第十九讲 距离和相似系数

2024.5.20

对于数据矩阵X (行代表个体、列代表变量)

XX^{T}：所有个体两两之间的相似度(内积)；

$X^{\mathsf{T}}X$：所有变量两两之间的相似度(内积)。

典则相关分析CCA(续)

Recap

协方差矩阵：$\Sigma = \mathrm{cov}\begin{pmatrix}\mathbf{x}_{p\times 1}\\ \mathbf{y}_{q\times 1}\end{pmatrix} = \begin{pmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{pmatrix}$,

$\Sigma_{\mathbf{xy}}$标准化：$A = \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}$,

A的SVD/对角化:

$$A_{p\times q} = \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2} = U_{p\times r}D_{r\times r}V_{q\times r}^{\top}$$
$$\Rightarrow U^{\top}AV = D$$

$r = rank(A)$

$D = diag(\sqrt{\lambda_1},...,\sqrt{\lambda_r})$,

$\lambda_1 \geq ... \geq \lambda_r > 0$, $A^{\top}A$的非0特征根，

$U = (\mathbf{u}_1,...,\mathbf{u}_r)$, $V = (\mathbf{v}_1,...,\mathbf{v}_r)$,

$U^{\top}U = V^{\top}V = I_r$.

总体CCA(第一典则相关)：

求线性组合$\xi = \mathbf{a}^{\top}\mathbf{x}$，$\eta = \mathbf{b}^{\top}\mathbf{y}$, 使得

$$corr(\xi,\eta) = max!$$

最大值在第一对典则变量$(\xi_1,\eta_1) = (\mathbf{u}_1^{\top}\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}, \mathbf{v}_1^{\top}\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y})$达到.

最大值$corr(\eta_1,\xi_1) = \sqrt{\lambda_1}$称为第一典则相关系数,

(总体)CCA：

- 对$k = 1, \ldots r$, 求线性组合$\xi = \mathbf{a}^{\mathsf{T}}\mathbf{x}$，$\eta = \mathbf{b}^{\mathsf{T}}\mathbf{y}$, 使得

$$corr(\xi, \eta) = \max! \quad 约束 \ \mathrm{var}(\xi) = \mathrm{var}(\eta) = 1$$

ξ与$\xi_1, \ldots, \xi_{k-1}$不相关,
η与$\eta_1, \ldots, \eta_{k-1}$不相关.

最大值在第k对典则变量达到:

$$(\xi_k, \eta_k) = \left(\mathbf{u}_k^{\mathsf{T}}\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}, \mathbf{v}_k^{\mathsf{T}}\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}\right),$$

最大值$corr(\eta_k, \xi_k) = \sqrt{\lambda_k}$称为第$k$典则相关系数。

- 所有典则变量：

$$\mathbf{x}_{cca} = (\xi_1, \ldots, \xi_r)^{\mathsf{T}} = U^{\mathsf{T}}\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x},$$

$$\mathbf{y}_{cca} = (\eta_1, \ldots, \eta_r)^{\mathsf{T}} = V^{\mathsf{T}}\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y},$$

$$\mathrm{cov}\begin{pmatrix}\mathbf{x}_{cca} \\ \mathbf{y}_{cca}\end{pmatrix} = \begin{pmatrix} I_r & D \\ D & I_r \end{pmatrix}$$

线性模型： $\mathbf{y} = B^{\mathrm{T}}\mathbf{x} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{x}$ ，

典则变量将模型对角化：

$\Leftrightarrow\ \mathbf{y}_{cca} = D\mathbf{x}_{cca} + \boldsymbol{\delta},\ \boldsymbol{\delta} \perp\!\!\!\perp \mathbf{x}_{cca}\ \Leftrightarrow\ \eta_i = \sqrt{\lambda_i}\xi_i + \delta_i, i = 1, \ldots, r$

特别地，$q = 1$情形：y是标量，$\sigma_y = \sqrt{\Sigma_{yy}}$，

$$\lambda_1 = \frac{\Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}}{\Sigma_{yy}} \triangleq R^2,$$

$$\mathbf{u}_1 = \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{x}y}\Sigma_{yy}^{-1/2}/\sqrt{\lambda_1},\ \ \mathrm{v}_1 = 1$$

$$(\xi_1, \eta_1) = \left(\Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}/\sqrt{\lambda_1},\ \ y\right)/\sigma_y \triangleq (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}/R,\ \ y)/\sigma_y \Rightarrow$$

一元线性模型 $y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon$，$\varepsilon \perp\!\!\!\perp \mathbf{x}$ 中

回归函数 $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = R\xi_1 = \sqrt{R^2} \times$ 典则变量

即回归函数 $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$是最能代表$y, \mathbf{x}$之间相关性的$\mathbf{x}$的线性组合。

样本CCA：

样本$X = (\mathbf{x}_1, \dots, \mathbf{x}_1)^\top, Y = (\mathbf{y}_1, \dots, \mathbf{y}_1)^\top,$

SVD: $A = S_{\mathbf{xx}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{yy}}^{-1/2} = UDV^\top$

$X_{cca} = X\Sigma_{\mathbf{xx}}^{-1/2} U = \left(\mathbf{x}_i^\top \Sigma_{\mathbf{xx}}^{-1/2} \mathbf{u}_j\right) = (\mathbf{x}_i$ 的第 j 个典则变量)

$Y_{cca} = Y\Sigma_{\mathbf{yy}}^{-1/2} V = \left(\mathbf{y}_i^\top \Sigma_{\mathbf{yy}}^{-1/2} \mathbf{v}_j\right) = (\mathbf{y}_i$ 的第 j 个典则变量)

例1. $\mathbf{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix},\mathbf{y}=\begin{pmatrix}x_3\\x_4\end{pmatrix}$, $x_1=\text{reading speed}, x_2=\text{reading power};$

$x_3=\text{arithmetic speed}, x_4=\text{arithmetic power}$

$$\Sigma=\begin{pmatrix}\Sigma_{xx}&\Sigma_{xy}\\\Sigma_{yx}&\Sigma_{yy}\end{pmatrix}=\begin{array}{c}\\x_1\\x_2\\x_3\\x_4\end{array}\begin{array}{c}\begin{array}{cccc}x_1&x_2&x_3&x_4\end{array}\\\begin{pmatrix}1.00&0.63&0.24&0.06\\&1.00&-0.06&0.07\\&&1.00&0.42\\&&&1.00\end{pmatrix}\end{array}$$

- Σ_{xy}标准化：$A=\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}=\begin{pmatrix}0.33&-0.03\\-0.2&0.11\end{pmatrix}$,
- 求奇异值分解：

$$\Phi=A^{\mathsf{T}}A=\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}=\begin{pmatrix}0.15&-0.03\\-0.03&0.01\end{pmatrix},$$

$$\Psi=AA^{\mathsf{T}}=\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}=\begin{pmatrix}0.11&-0.07\\-0.07&0.05\end{pmatrix},$$

$\Rightarrow$

$$\sqrt{\lambda_1}=0.40,$$
$$\mathbf{v}_1=(-0.98,0.21)^{\mathrm{T}},$$
$$\mathbf{u}_1=(-0.83,0.56)^{\mathrm{T}}$$

第一典则变量：$\xi_1=-1.25x_1+1.03x_2,\ \eta_1=-1.10y_1+0.46y_2.$

第一典则相关系数$\sqrt{\lambda_1}=0.40$，大于Σ_{xy}任何元素(为什么？)

例2. $\mathbf{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix},\mathbf{y}=\begin{pmatrix}y_1\\y_2\end{pmatrix}$，假设$\mathrm{cov}(y_1,x_2)=0.95$,其它都不相关，即

$$\Sigma=\mathrm{cov}\begin{pmatrix}x_1\\x_2\\y_1\\y_2\end{pmatrix}=\begin{pmatrix}100&0&0&0\\0&1&0.95&0\\0&0.95&1&0\\0&0&0&100\end{pmatrix},\ \Sigma_{\mathbf{xy}}=\begin{pmatrix}0&0\\0.95&0\end{pmatrix},\ A=\Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xy}}^{-1/2}=\Sigma_{\mathbf{xy}}$$

$$\Psi=\begin{pmatrix}0.95^2&0\\0&0\end{pmatrix},\Phi=\begin{pmatrix}0&0\\0&0.95^2\end{pmatrix},\quad \lambda_1=0.95^2,\ \mathbf{u}_1=(0,1)^{\mathsf{T}},\mathbf{v}_1=(1,0)^{\mathsf{T}}$$

第一（唯一）典则变量分别是：$\xi_1=\mathbf{u}_1^{\mathsf{T}}\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}=x_2$，$\eta_1=\mathbf{v}_1^{\mathsf{T}}\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}=y_1$,

典则相关系数为0.95。

- 在向量**x**中，第一主成分是x_1，方差贡献率为100/101，但它与**y**不相关；
  第二主成分是x_2，方差贡献率仅为1/101。
- 同样，在向量**y**中，第一主成分是y_2，但它与**x**不相关。第二主成分是y_1。

虽然x_2，y_1都是第二主成分，但它们最能代表描述**x**，**y**之间的相关性。

主成分与典则变量有相似之处，都是由原随机向量导出的能解释相关结构的变量，它们都可以看作是某种潜变量模型的特殊解。

主成分分析和因子分析

主成分$\mathbf{z} = V^{\mathsf{T}}\mathbf{x}$是随机向量$\mathbf{x}$分量的线性组合， 是因子模型的潜变量的预测，这些潜变量用于解释原始变量内部的相关性，并能保留原变量的（大部分）方差。

换言之，因子模型是主成分分析PCA的概率模型化，主成分是因子模型的特殊解（LS解）。

PCA：$\Sigma_{\mathbf{xx}} = V\Lambda V^{\mathsf{T}}$，主成分$\mathbf{z} = V^{\mathsf{T}}\mathbf{x}$

PCA反变换: $\mathbf{x} = V\mathbf{z} = V_m\mathbf{z}_m + \text{error},$

PCA是下述FA(factor analysis因子分析)的一个解

FA: $\mathbf{x} = L\mathbf{f} + \boldsymbol{\varepsilon}$

典则相关分析和结构方程模型

典则变量$\mathbf{x}_{cca}$，$\mathbf{y}_{cca}$是从随机向量$\mathbf{x}, \mathbf{y}$中各自提取的线性组合，可以认为是表述$\mathbf{x}, \mathbf{y}$之间相关性的某种潜变量的预测，这种用潜变量表征可测量随机变量的相关性的模型即为结构方程模型(SEM).

CCA：$\Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2} = UDV^{\mathrm{T}}$

$$\mathbf{x}_{cca} = U^{\mathrm{T}}\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x},\ \ \mathbf{y}_{cca} = V^{\mathrm{T}}\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}$$

CCA反变换：
$$\begin{cases} \mathbf{x} = \Sigma_{\mathbf{xx}}^{1/2}U\mathbf{x}_{cca} = \hat{L}_1\hat{\boldsymbol{f}}_1 + error \\ \mathbf{y} = \Sigma_{\mathbf{yy}}^{1/2}V\mathbf{y}_{cca} = \hat{L}_2\hat{\boldsymbol{f}}_2 + error \\ \mathrm{cov}\left(\mathbf{x}_{cca}, \mathbf{y}_{cca}\right) = D \end{cases}$$

$\hat{\boldsymbol{f}}_1$是$\mathbf{x}_{cca}$的前m个分量
$\hat{L}_1$是$\Sigma_{\mathbf{xx}}^{1/2}U$的前$m$ 列

CCA是下述SEM的一个解

SEM：
$$\begin{cases} \mathbf{x} = L_1\boldsymbol{f}_1 + \boldsymbol{\varepsilon}_1 \\ \mathbf{y} = L_2\boldsymbol{f}_2 + \boldsymbol{\varepsilon}_2 \\ \boldsymbol{f}_2 = B\boldsymbol{f}_1 + \boldsymbol{\delta} \end{cases}$$

距离和相似度

邻近程度 (proximity)度量：相似度和相异度

邻近程度度量(proximity)包括相似系数(similarity)和相异系数(dissimilarity)。

- 相似和相异是相反的概念，有的问题相似系数较容易定义，另外一些问题相异系数更容易定义。
- 有时临近性度量是主观评价(称为标度，scaling)，有时由计算而得。

基于相似/相异系数的方法：聚类分析、多维标度法、配列等。

相异系数/距离

相异系数通常以“距离”表示，距离越大，相异度越大。实际问题中定义的“距离”可能是主观感受的量化或标度，并不一定符合数学距离的定义。

相似系数/内积

相似系数（或相似度）通常以内积/相关系数表示，或定义为距离相反的概念（距离的减函数）。

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：$d_m(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{p} | x_i - y_i |^m \right)^{1/m},\ m > 0$

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

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

(非数学或主观)距离

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：距离未必满足三角不等式$d_{ab} + d_{bc} \geq d_{ac}$.

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

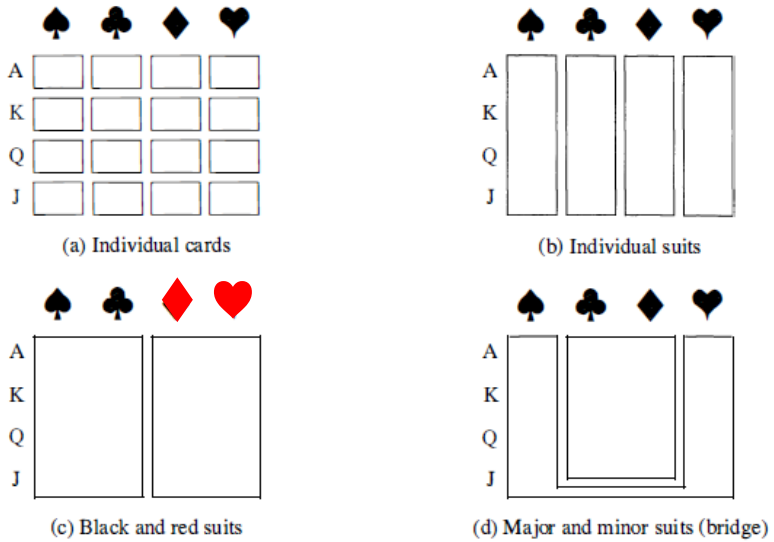
后面我们提到“距离”的时候，如果不加“数学”或“欧氏”等限制，一般指的是非数学距离或主观距离。

主观相似性系数

与距离定义类似，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（参见例1）或主观印象打分评定。通常具有对称性。

例1.（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，每列内4张牌具有相同的花式，认为是相似的；
(c) 根据颜色分为两个类，同色的都相似，不同色的不相似;
(d) 主牌、副牌各为一类



例2. 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
越南(汉越)	nhất	nhì	tam	tư	ngũ	lục	thất	bát	cửu	thập
日语(音读)	ichi	ni	san	shi,yon	go	roku	shichi,nana	hachi	ku	juu
粤语	ya	yi	sa	sei	ng	lao	cha	ba	gao	sa

相似度7/10

相似度4/10

内积相似性系数

对于可测量的情形，通常以内积方式定义。假设两个对象的测量为实数向量$\mathbf{x},\mathbf{y}$, 根据具体问题背景，定义（对称的）相似系数$s(\mathbf{x},\mathbf{y})$：

- 内积： $s(\mathbf{x},\mathbf{y}) = c\mathbf{x}^{\mathsf{T}}\mathbf{y}$
- 标准化内积/相关系数： $s(\mathbf{x},\mathbf{y}) = c\mathbf{x}^{\mathsf{T}}\mathbf{y}/\|\mathbf{x}\|\|\mathbf{y}\|$
- 距离的减函数，比如： $s(\mathbf{x},\mathbf{y}) = c\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|}$

对于数据矩阵X（行代表个体、列代表变量）

XX^{T}：所有个体两两之间的相似度；

$X^{\mathsf{T}}X$：所有变量两两之间的相似度。

例3. 假设 $\mathbf{x},\mathbf{y}$都是长度为p的0-1序列，

$$\mathbf{x} = 1\,1\,0\,1\,0\,0$$
$$\mathbf{y} = 0\,1\,0\,0\,1\,0$$

- $s(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathsf{T}}\mathbf{y}/p = 1$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).

附：0-1序列的各种相似系数

a =1-1 匹配个数；b =1-0匹配；
c = 0-1匹配；d =0-0匹配
$p = a + b + c + d$

TABLE 12.2 SIMILARITY COEFFICIENTS FOR CLUSTERING ITEMS*

Coefficient	Rationale
1. $\frac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No 0–0 matches in numerator.
5. $\frac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\frac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\frac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

相近性度量的欧氏表示

问题：给定n个对象(object)的两两之间的相近性度量，能否将物体映射为某个维数的欧氏空间中的n个点，使得这些点之间的数学距离或内积匹配原始的相近性度量？

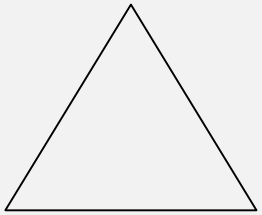
- 如果对象之间的相近度量是主观评价或经由非数学距离计算而得，那么欧氏表示称为多维标度法（MDS, Multi-dimensional scaling）.
- 如果对象之间的相近度量是高维空间的欧氏距离或内积，那么我们希望在低维欧氏空间表示这些高维对象， 这是一般的降维方法，比如PCA。

相似系数矩阵/距离矩阵

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

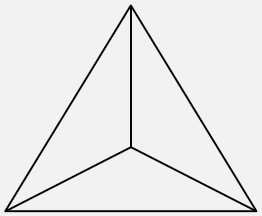
下面研究相似度矩阵和距离矩阵欧氏化的充要条件。需要注意的是，实际问题一般做不到完全欧氏化，但下面的讨论能够提供理论指导。

例4. 假设我们主观判定三个物体两两相似程度相同（两两之间的距离相同），在二维欧氏空间我们以等边三角形表示。

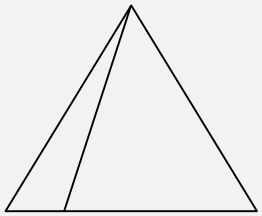


$$D = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

四个两两“距离”相同的物体如何在二维空间表示？下图是两种MDS表示



isoMDS



cMDS

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。

注：命题1中相似度矩阵如果是标准化的（对角元等于1），则命题1给出了何时S的元素的反余弦代表欧氏空间中n个单位向量两两之间的夹角。

反过来，命题1也给出了半正定矩阵的意义，即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。

例如，给定一个n阶相关系数矩阵$R \geq 0$，则我们一定可以在$k = rank(R)$维或k维以上欧氏空间中找到n个单位向量，它们两两之间的样本相关系数构成R矩阵（也一定能在概率空间中找到n个随机变量，它们两两之间的总体相关系数构成R矩阵）。

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$\mathbf{x}_1,\ldots,\mathbf{x}_n\in R^k$，使得

$$d_{ij}=||\mathbf{x}_i-\mathbf{x}_j||,$$

则称D是欧氏距离矩阵。

首先考虑距离矩阵D是欧氏距离矩阵的必要条件。

若$D_{n\times n}$是欧氏的，则存在p和向量$\mathbf{x}_1,\ldots,\mathbf{x}_n\in R^p$，$X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathsf{T}}$，使得

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

记$D_2=\left(d_{ij}^2\right)$，即矩阵$D_2$的$(i,j)$元为$d_{ij}^2$

$$D_2=\left(d_{ij}^2\right)=\left(\|\mathbf{x}_i\|^2\right)+\left(\|\mathbf{x}_j\|^2\right)-2\left(\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j\right)=\mathbf{1}_n\mathbf{a}^{\mathsf{T}}+\mathbf{a}\mathbf{1}_n^{\mathsf{T}}-2XX^{\mathsf{T}}\quad(*)$$

其中向量$\mathbf{a}=(\|\mathbf{x}_1\|^2,\ldots,\|\mathbf{x}_n\|^2)^{\mathsf{T}}$，前两项是秩1矩阵。
因此除了前两个次要的秩1矩阵之外，$-D_2$大致是半正定矩阵。

（*）式两端同时左、右乘投影矩阵$J_n = I_n - \mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$即可消掉前两个次要项：

$$-J_nD_2J_n = 2J_nXX^{\mathsf{T}}J_n = 2X_cX_c^{\mathsf{T}} \geq 0$$

其中$X_c = J_nX$是X的列中心化矩阵。因此我们得到了D是欧氏距离矩阵的一个必要条件：

D是欧氏距离矩阵$\Rightarrow -J_nD_2J_n \geq 0$（即欧氏相似度矩阵）。

下述命题2说明D_2的双向中心化矩阵的非负定性也是充分条件。

命题2. 假设$D_{n\times n} = (d_{ij})$是一个距离矩阵, 记$\tilde{D} = \left(-\frac{1}{2}d_{ij}^2\right)$, $J_n = I_n - \mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$,
则D是欧氏的当且仅当$S = J_n\tilde{D}J_n \geq 0$（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 (已正式出版).

证明：必要性前面已证。现证充分性。

假设 $S=(s_{ij})=J_n\tilde{D}J_n\geq 0,\ \mathrm{rank}(S)=k$,则存在$X_{n\times k}=(\mathbf{x}_1,...,\mathbf{x}_n)^{\mathsf{T}}$

使得$S=XX^{\mathsf{T}}$，即$s_{ij}=\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$.

下面证明对任何$1\leq i,j\leq n,\|\mathbf{x}_i-\mathbf{x}_j\|=d_{ij}$.

由$S=J_n\tilde{D}J_n$，记$a_{ij}=-\frac{1}{2}d_{ij}^2$，容易验证(后面引理1)：

$$s_{ij}=a_{ij}-\bar{a}_{i\bullet}-\bar{a}_{\bullet j}+\bar{a}_{\bullet\bullet}$$

注意到$a_{ii}=d_{ii}^2=0,\bar{a}_{i\bullet}=\bar{a}_{\bullet i}$,所以

$$\begin{aligned}\|\mathbf{x}_i-\mathbf{x}_j\|^2&=\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_i+\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j-2\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j=s_{ii}+s_{jj}-2s_{ij}\\&=\left(a_{ii}-\bar{a}_{i\bullet}-\bar{a}_{\bullet i}+\bar{a}_{\bullet\bullet}\right)+\left(a_{jj}-\bar{a}_{j\bullet}-\bar{a}_{\bullet j}+\bar{a}_{\bullet\bullet}\right)-2(a_{ij}-\bar{a}_{i\bullet}-\bar{a}_{\bullet j}+\bar{a}_{\bullet\bullet})\\&=-2a_{ij}=d_{ij}^2.\end{aligned}$$

注1：命题2说明了欧氏距离$d_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|$和中心化内积

$s_{ij}=(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}(\mathbf{x}_j-\bar{\mathbf{x}})$互相唯一确定：

$$d_{ij}^2=s_{ii}+s_{jj}-2s_{ij},\quad s_{ij}=-\frac{1}{2}\left(d_{ij}^2-\overline{d_{i\bullet}^2}-\overline{d_{\bullet j}^2}+\overline{d_{\bullet\bullet}^2}\right)$$

$S=(s_{ij})=\left((\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}(\mathbf{x}_j-\bar{\mathbf{x}})\right)$和$\tilde{D}=\left(-\frac{1}{2}d_{ij}^2\right)=\left(-\frac{1}{2}\|\mathbf{x}_i-\mathbf{x}_j\|^2\right)$的关系如下：

$$\tilde{D}=S-\frac{1}{2}\left(\mathbf{s}\mathbf{1}_n^{\mathsf{T}}+\mathbf{1}_n\mathbf{s}^{\mathsf{T}}\right),\quad S=J_n\tilde{D}J_n,$$

其中$\mathbf{s}=(s_{11},\ldots,s_{nn})^{\mathsf{T}}$, $J_n=I_n-\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n.$

引理1. 对任何$n\times p$矩阵$A=(a_{ij})$, 记行平均, 列平均和总平均分别为

$$\bar{a}_{i\bullet}=\sum_{j=1}^{p}a_{ij}/p,\quad \bar{a}_{\bullet j}=\sum_{i=1}^{n}a_{ij}/n,\quad \bar{a}_{\bullet\bullet}=\sum_{i=1}^{n}\sum_{j=1}^{p}a_{ij}/np$$

记$\bar{\mathbf{c}}=(\bar{a}_{\bullet 1},...,\bar{a}_{\bullet p})^{\mathsf{T}}$，$\bar{\mathbf{r}}=(\bar{a}_{1\bullet},...,\bar{a}_{n\bullet})^{\mathsf{T}}$，则$A$的双向中心化$C=J_nAJ_p$为

$$A_{\mathrm{cc}}=J_nAJ_p=A-\mathbf{1}_n\bar{\mathbf{c}}^{\mathsf{T}}-\bar{\mathbf{r}}\mathbf{1}_p{}^{\mathsf{T}}+\bar{a}_{\bullet\bullet}\mathbf{1}_n\mathbf{1}_p{}^{\mathsf{T}}=\left(a_{ij}-\bar{a}_{i\bullet}-\bar{a}_{\bullet j}+\bar{a}_{\bullet\bullet}\right)$$

验证：$J_nAJ_p=(I_n-\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n)A(I_p-\mathbf{1}_p\mathbf{1}_p^{\mathsf{T}}/p)=A-\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}A/n-A\mathbf{1}_p\mathbf{1}_p^{\mathsf{T}}/p+\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}A\mathbf{1}_p\mathbf{1}_p^{\mathsf{T}}/(np)$

注意$A^{\mathsf{T}}\mathbf{1}_n/n=\bar{\mathbf{c}}$，$A\mathbf{1}_p/p=\bar{\mathbf{r}}$，$\bar{a}_{\bullet\bullet}=\mathbf{1}_n^{\mathsf{T}}A\mathbf{1}_p/np$。

$A_{n\times p}$的各种中心化　$(\bar{\mathbf{r}}=A\mathbf{1}_p/p,\bar{\mathbf{c}}=A^{\mathsf{T}}\mathbf{1}_n/n)$

- 列中心化(最常用)：$J_nA=A-\mathbf{1}_n\bar{\mathbf{c}}^{\mathsf{T}}=\left(a_{ij}-\bar{a}_{\bullet j}\right)$
- 行中心化：$AJ_p=A-\bar{\mathbf{r}}\mathbf{1}_p{}^{\mathsf{T}}=\left(a_{ij}-\bar{a}_{i\bullet}\right)$
- 双向中心化(*anova*)：$J_nAJ_p=A-\mathbf{1}_n\bar{\mathbf{c}}^{\mathsf{T}}-\bar{\mathbf{r}}\mathbf{1}_p{}^{\mathsf{T}}+\bar{a}_{\bullet\bullet}\mathbf{1}_n\mathbf{1}_p{}^{\mathsf{T}}=\left(a_{ij}-\bar{a}_{i\bullet}-\bar{a}_{\bullet j}+\bar{a}_{\bullet\bullet}\right)$
- 双向中心化(列联表)：$A-\bar{\mathbf{r}}\,\bar{\mathbf{c}}^{\mathsf{T}}/\bar{a}_{\bullet\bullet}=\left(a_{ij}-N\bar{a}_{i\bullet}\bar{a}_{\bullet j}\right),\quad N=\sum a_{ij}.$

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2. 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，尝试随机取一个r.v.矩阵A，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$A = (a_{ij})$, 其元素iid 服从$N(0,1)$分布, 观察这种矩阵的效果。由 $EA^{\top}A = kI_p$, 有

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

所以平均来看， $A/\sqrt{k}$-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)

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

此时以大于0的概率（接近1的概率）有

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

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)

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