第十七讲 列联表与对应分析

2025.5.12

离散多元分析

多项分布

多项分布

假设每次随机试验有m种可能的互斥事件，概率各为$p_1,\dots,p_m$，满足$p_1+\cdots+p_m=1, p_i\geq 0$，记$n$次独立随机试验中各个事件发生的个数分别为$x_1,\dots,x_m$，则其联合概率函数

$$p(x_1,\dots,x_m)=\frac{n!}{\prod_{i=1}^m x_i!}\prod_{i=1}^m {p_i}^{x_i},\quad \sum_{i=1}^m x_i=n, x_i\geq 0$$

我们称$\mathbf{x}=(x_1,\dots,x_m)^\top$服从多项分布$M_m(n,(p_1,\dots,p_m))$。

容易证明:

命题1 . 若 $(x_1,\dots,x_m)\sim M_m(n,(p_1,\dots,p_m)), p_1+\cdots+p_m=1$, 则

(a) 合并一些格子计数之后(概率也同样合并)仍然服从多项分布.

(b) 对任何下标集$S=\{i_1,\dots,i_k\}\subset\{1,\dots,m\}$，$x_S=(x_{i_1},\dots,x_{i_k})$

$$x_S|_{x_{i_1}+\cdots+x_{i_k}=N}\sim M_k(N,(p_{i_1},\dots,p_{i_k})/(p_{i_1}+\cdots+p_{i_k})),$$

且x_S与x_{-S}条件独立.

(c) $(x_1,\dots,x_m)=(y_1,\dots,y_m)|y_1+\cdots+y_m=n$, 其中$y_1,\dots,y_m$独立，$y_i\sim Poisson(\lambda p_i)$.

(a)的说明： 比如5项分布 $(x_1,\dots,x_5)\sim M_5(n,(p_1,\dots,p_5))$ 合并3成项分布:

$(x_1+x_5,x_2,x_3+x_4)\sim M_3(n,(p_1+p_5,p_2,p_3+p_4))$。

命题2. 假设$\mathbf{x}=(x_1,\dots,x_m)^\top \sim M_m(n,\mathbf{p}),\mathbf{p}=(p_1,\dots,p_m)^\top$,则

(a) $E(\mathbf{x})=n\mathbf{p}$, 即$E(x_i)=np_i$;

(b) $\mathrm{var}(x_i)=np_i(1-p_i)$,

$\mathrm{cov}(x_i,x_j)=-np_ip_j,\ i\neq j$;

即$var(\mathbf{x})=n[diag(\mathbf{p})-\mathbf{p}\mathbf{p}^\top]\triangleq n\Sigma$,

$$\Sigma=\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_mp_1 & -p_mp_2 & \cdots & p_m(1-p_m) \end{pmatrix}$$

(c) $\frac{1}{\sqrt{n}}(\mathbf{x}-n\mathbf{p})\xrightarrow{d}N_m(0,\Sigma),n\to\infty$. （退化多元正态, $rank(\Sigma)=m-1$）

(d) $(\mathbf{x}-n\mathbf{p})^\top(n\Sigma)^-(\mathbf{x}-n\mathbf{p})=\sum_{i=1}^m\frac{(x_i-np_i)^2}{np_i}\xrightarrow{d}\chi^2_{m-1}$.

证明: (c)中心极限定理;

(d) 下面验证$A=\mathrm{diag}\left(\frac{1}{p_1},\dots,\frac{1}{p_m}\right)$是$\Sigma$的广义逆:

$$\mathrm{diag}\left(\frac{1}{p_1},\dots,\frac{1}{p_m}\right)diag(\mathbf{p})=I_m,\ \mathrm{diag}\left(\frac{1}{p_1},\dots,\frac{1}{p_m}\right)\mathbf{p}=\mathbf{1}_m$$

$$\Rightarrow\ A\Sigma=\mathrm{diag}\left(\frac{1}{p_1},\dots,\frac{1}{p_m}\right)[diag(\mathbf{p})-\mathbf{p}\mathbf{p}^\top]=I_m-\mathbf{1}_m\mathbf{p}^\top$$

$$\begin{aligned}\Rightarrow\ \Sigma A\Sigma &= [diag(\mathbf{p})-\mathbf{p}\mathbf{p}^\top](I_m-\mathbf{1}_m\mathbf{p}^\top)\\ &= diag(\mathbf{p})-\mathbf{p}\mathbf{p}^\top-diag(\mathbf{p})\mathbf{1}_m\mathbf{p}^\top+\mathbf{p}\mathbf{p}^\top\mathbf{1}_m\mathbf{p}^\top\\ &= diag(\mathbf{p})-\mathbf{p}\mathbf{p}^\top=\Sigma.\end{aligned}$$

$diag(\mathbf{p})\mathbf{1}_m=\mathbf{p}$
$\mathbf{p}^\top\mathbf{1}_m=1$

拟合优度检验: Pearson卡方

H_0: $p_i = p_{i0}$, p_{i0}已知. H_0成立时，

$$X^2 = \sum_{i=1}^{m} \frac{(x_i - np_{i0})^2}{np_{i0}} \triangleq \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{\mathrm{d}} \chi^2_{m-1},$$

称为Pearson拟合优度卡方检验 (Pearson,1900) 。

Pearson (1900) 提出了多项分布的拟合优度Pearson卡方检验，标志着现代统计的开端。Pearson(1904)提出列联表两个属性变量独立性的Pearson卡方检验.

- K. Pearson (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. 拟合优度
- K. Pearson (1904). On Contingency and its Relation to Association and Normal Correlation. 列联表

注意 $\sum_{i=1}^{m}\frac{(x_i-np_i)^2}{np_i}$ 加项分母是均值 $E_i=np_i$ 而不是方差$V_i=np_i(1-p_i)$，这是Pearson卡方令人疑惑的地方。 X^2为什么不是标准化后的平方和：

$$\sum_{i=1}^{m}\left(\frac{x_i-E_i}{\sqrt{V_i}}\right)^2=\sum_{i=1}^{m}\left(\frac{x_i-np_i}{\sqrt{np_i(1-p_i)}}\right)^2?$$

1. 简单理由：由命题1(c), $x_1,\dots,x_m$可看作是独立的 $x_i\sim Pois(np_i)$, 则

$$E(x_i)=var(x_i)=np_i,$$

故 $\frac{x_i-np_i}{\sqrt{np_i}}=\frac{x_i-E(x_i)}{\sqrt{var(x_i)}}$ 确是Poisson变量的标准化。

2. 复杂理由：约束 $\sum_{i=1}^{m}x_i=n$ 导致分母上是均值。

多项分布中x_i 的标准化确实应该是 $\frac{x_i-E_i}{\sqrt{V_i}}$, 但约束 $\sum_{i=1}^{m}x_i=n$导致不同 x_i 之间负相关， 需要将这些相关性考虑在内（即命题2 (d)中的Σ^-）。

例：二项分布（$m=2$情形）

假设 $(x_1,x_2)\sim M_2\big(n,(p_1,p_2)\big)$, 通常记作 $x_1\sim B(n,p_1)$,则Pearson卡方

$$\frac{(x_1-np_1)^2}{np_1}+\frac{(x_2-np_2)^2}{np_2}=\frac{(x_1-np_1)^2}{np_1}+\frac{(n-x_1-n(1-p_1))^2}{n(1-p_1)}$$

$$=(x_1-np_1)^2\left(\frac{1}{np_1}+\frac{1}{n(1-p_1)}\right)$$

$$=\frac{(x_1-np_1)^2}{np_1(1-p_1)}$$

$$=\left(\frac{x_1-np_1}{\sqrt{np_1(1-p_1)}}\right)^2\triangleq z^2$$

$x_2=n-x_1$
$p_2=1-p_1$

z是x_1的通常的中心标准化，而左端$\frac{(x_1-np_1)^2}{np_1}+\frac{(x_2-np_2)^2}{np_2}$中的

$$\frac{x_1-np_1}{\sqrt{np_1}},\frac{x_2-np_2}{\sqrt{np_2}}$$

是约束 $x_1+x_2=n$ 条件下x_1，x_2的同时中心标准化.

列联表与Pearson卡方

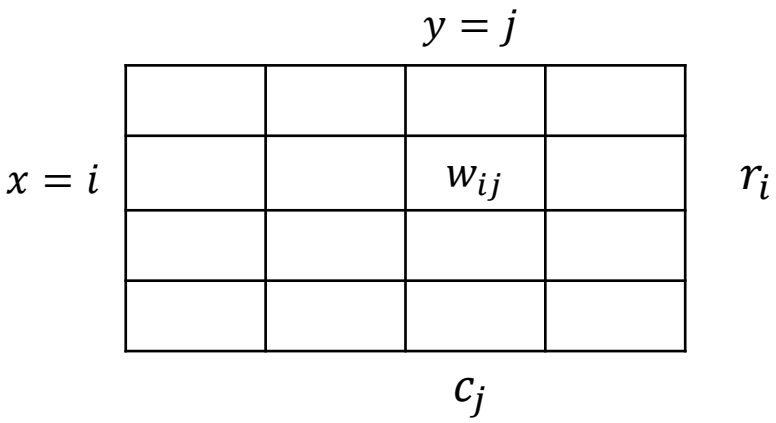
列联表

如果多项分布的格子由两个属性变量划分而成，则是列联表问题. 列联表分析主要考察两个属性变量的关系。为了与连续变量情形的相关系数概念区分，Pearson将属性变量之间的关联性度量称为contingency，将属性变量交叉分类得到的计数表格称为列联表contingency table 。

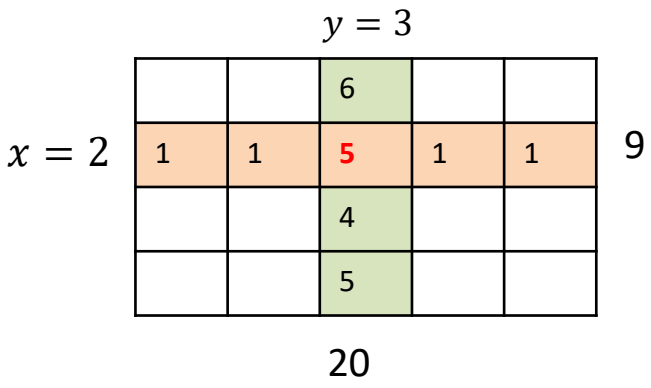
Contingency：一个事件可能引发的事件(依赖性)、应急、临时、附加条款。

假设属性变量x, y各有p, q个类别或水平。记 w_{ij}为 x 取水平i， y 取水平 j 的样本频数， $W_{p\times q} = (w_{ij})$称为$p \times q$ 列联表(contingency table) 或交叉分类表(cross tabulation)。关心的问题：

(1) x, y是否独立; (2) 如果不独立，x 与 y 哪个水平之间关联性最强?



例1. $x = i$ 与 $y = j$ 共同发生的次数 w_{ij} 代表了水平 i 与 j 的关联度，但考察计数w_{ij}的同时应参考其所在行、列的其它计数，特别地需要考虑各行各列的边际总数。



$x = i$：x取值为第i个水平，这里的代表水平/类别，不是通常的实数

表中$w_{23} = 5$， $x = 2$ 与$y = 3$ 的关联大还是小？

$x_{23} = 5$在第2行最大, 似乎$x = 2$ 与$y = 3$关联程度较高。但观察第 3 列， 5在第 3 列中并不大，因此 $x = 2$ 与$y = 3$ 关联程度可能并不高。

另一方面，$x_{23} = 5$在第3列中中等大小，这似乎说明$y = 3$与$x = 2$关联程度不高，但它在第2行是最大的计数。

如何综合地考察w_{ij}相对于所在行总和、列总和的大小并不是简单的问题，也没有唯一答案。

例2. $n = 781$件出土陶器按考古地点(0-6)和类型(a-d)两个属性变量交叉分类得到下述列联表。(i,j)格子元素x_{ij}为地点i出土的第j类陶器的个数/计数。

	a	b	c	d	总计
0	30	10	10	39	89
1	53	4	16	2	75
2	73	1	41	1	116
3	20	6	1	4	31
4	46	36	37	13	132
5	45	6	59	10	120
6	16	28	169	5	218
总计	283	91	333	74	781

不同考古地点代表不同的时期或文化，因此

- 陶器类型分布相似的考古地点，其年代可能接近；
- 通过研究陶器类型分布判断考古地点的年代次序关系和交流情况。

列联表矩阵与一般的数据矩阵略有不同，其元素非负，代表了关联大小，与此类似的是强度矩阵（intensity）、丰度矩阵(abundance)、网络的邻接矩阵等矩阵，矩阵(i,j)元素x_{ij}代表行标i与列标j的联系紧密程度，恰当的标准化对于此类矩阵分析尤其重要。

如上所述，考察w_{ij}大小的时候需要综合考虑其所在行、列的其它计算，需要考虑 w_{ij} 相对于行计数总和 r_i 与列计数总和 c_j 的大小，并作某种归一化或标准化。

记号

假设 $p \times q$ 列联表 $X = (w_{ij})$，定义

- 行和: $\mathbf{r} = W\mathbf{1}_q = (r_1, \ldots, r_p)^\top$
  列和: $\mathbf{c} = W^\top \mathbf{1}_p = (c_1, \ldots, c_q)^\top$
- $D_r = diag(\mathbf{r})$; $D_c = diag(\mathbf{c})$
- 行归一化: $P = (w_{ij}/r_i) = D_r^{-1}W$，其每一行的和都是1
  列归一化: $Q = (w_{ij}/c_j) = WD_c^{-1}$，其每一列的和都是1

W行归一化后， P的每一行的总和为1，各行之间具有可比性，我们认为 x 各水平的特征刻画，比如第 i 行

$$\mathbf{p}_i = (w_{i1}, \ldots, w_{iq})/r_i$$

是属性 x 的的水平 i 的特征刻画。同样Q的第 j 列

$$\mathbf{q}_j = (w_{1j}, \ldots, w_{pj})/c_j$$

是属性 y 的的水平 j 的特征刻画.

多项分布

假设列联表计数来自于 n次独立随机试验, 每次试验结果 $(x,y)=(i,j)$ 的概率p_{ij}， 所有概率为 $p\times q$ 矩阵：

$$P=\left(p_{ij}\right), \sum_{i,j} p_{ij}=1,$$

所有计数服从多项分布:

$$p\left(w_{ij},\, 1\le i\le p, 1\le j\le q\right)=\frac{n!}{\prod_{i,j} w_{ij}!}\prod_{i,j} {p_{ij}}^{w_{ij}},\quad \sum_{i,j} w_{ij}=n, w_{ij}\ge 0。$$

记作 $W_{p\times q}=\left(w_{ij}\right)\sim M_{pq}(n,P)$。

该分布与P2定义的多项分布无异， 只是这里增加了矩阵结构，
比如由命题1(b), 给定各个行总和, 各行服从多项分布：

$$\left(w_{i1},\dots,w_{iq}\right)|_{\sum_{j=1}^{q} w_{ij}=r_i}\sim M_q\left(r_i,\left(p_{i1},\dots,p_{iq}\right)/p_{i+}\right),$$

$$p_{i+}=p_{i1}+\cdots+p_{iq}$$

且各行$\left(w_{i1},\dots,w_{iq}\right)\ i=1,\dots,p$ 条件独立。同样给定列总和，
各列也服从独立的多项分布。

注: 因为这个原因， Pearson卡方也可用来检验多个多项分布的齐一性

列联表的Pearson卡方检验

Pearson卡方检验：随机独立试验下，两个属性变量 x, y 交叉分类得到 $p \times q$ 列联表 $W = (w_{ij})$。独立性零假设 $H_0: x \perp\!\!\!\perp y$ 成立时，Pearson卡方检验统计量

$$X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(w_{ij} - r_i c_j/n)^2}{r_i c_j/n} \xrightarrow{H_0} \chi^2_{(p-1)(q-1)}$$

其中 $r_i = \sum_{j=1}^{q} w_{ij}, c_j = \sum_{i=1}^{p} w_{ij}$。

由命题1(4): $\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(w_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(w_{ij}-np_{ij})^2}{np_{ij}} \xrightarrow{H_0} \chi^2_{pq-1}$ （*）

H_0: $x \perp\!\!\!\perp y$ 成立时，$p_{ij} = P(x=i, y=j) = P(x=i)P(y=j) = p_{i+}p_{+j}$，故 $Ew_{ij} = np_{ij} = np_{i+}p_{+j}$ 与未知参数有关，其估计:

$$\hat{E}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n \times \frac{r_i}{n} \times \frac{c_j}{n} = \frac{r_i c_j}{n},$$

代入（*）式得Pearson卡方检验统计量

$$X^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(w_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = n\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(w_{ij}-r_i c_j/n)^2}{r_i c_j} \xrightarrow{H_0} \chi^2_{(p-1)(q-1)}$$

注：估计共 $k = p + q - 2$ 个 p_{i+}，p_{+j} 后，原自由度 $pq-1$ 降为 $(pq-1)-k = (p-1)(q-1)$.

下面对X^2平方和中出现的“误差项”$\phi_{ij} \triangleq \frac{w_{ij}-r_i c_j/n}{\sqrt{r_i c_j}}$予以说明。

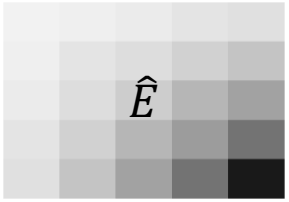
W的双向中心化

记独立下的期望估计 $\hat{E} = \left(\frac{r_i c_j}{n}\right) = \frac{\mathbf{r}\mathbf{c}^\top}{n}$，$W$偏离$\hat{E}$的部分

$$W_c = W - \hat{E} = W - \frac{\mathbf{r}\mathbf{c}^\top}{n} = \left(w_{ij} - \frac{r_i c_j}{n}\right)$$

是W的双向中心化(行和列和都为0)。

$\hat{E} = (r_i c_j/n)$ 是独立情形下W的期望（背景轮廓configuration，不包含关联信息)



我们将属性变量x, y分别用示性变量/哑变量/one-hot 表示为

$$\mathbf{x} = (x_1, \ldots, x_p)^\top, \mathbf{y} = (y_1, \ldots, y_q)^\top,$$

其中$x_i = 1_{(x=i)}, y_j = 1_{(y=j)}$.

One-hot:
x或**y**的分量仅有一个 1，其余全是0

x 水平 i 与 y 水平 j 之间的偏离独立性/关联性度量：

$$\begin{aligned} & P(x=i, y=j) - P(x=i)P(y=j) \\ &= P(x_i = 1, y_j = 1) - P(x_i = 1)P(y_j = 1) \\ &= E(x_i y_j) - E(x_i)E(y_j) = \mathrm{cov}(x_i, y_j) \end{aligned}$$

假设 n 个独立样本为 $(\mathbf{x}_k, \mathbf{y}_k), k = 1, \ldots, n,\ \mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^\top, \mathbf{y}_k = (y_{k1}, \ldots, y_{kq})^\top$ 分别是两个属性变量的示性表示, 其样本协方差矩阵为

$$S = \begin{pmatrix} S_{\mathbf{xx}} & S_{\mathbf{xy}} \\ S_{\mathbf{yx}} & S_{\mathbf{yy}} \end{pmatrix} = \frac{1}{n-1} \begin{pmatrix} D_r - \mathbf{r}\mathbf{r}^\top/n & W_c \\ W_c^\top & D_c - \mathbf{c}\mathbf{c}^\top/n \end{pmatrix}$$

其中 $S_{\mathbf{xy}} = W_c/(n-1)$, 所以中心化列联表 $W_c = (n-1)\, S_{\mathbf{xy}}$ 代表了 $\mathbf{x}_k$和 $\mathbf{y}_k, k = 1, \ldots, n,$ 的样本协方差矩阵.

与典则相关分析中的标准化完全相同，W_c或$S_{\mathbf{xy}}$的标准化：

$$W_s = S_{\mathbf{xx}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{yy}}^{-1/2} = [(D_r - \mathbf{r}\mathbf{r}^\top/n)^-]^{1/2} W_c [(D_c - \mathbf{c}\mathbf{c}^\top/n)^-]^{1/2}$$
$$= D_r^{-1/2} W_c D_c^{-1/2} = \left(\frac{w_{ij} - r_i c_j/n}{\sqrt{r_i c_j}} \right)$$

W的中心标准化

W的中心标准化矩阵:$W_s = \left(\frac{w_{ij} - r_i c_j/n}{\sqrt{r_i c_j}} \right) = D_r^{-1/2} W_c\, D_c^{-1/2}$

该矩阵消除了独立情形下的背景, 并校正了行和、列和的大小。

Pearson卡方：$X^2 = n\|W_s\|^2$, $\|W_s\|^2$称为contingency度量.

中心标准化矩阵 W_s 的 (i,j) 元

$$\phi_{ij} = \frac{w_{ij} - r_i c_j / n}{\sqrt{r_i c_j}}$$

是w_{ij}的标准化残差，描述了偏离独立性的大小， 独立性假设 $H_0: x \perp\!\!\!\perp y$ 的Pearson卡方检验统计量

$$X^2 = n\|W_s\|^2 = n\sum \phi_{ij}^2$$

该检验是omnibus检验，总体上检验 x, y 的各水平之间的关联性。

如果x, y不独立，我们希望了解关联的模式 - 具体哪些水平之间呈现了关联性？这称为事后分析(post hoc analysis), 可以通过观察残差ϕ_{ij}的大小来发现。

对应分析利用$W_s = (\phi_{ij})$ 的奇异值分解将两个属性变量水平之间的关联性用双标图展示。

对应分析

对应分析简介

对列联表应用SVD和双标图可视化方法，研究行标和列标之间的对应或关联性的方法称为对应分析（correspondence analysis，课本12.7），由法国学者Jean-Paul Benzécri（1973）提出。对应分析是法国统计学派的代表性方法。

有一种观点认为，统计学作为一个学科是从法国人Pierre-Simon Laplace（1749-1827）和比利时-法国人Adolphe Quetelet(凯特勒1796-1896)开始的，后者提出了BMI。

Johnson&Wichern 原文：Developed by the French, correspondence analysis is a graphical procedure for representing associations in a table of frequencies or counts.

中译版P557：对应分析是一种将频数或计数表中的各种联系用图来表示的方法，由弗伦奇所提出.

French应译作“法国人”！

对应分析

假设标准化的列联表 $W_s = D_r^{-1/2} W_c D_c^{-1/2}$ 有如下奇异值分解：

$$W_s = D_r^{-1/2} W_c D_c^{-1/2} = UD\,V^{\top}$$

其中 $U = (\mathbf{u}_1, \ldots, \mathbf{u}_k), V = (\mathbf{v}_1, \ldots, \mathbf{v}_k), D = diag\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\right),$ $\lambda_1 \geq \cdots \geq \lambda_k > 0, k = rank(W_c)$。

UD 是列的主成分（用之表示行），VD 是行的主成分（用之表示列）。由于 $W_c = D_r^{1/2} UD\,V^{\top} D_c^{1/2}$，故我们以 $D_r^{1/2} UD$ 表示 W的行属性 x，进而，以 $D_r^{-1}\left(D_r^{1/2} UD\right) = D_r^{-1/2} UD$ 表示 $P = D_r^{-1} W$ 的行属性 x。同理以 $D_c^{-1/2} VD$ 表示 $Q = W D_c^{-1}$ 的 列属性 y

定义

- 行属性的主坐标（Principal coordinate of row）为 $P = D_r^{-1} W$ 的行表示：

$$F_{p\times k} = D_r^{-1/2} UD = (D_r^{-1/2}\mathbf{u}_1\sqrt{\lambda_1}, \ldots, D_r^{-1/2}\mathbf{u}_k\sqrt{\lambda_k})$$

- 列属性的主坐标（Principal coordinate of col）为 $Q = W\,D_c^{-1}$ 的列表示：

$$G_{q\times k} = D_c^{-1/2} VD = (D_c^{-1/2}\mathbf{v}_1\sqrt{\lambda_1}, \ldots, D_c^{-1/2}\mathbf{v}_k\sqrt{\lambda_k})$$

双标图

取 F 的前两列作为行属性 p 个水平的2维欧氏表示，
取 G 的前两列作为列属性 q 个水平的2维欧氏表示，
双标图：将这些2维坐标在同一个坐标系中画出散点图，
两个点与点之间的距离近似代表属性水平之间的关联性。

下面解释为什么双标图点之间的距离代表关联性

引理1. 若矩阵$A = (a_{ij})$是双中心化的（行和、列和为0），则

$$\mathbf{u}^{\top} A\mathbf{v} = -\frac{1}{2}\sum_{i,j} a_{ij}(u_i - v_j)^2$$

证明：因为$\sum_i a_{ij} = \sum_j a_{ij} = 0$, 所以

$$\sum_{i,j} a_{ij}u_i^2 = \sum_i u_i^2 \sum_j a_{ij} = 0, \quad \sum_{i,j} a_{ij}v_j^2 = 0.$$

从而

$$\sum_{i,j} a_{ij}(u_i - v_j)^2 = \sum_{i,j} a_{ij}u_i^2 + \sum_{i,j} a_{ij}v_j^2 - 2\sum_{i,j} a_{ij}u_i v_j$$
$$= -2\sum_{i,j} a_{ij}u_i v_j.$$

为简便计，我们只考虑一维坐标。假设$f_1,\dots,f_p$是属性变量x的p个水平的一维数值表示，$g_1,\dots,g_q$是属性变量y的q个水平的数值表示。我们希望$w_{ij}>\frac{r_ic_j}{n}$时，f_i与g_j接近。

命题3. 考虑优化问题

$$\min\sum_{i,j}\left(w_{ij}-\frac{r_ic_j}{n}\right)(f_i-g_j)^2\text{，s.t. }||D_r^{1/2}\mathbf{f}||=||D_c^{1/2}\mathbf{g}||=1,$$

其中$\mathbf{f}=(f_1,\dots,f_p)^\top\in R^p,\mathbf{g}=(g_1,\dots,g_q)^\top\in R^q$，则最优解为$\mathbf{f}=D_r^{-1/2}\mathbf{u}_1,\mathbf{g}=D_c^{-1/2}\mathbf{v}_1$，其中$\mathbf{u}_1$，$\mathbf{v}_1$是$W_s$的第一奇异特征向量。

证明：由引理1，

$$\sum_{i,j}\left(w_{ij}-\frac{r_ic_j}{n}\right)(f_i-g_j)^2=-2\mathbf{f}^\top W_c\mathbf{g}$$

$$=-2\mathbf{u}^\top D_r^{-\frac{1}{2}}W_cD_c^{-\frac{1}{2}}\mathbf{v}=-2\mathbf{u}^\top W_s\,\mathbf{v}$$

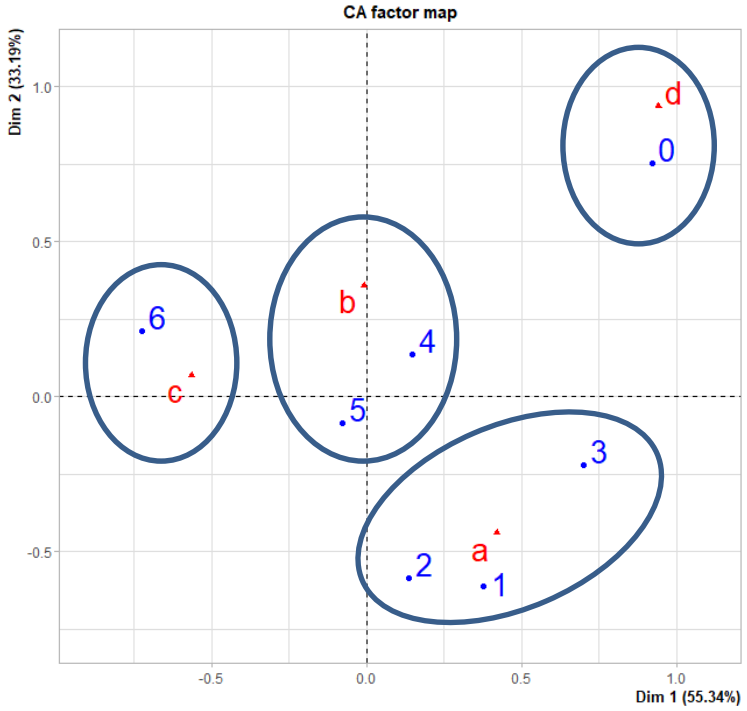
令$\mathbf{u}=D_r^{1/2}\mathbf{f}$，$\mathbf{v}=D_c^{1/2}\mathbf{g}$

由16讲定理3，$\mathbf{u}=\mathbf{u}_1$，$\mathbf{v}=\mathbf{v}_1$时达到最小。

此时$\mathbf{f}=D_r^{-1/2}\mathbf{u}_1$，$\mathbf{g}=D_c^{-1/2}\mathbf{v}_1$。

类似地，对应分析的前2个主坐标也有类似的最优性。

例2（续）对应分析。双标图显示四种陶器差别较大（b，c比较接近），地点1，2，3很相似，它们出土较多的a类型陶器。（d，0）联系密切，与其他遗址和陶器差别较大。



陶器在第一主坐标方向方向的相邻次序为 c-b-a-d,我们认为可能代表不同年代风格连续变化的次序，考古地点的时间次序可能是

6-5-4-2-1-3-0

c---b----a----d

```
library(FactoMineR)
CA(contingency_table)
```

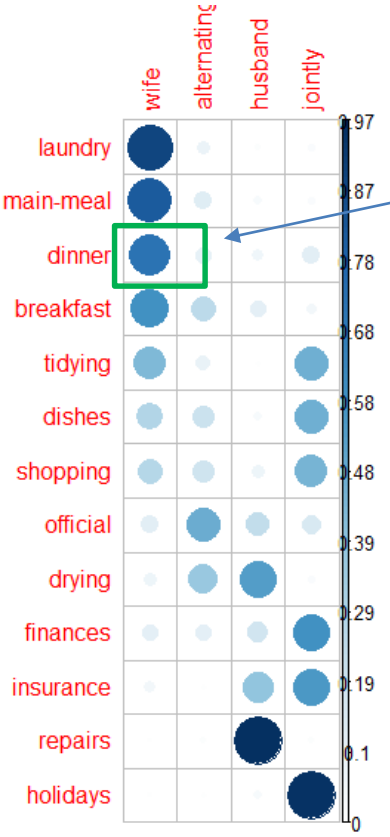
例3（家务分配）一项社会问卷调查询问了1774个家庭的13项家务活分配情况，共有4种分配方案：wife，alternating，husband，jointly. 原始数据如下：

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

比如，有156个家庭的妻子负责洗衣(laundry)，14个家庭轮流(alternating).

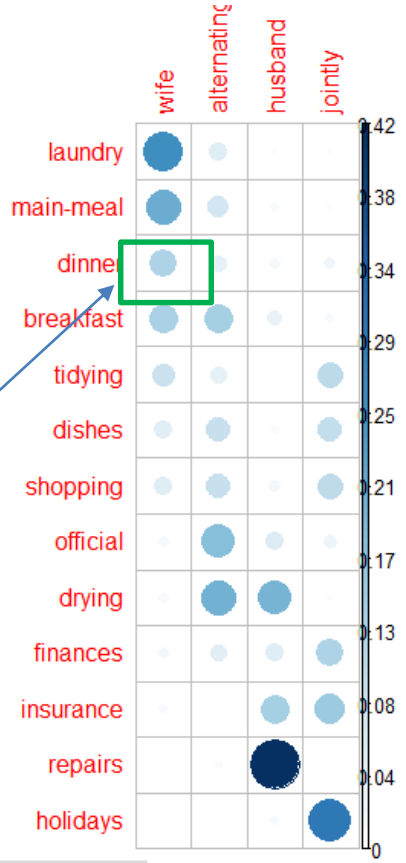
行归一化（行和为1），前4行类似，都是同一类型的家务(圆圈大小表示概率），与马赛克图类似，更直观：

列归一化（列和为1）：

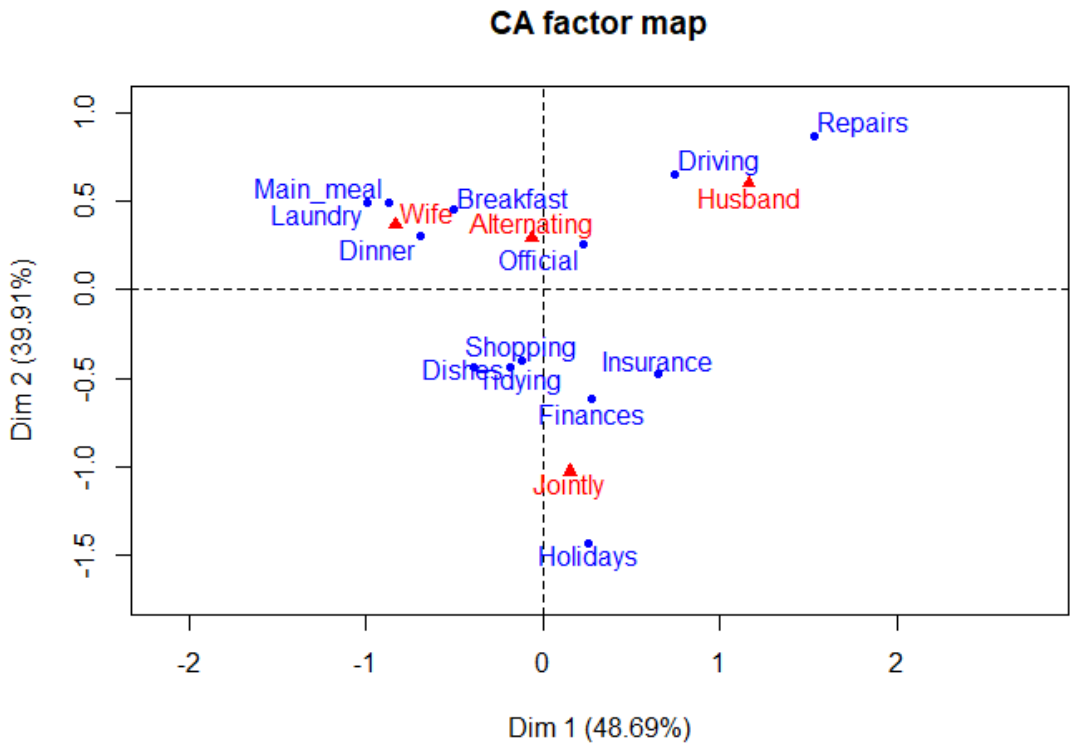


(dinner, wife):
左图第三行， 在dinner的4种完成方式中，wife完成的比例最大

而右图圆圈较小,在wife的家务活中，dinner占有稍小的（第4大）比例。



Key：对应关系/圆圈的大小要结合左右两个图来看，这正是对应分析方法的关键，即同时考虑行和与列和。



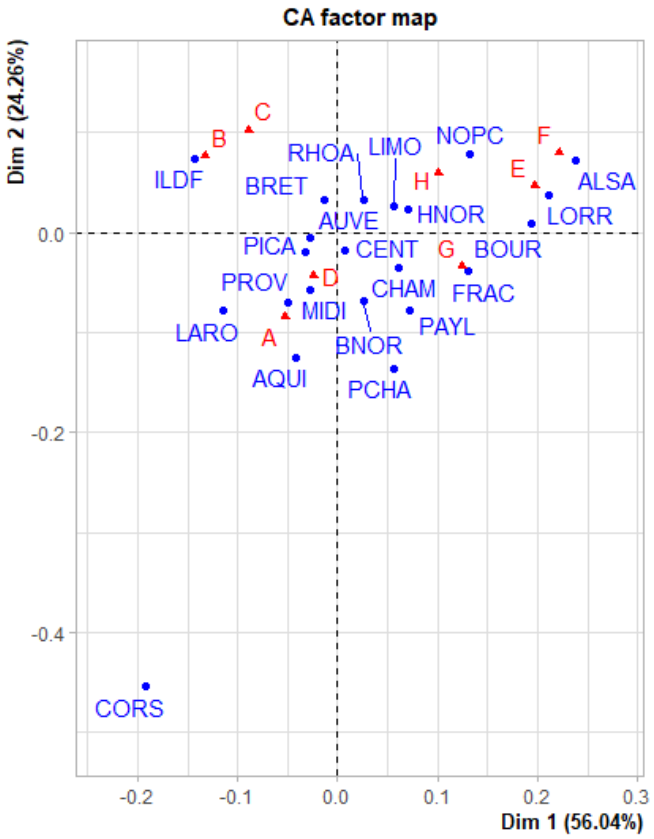
双标图表明： 妻子主要负责做饭洗衣，丈夫负责驾驶和修理房子，一起做的家务主要是购物、财务、保险和节日活动。

例4. 1976年法国本土22个大区的202100位大学毕业生的专业统计如下，8个专业(A-H)如下：

	A	B	C	D	E	F	G	H
ILDF	9724	5650	8679	9432	839	3353	5355	83
CHAM	924	464	567	984	132	423	736	12
PICA	1081	490	830	1222	118	410	743	13
HNOR	1135	587	686	904	83	629	813	13
CENT	1482	667	1020	1535	173	629	989	26
BNOR	1033	509	553	1063	100	433	742	13
BOUR	1272	527	861	1116	219	769	1232	13
NOPC	2549	1141	2164	2752	587	1660	1951	41
LORR	1828	681	1364	1741	302	1289	1683	15
ALSA	1076	443	880	1121	145	917	1091	15
FRAC	827	333	481	892	137	451	618	18
PAYL	2213	809	1439	2623	269	990	1783	14
BRET	2158	1271	1633	2352	350	950	1509	22
PCHA	1358	503	639	1377	164	495	959	10
AQUI	2757	873	1466	2296	215	789	1459	17
MIDI	2493	1120	1494	2329	254	855	1565	28
LIMO	551	297	386	663	67	334	378	12
RHOA	3951	2127	3218	4743	545	2072	3018	36
AUVE	1066	579	724	1239	126	476	649	12
LARO	1844	816	1154	1839	156	469	993	16
PROV	3944	1645	2415	3616	343	1236	2404	22
CORS	327	31	85	178	9	27	79	0

A: Philosophy-Letters,
B: Economics and Social Sciences,
C: Mathematics and Physics,
D: Mathematics and Natural Sciences,
E: Mathematics and Techniques,
F: Industrial Techniques,
G: Economic Techniques,
H: Computer Techniques.

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：法兰西岛(大巴黎)
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 comté:弗朗什孔泰
PCHA: Poitou Charentes：普瓦图-夏朗德
LIMO: Limousin：利穆赞
AUVE: Auvergne：奥弗涅
RHOA: Rhone Alpes：罗讷-阿尔卑斯
Alpes Pays de savoie：阿尔卑斯 - 萨瓦地区
AQUI: Aquitaine：阿基坦
MIDI: Midi pyrénées：南部-比利牛斯
PROV: Provence Alpes Cote d'azur：普罗旺斯 - 蓝色海岸
LARO: Languedoc Roussilon：朗格多克-鲁西永
CORS: corse：科西嘉



科西嘉(CORS)远离其它地区。大巴黎区也较为特殊，与B,C关系密切，有较多的经济学、数理专业学生。

阿尔萨斯、洛林地区工业技术（E，F）学生较多。普罗旺斯地区较多文学艺术(A)专业学生。

香槟亚丁？农业区

普氏分析（SVD的另一个应用）

古希腊神话人物Procrustes普罗克路斯忒斯是一个强盗，客栈老板，他把每个入住的客人拉伸或将腿截断，以使其与他的铁床对齐。

有一种说法：他实际上有两个大小不同的铁床。。

Procrustean bed：不合理但需要严格遵守的规则或标准.
Procrustean solution：先定目标模型和方法，再寻找数据或删减数据以符合模型的欺骗做法。

普氏分析（Procustes analysis）将两个数据集对齐，假设有两个$n\times p$数据集X,Y,普氏分析寻找旋转、平移变换使得变换之后两者尽量相似。

目标函数

设$X_{n\times p}=(\mathbf{x}_1,...,\mathbf{x}_n)^{\mathsf{T}}, Y_{n\times p}=(\mathbf{y}_1,...,\mathbf{y}_n)^{\mathsf{T}}$, 普氏分析求解$p\times p$正交矩阵$Q$和$\mathbf{b}\in R^p$使得

$$\sum_{i=1}^{n}\|\mathbf{x}_i-(Q\mathbf{y}_i+\mathbf{b})\|^2=\|X-(YQ^{\mathsf{T}}+\mathbf{1}\mathbf{b}^{\mathsf{T}})\|_F^2=\min!$$

对于给定的正交矩阵Q，$\mathbf{b}$的最优解为$\hat{\mathbf{b}}=\bar{\mathbf{x}}-Q\bar{\mathbf{y}}$,此时目标函数

$\| X-(YQ^{\mathsf{T}}+\mathbf{1}\mathbf{b}^{\mathsf{T}})\|_F^2=\| X-\mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}-(Y-\mathbf{1}\bar{\mathbf{y}}^{\mathsf{T}})Q^{\mathsf{T}}\|_F^2$ 所以我们不妨

假设X,Y都是中心化的。因为

$$\| X-YQ^{\mathsf{T}}\|_F^2=tr(X^{\mathsf{T}}X)+tr(Y^{\mathsf{T}}Y)-2tr(X^{\mathsf{T}}YQ^{\mathsf{T}})$$

所以原极小化问题等价于极大化问题：

$$\max_{Q^{\mathsf{T}}Q=QQ^{\mathsf{T}}=I_p} tr(X^{\mathsf{T}}YQ^{\mathsf{T}})$$

其中$X^{\mathsf{T}}Y/(n-1)$为样本协方差矩阵。

命题4：若$A_{p\times p}=X^{\mathsf{T}}Y$的$SVD$分解为$A=UDV^{\mathsf{T}}$,其中$U,V$是$p\times p$正交矩阵，则上述优化问题最优的正交变换为$A$的“方向”矩阵$Q=UV^{\mathsf{T}}$，即$YQ^{\mathsf{T}}=YVU^{\mathsf{T}}$与$X$最接近。

证明：记矩阵$A=X^{\mathsf{T}}Y$的SVD分解为$A=U_{p\times p}D_{p\times p}V^{\mathsf{T}}{}_{p\times p}=\sum_{i=1}^{p}\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i{}^{\mathsf{T}}$,

$$\Rightarrow tr(X^{\mathsf{T}}YQ^{\mathsf{T}})=tr\left(\sum_{i=1}^{p}\sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i{}^{\mathsf{T}}Q^{\mathsf{T}}\right)=\sum_{i=1}^{p}\sqrt{\lambda_i}tr\left(\mathbf{u}_i\mathbf{v}_i{}^{\mathsf{T}}Q^{\mathsf{T}}\right)=\sum_{i=1}^{p}\sqrt{\lambda_i}\mathbf{v}_i{}^{\mathsf{T}}Q^{\mathsf{T}}\mathbf{u}_i,$$

由Cauchy不等式，$\mathbf{v}_i^{\mathsf{T}} Q^{\mathsf{T}} \mathbf{u}_i \le \| Q\mathbf{v}_i \| \cdot \| \mathbf{u}_i \| = 1$, 当$\mathbf{u}_i = Q\mathbf{v}_i$时达到最大，

此时$U = QV \Rightarrow \quad Q = UV^{\mathsf{T}}$,

$$\max tr(X^{\mathsf{T}} Y Q^{\mathsf{T}}) = tr(X^{\mathsf{T}} Y V U^{\mathsf{T}}) = tr(U^{\mathsf{T}} X^{\mathsf{T}} Y V) = tr(D) = \sqrt{\lambda_1} + ... + \sqrt{\lambda_p}$$

任何复数有极分解（polarvdecomposition）: $z = re^{i\theta}$ 。

任何向量$\mathbf{x} \in R^p$有极分解: $\mathbf{x} = \boldsymbol{\theta} r$, 其中$\boldsymbol{\theta} = \mathbf{x} / \| \mathbf{x} \|, r = \| \mathbf{x} \|$。

任何矩阵也有极分解: $A = \Theta R_1 = R_2 \Theta$

命题5. 假设矩阵A的SVD为$A = UDV^{\mathsf{T}}$, 则A有极分解表示

$$A = \Theta R_1 = R_2 \Theta,$$

其中$R_1 = (A^{\mathsf{T}} A)^{1/2}$和$R_2 = (AA^{\mathsf{T}})^{1/2}$代表$A$的模长,

$\Theta = UV^{\mathsf{T}}$代表A的"方向",

证: 因为$A^{\mathsf{T}} A = VDU^{\mathsf{T}} UDV^{\mathsf{T}} = VD^2 V^{\mathsf{T}}$, 所以

$$A = UDV^{\mathsf{T}} = UV^{\mathsf{T}} VDV^{\mathsf{T}} = UV^{\mathsf{T}} \left(VD^2 V^{\mathsf{T}} \right)^{1/2} = UV^{\mathsf{T}} (A^{\mathsf{T}} A)^{1/2} \triangleq \Theta (A^{\mathsf{T}} A)^{1/2}$$

其中$\Theta = UV^{\mathsf{T}}$. 同理也有 $A = (AA^{\mathsf{T}})^{1/2} \Theta$