^{\mathrm{T}} \boldsymbol{b} = 0 \end{cases}$$

  - 它们的方差为:

$$\begin{cases} \operatorname{Var}\left(\boldsymbol{r}_k\right) = \dfrac{1}{x_{\bullet\bullet}} \sum_{i=1}^{n} \boldsymbol{x}_{i\bullet} r_{ki}^2 = \dfrac{1}{x_{\bullet\bullet}} \boldsymbol{r}_k^{\mathrm{T}} \mathscr{A} \boldsymbol{r}_k = \dfrac{\lambda_k}{x_{\bullet\bullet}} \\ \operatorname{Var}\left(\boldsymbol{s}_k\right) = \dfrac{1}{x_{\bullet\bullet}} \sum_{j=1}^{p} \boldsymbol{x}_{\bullet j} s_{kj}^2 = \dfrac{1}{x_{\bullet\bullet}} \boldsymbol{s}_k^{\mathrm{T}} \mathscr{B} \boldsymbol{s}_k = \dfrac{\lambda_k}{x_{\bullet\bullet}} \end{cases}$$

  - 因此，χ^2 统计量 t 分解后的第 k 个因子的比例 $\dfrac{\lambda_k}{\sum_{j=1}^{R} \lambda_j}$，也可以理解为方差由因子 k 解释的比例.

Chi-Square Decomposition χ^2 分解

- 如同第 10 章中一样，向量 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 也可以看作是因子 (分别为行因子和列因子).
  - 定义第 i 行对于因子 $\boldsymbol{r}_k$ 的绝对贡献为:

$$C_a\left(i, \boldsymbol{r}_k\right) = \frac{x_{i\bullet}\, r_{ki}^2}{\lambda_k}, \quad i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, R$$

它们表示了哪些行的状态在第 k 个行因子的变化当中最为重要.

  - 类似地，定义第 j 列对于列因子 $\boldsymbol{s}_k$ 的绝对贡献为:

$$C_a\left(j, \boldsymbol{s}_k\right) = \frac{x_{\bullet j}\, s_{kj}^2}{\lambda_k}, \quad j = 1, 2, \ldots, p; \quad k = 1, 2, \ldots, R$$

  - 对于利用对应分析获得的图形，这些绝对贡献可能有助于我们进行具体解释.

Correspondence Analysis in Practice 实践中的对应分析

- $\mathcal{X}$ 的 n 行和 p 列在坐标轴 $k = 1, 2, \ldots, R$ 上的图形表示由 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 的元素给出.
  - 如果前两个因子解释方差的累积比例 $\Psi_2 = \dfrac{\lambda_1 + \lambda_2}{\sum_{k=1}^{R} \lambda_k}$ 足够大，那么二维图形的表现就足够好了.
- 图形的解释可总结如下:

① 接近的两行 (两列) 表示这两行 (两列) 的轮廓相似，其中 “轮廓” 是指一行 (列) 的条件频率分布；这两行 (列) 几乎成比例. 当两行 (两列) 相距较远时，则需要做出相反的解释.

Correspondence Analysis in Practice 实践中的对应分析

- $\mathcal{X}$ 的 n 行和 p 列在坐标轴 $k = 1, 2, \ldots, R$ 上的图形表示由 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 的元素给出.
  - 如果前两个因子解释方差的累积比例 $\Psi_2 = \dfrac{\lambda_1 + \lambda_2}{\sum_{k=1}^{R} \lambda_k}$ 足够大，那么二维图形的表现就足够好了.

- 图形的解释可总结如下:

② 某一特定的行接近于某一特定的列，表明该行 (列) 在该列 (行) 中具有特别重要的权重. 与此相反，与某一特定的列相距甚远的一行，则表示该列中几乎没有该行的观测结果 (反之亦然). 当然，当这些点远离 0 时，上述结论尤为正确.

③ 原点代表因子 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 的平均值. 于是，某一特定的点 (行或列) 的投影接近于原点，说明其表现具有平均的效果.

Correspondence Analysis in Practice 实践中的对应分析

- $\mathcal{X}$ 的 n 行和 p 列在坐标轴 $k = 1, 2, \ldots, R$ 上的图形表示由 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$ 的元素给出.
  - 如果前两个因子解释方差的累积比例 $\Psi_2 = \dfrac{\lambda_1 + \lambda_2}{\sum_{k=1}^{R} \lambda_k}$ 足够大，那么二维图形的表现就足够好了.

- 图形的解释可总结如下:

④ 绝对贡献用于评价每一行 (列) 在因子方差中所占的权重.

⑤ 在进行评价时，上述所有的解释都必须考虑到图形表示的质量，如同主成分分析一样，我们使用累积方差的百分比来进行评价.

Correspondence Analysis in Practice 实践中的对应分析

- **备注**: 对应分析可应用于更一般的 $n \times p$ 的表 $\mathscr{X}$，并不 “严格要求” 是列联表.
  - 只要表中各行的和、各列的和具有统计 (或自然) 意义，上述备注即成立.
  - 特别，这意味着所有的变量都具有相同的度量单位.
  - 此时，$x_{\bullet\bullet}$ 构成了观察到的现象的总频数，并由个体 (n 行) 和变量 (p 列) 之间共享.

  - 表 $\mathscr{X}$ 的行以及列的表示，即 $\boldsymbol{r}_k$ 与 $\boldsymbol{s}_k$，满足基本性质

$$\boldsymbol{r}_k = \frac{1}{\sqrt{\lambda_k}} \mathscr{A}^{-1/2} \mathscr{C} \mathscr{B}^{1/2} \boldsymbol{s}_k$$

$$\boldsymbol{s}_k = \frac{1}{\sqrt{\lambda_k}} \mathscr{B}^{-1/2} \mathscr{C}^{\mathrm{T}} \mathscr{A}^{1/2} \boldsymbol{r}_k$$

    并显示出了哪些变量对每个个体具有重要的权重，以及哪些个体对每个变量具有重要的权重.

Correspondence Analysis in Practice 实践中的对应分析

- **备注**: 对应分析可应用于更一般的 $n \times p$ 的表 $\mathcal{X}$，并不 “严格要求” 是列联表.
  - 对应分析可以用作主成分分析的一种替代方法.
  - 主成分分析主要考虑的是协方差和相关性.
  - 而对应分析讨论的是一种更为普遍的关联性.

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

  - 问题 1: 居住地区？

$$\begin{cases} X_1 : \text{WaBr} & \text{Walloon Brabant} \\ X_2 : \text{Brar} & \text{Brussels area} \\ X_3 : \text{Antw} & \text{Antwerp} \\ X_4 : \text{FlBr} & \text{Flemish Brabant} \\ X_5 : \text{OcFl} & \text{Occidental Flanders} \\ X_6 : \text{OrFl} & \text{Oriental Flanders} \\ X_7 : \text{Hain} & \text{Hainaut} \\ X_8 : \text{Lieg} & \text{Liege} \\ X_9 : \text{Limb} & \text{Limburg} \\ X_{10} : \text{Luxe} & \text{Luxembourg} \end{cases}$$

  - 问题 2: 定期阅读哪种报纸？

15 个可选答案：va, vb, vc, vd, ve, ff, fg, fh, fi, bj, bk, bl, vm, fn, fo

首字母 v

首字母 f

首字母 b

归为三大类

- v: Flemish newspapers
- f: French newspapers
- b: Both languages

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
rm(list = ls(all = TRUE))
graphics.off()
library(readr)
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with data/Data")
x = read_csv("journaux.csv", col_names = FALSE) # 读入数据
x = as.data.frame(x)
x
```

```
> x
    X1   X2   X3   X4   X5   X6  X7  X8  X9 X10
1  1.8  7.8  9.1  3.0  4.3  3.9 0.1 0.3 3.3 0.0
2  0.1  3.4 17.8  1.0  0.7  4.1 0.0 0.0 0.2 0.0
3  0.1  9.4  4.6  7.7  4.4  5.8 1.6 0.1 1.4 0.0
4  0.5 15.6 16.1 12.0 10.5 10.2 0.7 0.3 5.4 0.0
5  0.1  5.2  3.3  4.8  1.6  1.4 0.1 0.0 3.5 0.0
6  5.6 13.7  3.1  2.4  0.5  1.7 1.9 2.3 0.2 0.2
7  4.1 16.5  1.9  1.0  1.0  0.9 2.4 3.2 0.1 0.3
8  8.3 29.5  1.8  7.3  0.8  0.4 5.1 3.2 0.2 0.3
9  0.9  7.8  0.2  2.6  0.1  0.1 5.6 3.8 0.1 0.8
10 6.1 18.2 10.8  4.1  4.5  5.3 2.0 2.6 3.4 0.2
11 8.3 35.4  6.2 11.0  5.0  6.1 5.5 3.3 1.5 0.3
12 4.4  9.9  6.7  3.4  1.1  3.9 2.1 1.5 2.1 0.0
13 0.3 11.6 14.2  4.7  5.1  7.9 0.3 0.5 3.0 0.0
14 5.1 21.0  1.3  3.4  0.2  0.2 2.3 4.4 0.0 0.4
15 2.2  9.8  0.1  0.3  0.0  0.7 2.3 3.0 0.3 1.0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
a  = rowSums(x) # 每一行之和构成的向量
a
b  = colSums(x) # 每一列之和构成的向量
b
e  = matrix(a) %*% b / sum(a) # 各单元 (交叉位置) 的期望值 (独立性假设下)
round(e, digits = 2)
```

```
> a
 [1] 33.6 27.3 35.1 71.3 20.0 31.6 31.4 56.9 22.0 57.2 82.6 35.1 47.6 38.3 19.7
```

```
> b
   X1    X2    X3    X4    X5    X6    X7    X8    X9   X10
 47.9 214.8  97.2  68.7  39.8  52.6  32.0  28.5  24.7   3.5
```

$$a = \begin{pmatrix} 33.6 \\ 27.3 \\ 35.1 \\ 71.3 \\ 20.0 \\ 31.6 \\ 31.4 \\ 56.9 \\ 22.0 \\ 57.2 \\ 82.6 \\ 35.1 \\ 47.6 \\ 38.3 \\ 19.7 \end{pmatrix}$$

$$b = \begin{pmatrix} 47.9 \\ 214.8 \\ 97.2 \\ 68.7 \\ 39.8 \\ 52.6 \\ 32.0 \\ 28.5 \\ 24.7 \\ 3.5 \end{pmatrix}$$

```
> round(e, digits = 2)
      [,1]  [,2]  [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] 2.64 11.84  5.36 3.79 2.19 2.90 1.76 1.57 1.36  0.19
 [2,] 2.14  9.62  4.35 3.08 1.78 2.36 1.43 1.28 1.11  0.16
 [3,] 2.76 12.37  5.60 3.96 2.29 3.03 1.84 1.64 1.42  0.20
 [4,] 5.60 25.12 11.37 8.03 4.65 6.15 3.74 3.33 2.89  0.41
 [5,] 1.57  7.05  3.19 2.25 1.31 1.73 1.05 0.93 0.81  0.11
 [6,] 2.48 11.13  5.04 3.56 2.06 2.73 1.66 1.48 1.28  0.18
 [7,] 2.47 11.06  5.01 3.54 2.05 2.71 1.65 1.47 1.27  0.18
 [8,] 4.47 20.05  9.07 6.41 3.71 4.91 2.99 2.66 2.31  0.33
 [9,] 1.73  7.75  3.51 2.48 1.44 1.90 1.15 1.03 0.89  0.13
[10,] 4.49 20.15  9.12 6.45 3.73 4.93 3.00 2.67 2.32  0.33
[11,] 6.49 29.10 13.17 9.31 5.39 7.13 4.34 3.86 3.35  0.47
[12,] 2.76 12.37  5.60 3.96 2.29 3.03 1.84 1.64 1.42  0.20
[13,] 3.74 16.77  7.59 5.36 3.11 4.11 2.50 2.23 1.93  0.27
[14,] 3.01 13.49  6.11 4.32 2.50 3.30 2.01 1.79 1.55  0.22
[15,] 1.55  6.94  3.14 2.22 1.29 1.70 1.03 0.92 0.80  0.11
```

$$E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}, \quad i = 1, 2, \ldots, 15; \ j = 1, 2, \ldots, 10$$

E_{ij} 是在独立假设下 (i, j) 单元相应期望频数的估计值

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
cc = (x - e)/sqrt(e) # 计算拟进行奇异值分解的 Chi 矩阵
round(cc, digits = 2)
```

```
> round(cc, digits = 2)
      X1    X2    X3    X4    X5    X6    X7    X8    X9   X10
1  -0.52 -1.17  1.62 -0.40  1.42  0.59 -1.25 -1.01  1.66 -0.44
2  -1.40 -2.00  6.45 -1.18 -0.81  1.14 -1.20 -1.13 -0.86 -0.40
3  -1.60 -0.84 -0.42  1.88  1.39  1.59 -0.18 -1.20 -0.02 -0.45
4  -2.16 -1.90  1.40  1.40  2.71  1.63 -1.57 -1.66  1.48 -0.64
5  -1.17 -0.70  0.06  1.70  0.26 -0.25 -0.93 -0.97  2.99 -0.34
6   1.98  0.77 -0.86 -0.62 -1.09 -0.62  0.19  0.68 -0.95  0.04
7   1.04  1.63 -1.39 -1.35 -0.73 -1.10  0.59  1.43 -1.04  0.28
8   1.81  2.11 -2.41  0.35 -1.51 -2.04  1.22  0.33 -1.39 -0.05
9  -0.63  0.02 -1.77  0.08 -1.11 -1.31  4.14  2.73 -0.84  1.90
10  0.76 -0.43  0.56 -0.92  0.40  0.16 -0.58 -0.05  0.71 -0.22
11  0.71  1.17 -1.92  0.55 -0.17 -0.38  0.56 -0.29 -1.01 -0.25
12  0.99 -0.70  0.47 -0.28 -0.79  0.50  0.19 -0.11  0.57 -0.45
13 -1.78 -1.26  2.40 -0.29  1.13  1.87 -1.39 -1.16  0.77 -0.52
14  1.21  2.04 -1.94 -0.44 -1.45 -1.71  0.20  1.95 -1.25  0.38
15  0.52  1.09 -1.72 -1.29 -1.13 -0.77  1.25  2.17 -0.56  2.64
```

$$\mathscr{C} = \left(c_{ij}\right)_{n\times p}$$

$$c_{ij} = \frac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}}, \quad i = 1, 2, \ldots, 15 \; ; \; j = 1, 2, \ldots, 10$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
?svd
sv = svd(cc) # 奇异值分解
g = sv$u
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

$$\Gamma = (\gamma_1, \gamma_2, \ \cdots, \gamma_{10})$$

```
> round(g, digits = 2)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] -0.24  0.03 -0.06  0.35 -0.09 -0.11 -0.36 -0.26 -0.01  0.39
 [2,] -0.39 -0.75  0.08 -0.24  0.26  0.22  0.04 -0.11 -0.11  0.12
 [3,] -0.16  0.34  0.16 -0.46 -0.12 -0.03  0.31  0.14 -0.04  0.26
 [4,] -0.37  0.31  0.13 -0.05 -0.20  0.09 -0.20 -0.04 -0.47  0.25
 [5,] -0.16  0.35  0.04  0.43  0.61  0.34  0.17  0.04  0.12  0.12
 [6,]  0.18 -0.14 -0.24  0.01  0.00 -0.26  0.20  0.08 -0.39  0.43
 [7,]  0.24 -0.13 -0.11  0.08 -0.25  0.13 -0.32  0.19  0.37  0.15
 [8,]  0.33  0.05 -0.26 -0.25  0.34  0.08 -0.13 -0.43  0.08  0.41
 [9,]  0.32 -0.05  0.80 -0.05  0.22 -0.19 -0.31  0.12  0.00  0.20
[10,] -0.05 -0.07 -0.14  0.31 -0.13 -0.30 -0.17 -0.01 -0.09  0.06
[11,]  0.15  0.17 -0.15 -0.33 -0.03 -0.05  0.04 -0.22  0.22  0.10
[12,] -0.02 -0.09 -0.08  0.14  0.24 -0.56  0.34  0.31  0.12  0.16
[13,] -0.32 -0.06  0.07 -0.01 -0.25  0.10  0.11  0.19  0.56  0.43
[14,]  0.31 -0.08 -0.16  0.05 -0.04  0.47 -0.03  0.55 -0.29  0.21
[15,]  0.28 -0.14  0.30  0.35 -0.36  0.20  0.54 -0.41 -0.06  0.13
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
l = sv$d
round(l, digits = 2)
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

$$\Lambda = \mathrm{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_{10}}\right)$$

```
> l = sv$d
> round(l, digits = 2)
 [1] 13.54  6.61  5.02  3.43  2.84  2.16  1.46  1.09  0.90  0.00
```

$$\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}, \sqrt{\lambda_4}, \sqrt{\lambda_5}, \sqrt{\lambda_6}, \sqrt{\lambda_7}, \sqrt{\lambda_8}, \sqrt{\lambda_9}, \sqrt{\lambda_{10}}$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
d = sv$v
round(d, digits = 2)
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

$$\Delta = (\delta_1, \delta_2, \ldots, \delta_{10})$$

```
> round(d, digits = 2)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,]  0.30 -0.14 -0.53  0.16  0.06 -0.57  0.10 -0.14 -0.37 -0.28
 [2,]  0.35  0.03 -0.31 -0.08 -0.12  0.47 -0.04 -0.10  0.42 -0.59
 [3,] -0.55 -0.66  0.05 -0.02  0.22  0.11 -0.15 -0.08 -0.11 -0.40
 [4,] -0.08  0.49  0.09 -0.36  0.46  0.16  0.22  0.10 -0.45 -0.34
 [5,] -0.27  0.38  0.06  0.00 -0.54 -0.10 -0.56 -0.20 -0.25 -0.26
 [6,] -0.32  0.05  0.09 -0.18 -0.41 -0.36  0.57  0.27  0.27 -0.29
 [7,]  0.32 -0.04  0.56 -0.22  0.26 -0.47 -0.26 -0.17  0.31 -0.23
 [8,]  0.34 -0.19  0.32  0.28 -0.19  0.11 -0.11  0.68 -0.32 -0.22
 [9,] -0.24  0.34  0.05  0.78  0.32 -0.07  0.02  0.03  0.25 -0.20
[10,]  0.17 -0.10  0.42  0.25 -0.23  0.20  0.45 -0.59 -0.27 -0.08
```

```
ll = l * l # Chi 矩阵 cc 的特征值
round(ll, digits = 2)
```

```
> round(ll, digits = 2)
 [1] 183.40  43.76  25.21  11.74   8.04   4.68   2.13   1.20   0.82   0.00
```

$$\lambda_1 \ , \ \lambda_2 \ , \ \lambda_3 \ , \ \lambda_4 \ , \ \lambda_5 \ , \ \lambda_6 \ , \ \lambda_7 \ , \ \lambda_8 \ , \ \lambda_9 \ , \ \lambda_{10}$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.

```
# cumulated percentage of the variance 方差的百分比与累积百分比
rat = ll / sum(ll) # 方差的百分比
aux  = cumsum(ll) / sum(ll) # 方差的累积百分比
perc = cbind(Lambda = ll, Perc_Var = rat, Cum_Perc = aux) # 合并在一起
round(perc, digits = 3)
```

- 由于前两个特征值的累积贡献率达到81%，所以二维图形表示已足够好.

```
> round(perc, digits = 3)
       Lambda Perc_Var Cum_Perc
 [1,] 183.402    0.653    0.653
 [2,]  43.758    0.156    0.808
 [3,]  25.214    0.090    0.898
 [4,]  11.740    0.042    0.940
 [5,]   8.043    0.029    0.969
 [6,]   4.681    0.017    0.985
 [7,]   2.127    0.008    0.993
 [8,]   1.198    0.004    0.997
 [9,]   0.817    0.003    1.000
[10,]   0.000    0.000    1.000
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 计算行权重 $\boldsymbol{r}_k = \sqrt{\lambda_k}\mathscr{A}^{-1/2}\boldsymbol{\gamma}_k$, $k = 1, 2, \ldots, 15$

```
# 计算行 (个体的) 权重
r1 = matrix(l, nrow = nrow(g), ncol = ncol(g), byrow = T) * g  # 计算 sqrt(λ_k) γ_k
r = r1/matrix(sqrt(a), nrow = nrow(g), ncol = ncol(g), byrow = F) # 计算 r_k = A^{-1/2}(sqrt(λ_k) γ_k)
round(r, digits = 3)
```

```
> round(r, digits = 3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9] [,10]
 [1,] -0.554  0.032 -0.052  0.206 -0.042 -0.040 -0.089 -0.049 -0.002     0
 [2,] -1.022 -0.945  0.079 -0.156  0.139  0.091  0.011 -0.023 -0.018     0
 [3,] -0.357  0.383  0.138 -0.266 -0.060 -0.011  0.077  0.026 -0.006     0
 [4,] -0.590  0.242  0.076 -0.018 -0.068  0.022 -0.034 -0.005 -0.050     0
 [5,] -0.482  0.511  0.041  0.331  0.387  0.165  0.057  0.011  0.024     0
 [6,]  0.427 -0.159 -0.218  0.004  0.001 -0.102  0.051  0.015 -0.062     0
 [7,]  0.585 -0.150 -0.099  0.051 -0.129  0.049 -0.084  0.037  0.059     0
 [8,]  0.592  0.043 -0.170 -0.113  0.129  0.023 -0.025 -0.062  0.009     0
 [9,]  0.914 -0.069  0.855 -0.038  0.132 -0.089 -0.095  0.027 -0.001     0
[10,] -0.096 -0.065 -0.091  0.141 -0.047 -0.087 -0.033 -0.002 -0.011     0
[11,]  0.229  0.121 -0.085 -0.124 -0.008 -0.011  0.006 -0.027  0.022     0
[12,] -0.054 -0.106 -0.068  0.081  0.117 -0.206  0.083  0.057  0.018     0
[13,] -0.621 -0.059  0.050 -0.003 -0.105  0.032  0.023  0.031  0.073     0
[14,]  0.680 -0.082 -0.133  0.026 -0.018  0.165 -0.007  0.098 -0.042     0
[15,]  0.869 -0.204  0.339  0.272 -0.228  0.096  0.179 -0.101 -0.012     0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 计算列权重 $\boldsymbol{s}_k = \sqrt{\lambda_k}\mathcal{B}^{-1/2}\boldsymbol{\delta}_k$, $k = 1, 2, \ldots, 10$

```
# 计算列 (变量的) 权重
s1 = matrix(l, nrow = nrow(d), ncol = ncol(d), byrow = T) * d # 计算 √λ_k δ_k
s = s1/matrix(sqrt(b), nrow = nrow(d), ncol = ncol(d), byrow = F) # 计算 s_k = B^{-1/2}(√λ_k δ_k)
round(s, digits = 3)
```

```
> round(s, digits = 3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9] [,10]
 [1,]  0.583 -0.139 -0.388  0.080  0.025 -0.179  0.020 -0.022 -0.048     0
 [2,]  0.328  0.014 -0.106 -0.019 -0.023  0.069 -0.004 -0.007  0.026     0
 [3,] -0.752 -0.442  0.027 -0.008  0.064  0.024 -0.022 -0.009 -0.010     0
 [4,] -0.131  0.389  0.057 -0.150  0.158  0.043  0.039  0.013 -0.049     0
 [5,] -0.579  0.394  0.045 -0.001 -0.243 -0.036 -0.129 -0.035 -0.036     0
 [6,] -0.590  0.044  0.061 -0.085 -0.161 -0.107  0.115  0.041  0.034     0
 [7,]  0.774 -0.041  0.497 -0.133  0.133 -0.179 -0.067 -0.033  0.049     0
 [8,]  0.867 -0.233  0.301  0.178 -0.099  0.045 -0.029  0.139 -0.054     0
 [9,] -0.646  0.454  0.053  0.540  0.181 -0.031  0.007  0.006  0.046     0
[10,]  1.229 -0.355  1.126  0.463 -0.351  0.232  0.350 -0.345 -0.129     0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 计算行的绝对贡献：$C_a\left(i, \boldsymbol{r}_k\right)=\dfrac{x_{i\bullet}\, r_{ki}^2}{\lambda_k}$，　$i=1, 2, \ldots, n$；　$k=1, 2, \ldots, R$

```
# 计算行的绝对贡献
car = matrix(matrix(a), nrow = nrow(r), ncol = ncol(r), byrow = F) * r^2 / matrix(l^2, nrow = nrow(r), ncol = ncol(r), byrow = T)
round(car, digits = 3) # 显示结果
```

```
> round(car, digits = 3) # 显示结果
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 0.056 0.001 0.004 0.122 0.007 0.011 0.126 0.069 0.000 0.148
 [2,] 0.155 0.557 0.007 0.056 0.066 0.048 0.001 0.012 0.011 0.013
 [3,] 0.024 0.118 0.027 0.212 0.015 0.001 0.098 0.019 0.002 0.067
 [4,] 0.135 0.095 0.016 0.002 0.041 0.008 0.039 0.002 0.216 0.060
 [5,] 0.025 0.119 0.001 0.187 0.373 0.117 0.030 0.002 0.014 0.016
 [6,] 0.031 0.018 0.060 0.000 0.000 0.070 0.038 0.006 0.149 0.182
 [7,] 0.058 0.016 0.012 0.007 0.065 0.016 0.104 0.035 0.134 0.024
 [8,] 0.109 0.002 0.066 0.062 0.117 0.007 0.017 0.185 0.006 0.166
 [9,] 0.100 0.002 0.638 0.003 0.047 0.038 0.094 0.014 0.000 0.041
[10,] 0.003 0.005 0.019 0.097 0.016 0.093 0.030 0.000 0.008 0.004
[11,] 0.024 0.028 0.024 0.107 0.001 0.002 0.001 0.049 0.047 0.009
[12,] 0.001 0.009 0.006 0.019 0.059 0.317 0.113 0.094 0.013 0.025
[13,] 0.100 0.004 0.005 0.000 0.065 0.011 0.012 0.038 0.313 0.182
[14,] 0.097 0.006 0.027 0.002 0.002 0.224 0.001 0.307 0.082 0.045
[15,] 0.081 0.019 0.090 0.124 0.127 0.039 0.296 0.169 0.003 0.018
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 计算列的绝对贡献：$C_a\left(j, \boldsymbol{s}_k\right)=\dfrac{x_{\bullet j}\ s_{kj}^2}{\lambda_k}$，　$j=1,\ 2,\ \ldots,\ p$；　$k=1,\ 2,\ \ldots,\ R$

```
# 计算列的绝对贡献
cas = matrix(matrix(b), nrow = nrow(s), ncol = ncol(s), byrow = F) * s^2 / matrix(l^2, nrow = nrow(s), ncol = ncol(s), byrow = T)
round(cas, digits = 3) # 显示结果
```

```
> round(cas, digits = 3) # 显示结果
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 0.089 0.021 0.286 0.026 0.004 0.329 0.009 0.020 0.138 0.079
 [2,] 0.126 0.001 0.096 0.006 0.014 0.217 0.001 0.009 0.177 0.352
 [3,] 0.300 0.435 0.003 0.001 0.050 0.012 0.022 0.007 0.012 0.159
 [4,] 0.006 0.237 0.009 0.132 0.214 0.027 0.048 0.009 0.205 0.113
 [5,] 0.073 0.141 0.003 0.000 0.292 0.011 0.312 0.042 0.062 0.065
 [6,] 0.100 0.002 0.008 0.032 0.169 0.128 0.325 0.075 0.075 0.086
 [7,] 0.105 0.001 0.314 0.048 0.070 0.219 0.068 0.028 0.094 0.052
 [8,] 0.117 0.035 0.102 0.077 0.035 0.012 0.012 0.461 0.101 0.047
 [9,] 0.056 0.116 0.003 0.613 0.100 0.005 0.001 0.001 0.064 0.041
[10,] 0.029 0.010 0.176 0.064 0.054 0.040 0.202 0.348 0.072 0.006
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 分别取前两个因子

```
# 分别取前两个因子
rr = r[, 1:2]
round(rr, digits = 3) # 显示结果
ss = s[, 1:2]
round(ss, digits = 3) # 显示结果
```

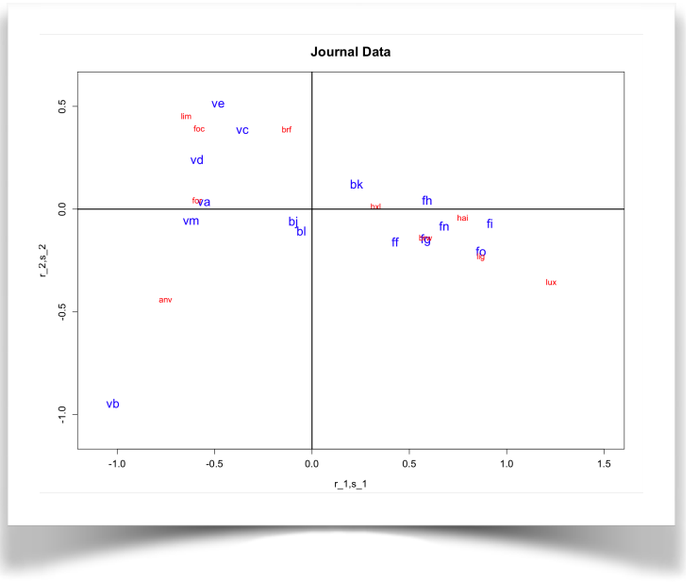
```
> round(rr, digits = 3)
       [,1]   [,2]
 [1,] -0.554  0.032
 [2,] -1.022 -0.945
 [3,] -0.357  0.383
 [4,] -0.590  0.242
 [5,] -0.482  0.511
 [6,]  0.427 -0.159
 [7,]  0.585 -0.150
 [8,]  0.592  0.043
 [9,]  0.914 -0.069
[10,] -0.096 -0.065
[11,]  0.229  0.121
[12,] -0.054 -0.106
[13,] -0.621 -0.059
[14,]  0.680 -0.082
[15,]  0.869 -0.204
```

```
> round(ss, digits = 3)
       [,1]   [,2]
 [1,]  0.583 -0.139
 [2,]  0.328  0.014
 [3,] -0.752 -0.442
 [4,] -0.131  0.389
 [5,] -0.579  0.394
 [6,] -0.590  0.044
 [7,]  0.774 -0.041
 [8,]  0.867 -0.233
 [9,] -0.646  0.454
[10,]  1.229 -0.355
```

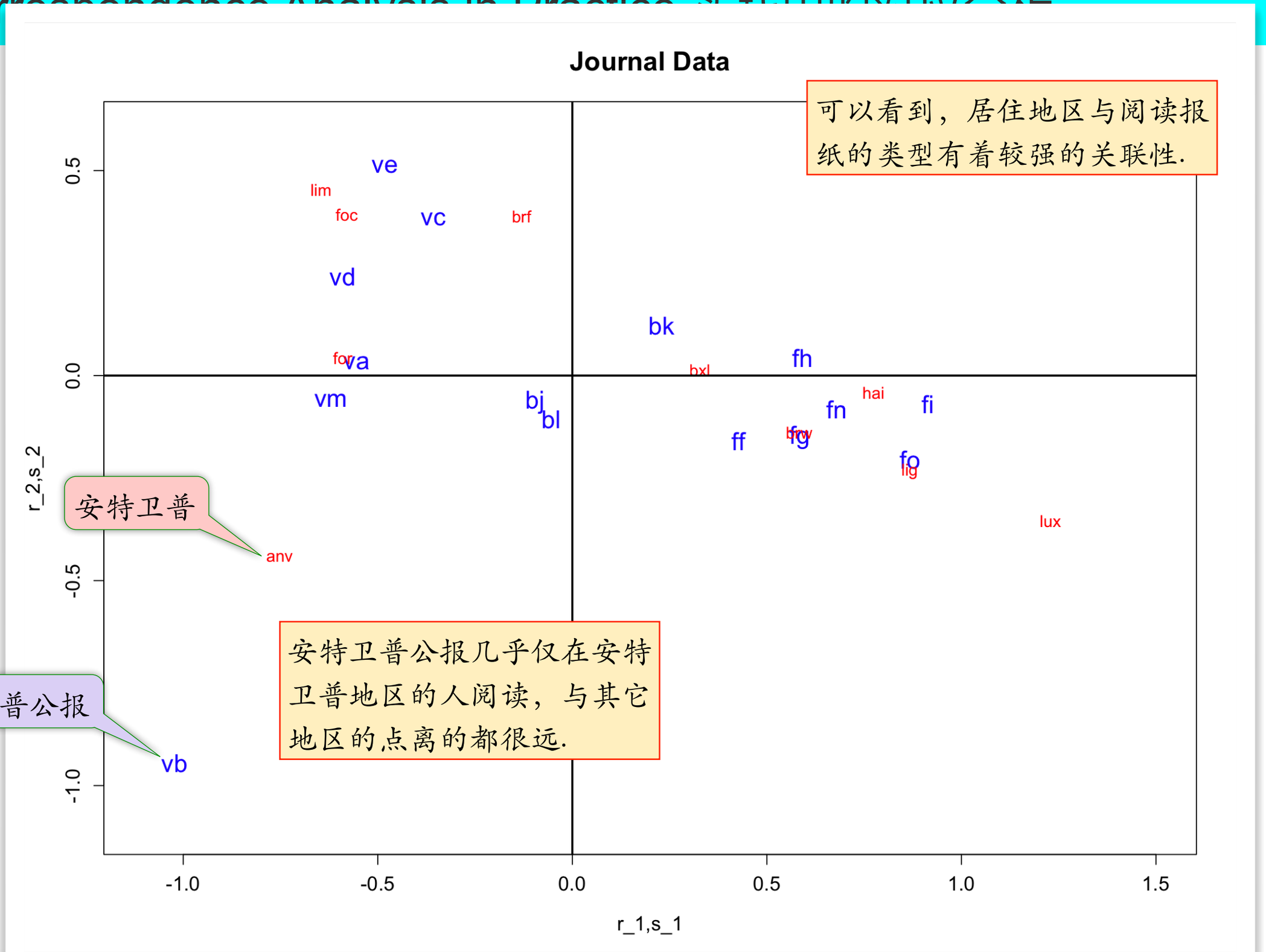
Correspondence Analysis in Practice 实践中的对应分析

- Example: 1980 年代比利时人的阅读习惯调查.
  - 绘图

```
# labels for journals 报纸的标志
types    = c("va", "vb", "vc", "vd", "ve", "ff", "fg", "fh", "fi", "bj", "bk", "bl", "vm", "fn", "fo")
# labels for regions 地区的标志
regions  = c("brw", "bxl", "anv", "brf", "foc", "for", "hai", "lig", "lim", "lux")
# plot 绘图
plot(rr, type = "n", xlim = c(-1.1, 1.5), ylim = c(-1.1, 0.6), xlab = "r_1,s_1",
    ylab = "r_2,s_2", main = "Journal Data", cex.axis = 1.2, cex.lab = 1.2, cex.main = 1.6)
points(ss, type = "n")
text(rr, types, cex = 1.5, col = "blue")
text(ss, regions, col = "red")
abline(h = 0, v = 0, lwd = 2)
```



Correspondence Analysis in Practice 实践中的对应分析



可以看到，居住地区与阅读报纸的类型有着较强的关联性.

安特卫普

安特卫普公报

安特卫普公报几乎仅在安特卫普地区的人阅读，与其它地区的点离的都很远.

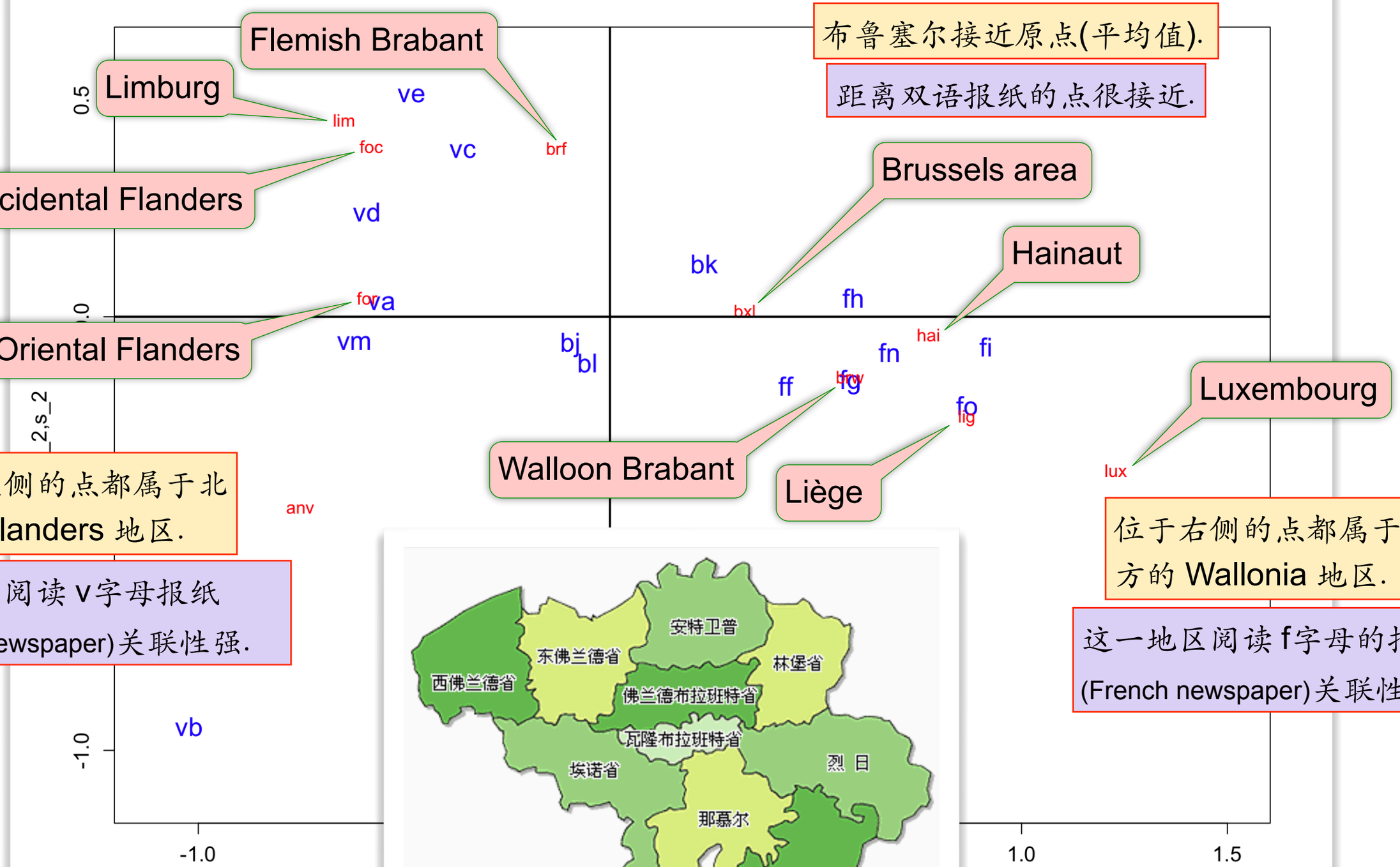
Correspondence A…础中的对应分析

还值得注意的一点是: Walloon Brabant 与Flemish Brabant 距离布鲁塞尔地区不远.

因说法语的人多，所以该点稍微偏右一些.

布鲁塞尔接近原点(平均值).

距离双语报纸的点很接近.



位于左侧的点都属于北方的 Flanders 地区.

这一地区阅读 v字母报纸(Flemish newspaper)关联性强.

位于右侧的点都属于南方的 Wallonia 地区.

这一地区阅读 f字母的报纸(French newspaper)关联性强.

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

22 个大区:

NOPC	:	Nord-Pas-de-Calais 北部加莱海峡
PICA	:	Picardie 皮卡第
HNOR	:	Normandie 诺曼底
ILDF	:	Ile de France 巴黎大区
CHAM	:	Champagne Ardenne 香槟-阿登
LORR	:	Lorraine 洛林
ALSA	:	Alsace 阿尔萨斯
BRET	:	Bretagne 布列塔尼
PAYL	:	Pays de loire 卢瓦尔河地区
CENT	:	Centre Val de loire 中央-卢瓦河山谷
BOUR	:	Bourgogne 勃艮第
FRAC	:	Franche comte 佛朗什孔泰
PCHA	:	Poitou Charentes 普瓦图-夏郎德
LIMO	:	Limousin 利木赞
AUVE	:	Auvergne 奥佛涅
RHOA	:	Rhone Alpes 罗讷-阿尔卑斯
BNOR	:	Basse-Normandie 下诺曼底
AQUI	:	Aquitaine 阿基坦
MIDI	:	Midi pyrenees 南部-比利牛斯
PROV	:	Provence Alpes cote d'azur 普罗旺斯-阿尔卑斯南部，蓝色海岸
LARO	:	Languedoc Roussilon 朗格多克-鲁西永
CORS	:	corse 科西嘉

报考专业:

X_1	: A	Philosophy-Letters
X_2	: B	Economics and Social Sciences
X_3	: C	Mathematics and Physics
X_4	: D	Mathematics and Natural Sciences
X_5	: E	Mathematics and Techniques
X_6	: F	Industrial Techniques
X_7	: G	Economic Techniques
X_8	: H	Computer Techniques

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# 读入数据
library(readr)
setwd("~/Desktop/2023_Applied Multivariate Statistical Analysis/R Codes with data/Data")
bac = read_csv("bac.cs
x1 = as.data.frame(bac)
colnames(x1) = c("Regi
x1
x1 = x1[, 2:ncol(x1)] # 易
x1
```

```
> x1
      A    B    C    D   E    F    G  H
1  9724 5650 8679 9432 839 3353 5355 83
2   924  464  567  984 132  423  736 12
3  1081  490  830 1222 118  410  743 13
4  1135  587  686  904  83  629  813 13
5  1482  667 1020 1535 173  629  989 26
6  1033  509  553 1063 100  433  742 13
7  1272  527  861 1116 219  769 1232 13
8  2549 1141 2164 2752 587 1660 1951 41
9  1828  681 1364 1741 302 1289 1683 15
10 1076  443  880 1121 145  917 1091 15
11  827  333  481  892 137  451  618 18
12 2213  809 1439 2623 269  990 1783 14
13 2158 1271 1633 2352 350  950 1509 22
14 1358  503  639 1377 164  495  959 10
15 2757  873 1466 2296 215  789 1459 17
16 2493 1120 1494 2329 254  885 1565 28
17  551  297  386  663  67  334  378 12
18 3951 2127 3218 4743 545 2072 3018 36
19 1066  579  724 1239 126  476  649 12
20 1844  816 1154 1839 156  469  993 16
21 3944 1645 2415 3616 343 1236 2404 22
22  327   31   85  178   9   27   79  0
```

```
> x1
   Region    A    B    C    D   E    F    G  H
1    ILDF 9724 5650 8679 9432 839 3353 5355 83
2    CHAM  924  464  567  984 132  423  736 12
3    PICA 1081  490  830 1222 118  410  743 13
4    HNOR 1135  587  686  904  83  629  813 13
5    CENT 1482  667 1020 1535 173  629  989 26
6    BNOR 1033  509  553 1063 100  433  742 13
7    BOUR 1272  527  861 1116 219  769 1232 13
8    NOPC 2549 1141 2164 2752 587 1660 1951 41
9    LORR 1828  681 1364 1741 302 1289 1683 15
10   ALSA 1076  443  880 1121 145  917 1091 15
11   FRAC  827  333  481  892 137  451  618 18
12   PAYL 2213  809 1439 2623 269  990 1783 14
13   BRET 2158 1271 1633 2352 350  950 1509 22
14   PCHA 1358  503  639 1377 164  495  959 10
15   AQUI 2757  873 1466 2296 215  789 1459 17
16   MIDI 2493 1120 1494 2329 254  885 1565 28
17   LIMO  551  297  386  663  67  334  378 12
18   RHOA 3951 2127 3218 4743 545 2072 3018 36
19   AUVE 1066  579  724 1239 126  476  649 12
20   LARO 1844  816 1154 1839 156  469  993 16
21   PROV 3944 1645 2415 3616 343 1236 2404 22
22   CORS  327   31   85  178   9   27   79  0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# set to 0/1 to ex/include Corsica 设置 0/1 分别对应 不含/包含 Corsica 大区
wcors = 0
wcorsica = c(rep(1, nrow(x1) - 1), wcors)
wcorsica

x = subset(x1, wcorsica == 1) # 不含 Corsica 大区的子数据集
x
a = rowSums(x) # 计算每一行之和
a
b = colSums(x) # 计算每一列之和
b
```

```
> wcorsica
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
```

```
> x
      A    B    C    D   E    F    G  H
1  9724 5650 8679 9432 839 3353 5355 83
2   924  464  567  984 132  423  736 12
3  1081  490  830 1222 118  410  743 13
4  1135  587  686  904  83  629  813 13
5  1482  667 1020 1535 173  629  989 26
6  1033  509  553 1063 100  433  742 13
7  1272  527  861 1116 219  769 1232 13
8  2549 1141 2164 2752 587 1660 1951 41
9  1828  681 1364 1741 302 1289 1683 15
10 1076  443  880 1121 145  917 1091 15
11  827  333  481  892 137  451  618 18
12 2213  809 1439 2623 269  990 1783 14
13 2158 1271 1633 2352 350  950 1509 22
14 1358  503  639 1377 164  495  959 10
15 2757  873 1466 2296 215  789 1459 17
16 2493 1120 1494 2329 254  885 1565 28
17  551  297  386  663  67  334  378 12
18 3951 2127 3218 4743 545 2072 3018 36
19 1066  579  724 1239 126  476  649 12
20 1844  816 1154 1839 156  469  993 16
21 3944 1645 2415 3616 343 1236 2404 22
```

```
> a
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20    21
43115  4242  4907  4850  6521  4446  6009 12845  8903  5688  3757 10140 10245  5505  9872 10168  2688 19710  4871  7287 15625
```

```
> b
    A     B     C     D     E     F     G     H
45266 21532 32653 45839  5324 19659 30670   451
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
e = matrix(a) %*% b / sum(a) # 计算各单元 (交叉位置) 的期望值 (独立
round(e, digits = 2)

# chi-matrix 计算拟进行奇异值分解的 Chi 矩阵
cc = (x - e) / sqrt(e)
round(cc, digits = 2)
```

$$E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}, \quad i = 1, 2, \ldots, 21; \ j = 1, 2, \ldots, 8$$

```
> round(e, digits = 2)
       [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]  [,8]
 [1,] 9690.67 4609.63 6990.45 9813.34 1139.78 4208.65 6565.92 96.55
 [2,]  953.45  453.53  687.78  965.52  112.14  414.08  646.01  9.50
 [3,] 1102.91  524.63  795.60 1116.88  129.72  478.99  747.28 10.99
 [4,] 1090.10  518.54  786.35 1103.90  128.21  473.43  738.60 10.86
 [5,] 1465.68  697.19 1057.28 1484.24  172.39  636.54  993.07 14.60
 [6,]  999.30  475.34  720.85 1011.95  117.53  433.99  677.07  9.96
 [7,] 1350.60  642.45  974.27 1367.70  158.85  586.57  915.10 13.46
 [8,] 2887.09 1373.32 2082.62 2923.63  339.57 1253.86 1956.15 28.76
 [9,] 2001.07  951.86 1443.49 2026.40  235.36  869.06 1355.82 19.94
[10,] 1278.45  608.13  922.22 1294.64  150.37  555.23  866.22 12.74
[11,]  844.44  401.68  609.14  855.13   99.32  366.74  572.15  8.41
[12,] 2279.10 1084.12 1644.05 2307.95  268.06  989.81 1544.21 22.71
[13,] 2302.70 1095.34 1661.07 2331.85  270.83 1000.06 1560.20 22.94
[14,] 1237.32  588.57  892.55 1252.99  145.53  537.37  838.35 12.33
[15,] 2218.86 1055.46 1600.60 2246.95  260.97  963.65 1503.39 22.11
[16,] 2285.39 1087.11 1648.59 2314.32  268.80  992.55 1548.47 22.77
[17,]  604.16  287.39  435.82  611.81   71.06  262.39  409.35  6.02
[18,] 4430.09 2107.29 3195.68 4486.16  521.05 1923.98 3001.61 44.14
[19,] 1094.82  520.78  789.76 1108.68  128.77  475.48  741.80 10.91
[20,] 1637.85  779.09 1181.48 1658.58  192.64  711.32 1109.73 16.32
[21,] 3511.93 1670.54 2533.36 3556.38  413.06 1525.23 2379.51 34.99
```

```
> round(cc, digits = 2)
       A     B     C     D     E     F     G     H
1   0.34 15.32 20.20 -3.85 -8.91 -13.19 -14.94 -1.38
2  -0.95  0.49 -4.61  0.59  1.88   0.44   3.54  0.81
3  -0.66 -1.51  1.22  3.15 -1.03  -3.15  -0.16  0.61
4   1.36  3.01 -3.58 -6.02 -3.99   7.15   2.74  0.65
5   0.43 -1.14 -1.15  1.32  0.05  -0.30  -0.13  2.98
6   1.07  1.54 -6.25  1.60 -1.62  -0.05   2.50  0.96
7  -2.14 -4.55 -3.63 -6.81  4.77   7.53  10.48 -0.12
8  -6.29 -6.27  1.78 -3.17 13.43  11.47  -0.12  2.28
9  -3.87 -8.78 -2.09 -6.34  4.34  14.24   8.89 -1.11
10 -5.66 -6.70 -1.39 -4.83 -0.44  15.35   7.64  0.63
11 -0.60 -3.43 -5.19  1.26  3.78   4.40   1.92  3.31
12 -1.38 -8.36 -5.06  6.56  0.06   0.01   6.08 -1.83
13 -3.02  5.31 -0.69  0.42  4.81  -1.58  -1.30 -0.20
14  3.43 -3.53 -8.49  3.50  1.53  -1.83   4.17 -0.66
15 11.42 -5.62 -3.36  1.03 -2.85  -5.63  -1.14 -1.09
16  4.34  1.00 -3.81  0.31 -0.90  -3.41   0.42  1.10
17 -2.16  0.57 -2.39  2.07 -0.48   4.42  -1.55  2.44
18 -7.20  0.43  0.39  3.83  1.05   3.37   0.30 -1.22
19 -0.87  2.55 -2.34  3.91 -0.24   0.02  -3.41  0.33
20  5.09  1.32 -0.80  4.43 -2.64  -9.09  -3.50 -0.08
21  7.29 -0.62 -2.35  1.00 -3.45  -7.41   0.50 -2.20
```

$$c_{ij} = \frac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}}, \quad i = 1, 2, \ldots, 21; \ j = 1, 2, \ldots, 8$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# singular value decomposition 矩阵 cc 的奇异值分解
sv   = svd(cc)
g    = sv$u
round(g, digits = 2)
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

$$\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_8)$$

```
> round(g, digits = 2)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
 [1,] -0.62  0.47  0.18 -0.09 -0.09 -0.06  0.09  0.00
 [2,]  0.08 -0.09 -0.07  0.20  0.13 -0.22  0.15 -0.27
 [3,] -0.05 -0.06 -0.12 -0.08 -0.23 -0.04  0.34 -0.19
 [4,]  0.10  0.07  0.45  0.42  0.17  0.19 -0.15 -0.12
 [5,]  0.01 -0.05 -0.07  0.01  0.02  0.17  0.52  0.02
 [6,]  0.04 -0.17  0.02  0.36  0.07 -0.01  0.08 -0.18
 [7,]  0.30  0.05  0.25 -0.06  0.18 -0.49  0.23 -0.25
 [8,]  0.30  0.32 -0.36 -0.44  0.37  0.20 -0.07 -0.13
 [9,]  0.40  0.16  0.25 -0.19 -0.09 -0.05 -0.20  0.26
[10,]  0.36  0.21  0.28  0.09 -0.34  0.19  0.09 -0.08
[11,]  0.16 -0.07 -0.15  0.05  0.17  0.24  0.32  0.06
[12,]  0.15 -0.27 -0.19 -0.10 -0.52 -0.17  0.00  0.21
[13,] -0.03  0.08 -0.27  0.18  0.34 -0.31 -0.21  0.01
[14,]  0.09 -0.34 -0.07  0.05  0.09 -0.11 -0.14  0.16
[15,] -0.07 -0.39  0.25 -0.40  0.09  0.34 -0.20 -0.45
[16,] -0.04 -0.19  0.08  0.08  0.22  0.01  0.16  0.27
[17,]  0.06  0.04 -0.12  0.26 -0.03  0.40  0.08  0.00
[18,]  0.06  0.13 -0.30  0.21 -0.33 -0.08 -0.27 -0.52
[19,] -0.04 -0.03 -0.22  0.25  0.01  0.25 -0.30  0.16
[20,] -0.20 -0.24 -0.08 -0.06  0.01  0.05  0.07 -0.18
[21,] -0.12 -0.30  0.20 -0.12  0.02 -0.12 -0.18  0.03
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# singular value decomposition 矩阵 cc 的奇异值分解
sv   = svd(cc)
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

```
l = sv$d
round(l, digits = 2)
```

$$\Lambda = \mathrm{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_8}\right)$$

```
> round(l, digits = 2)
[1] 49.08 30.16 17.85 14.00 12.22  9.80  4.78  0.00
```

$$\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}, \sqrt{\lambda_4}, \sqrt{\lambda_5}, \sqrt{\lambda_6}, \sqrt{\lambda_7}, \sqrt{\lambda_8}$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# singular value decomposition 矩阵 cc 的奇异值分解
sv   = svd(cc)
```

$$\mathscr{C} = \Gamma \Lambda \Delta^{\mathrm{T}}$$

```
d = sv$v
round(d, digits = 2)
```

$$\Delta = (\delta_1, \delta_2, \ \ldots, \delta_8)$$

```
> round(d, digits = 2)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7] [,8]
[1,] -0.19 -0.48  0.44 -0.29  0.36  0.29 -0.12 0.47
[2,] -0.42  0.27  0.01  0.67  0.37 -0.21 -0.11 0.33
[3,] -0.35  0.60  0.08 -0.49 -0.29 -0.08  0.16 0.40
[4,] -0.10 -0.35 -0.64  0.14 -0.43  0.18 -0.06 0.48
[5,]  0.29  0.12 -0.54 -0.37  0.61 -0.25 -0.10 0.16
[6,]  0.62  0.40  0.15  0.20 -0.04  0.48 -0.25 0.31
[7,]  0.44 -0.21  0.27  0.10 -0.18 -0.64  0.29 0.39
[8,]  0.04  0.03 -0.11  0.12  0.23  0.36  0.89 0.05
```

```
# eigenvalues 特征值
ll   = l * l
round(ll, digits = 2)
```

```
> round(ll, digits = 2)
[1] 2408.55  909.48  318.46  195.88  149.27   96.07   22.83    0.00
```

$$\lambda_1 \ , \ \lambda_2 \ , \ \lambda_3 \ , \ \lambda_4 \ , \ \lambda_5 \ , \ \lambda_6 \ , \ \lambda_7 \ , \ \lambda_8$$

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.

```
# cumulated percentage of the variance 方差的百分比与累积百分比
rat = ll / sum(ll) # 方差的百分比
aux  = cumsum(ll) / sum(ll) # 方差的累积百分比
perc = cbind(Lambda = ll, Perc_Var = rat, Cum_Perc = aux) # 合并在一起
round(perc, digits = 3)
```

```
> round(perc, digits = 3)
       Lambda Perc_Var Cum_Perc
[1,] 2408.554    0.587    0.587
[2,]  909.479    0.222    0.809
[3,]  318.460    0.078    0.887
[4,]  195.877    0.048    0.935
[5,]  149.275    0.036    0.971
[6,]   96.067    0.023    0.994
[7,]   22.827    0.006    1.000
[8,]    0.000    0.000    1.000
```

- 由于前两个特征值的方差累积贡献率已经达到了 80.9%，所以二维图形表示已足够好.

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 计算行权重 $\boldsymbol{r}_k = \sqrt{\lambda_k}\mathscr{A}^{-1/2}\boldsymbol{\gamma}_k$, $k = 1, 2, \ldots, 21$

```
# 计算行 (个体的) 权重
r1 = matrix(l, nrow = nrow(g), ncol = ncol(g), byrow = T) * g  # 计算 sqrt(λ_k) γ_k
r = r1/matrix(sqrt(a), nrow = nrow(g), ncol = ncol(g), byrow = F) # 计算 r_k = 𝒜^{-1/2} (sqrt(λ_k) γ_k)
round(r, digits = 3)
```

```
> round(r, digits = 3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7] [,8]
 [1,]  0.146 -0.068 -0.016  0.006  0.005  0.003 -0.002    0
 [2,] -0.060  0.041  0.019 -0.043 -0.025  0.033 -0.011    0
 [3,]  0.032  0.026  0.032  0.017  0.040  0.005 -0.023    0
 [4,] -0.069 -0.029 -0.116 -0.084 -0.030 -0.027  0.010    0
 [5,] -0.007  0.021  0.015 -0.001 -0.003 -0.021 -0.031    0
 [6,] -0.027  0.076 -0.006 -0.075 -0.014  0.002 -0.006    0
 [7,] -0.192 -0.019 -0.058  0.010 -0.029  0.062 -0.014    0
 [8,] -0.128 -0.086  0.057  0.054 -0.040 -0.017  0.003    0
 [9,] -0.208 -0.051 -0.047  0.028  0.012  0.005  0.010    0
[10,] -0.233 -0.084 -0.066 -0.017  0.056 -0.025 -0.006    0
[11,] -0.130  0.037  0.044 -0.011 -0.034 -0.038 -0.025    0
[12,] -0.074  0.082  0.034  0.014  0.063  0.016  0.000    0
[13,]  0.016 -0.025  0.047 -0.025 -0.040  0.030  0.010    0
[14,] -0.061  0.139  0.018 -0.009 -0.014  0.015  0.009    0
[15,]  0.037  0.118 -0.046  0.056 -0.011 -0.034  0.010    0
[16,]  0.021  0.057 -0.014 -0.011 -0.027 -0.001 -0.008    0
[17,] -0.054 -0.022  0.043 -0.071  0.008 -0.075 -0.007    0
[18,] -0.023 -0.027  0.039 -0.021  0.029  0.006  0.009    0
[19,]  0.029  0.014  0.055 -0.049 -0.002 -0.035  0.021    0
[20,]  0.113  0.086  0.018  0.009 -0.001 -0.006 -0.004    0
[21,]  0.047  0.072 -0.028  0.013 -0.002  0.010  0.007    0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 计算列权重 $\boldsymbol{s}_k = \sqrt{\lambda_k}\mathcal{B}^{-1/2}\boldsymbol{\delta}_k$, $k = 1, 2, \ldots, 10$

```
# 计算列 (变量的) 权重
s1 = matrix(l, nrow = nrow(d), ncol = ncol(d), byrow = T) * d # 计算 √λ_k δ_k
s = (s1/matrix(sqrt(b), nrow = nrow(d), ncol = ncol(d), byrow = F)) * (-1) # 计算 s_k = B^{-1/2}(√λ_k δ_k)
round(s, digits = 3)
```

```
> round(s, digits = 3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7] [,8]
[1,]  0.045  0.068 -0.037  0.019 -0.021 -0.014  0.003    0
[2,]  0.139 -0.056 -0.001 -0.064 -0.031  0.014  0.004    0
[3,]  0.094 -0.100 -0.008  0.038  0.020  0.004 -0.004    0
[4,]  0.023  0.050  0.053 -0.009  0.024 -0.008  0.001    0
[5,] -0.193 -0.049  0.132  0.072 -0.103  0.033  0.007    0
[6,] -0.216 -0.086 -0.019 -0.020  0.004 -0.034  0.009    0
[7,] -0.124  0.035 -0.028 -0.008  0.012  0.036 -0.008    0
[8,] -0.095 -0.044  0.089 -0.079 -0.134 -0.164 -0.200    0
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 计算行的绝对贡献：$C_a\left(i, \boldsymbol{r}_k\right)=\dfrac{x_{i\bullet}\, r_{ki}^2}{\lambda_k}$，　$i=1,2,\ldots,n$；　$k=1,2,\ldots,R$

```
# 计算行的绝对贡献
car  = matrix(matrix(a), nrow = nrow(r), ncol = ncol(r), byrow = F) * r^2 / matrix(l^2, nrow = nrow(r), ncol = ncol(r), byrow = T)
round(car, digits = 3) # 显示结果
```

```
> round(car, digits = 3) # 显示结果
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
 [1,] 0.384 0.217 0.033 0.008 0.008 0.003 0.008 0.000
 [2,] 0.006 0.008 0.005 0.040 0.018 0.048 0.024 0.075
 [3,] 0.002 0.004 0.016 0.007 0.053 0.001 0.117 0.038
 [4,] 0.010 0.004 0.203 0.175 0.030 0.038 0.022 0.013
 [5,] 0.000 0.003 0.004 0.000 0.000 0.029 0.272 0.000
 [6,] 0.001 0.028 0.001 0.126 0.005 0.000 0.006 0.032
 [7,] 0.092 0.002 0.063 0.003 0.033 0.243 0.055 0.064
 [8,] 0.087 0.105 0.131 0.193 0.136 0.039 0.005 0.017
 [9,] 0.161 0.026 0.061 0.034 0.009 0.002 0.041 0.068
[10,] 0.128 0.044 0.077 0.008 0.119 0.038 0.009 0.007
[11,] 0.027 0.006 0.023 0.002 0.028 0.057 0.104 0.003
[12,] 0.023 0.074 0.037 0.011 0.272 0.027 0.000 0.042
[13,] 0.001 0.007 0.071 0.033 0.112 0.099 0.043 0.000
[14,] 0.009 0.117 0.005 0.002 0.008 0.013 0.020 0.026
[15,] 0.006 0.152 0.064 0.161 0.008 0.116 0.041 0.206
[16,] 0.002 0.036 0.006 0.007 0.049 0.000 0.027 0.072
[17,] 0.003 0.001 0.015 0.069 0.001 0.159 0.006 0.000
[18,] 0.004 0.016 0.092 0.042 0.110 0.007 0.074 0.275
[19,] 0.002 0.001 0.047 0.060 0.000 0.062 0.091 0.025
[20,] 0.038 0.059 0.007 0.003 0.000 0.003 0.004 0.034
[21,] 0.014 0.088 0.038 0.014 0.001 0.015 0.031 0.001
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 计算列的绝对贡献：$C_a\left(j, s_k\right)=\dfrac{x_{\bullet j}\ s_{kj}^2}{\lambda_k}$，　$j=1,\ 2,\ \ldots,\ p$；　$k=1,\ 2,\ \ldots,\ R$

```
# 计算列的绝对贡献
cas  = matrix(matrix(b), nrow = nrow(s), ncol = ncol(s), byrow = F) * s^2 / matrix(l^2, nrow = nrow(s), ncol = ncol(s), byrow = T)
round(cas, digits = 3) # 显示结果
```

```
> round(cas, digits = 3) # 显示结果
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
[1,] 0.038 0.229 0.192 0.083 0.132 0.087 0.015 0.225
[2,] 0.172 0.073 0.000 0.452 0.137 0.046 0.012 0.107
[3,] 0.120 0.356 0.006 0.239 0.086 0.006 0.025 0.162
[4,] 0.010 0.124 0.404 0.019 0.181 0.031 0.004 0.228
[5,] 0.083 0.014 0.290 0.140 0.376 0.060 0.011 0.026
[6,] 0.379 0.161 0.022 0.041 0.002 0.234 0.063 0.098
[7,] 0.197 0.042 0.075 0.010 0.031 0.410 0.082 0.152
[8,] 0.002 0.001 0.011 0.014 0.054 0.126 0.789 0.002
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 分别取前两个因子

```
# 分别取前两个因子
rr = r[, 1:2]
round(rr, digits = 3) # 显示结果
ss = s[, 1:2]
round(ss, digits = 3) # 显示结果
```

```
> round(rr, digits = 3) # 显示结果
        [,1]   [,2]
 [1,]  0.146 -0.068
 [2,] -0.060  0.041
 [3,]  0.032  0.026
 [4,] -0.069 -0.029
 [5,] -0.007  0.021
 [6,] -0.027  0.076
 [7,] -0.192 -0.019
 [8,] -0.128 -0.086
 [9,] -0.208 -0.051
[10,] -0.233 -0.084
[11,] -0.130  0.037
[12,] -0.074  0.082
[13,]  0.016 -0.025
[14,] -0.061  0.139
[15,]  0.037  0.118
[16,]  0.021  0.057
[17,] -0.054 -0.022
[18,] -0.023 -0.027
[19,]  0.029  0.014
[20,]  0.113  0.086
[21,]  0.047  0.072
```

```
> round(ss, digits = 3) # 显示结果
        [,1]   [,2]
[1,]  0.045  0.068
[2,]  0.139 -0.056
[3,]  0.094 -0.100
[4,]  0.023  0.050
[5,] -0.193 -0.049
[6,] -0.216 -0.086
[7,] -0.124  0.035
[8,] -0.095 -0.044
```

Correspondence Analysis in Practice 实践中的对应分析

- Example: 1976 年法国中学会考的 202,100 个观测数据的对应分析.
  - 绘图

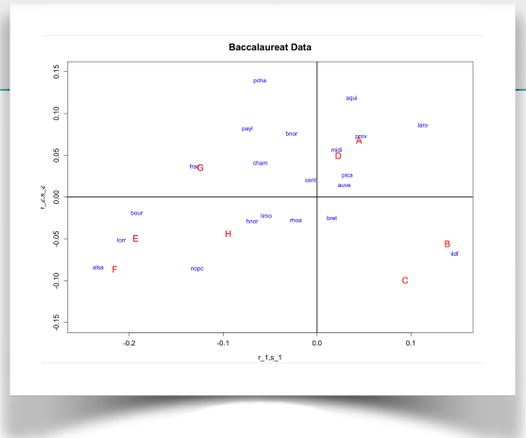
```
if (wcors == 0) {
  # labels for modalities
  types   = c("A", "B", "C", "D", "E", "F", "G", "H")

  # labels for regions
  regions = c("ildf", "cham", "pica", "hnor", "cent", "bnor", "bour", "nopc", "lorr", "alsa", "frac", "payl", "bret", "pcha", "aqui", "midi", "limo", "rhoa", "auve", "laro", "prov")

  # plot 1
  plot(rr, type = "n", xlim = c(-0.25, 0.15), ylim = c(-0.15, 0.15), xlab = "r_1,s_1", ylab = "r_2,s_2", main = "Baccalaureat Data", cex.axis = 1.2, cex.lab = 1.2, cex.main = 1.6)
  points(ss, type = "n")
  text(rr, regions, col = "blue")
  text(ss, types, cex = 1.5, col = "red")
  abline(h = 0, v = 0, lwd = 2)
} else {
  # labels for modalities
  types   = c("A", "B", "C", "D", "E", "F", "G", "H")

  # labels for regions
  regions = c("ildf", "cham", "pica", "hnor", "cent", "bnor", "bour", "nopc", "lorr", "alsa", "frac", "payl", "bret", "pcha", "aqui", "midi", "limo", "rhoa", "auve", "laro", "prov", "cors")

  # plot 2
  plot(rr, type = "n", xlim = c(-0.2, 0.25), ylim = c(-0.5, 0.15), xlab = "r_1,s_1", ylab = "r_2,s_2", main = "Baccalaureat Data", cex.axis = 1.2, cex.lab = 1.2, cex.main = 1.6)
  points(ss, type = "n")
  text(rr, regions, col = "blue")
  text(ss, types, cex = 1.5, col = "red")
  abline(h = 0, v = 0, lwd = 2)
}
```



ysis in Practice 实践中的对应分析

Economics and Social Sciences

Industrial Techniques

在首个轴上来看，学位 B 在最右端，F 在最左端，第一个轴主要反映这两者之间的变化.

ysis in Practice 实践中的对应分析

ysis in Practice 实践中的对应分析

从大区分析，第二个轴上与巴黎大区相对的是普瓦图-夏郎德、阿基坦.

从大区分析，第一个轴上巴黎大区在最右侧，与之相对的是阿尔萨斯、洛林.

ysis in Practice 实践中的对应分析

图右边的四个学位是经典的学士学位.

Philosophy-Letters

Mathematics and Natural Sciences

Economic Techniques

Computer Techniques

Mathematics and Techniques

Economics and Social Sciences

Industrial Techniques

Mathematics and Physics

图左边的四个学位则是技术性更强的学士学位.

ysis in Practice 实践中的对应分析

ysis in Practice 实践中的对应分析

Correspondence Analysis in Practice 实践中的对应分析

- Example: MASS 包中 caith 数据集的对应分析
- Example: 收入与品牌选择数据的对应分析