第十八讲 距离和相似系数

2025.5.19

$$s_{ij} = \exp\left(-\frac{1}{2}d_{ij}^2\right)$$

d: 距离，s: 相似系数

内积：相似度

向量内积

向量内积

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\top}\mathbf{y} = \sum x_i y_i$$

衡量两个向量 $\mathbf{x}, \mathbf{y}$ 的相似性，当$\mathbf{y} \propto \mathbf{x}$时相似度最大。

$X^{\top}Y$

假设矩阵 $X_{n\times p}, Y_{n\times q}$ 的行代表样本，列代表变量，则

$$X = \left(\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(p)}\right),\ Y = \left(\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(q)}\right),\ X^{\top}Y = \left(\mathbf{x}_{(i)}^{\top}\mathbf{y}_{(j)}\right)$$

描述了X, Y列向量/变量之间相似性，其 (i, j) 元 $\mathbf{x}_{(i)}^{\top}\mathbf{y}_{(j)}$ 代表 $\mathbf{x}_{(i)}, \mathbf{y}_{(j)}$之间的相似性。 如果$X, Y$是列中心化的，则 $\frac{X^{\top}Y}{n-1} = S_{\mathbf{xy}}$ 是样本协方差矩阵。所以协方差或相关系数都是某种相似度量。

$Y = X$时，$X = \left(\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(p)}\right) = (\mathbf{x}_1, \dots, \mathbf{x}_n)^{\top}$

- $X^{\top}X = \left(\mathbf{x}_{(i)}^{\top}\mathbf{x}_{(j)}\right)$描述$X$各列(变量)之间的相似性；
- $XX^{\top} = \left(\mathbf{x}_i^{\top}\mathbf{x}_j\right)$描述了样本 ($X$的行)之间的相似性。

这里的“因子”与因子分析中的“因子”含义不同。

例1. (1) n个p-水平属性变量(因子变量)的 one-hot embedding $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^p$, 其中$\mathbf{x}_i$分量中仅有一个1，其余全是0。记

$$X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top = (x_{ij}),\ X\mathbb{1}_p = \mathbb{1}_p$$

$x_{ij} = 1$，若样本 i取第 j 个水平

记X第j列的总和

$$r_j = \sum_{i=1}^n x_{ij} = \text{水平 } j \text{ 的样本个数},$$

水平之间互斥，X不同列内积为0（不相似）：

$$X^\top X = \begin{pmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_p \end{pmatrix},$$

例如因子变量 size 取值大、中、小， $n = 5$ 次观测为大、小、小、中、大，以示性变量表示，样本矩阵X如下

$$X = \begin{array}{c} \text{大}\ \text{中}\ \text{小} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array},\ X^\top X = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

$$XX^\top = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(2) 假设n个样本，每个样本有两个因子变量x, y，各有 p, q个水平x_1,…, x_p和y_1,…, y_q，则数据one-hot表示为$n \times (p+q)$矩阵

$$Z = (X, Y),$$

$X\mathbb{1}_p = \mathbb{1}_p$
$Y\mathbb{1}_q = \mathbb{1}_q$

其中$X = (x_{ij}), Y = \left(y_{ij}\right)$分别是第一、二个因子的one-hot表示 。

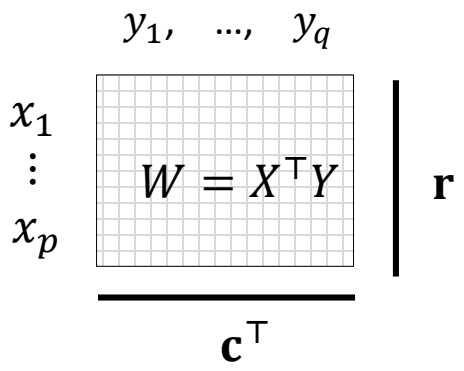
记 $\mathbf{r} = X^\top\mathbb{1}, \mathbf{c} = Y^\top\mathbb{1}$分别为$X$， Y的列边际总和，则同(1),

$$X^\top X = diag(\mathbf{r}) \triangleq D_{\mathbf{r}},\ Y^\top Y = diag(\mathbf{c}) \triangleq D_{\mathbf{c}}$$

而$W = X^\top Y = (w_{jk})$, 其 (j,k) 元

$$w_{jk} = \sum_{i=1}^{n} x_{ij}y_{ik} = \#\{x, y \text{ 水平各为} j, k \text{ 的样本个数}\}$$

则$W = X^\top Y$是两个因子x, y交叉分类得到的列联表：



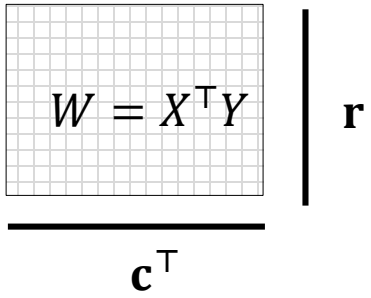
$$\mathbf{r} = X^\top\mathbb{1}_n = X^\top Y\mathbb{1}_q = W\mathbb{1}_q$$

$$\mathbf{c} = Y^\top\mathbb{1}_n = Y^\top X\mathbb{1}_p = W^\top\mathbb{1}_p$$

综上，

$$Z^\top Z = \begin{pmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{pmatrix} = \begin{pmatrix} D_{\mathbf{r}} & W \\ W^\top & D_{\mathbf{c}} \end{pmatrix}$$

$Z^\top Z \in R^{(p+q)\times(p+q)}$ 是所有 $x_1,\ldots,x_p$；$y_1,\ldots,y_q$的内积相似度矩阵，其中我们主要关心W。



例如，$n = 5$，每个个体有两个因子变量：x (size: 大中小)，y (weight: 轻、重)

$$Z = (X, Y) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \text{size} \quad\quad \text{weight} \\ \text{大}\ \text{中}\ \text{小}\quad \text{重}\ \text{轻} \end{array} \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$Z^\top Z = \begin{pmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{pmatrix} = \begin{array}{c} \\ \text{大} \\ \text{中} \\ \text{小} \\ \text{重} \\ \text{轻} \end{array} \begin{array}{c} \text{大}\ \text{中}\ \text{小}\ \text{重}\ \text{轻} \\ \end{array} \begin{pmatrix} 2 & 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 2 & 1 & 1 \\ 2 & 0 & 1 & 2 & 0 \\ 0 & 1 & 1 & 0 & 3 \end{pmatrix} \text{列联表}$$

记$Z_c = (X_c, Y_c) = (X - \mathbb{1}\mathbf{r}^\top/n, Y - \mathbb{1}\mathbf{c}^\top/n)$为 Z 的列中心化，则

$$(n-1)\begin{pmatrix} S_{\mathbf{xx}} & S_{\mathbf{xy}} \\ S_{\mathbf{yx}} & S_{\mathbf{yy}}) \end{pmatrix} = \begin{pmatrix} X_c^\top X_c & X_c^\top Y_c \\ Y_c^\top X_c & Y_c^\top Y_c \end{pmatrix} = \begin{pmatrix} D_{\mathbf{r}} - \mathbf{r}\mathbf{r}^\top/n & W - \mathbf{r}\mathbf{c}^\top/n \\ W^\top - \mathbf{c}\mathbf{r}^\top/n & D_{\mathbf{c}} - \mathbf{c}\mathbf{c}^\top/n \end{pmatrix}$$

其中 $W - \mathbf{r}\mathbf{c}^\top/n \triangleq W_c$是$W$的双向中心化，而$W$的中心标准化为

$$W_s = D_{\mathbf{r}}^{-1/2} W_c D_{\mathbf{c}}^{-1/2}$$

- Pearons独立性/齐一性检验的卡方统计量

$$X^2 = n\|W_s\|^2$$

$|W_s|, \lambda_{max}(W_s)$?

- 对应分析：

  W_s奇异值分解：$W_s = D_{\mathbf{r}}^{-1/2} W_c D_{\mathbf{c}}^{-1/2} = UDV^\top$

  $F = D_{\mathbf{r}}^{-1/2} UD,\ G = D_{\mathbf{c}}^{-1/2} VD,$

  biplot: 取 $F[,1:2], G[,1:2]$ 分别作为因子 x, y 的二维表示，

  ```
  plot(F[,1:2])
  points(G[,1:2])
  ```

通过例1的分析，我们可以这样理解对应分析：

- 列联表的欧氏表示：通过 one-hot embedding 把 列联表 W 中的所有个体用欧氏向量表示出来。
- 对欧氏表示的方差-协方差矩阵应用主成分分析

对于一般的相似度矩阵或丰度矩阵可以类似地处理。

相似度和距离

相似度或邻近程度：proximity， similarity，closeness
距离或相异度：distance，dissimilarity
有些问题相似度容易定义，另一些问题距离可能更容易确定，两者是相反的概念。

相似度和距离

距离表示两个物体的相异程度。满足数学定义的距离比较容易定义，最常用的是欧氏距离。但在某些问题中，“距离”代表主观感知的远近程度，未必满足距离的数学定义。

相似度或相似系数代表两个物体的相似程度，在数学上没有严格定义。在实际应用中通常以欧氏内积或相关系数作为相似度，有时以距离的减函数定义。凭主观感知打分的相似度也很常见，尤其是研究对象不可测量的时候。

无论哪种方式定义，距离或相似度一般都是对称的。基于相似/相异系数的方法：聚类分析、多维标度法、配列等。

相似系数

对于可测量的情形，通常以内积或者相关概念定义相似度。对于没有具体测量的情形，两个研究对象的相似系数通常根据具体问题给出或主观印象打分评定。

假设两个对象的测量为向量$\mathbf{x},\mathbf{y}$, 根据具体问题背景，相似系数$s(\mathbf{x},\mathbf{y})$可以定义为与内积或距离有关的函数：

- 内积： $s(\mathbf{x},\mathbf{y})=c\mathbf{x}^{\mathsf{T}}\mathbf{y}$, $s(\mathbf{x},\mathbf{y})=c\mathbf{x}^{\mathsf{T}}\mathbf{y}/\|\mathbf{x}\|\|\mathbf{y}\|$
- 距离的减函数，比如高斯核函数将距离转化为相似系数：

$$s(\mathbf{x},\mathbf{y})=\exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right)$$

例2. 假设 $\mathbf{x},\mathbf{y}$都是长度为p的0-1序列，

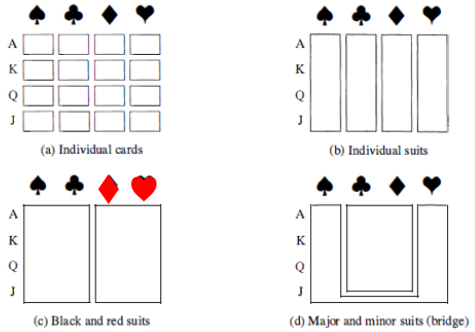
$$\mathbf{x}=110100$$
$$\mathbf{y}=010010$$

下述相似系数都与

- $s(\mathbf{x},\mathbf{y})=\mathbf{x}^{\mathsf{T}}\mathbf{y}/p=1$匹配的比例=1/6.
- $s(\mathbf{x},\mathbf{y})=[\mathbf{x}^{\mathsf{T}}\mathbf{y}+(\mathbf{1}-\mathbf{x})^{\mathsf{T}}(\mathbf{1}-\mathbf{y})]/p=1$或0匹配的比例=3/6.
- $s(\mathbf{x},\mathbf{y})=\mathbf{x}^{\mathsf{T}}\mathbf{y}/\sum 1(x_i+y_i>0)=1/4$ (都是0的位置不统计在内).

例3.（不同“相似性”的定义） 下面图(a)中的16张扑克牌。

(b) 根据花式区分，相同的花式是相似的；
(c) 根据颜色区分，同色的相似；
(d) 主牌相似、副牌相似。



例4. 各种语言在历史上不断演变或相互影响，研究语言之间的关系有助于了解历史上文化发展融合过程。语言的相似性可以体现在多方面，主要体现在发音，其中数字1,2,…,9, 10的读音颇具代表性。汉字在日本分音读和训读（本地）两种读法，在越南分汉语词和纯越词两种读法。下表是粤语（南越、唐话）、越南（汉越）、日语（音读）的数字读音：

数字	1	2	3	4	5	6	7	8	9	10	
粤语	ya	yi	sa	sei	ng	lao	cha	ba	gao	sa	相似度7/10
越南(汉越)	nhất	nhì	tam	tư	ngũ	lục	thất	bát	cửu	thập	

粤语	ya	yi	sa	sei	ng	lao	cha	ba	gao	sa	相似度4/10
日语(音读)	ichi	ni	san	shi,yon	go	roku	shichi,nana	hachi	ku	juu	

数学距离

数学距离的定义：$d = d(\cdot,\cdot)$称为是R^p的距离，如果对任何$\mathbf{x},\mathbf{y},\mathbf{z} \in R^p$

- 非负性：$d(\mathbf{x},\mathbf{y}) \geq 0,\ d(\mathbf{x},\mathbf{y}) = 0$当且仅当$\mathbf{x} = \mathbf{y}$
- 对称性：$d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$
- 三角不等式：$d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) \geq d(\mathbf{x},\mathbf{z})$

向量$\mathbf{x} = (x_1,...,x_p)^{\mathsf{T}}$，$\mathbf{y} = (y_1,...,y_p)^{\mathsf{T}} \in \mathrm{R}^p$的距离(distance, dissimilarity)

- 欧氏距离：$d(\mathbf{x},\mathbf{y}) = \| \mathbf{x}-\mathbf{y} \| = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}(\mathbf{x}-\mathbf{y})}$
- 马氏距离：$d_{\mathrm{A}}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}} A (\mathbf{x}-\mathbf{y})},\ A > 0$

  通常$A = S^{-1}$, S为协方差矩阵.
- 闵科夫斯基距离：$d_m(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{p} | x_i - y_i |^m \right)^{1/m},\ m > 0$

  $m \to 0\ (l_0)$：$d_0(\mathbf{x},\mathbf{y}) = \#\{i : x_i \neq y_i\}$;

  $m = 1\ \ (l_1)$：$d_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} | x_i - y_i |$ (R : manhattan)

  $m \to \infty\ (l_\infty)$：$d_\infty(\mathbf{x},\mathbf{y}) = \max_{1 \leq i \leq p} \{ | x_i - y_i | \}$ (R : maximum)
- Hamming 距离：$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} 1_{(x_i \neq y_i)}$, 主要用于0-1序列或字符串

非数学距离

个体a,b之间的非数学距离d_{ab}满足条件：

非负：$d_{ab} \geq 0, d_{aa} = 0$；　对称：$d_{ab} = d_{ba}$.

注意非数学距离不一定满足三角不等式.

向量$\mathbf{x} = (x_1,...,x_p)^{\mathsf{T}}$，$\mathbf{y} = (y_1,...,y_p)^{\mathsf{T}} \in R^p$，下述两种距离不满足数学距离的定义，但蕴含了“差距相对于总量”的相对性概念：

- Canberra距离：$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}, x_i, y_i > 0$
- Czekanowski距离：$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{p} |x_i - y_i|}{\sum_{i=1}^{p} (x_i + y_i)}$，　$x_i, y_i > 0$

后面我们提到“距离”的时候，如果不加“数学”或“欧氏”等限制，一般指的是非数学距离或主观距离。

可欧氏化的相似度和距离

问题：给定n个物体两两之间的主观相近性度量（距离/相似度量），能否将物体映射为某个欧氏空间中的n个点，使得这些点之间的数学距离或内积匹配主观的相近性度量？

如果能完全匹配，称为可欧氏化的。

相似系数矩阵/距离矩阵

- 相似系数矩阵 $S=(s_{ij})$，s_{ij} 是对象 i,j 的相似系数(可正可负)，对角元最大，假设S对称。
- 距离矩阵$D=(d_{ij}), d_{ij}\geq 0, d_{ii}=0$，假设$D$对称。

下面研究相似度和距离矩阵欧氏化的充要条件。例如，下述两个距离矩阵是否是欧氏的？

$$D=\begin{pmatrix} 0 & 1 & 1 & 3 \\ & 0 & 1 & 3 \\ & & 0 & 3 \\ & & & 0 \end{pmatrix},\quad D=\begin{pmatrix} 0 & 1 & 3 & 3 \\ & 0 & 1 & 3 \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}$$

注意，实际问题一般做不到完全欧氏化，但下面的讨论能够提供理论指导。

欧氏相似系数矩阵

一个相似度矩阵$S=(s_{ij})$称为是欧氏的（Euclidean），如果存在某个k以及n个向量$\mathbf{x}_1,\ldots,\mathbf{x}_n\in R^k$，使得$s_{ij}=\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$，即$S=XX^{\mathsf{T}}$，$X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathsf{T}}$.

命题1. 对称的相似系数矩阵S为欧氏的 $\Leftrightarrow S\geq 0$(半正定)。

$\Leftrightarrow$ 代数定理：$B\geq 0$当且仅当存在X使得$B=XX^{\mathsf{T}}$.

证明：若S为欧氏的，存在$p,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\in R^p$，使得$s_{ij}=\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$，

记$X=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)^{\mathsf{T}}$,则$S=XX^{\mathsf{T}}\Rightarrow S\geq 0.$

反之，若相似度矩阵S为半正定的，则S有谱分解

$$S=U\Lambda U^{\mathsf{T}}=\left(U\Lambda^{1/2}\right)\left(U\Lambda^{1/2}\right)^{\mathsf{T}},U^{\mathsf{T}}U=UU^{\mathsf{T}}=I_n$$

这说明S可由$X=U\Lambda^{1/2}$的各行之间的内积构成。

半正定矩阵：可以表示为某个欧氏空间n个向量两两之间的内积

欧氏距离矩阵

假设 $D=(d_{ij})$ 是 $n\times n$ 距离矩阵，即 $d_{ij}=d_{ji}\geq 0$，$d_{ii}=0$。如果存在某个 k 以及 n 个向量 $\mathbf{x}_1,\dots,\mathbf{x}_n\in R^k$，使得 $d_{ij}=||\mathbf{x}_i-\mathbf{x}_j||$，则称 D 是欧氏距离矩阵或 D 是可欧氏化的。

首先考虑 D 是欧氏距离矩阵的必要条件。

假设D是欧氏的，则存在某个k和向量$\mathbf{x}_1,\dots,\mathbf{x}_n\in R^k$，使得

$$d_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|\Rightarrow d_{ij}^2=\|\mathbf{x}_i\|^2+\|\mathbf{x}_j\|^2-2\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$$

记$\boldsymbol{\xi}=(\|\mathbf{x}_1\|^2,\dots,\|\mathbf{x}_n\|^2)^{\mathsf{T}}/2$，则

$$A\triangleq\left(-d_{ij}^2/2\right)=\left(-\|\mathbf{x}_i\|^2/2\right)+\left(-\|\mathbf{x}_j\|^2/2\right)+\left(\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j\right)$$

$$=-\mathbb{1}\boldsymbol{\xi}^{\mathsf{T}}-\boldsymbol{\xi}\mathbb{1}^{\mathsf{T}}+X\,X^{\mathsf{T}}\qquad(*)$$

其中$X=(\mathbf{x}_1,\dots,\mathbf{x}_n)^{\mathsf{T}}$。

注意$(*)$式的主要部分 $XX^{\mathsf{T}}\geq 0$ (半正定)，$(*)$式两边同时左、右乘 $J_n=I_n-P_{\mathbb{1}}=I_n-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{n}$，消去秩1项，即双向中心化：

$$S\triangleq J_nAJ_n=J_nXX^{\mathsf{T}}J_n=X_cX_c^{\mathsf{T}}\geq 0$$

其中$X_c=J_nX$是X的中心化。

$S\geq 0\Leftrightarrow S$是欧氏相似度矩阵

故 $S \geq 0$ 是必要条件，下面证明它也是D可欧氏化的充分条件。

假设$S = (s_{ij}) = J_n A J_n \geq 0$，则存在某个 k 和$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^k$， 使得 $s_{ij} = \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j$，即$S = XX^{\mathsf{T}}$，$X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathsf{T}}$。需要证明$\|\mathbf{x}_i - \mathbf{x}_j\| = d_{ij}$

记$A = (-d_{ij}^2/2) \triangleq (a_{ij}), a_{ij} = -\frac{d_{ij}^2}{2}$， 容易验证$S = J_n A J_n$ 的 (i,j) 元

$$\mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j = s_{ij} = a_{ij} - \bar{a}_{i\bullet} - \bar{a}_{\bullet j} + \bar{a}_{\bullet\bullet}$$

其中$\bar{a}_{i\bullet}$，$\bar{a}_{\bullet j}$是A的第i行的平均.

由于A对称且对角元为0，$a_{ii} = -\frac{d_{ii}^2}{2} = 0, \bar{a}_{\bullet i} = \bar{a}_{i\bullet}$，

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_i + \mathbf{x}_j^{\mathsf{T}} \mathbf{x}_j - 2\mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j = s_{ii} + s_{jj} - 2s_{ij}$$

$$= (a_{ii} - \bar{a}_{i\bullet} - \bar{a}_{\bullet i} + \bar{a}_{\bullet\bullet}) + (a_{jj} - \bar{a}_{i\bullet} - \bar{a}_{\bullet i} + \bar{a}_{\bullet\bullet}) - 2(a_{ij} - \bar{a}_{i\bullet} - \bar{a}_{\bullet j} + \bar{a}_{\bullet\bullet})$$

$$= -2a_{ij} = d_{ij}^2$$

至此，我们证明了欧氏距离矩阵的刻画

命题2. 假设 $D=(d_{ij})$ 是 $n\times n$ 距离矩阵 ($d_{ij}=d_{ji}\geq 0,\ d_{ii}=0$), 记

$$A=\left(-d_{ij}^2/2\right),\ J_n=I_n-\frac{\mathbb{1}\mathbb{1}^{\mathsf{T}}}{n},$$

则 D 是欧氏距离矩阵当且仅当 $S=J_nAJ_n\geq 0$（即S是欧氏内积矩阵)。

参见： K.V.Mardia, J.T.Kent, J.M.Bibby (2024) Multivariate Analysis, Academic Press。

从前述证明过程，我们有

- 给定欧氏内积矩阵$S=(s_{ij})$, $D=\left(\sqrt{s_{ii}+s_{jj}-2s_{ij}}\right)$ 是欧氏距离矩阵。
- 给定欧氏距离矩阵$D=(d_{ij})$, $S=\left(-\frac{1}{2}\left(d_{ij}^2-\overline{d_{i\bullet}^2}-\overline{d_{\bullet j}^2}+\overline{d_{\bullet\bullet}^2}\right)\right)$ 是欧氏内积矩阵。

附录：Johnson–Lindenstrauss lemma

前面讨论的距离、相似度欧氏表示一般应用于主观度量。如果距离或相似度是高维数据计算得到的数学距离或内积，我们关心的是它们能否在低维欧氏空间近似表示？

Johnson–Lindenstrauss引理证明了高维欧氏空间R^p的 n 个点可以线性映射到某个低维空间R^k, $k \sim \log(n)$, 并且大致保持n个点两两之间的欧氏距离。

Johnson–Lindenstrauss lemma

引理A1. 对任何给定$0 < \varepsilon < 1$和n个点$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, 若正整数

$$k \geq \frac{24}{3\varepsilon^2 - 2\varepsilon^3} \log(n)$$

则一定存在$k \times p$矩阵A, 使得

$$(1-\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \leq \left\|A\mathbf{x}_i - A\mathbf{x}_j\right\|^2 \leq (1+\varepsilon)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2, \forall i, j. \quad (*)$$

例如，$p = 10000, \varepsilon = 0.5$，则$k \approx 442$.

我们不知道如何选取$A_{k\times p}$，尝试随机取一个r.v.矩阵B，只要能证明（*）式成立的概率大于0即证明了存在性。这是Erdös发明的存在性的概率化证明方法。

随机选取一个$k\times p$矩阵$B=(b_{ij})$, 其元素iid 服从$N(0,1)$分布, 由$EB^{\top}B=kI_p$, 有

$$E\left\|\frac{1}{\sqrt{k}}B\mathbf{x}_i-\frac{1}{\sqrt{k}}B\mathbf{x}_j\right\|^2=\frac{1}{k}(\mathbf{x}_i-\mathbf{x}_j)^{\top}EB^{\top}B(\mathbf{x}_i-\mathbf{x}_j)=\|\mathbf{x}_i-\mathbf{x}_j\|^2.$$

所以平均来看，$A=B/\sqrt{k}$-变换是保距的，但这不能保证去掉期望后两边大致相等。如果k也较大，则由切比雪夫型集中不等式(concentration inequality, k,p都很大)

$$\frac{1}{k}B^{\top}B\xrightarrow{\mathrm{P}}I_p,$$

此时以大于0的概率（接近1的概率）有

$$\|A\mathbf{x}_i-A\mathbf{x}_j\|^2=\|B\mathbf{x}_i/\sqrt{k}-B\mathbf{x}_j/\sqrt{k}\|^2\approx\|\mathbf{x}_i-\mathbf{x}_j\|^2,$$

该事件的概率大于0，故一定存在一个A,其元素iid 服从$N(0,1/k)$使得上式成立。

令人意外的是，

- 上述观察几乎就是JL引理存在性的概率化证明的全部过程，证明过程中只需把极限收敛一步换成集中不等式（切比雪夫）。

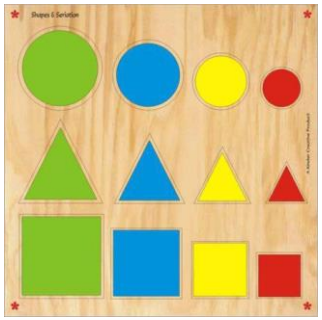
  Sanjoy Dasgupta, Anupam Gupta (2003) An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures & Algorithms 22: 60-65.

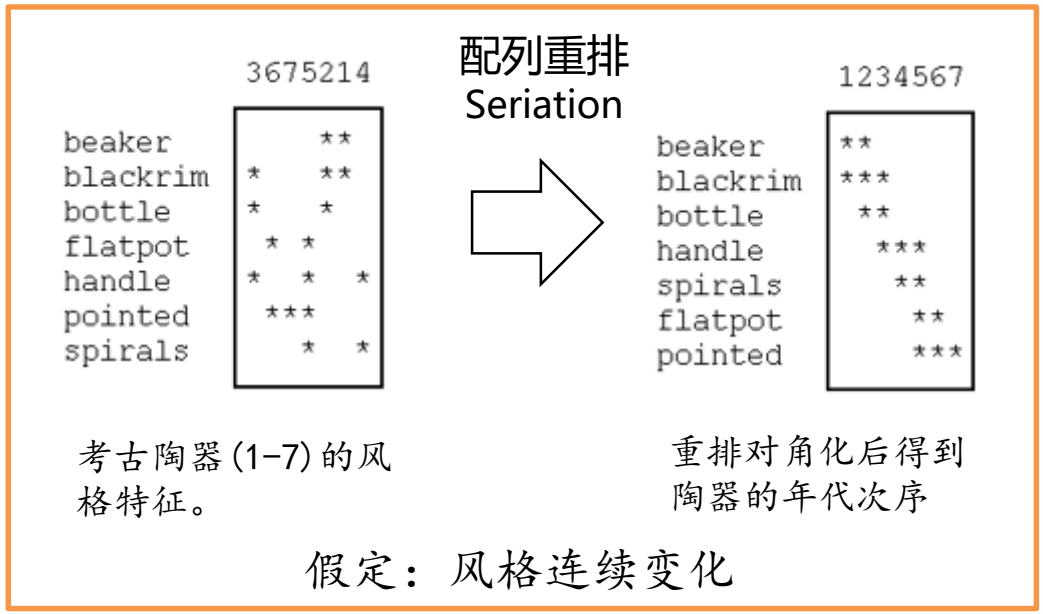
- 证明过程中的随机矩阵变换方法产生了一类被称为“随机投影random projection”的降维方法 (注意名字的歧义性，这里的投影不是严格意义上的正交投影)

$$\mathbf{x}_i \in R^p \to A\mathbf{x}_i \in R^k, i = 1, \dots, n$$
$$X \to XA^\top$$
$$A_{k\times p} = (a_{ij}),\ a_{ij}\ iid \sim N(0, 1/k)$$

配列（Seriation）

秩序：人类天然有寻求秩序的倾向。
配列排序：将物体进行置换排序，使相似的物体彼此靠近。

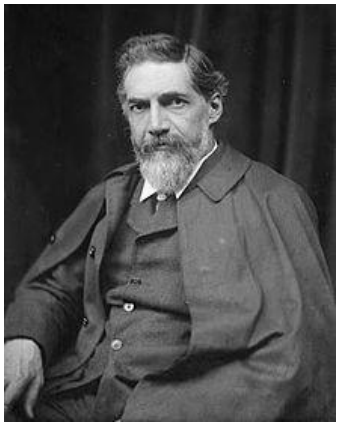




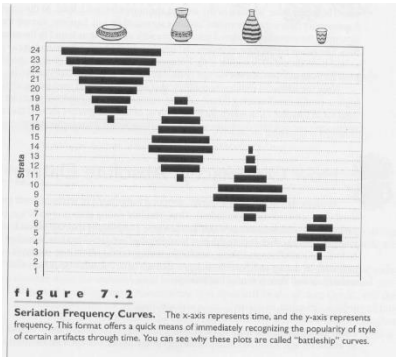
- One-mode seriation：按一个指标排序；
- Two-mode seriation：按两个指标排序

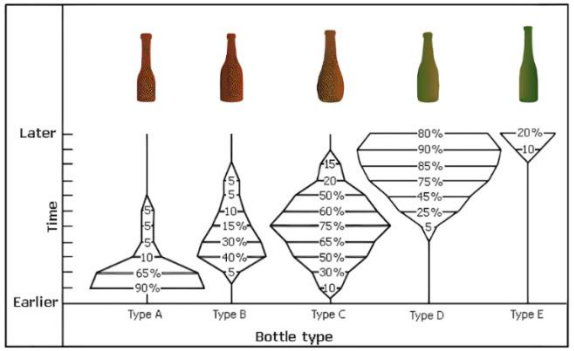
Petrie和考古学年代排序

Flinders Petrie（弗林德斯·皮特里1853–1942），英国埃及学家，在考古学和文物保存方面首创了成体系的方法，包括seriation方法。中国考古学奠基人夏鼐深受其影响。



By linking styles of pottery with periods, Petrie was the first to use seriation in Egyptology, a new method for establishing the chronology of a site。





Petrie, F. W. M. (1899). Sequences in prehistoric remains. *Journal of the Anthropological Institute* **29**:295–301

单向配列（One-mode Seriation）

依据物体的单个属性的相似度，比如按尺寸大小（右图），将物体从小到大排列。

单向配列问题

One-mode seriation： 给定n个物体的相似度矩阵$S = (s_{ij})$或距离矩阵$D = (d_{ij})$，seriation寻找$\{1,2,\dots,n\}$的一个置换π, 使得当s_{ij}较大时（或d_{ij}较小时），$|\pi(i) - \pi(j)|$较小。

组合优化

将单向配列问题设定为二次组合优化问题，极小化目标函数

$$\min \sum s_{ij}[\pi(i) - \pi(j)]^2 \qquad (*)$$

计算复杂度： $n!$

注：其它各种合理的目标设定都是允许的：

$$\min \sum (d_{ij} - |\pi(i) - \pi(j)|)^2, \quad \min \sum \frac{1}{1+d_{ij}} (\pi(i) - \pi(j))^2$$

等等。我们下面将主要讨论（*）

松弛为二次优化

$(*)$ 中放松π：$\{1,2,\dots,n\}\to\{1,2,\dots,n\}$是置换的要求，并改用记号$x$替代$\pi$，假设 x：$\{1,2,\dots,n\}\to R$, 记$x_i=x(i)$, 改写$(*)$

$$\min\sum s_{ij}(x(i)-x(j))^2 \triangleq \min\sum s_{ij}(x_i-x_j)^2$$

记 $d_i=\sum_j s_{ij}$, 矩阵$D=\mathrm{diag}(d_1,\dots.,d_n)$,改写目标函数

$$\sum_{i,j} s_{ij}(x_i-x_j)^2 \quad = 2\sum_j d_i\,x_i^2-2\sum_{i,j} s_{ij}x_i\,x_j$$

$$= 2\mathbf{x}^{\mathsf{T}}D\,\mathbf{x}-2\mathbf{x}^{\mathsf{T}}S\mathbf{x}=2\mathbf{x}^{\mathsf{T}}(D-S)\,\mathbf{x} \quad \triangleq 2\mathbf{x}^{\mathsf{T}}L\mathbf{x}\geq 0$$

其中拉普拉斯(Laplacian)矩阵

$$L=D-S\geq 0$$

显然当$\mathbf{x}\propto\mathbb{1}$，二次型达到极小值0，为了避免得到这个解（不可排序），我们约束$\mathbf{x}$分量不全相同（约束$\mathbf{x}\perp\mathbb{1}$）。

在上述约束下， 二次型$\mathbf{x}^{\mathsf{T}}L\mathbf{x}$的极小值为最小非0特征根，并在相应的特征向量处达到极小值。

综合上述讨论，我们有

单向谱配列

假设$S=(s_{ij})$为n个物体的相似度矩阵(对称), 假设所有 $s_{ij}\geq 0$，记Laplacian矩阵$L=D-S$, $D=\mathrm{diag}(\mathbf{d})$, $\mathbf{d}=S\mathbb{1}$

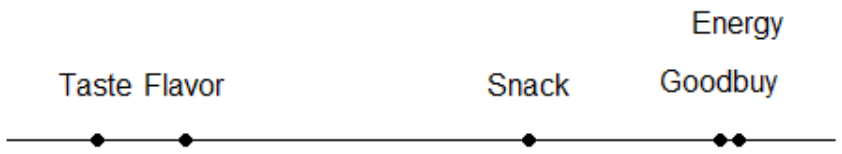
$$\min\frac{1}{2}\sum s_{ij}(x_i-x_j)^2=\min\mathbf{x}^{\mathsf{T}}L\mathbf{x}\text{，}\ \text{s.t.}\ \ \|\mathbf{x}\|=1\text{，}\ \mathbf{x}\perp\mathbb{1}$$

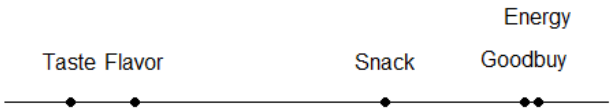
当$\mathbf{x}$是 L的最小非0特征根对应的特征向量时达到极小。将该特征向量排序即得到谱配列（spectral seriation）

例5（食品评价）右表是相关系数矩阵，当作相似度矩阵，L 的最小非0特征根0.807，对应的特征向量：

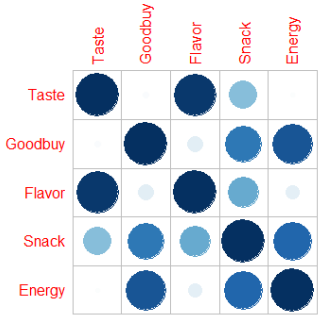
	Taste	Goodbuy	Flavor	Snack	Energy
Taste	1	0.02	0.96	0.42	0.01
Goodbuy	0.02	1	0.13	0.71	0.85
Flavor	0.96	0.13	1	0.50	0.11
Snack	0.42	0.71	0.50	1	0.79
Energy	0.01	0.85	0.11	0.79	1

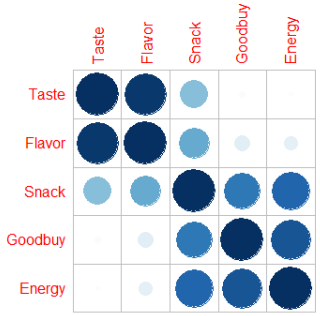
Taste	Flavor	Snack	Goodbuy	Energy
−0.600	− 0.449	0.131	0.458	0.480

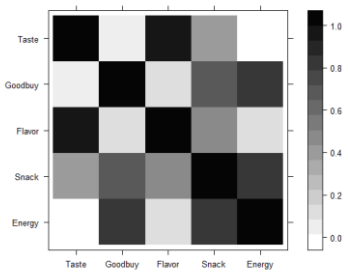


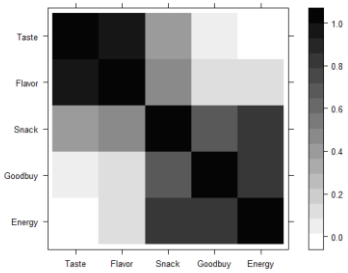


按照得到次序重新排列相关系数矩阵，
下图左列：未排序的相关系数矩阵；右列：排序之后



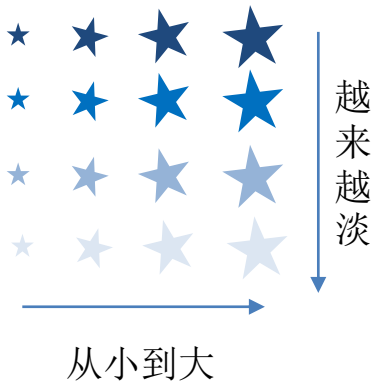






双向配列（Two-mode Seriation）

依据物体的两个属性，比如按尺寸大小或接近于圆的程度（右图），将物体按两个属性排列。



双向配列问题

假设 $W_{p\times q}=(w_{ij})$ 的 (i,j) 元 w_{ij} 代表行标 x_i 与列标 y_j 之间的联系紧密程度，比如 W 可以是列联表、联系矩阵、丰度/含量矩阵（但不是相似系数方阵）。Two-mode seriation 置换行、列使得对角线附近的值较大。

	y_1	$\cdots$	y_q
x_1			
$\vdots$		W	
x_p			

我们将采用两种方法求解：

(1) 二次优化 (2)转化为单向配列问题

二次优化

组合优化：求解行标号和列标号的置换$u_i = u(i), v_j = v(j)$,使得若a_{ij}较大，则$|u_i - v_j|$较小，为此我们极小化

$$\sum w_{ij}(u_i - v_j)^2 = \min!$$

二次 优化: 放松对 u, v 的要求,假设 $\mathbf{u} = (u_i) \in R^p, \mathbf{v} = (v_j) \in R^q$, $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, 双向配列问题设定为求解二次优化问题

$$\min \sum w_{ij}(u_i - v_j)^2 \qquad (**)$$

求解:记 $\mathbf{r} = W\mathbb{1}$, $\mathbf{c} = W^\top\mathbb{1}$, $D_r = \mathrm{diag}(\mathbf{r})$, $D_c = \mathrm{diag}(\mathbf{c})$, 则容易验证

$$\begin{aligned}\sum w_{ij}(u_i - v_j)^2 &= \sum_i r_i u_i^2 + \sum_j c_j v_j^2 - 2\sum_{i,j} w_{ij}\, u_i v_j \\ &= \mathbf{u}^\top D_r \mathbf{u} + \mathbf{v}^\top D_c \mathbf{v} - 2\mathbf{u}^\top W\, \mathbf{v} \\ &= (\mathbf{u}^\top, \mathbf{v}^\top)\begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix}\begin{pmatrix}\mathbf{u} \\ \mathbf{v}\end{pmatrix} - (\mathbf{u}^\top, \mathbf{v}^\top)\begin{pmatrix} 0 & W \\ W^\top & 0 \end{pmatrix}\begin{pmatrix}\mathbf{u} \\ \mathbf{v}\end{pmatrix} \\ &= (\mathbf{u}^\top, \mathbf{v}^\top)\left\{\begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix} - \begin{pmatrix} 0 & W \\ W^\top & 0 \end{pmatrix}\right\}\begin{pmatrix}\mathbf{u} \\ \mathbf{v}\end{pmatrix} \triangleq \mathbf{x}^\top L\mathbf{x}\end{aligned}$$

其中Laplacian矩阵$L = D - A$，$A = \begin{pmatrix} 0 & W \\ W^\top & 0 \end{pmatrix}$，$D = \begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix}$,

同单向谱配列问题,L的最小非0 特征根的特征向量 $\mathbf{z}_{\min}=\begin{pmatrix}\mathbf{u}_{\min}\\ \mathbf{v}_{\min}\end{pmatrix}$ 是(**)的最优解。综上，我们有如下谱配列方法：

双向谱配列

假设 $W=(w_{ij})$ 为 $p\times q$ 联系/丰度矩阵, 假设所有 $w_{ij}\geq 0$，记

$$A=\begin{pmatrix}0 & W\\ W^{\top} & 0\end{pmatrix}$$

$\mathbf{d}=A\mathbb{1}=\begin{pmatrix}\mathbf{r}\\ \mathbf{c}\end{pmatrix}$, $D=\mathrm{diag}(\mathbf{d})$, 令Laplacian矩阵

$$L=D-A=\begin{pmatrix}D_r & -W\\ -W^{\top} & D_c\end{pmatrix},$$

假设 $\mathbf{u}=(u_i)\in R^p$，$\mathbf{v}=(v_j)\in R^q$，则

$$\sum w_{ij}(u_i-v_j)^2=\mathbf{z}^{\top}L\mathbf{z},\ \mathbf{z}=\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$$

当 $\mathbf{z}$ 是 L 的最小非 0 特征根对应的特征向量时达到极小。将该特征向量的 $\mathbf{u},\mathbf{v}$ 分别排序即得到双向谱配列。

注：若 W 是双向中心化的，则 $\sum w_{ij}(u_i-v_j)^2=-\frac{1}{2}\mathbf{u}^{\top}W\,\mathbf{v}$，此时最优$\mathbf{u}$，$\mathbf{v}$ 是 W 的最大奇异特征向量

W欧氏化

另一种求双向配列问题的方法是：

类似于例1(2), 设想$W_{p\times q}$是可欧氏化的，即假设W是某些原始欧氏one-hot数据生成的（即使W 元素不是计数！），利用欧氏表示计算内积相似度，再利用单向配列结果。

假设 W 由 n-维欧氏空间中one-hot矩阵$X=(\mathbf{x}_{(1)},\dots,\mathbf{x}_{(p)})\in R^{n\times p}$, $Y=(\mathbf{y}_{(1)},\dots,\mathbf{y}_{(q)})\in R^{n\times q}$ 生成：

$$w_{jk}=\mathbf{x}_{(j)}^{\top}\mathbf{y}_{(k)}\,,\ W=X^{\top}Y$$

记 $Z=(X,Y)$, 则

$$S=Z^{\top}Z=\begin{pmatrix}X^{\top}X & X^{\top}Y\\ Y^{\top}X & Y^{\top}Y\end{pmatrix}=\begin{matrix}\\ \mathbf{x}\\ \mathbf{y}\end{matrix}\begin{matrix}\begin{matrix}\mathbf{x} & \mathbf{y}\end{matrix}\\ \begin{pmatrix}D_{\mathbf{r}} & W\\ W^{\top} & D_{\mathbf{c}}\end{pmatrix}\end{matrix}$$

这是所有行标、列标$x_1,\cdots,x_p\,,y_1,\cdots,y_q$的内积相似系数矩阵。

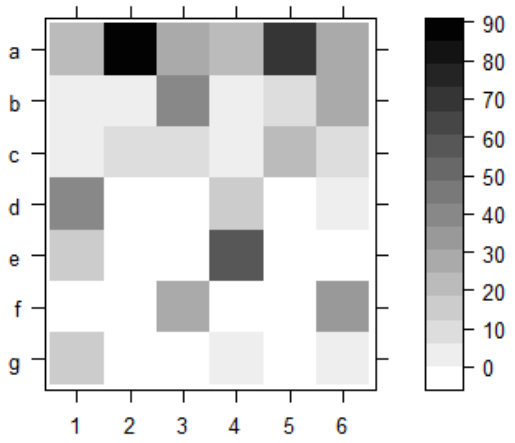
由 P27 单向谱配列方法，记 $D=\mathrm{diag}(S\mathbb{1})=\begin{pmatrix}2D_{\mathbf{r}} & 0\\ 0 & 2D_{\mathbf{c}}\end{pmatrix}$,

Laplacian： $L=D-S=\begin{pmatrix}D_{\mathbf{r}} & -W\\ -W^{\top} & D_{\mathbf{c}}\end{pmatrix}$,

对L的最小非 0 特征根对应的特征向量排序即可，这与上页二次优化结果完全相同。

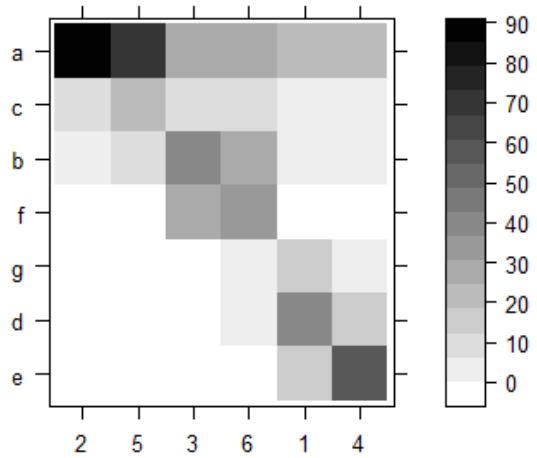
例6. 6个考古地点出土7种燧石，个数统计如下

	a	b	c	d	e	f	g
1	20	3	4	42	18	0	13
2	85	3	12	0	0	0	0
3	26	40	8	0	0	26	0
4	20	1	4	13	58	0	4
5	67	10	23	0	0	0	0
6	26	29	8	3	0	33	1



右图：双向中心化后，应用1阶奇异值分解分别重新排列行和列。考古点排序为

2 5 3 6 1 4



R：library(seriation)
seriate(x)