典型相关分析与多维标度法

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 典型相关分析

- 典型相关分析 (Canonical correlation analysis, CCA) 研究多个变量与多个变量之间的相关性
- 工厂对原料的主要质量指标 $\mathbf{X} = (X_1, \ldots, X_p)'$ 和产品质量的主要指标 $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ 之间的关系很感兴趣
- 婚姻研究中, 小伙子对他所追求姑娘的主要指标 $\mathbf{X}$ 和姑娘向往的主要指标 $\mathbf{Y}$ 之间的关系
- 直接使用 $Cov(\mathbf{X}, \mathbf{Y})$(或者相关系数矩阵) 在多元场合无法从整体上合适解释两者之间相关性
- Hotelling (1935,1936) 最早提出使用它们的线性组合变量 (典型变量)$a'\mathbf{X}$ 和 $b'\mathbf{Y}$ 之间的相关性来度量 $\mathbf{X}$ 和 $\mathbf{Y}$ 之间的相关性. 什么样的 a, b 合适呢?

- 选择 a, b, 使得相关性最大:

$$(\hat{a}, \hat{b}) = \arg\max_{a,b\neq 0} corr(a'\mathbf{X}, b'\mathbf{Y})$$
$$= \arg\max_{a,b\neq 0} \frac{Cov(a'\mathbf{X}, b'\mathbf{Y})}{\sqrt{a'Cov(\mathbf{X})ab'Cov(\mathbf{Y})b}}$$

注意到 $corr(ca'\mathbf{X}, cb'\mathbf{Y}) = corr(a'\mathbf{X}, b'\mathbf{Y}), \forall c \neq 0$, 因此上述 $(\hat{a}, \hat{b})$ 不唯一. 为此, 可施加适当的限制条件使解唯一. 自然的限制条件为

$$a'Cov(\mathbf{X})a = Var(a'\mathbf{X}) = 1, b'Cov(\mathbf{Y})b = Var(b'\mathbf{Y}) = 1$$

- 记 $\Sigma_{XX} = Cov(\mathbf{X}), \Sigma_{YY} = Cov(\mathbf{Y}), \Sigma_{XY} = Cov(\mathbf{X}, \mathbf{Y})$, 则问题转换为

$$\text{最大化} \quad a'\Sigma_{XY}b$$
$$s.t. \quad a'\Sigma_{XX}a = 1, \quad b'\Sigma_{YY}b = 1$$

- 假设 $\Sigma_{XX} > 0, \Sigma_{YY} > 0$, 则使用 Lagrange 乘子法

$$G(a,b) = a'\Sigma_{XY}b - \frac{1}{2}\lambda_1(a'\Sigma_{XX}a - 1) - \frac{1}{2}\lambda_2(b'\Sigma_{YY}b - 1)$$

分别对 a, b 求偏导并令为零, 得到

$$\boxed{\begin{cases} \Sigma_{XY}b - \lambda_1\Sigma_{XX}a = 0 \\ \Sigma_{YX}a - \lambda_2\Sigma_{YY}b = 0 \end{cases}}$$

广义特征根问题

由此得到

$$\lambda_1 = \lambda_1 a'\Sigma_{XX}a = a'\Sigma_{XY}b = \lambda_2$$

因此记 $\lambda = \lambda_1 = \lambda_2$, 将 $\lambda b = \Sigma_{YY}^{-1}\Sigma_{YX}a$ 带入得到

$$\begin{cases} \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}a - \lambda^2\Sigma_{XX}a = 0 \\ \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}b - \lambda^2\Sigma_{YY}b = 0 \end{cases}$$

即 a, b 分别为矩阵

$$M_1 = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$$
$$M_2 = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

的特征根为 λ^2 所对应的特征向量.

- 若记 $K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, $\alpha = \Sigma_{XX}^{1/2}a, \beta = \Sigma_{YY}^{1/2}b$, 则

$$KK'\alpha = \lambda^2\alpha$$
$$K'K\beta = \lambda^2\beta$$

即 α, β 分别为矩阵 KK' 和 $K'K$ 的特征根 λ^2 所对应的特征向量.

- 因此**第一典则方向**为

$$(a_1, b_1) = \arg\max_{a,b} a'\Sigma_{XY}b \quad s.t. \quad a'\Sigma_{XX}a = 1, \quad b'\Sigma_{YY}b = 1$$

此时最大的相关系数为 $\rho_1 = corr(a_1'\mathbf{X}, b_1'\mathbf{Y})$.

- 而给定前 $k-1$ $(k>1)$ 个典则方向 $(a_1, b_1), \ldots, (a_{k-1}, b_{k-1})$ 后, 第k **个典则方向** 为

$$(a_k, b_k) = \underset{\substack{a'\Sigma_{XX}a=1 \\ b'\Sigma_{YY}b=1 \\ corr(a'X, a_i'X)=0, i=1,\ldots,k-1 \\ corr(b'Y, b_i'Y)=0, i=1,\ldots,k-1}}{\operatorname{argmax}} a'\Sigma_{XY}b$$

- 所有典则方向可以通过广义特征根方程得到. 令 $\lambda_1^2 \geq \cdots \geq \lambda_s^2 > 0$ 为 KK' 和 $K'K$ 的全部非零特征根, 其中 $s \leq min\{p, q\}$. 对应的 KK' 的特征向量为 α_i, $K'K$ 的特征向量为 β_i, 则可以得到

$$a_i = \Sigma_{XX}^{-1/2}\alpha_i, b_i = \Sigma_{YY}^{-1/2}\beta_i, i = 1, \ldots, s$$

◇ 称 $(a_1, b_1), \ldots, (a_s, b_s)$ 为**典则方向**(canonical directions), 而称

$$U_i = a_i'\mathbf{X}, V_i = b_i'\mathbf{Y}$$

为第 i 对**典型相关变量**(canonical variates), 其满足

$$corr(U_i, V_i) = \lambda_i^2, i = 1, \ldots, s$$

$$corr(U_i, U_j) = 0, corr(V_i, V_j) = 0, i \neq j$$

◇ 从上面可以看出, 第二对典型相关变量应不包含第一对典型相关变量的信息 (相关系数为零). 以此类推. 第 k 对典型相关变量应和之前的 $k-1$ 对典型相关变量不相关.

定理 1. 设 $X^* = A'X + \mathbf{u}, Y^* = B'Y + \mathbf{v}$, 其中 $A : p \times p, B : q \times q$ 为可逆方阵, $\mathbf{u} : p \times 1, \mathbf{v} : q \times 1$ 为实常数向量, 则

(1) X^* 和 Y^* 的典型相关变量为 $a^{*\prime}_i X^*$ 和 $b^{*\prime}_i Y$, 其中 $a^*{}_i = A^{-1}a_i, b^*{}_i = B^{-1}b_i$, a_i, b_i 为 X, Y 的第 i 对典型相关变量的系数.

(2) $corr(a^{*\prime}_i X^*, b^{*\prime}_i Y) = corr(a_i' X, b_i' Y)$, 即线性变换不改变相关性.

注: 若在定理中取 $A = (diag\Sigma_{XX})^{1/2}$, $B = (diag\Sigma_{YY})^{-1/2}$, 则前面关于协方差矩阵的结果都可以应用到相关系数矩阵下.

样本典型相关 (classical CCA)

- 当总体协方差 $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY}$ 未知时候, 设 $\begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}_{,i=1,\ldots,n}$ 为总体 $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ 的一组样本, $n \geq p, n \geq q$.

- 则由样本协方差矩阵得到 $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY}$ 的估计

$$\hat{\Sigma}_{XX} = S_{XX} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{n-1}A_{XX}$$

$$\hat{\Sigma}_{YY} = S_{YY} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n-1}A_{YY}$$

$$\hat{\Sigma}_{XY} = S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n-1}A_{XY}$$

- 使用估计 $\hat{\Sigma}_{XX}, \hat{\Sigma}_{YY}, \hat{\Sigma}_{XY}$ 代替得到**样本典型相关变量**$(\hat{U}_i, \hat{V}_i)$ 和典则方向 $(\hat{a}_i, \hat{b}_i)$

- 这等价于使样本相关最大化: 记 $\mathbf{x}, \mathbf{y}$ 为中心化的 $n \times p, n \times q$ 样本矩阵, 则

$$(\hat{a}_1, \hat{b}_1) = \underset{\|\mathbf{x}a\|=1, \|\mathbf{y}b\|=1}{\operatorname{argmax}} \; a'\mathbf{x}'\mathbf{y}b$$

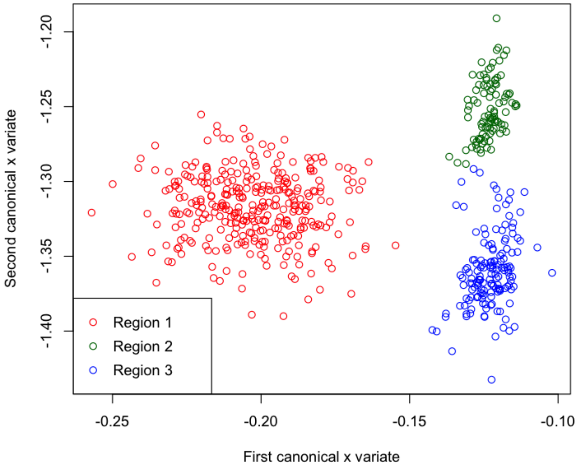
例: 橄榄油数据

- R 包 *classifly* 中的数据集 *olives* 记录了 $n = 572$ 种橄榄油的 $p = 9$ 特征变量值, 其中变量 1 取值 $\{1, 2, 3\}$, 表示意大利的三个地区. 其他变量为 8 种脂肪酸含量测量值.

- 我们感兴趣的是三个地区与脂肪酸测量之间的相关性. 因此取 $\mathbf{x} \in R^{572\times 8}$, $\mathbf{y} \in R^{572\times 3}$ 为三个地区的示性变量矩阵, 每行表示第一个地区:

$$\mathbf{y} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ & \cdots & \end{pmatrix}$$

- 此时, 典则相关分析与线性判别分析等价.



-1.20
-1.25
-1.30
-1.35
-1.40
Second canonical x variate
-0.25
-0.20
-0.15
-0.10
First canonical x variate
Region 1
Region 2
Region 3

Regularized CCA

- CCA 假定了样本量 $n > \max\{p, q\}$, 因此当 $n \leq \{p, q\}$ 时候就不能使用 (样本协方差矩阵不可逆)

- 另外, 当 $\mathbf{X}$ 或者 $\mathbf{Y}$ 的分量之间高度相关时候, 样本协方差矩阵 S_{XX} 和/或 S_{YY} 也倾向于病态

- 因此, 经典的 CCA 的一个标准条件为 $n \geq p + q + 1$(Eaton and Perlman 1973)

- 一种解决方法就是使用 $\Sigma_{XX}(\lambda_1)$ 和 $\Sigma_{YY}(\lambda_2)$ 来代替 S_{XX} 和 S_{YY}:

$$\Sigma_{XX}(\lambda_1) = S_{XX} + \lambda_1 I_p$$
$$\Sigma_{YY}(\lambda_2) = S_{YY} + \lambda_2 I_q$$

- 最优的 λ_1, λ_2 使用交叉验证方法来估计: 记 $\lambda = (\lambda_1, \lambda_2)$, $(a_\lambda^{(-i)}, b_\lambda^{(-i)})$ 为去掉第 i 组观测 $(\mathbf{x}_i, \mathbf{y}_i)$ 后得到的样本第一点则变量方向, 令

$$CV(\lambda_1, \lambda_2) = corr(\{a_\lambda^{(-i)\prime}\mathbf{x}_i\}_{i=1}^n, \{b_\lambda^{(-i)\prime}\mathbf{y}_i\}_{i=1}^n)$$

则最优的 λ 为

$$\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2) = \arg\max_{\lambda_1, \lambda_2} CV(\lambda_1, \lambda_2)$$

CCA in R

R base 里面的 **cancor**函数用于典则相关分析:

```
cc = cancor(x,y) #x:nxp; y:nxq
a = cc$xcoef
b = cc$ycoef
rho = cc$cor
xvars = x %*% alpha
yvars = y %*% beta
```

cca包的函数 *cc* 还可以进行正则化的 CCA.

1.1.1 CCA-LDA

- 设从 k 个总体 $G_1, \ldots, G_k$ 中各抽取 $n_1, \ldots, n_k$ 个样本点: $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}; \ldots; \mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_k}^{(k)}$, 每个均是 p 维向量.

- 令

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_1^{(i)\prime} \\ \vdots \\ \mathbf{x}_{n_i}^{(i)\prime} \end{pmatrix}, i = 1, \ldots, k; \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_k \end{pmatrix}; n = n_1 + \cdots + n_k$$

- 由 Fisher 线性判别分析知判别函数是矩阵 $W^{-1}B$ 的特征向量:

$$B = \sum_{i=1}^{k} n_i (\bar{\mathbf{x}}^{(i)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(i)} - \bar{\mathbf{x}})'$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)})'$$

特别第一判别函数的方向为

$$\hat{a_1} = \underset{\|a\|=1}{\text{argmax}} \frac{a'Ba}{a'Wa}$$

- 我们使用 0-1 元素方法来建立伪变量:

$$\mathbf{y}_i = \mathbf{1}_{n_i} e'_k(i), i = 1, \ldots, k; \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k \end{pmatrix}$$

- 考虑 $\mathbf{x} \in R^{n \times p}$ 和 $\mathbf{y} \in R^{n \times k}$ 之间的样本典型相关. 由前面的讨论知样本典则方向 $\hat{a}$ 为矩阵

$$A_{XY} A_{YY}^{-1} A_{YX} a - \lambda^2 A_{XX} a = 0$$

的解.

– 由 $\mathbf{y}$ 的构造过程知 $A_{YY} = \mathbf{y}'(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{y}$ 不可逆, 因此上式中应使用广义逆 A_{YY}^-. 注意到上式与 A_{YY}^- 的取法无关, 而 $(\mathbf{y}'\mathbf{y})^{-1}$ 为一个广义逆 A_{YY}^-, 因此

$$A_{XY}(\mathbf{y}'\mathbf{y})^{-1}A_{YX} = B$$

– 又由 $A_{XX} = B + W$, 得到样本典则方向 $\hat{a}$ 为矩阵

$$Ba - \lambda^2(B+W)a = 0$$

的解. 等价于

$$\frac{\lambda^2}{1-\lambda^2}Wa - Ba = 0$$

即 $\hat{a}$ 为 B 相对于 W(即 $W^{-1}B$) 的特征向量, 与 FLDA 的解相同.

1.1.2 PCA-CCA-PLS

- 主成分分析 (PCA), 典型相关分析 (CCA) 和偏最小二乘 (PLS) 均使用投影方法. 它们之间的关系可以通过它们定义投影方向的优化准则来体现.

- CCA: 两个中心化的数据阵 $\mathbf{x}$: $n \times p$ 和 $\mathbf{y}$: $n \times q$, 寻找方向 a, b 使得

$$\max_{\|\mathbf{x}a\|=\|\mathbf{y}b\|=1} cov(\mathbf{x}a, \mathbf{y}b)$$

其中 $corr, cov, var$ 均指样本相关系数, 样本协方差和样本方差, 下同.

- PCA: 一个中心化的数据矩阵 $\mathbf{x}$: $n \times p$, 寻找样本方差最大的方向 a:

$$\max_{\|a\|=1} var(\mathbf{x}a) = a'\mathbf{x}'\mathbf{x}a$$

第 i 个主成分方向通过给定与前 $i-1$ 个主成分方向正交条件下, 再最大化上式得到. 记 $V_{p\times k}=[a_1,\ldots,a_k]$ 为前 k 个主成分方向组成的投影矩阵, 则主成分得分为 $T=\mathbf{x}V$.

– 在回归背景下, 考虑回归模型

$$\mathbf{y}_{n\times 1}=\mathbf{x}_{n\times p}\beta+e$$

在预测 Y 时候, 当 $p>n$ 时候, OLS 方法失效. PCA 回归给出了一种方法.

– PCR: 使用前 k 个主成分得分作为新的设计阵, 则有

$$\mathbf{y}_{n\times 1}=T_{n\times k}c+\epsilon$$

– 但是, PCR 方法仅仅考虑 $\mathbf{x}$ 的相关结构, 而不管这些方向是否对预测 Y 有帮助.

- PLS1: 偏最小二乘也是寻找投影 $W = [w_1, \ldots, w_k]$, 但是要求 X 在每个方向上的投影 (因子) 与 Y 相关性最大, 即

$$w_1 = \operatorname*{argmax}_{\|w\|=1} (cov(\mathbf{x}w, \mathbf{y}))^2 = w'\mathbf{x}'\mathbf{y}\mathbf{y}'\mathbf{x}w$$

由于 $\mathbf{x}'\mathbf{y}\mathbf{y}'\mathbf{x}$ 是秩为 1 的矩阵, 因此需要引入限制条件来得到其它投影方向

$$w_i = \operatorname*{argmax}_{w} w'\mathbf{x}'\mathbf{y}\mathbf{y}'\mathbf{x}w$$

$$s.t. \quad \|w\| = 1, w'\mathbf{x}'\mathbf{x}w_j = 0, 1 \leq j < i$$

最后使用回归模型

$$\mathbf{y}_{n\times 1} = XWc + \epsilon$$

进行预测. 当向量变量为多个时候, PLS1 的推广称为 PLS2.

1.2 多维标度法

- 多维标度法（Multidimensional Scaling, MDS）是一种在低维空间展示“距离”数据结构的多元数据分析技术.
- 多维标度法起源于心理测度学，用于理解人们判断的相似性. 多维标度法现在已经成为一种广泛用于心理学、市场调查、社会学、物理学、政治科学及生物学等领域的数据分析方法。
- 给你若干城市的距离，你能否确定这些城市之间的相对位置呢？假定你知道只是哪两个城市最近，哪两个城市次近等等，你是否还能确定它们之间的相对位置呢？
- 假定通过调查了解了 10 种饮料产品在消费者心中的相似程度，你能否确定这些产品在消费者心理空间中的相对位置呢？
- **多维标度法的目标**：当 n 个对象中各对对象之间的相似性（或

距离）给定时，确定这些对象在低维 (欧式) 空间中的表示（称为**感知图**, Perceptual Mapping)，并使其尽可能与原先的相似性（或距离）“大体匹配”，使得由降维所引起的任何变形达到最小。

- 低维 (欧式) 空间中排列的每一个点代表一个对象，因此点间的距离与对象间的相似性高度相关。也就是说，两个相似的对象由低维 (欧式) 空间中两个距离相近的点表示，而两个不相似的对象则由低维 (欧式) 空间两个距离较远的点表示。低维空间通常为二维或三维的欧氏空间，但也可以是非欧氏三维以上空间.

- n 个对象中各对对象之间的相似性常使用相异度或者相似度来反应对象的相近程度, 这种度量是一种广义距离.

一个 $n \times n$ 矩阵 $D = (d_{ij})$ 称为距离阵 (proximity)，如果满足

(1) $D = D'$

(2) $d_{ij} \geq 0, d_{ii} = 0, i, j = 1, \ldots, n$

Definition

- 给定距离阵 D 后, MDS 目的就是寻找一个整数 k, 以及 R^k 空间里的 n 个点 $x_1, \ldots, x_n$, 用 $\hat{D} = (\hat{d}_{ij})$ 表示它们两两间的欧式距离, 使得 $\hat{D}$ 和 D 在某种意义下相近.

- $\hat{D}$ 称为是 D 的**拟合距离阵**, $x_1, \ldots, x_n$ 称为 D 的**拟合构图**. 如果 $\hat{D} = D$, 则称 $x_1, \ldots, x_n$ 为 D 的一个**构图**.

- MDS 按照相似性 (或距离) 数据测量尺度的不同, 常分为度量 MDS(metric MDS) 和非度量 MDS(nonmetric MDS).

- 当相似性（距离）的实际数值为间隔尺度和比例尺度时称为

度量 MDS，当原始相似性（距离）为有序尺度时称为非度量 MDS。

- 度量 MDS 假设描述 p 维对象相似度 (或距离)S_{ij} 和低维空间下的距离 d_{ij} 之间有函数关系

$$d_{ij} = f(S_{ij})$$

其中 f 为连续的单调函数. 比如

$$f(S_{ij}) = bS_{ij} \qquad \textit{Ratio MDS}$$

$$f(S_{ij}) = a + bS_{ij} \qquad \textit{Interval MDS}$$

其中 a, b 为参数. 其他形式包括高阶多项式, 对数和指数等等.

- 非度量 MDS 仅要求函数 f 保持相似性 (或距离) 之间的大小序关系:

$$S_{ij} < S_{rs} \Longrightarrow f(S_{ij}) < f(S_{rs}) \qquad \textit{Ordinal MDS}$$

1.2.1 度量 MDS

- 度量 MDS 需要寻找映射函数 f. 常用的方法是构造合适的目标函数, 然后优化其得到 f.

- 比如一种度量 d_{ij} 和 $f(S_{ij})$ 之间差异性的目标函数为

$$Stress = \sqrt{\frac{\sum_i \sum_j [f(S_{ijj}) - d_{ij}]^2}{scale\ factor}}$$

- 这类目标函数称为**应力**(Stress), 不同形式的 scale factor 构成不同的应力和 MDS 类型.

- Krustal (1964) 提出一种 scale factor 为

$$\sum_i \sum_j d_{ij}^2 \qquad Stress-1$$

- 对应力函数进行最小化可以使用各种数值方法进行.

Classical MDS

- 原始空间下的距离阵和低维空间下的距离阵都采用欧式距离阵
- 距离阵 D 为欧式的, 即存在某个正整数 p 以及 R^p 空间的 n 个点 $x_1, \ldots, x_n$, 使得

$$d_{ij}^2 = \|x_i - x_j\|^2, i, j = 1, \ldots, n$$

- 目标在于: 寻找 D 的 (拟合) 构图 $x_1, \ldots, x_n$, 其想法为
  - 将平方的欧式距离阵 $D = (d_{ij}^2)$ 变换为一个非负定矩阵 B
  - 由 B 的特征根和特征向量得到构图 X, X 的每一行表示低维空间的点.
- 为此, 记原始的 p 维对象 (观测点) 为 $\mathbf{x}_1, \ldots, \mathbf{x}_n$(一般是未知的), 两两之间的距离平方为

$$d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j, i, j = 1, \ldots, n$$

- 再记 $X_{n\times p}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]'$, $K=XX'$ 和列向量 $\kappa=diag(K)$, 则上述矩阵 D 可以表示为

$$D=\kappa\mathbf{1}'+\mathbf{1}\kappa'-2K$$

- 令 $B=-\frac{1}{2}HDH, H=I_n-\frac{1}{n}\mathbf{1}\mathbf{1}'$, 则

$$B=-\frac{1}{2}H(\kappa\mathbf{1}'+\mathbf{1}\kappa'-2K)H=(HX)(HX)'\geq 0$$

- **可以证明：距离阵 D 为欧式的，当且仅当上式定义的 $B\geq 0$.**

Classical MDS **算法**: 给定平方距离矩阵 $D = (d_{ij}^2)$,

1. 计算矩阵 H 和 B.

2. 对 B 进行正交分解: $B = U\Lambda U'$, 选择 r 个最大的特征根和相应的特征向量, 得到 r 维空间下的拟合构图 $\hat{X} = \Lambda_r^{1/2} U_r'$, 其中 $\Lambda_r = diag(\lambda_1, \ldots, \lambda_r)$ 为 B 前 r 个特征根; $U_r = [u_1, \ldots, u_r]$ 为相应的特征向量组成的矩阵, $\hat{X}$ 称为 X 的古典解.

- 其中, r 的确定: 事先确定 $r = 1, 2$ 或 3; 或者通过计算前面特征根占全体特征根的比例确定.

- 预先给定变差贡献比例

$$k = \frac{\lambda_1 + \cdots + \lambda_r}{\text{正特征根和}} \geq k_0$$

- **注意:** 并不是所有的距离阵都是欧式距离阵.
- 当距离阵为欧氏时, 可求得一个 D 的构图 X, 当距离阵不是欧氏时, 只能求得 D 的拟合构图。在实际应用中, 即使 D 为欧氏, 一般也只求 $r = 2$ 或 3 的低维拟合构图。
- 值得注意的是, 由于多维标度法求解的 n 个点仅仅要求它们的相对欧氏距离与 D 相近, 也就是说, 只与相对位置相近而与绝对位置无关, 根据欧氏距离在正交变换和平移变换下的不变性, 显然所求得解并不唯一。

相似系数矩阵

- 某些问题中, 已知的是 n 个对象之间的相似系数矩阵 C, 而不是距离阵 D
- 相似系数矩阵 C 即为: 称 n 阶方阵 $C = (c_{ij})$ 为相似系数矩阵, 如果 (1) $C' = C$. (2) $c_{ij} \le c_{ii}, \forall i, j$

- 相似系数矩阵 C 和距离阵 D 之间有关系

$$d_{ij} = (c_{ii} + c_{jj} - 2c_{ij})^{1/2}$$

- 可以证明: 若相似系数矩阵 $C \geq 0$, 则由上式定义的距离阵 D 为欧式的.

古典解的性质

- 由奇异值分解知

$$HX = U\Lambda V'$$

其中 $U_{n\times p}$ 的列向量为 $(HX)(HX)' = B$ 的特征向量, $V_{p\times p}$ 的行为 $(HX)'(HX) = X'HX$ 的特征向量.

- 从而 $\hat{X} = HXV_k$, 其中 V_k 为 V' 的前 k 列. 其中 H 仅起中心化的作用, 与 X 无关, 因此拟合构图 $\hat{X}$ 等价于 XV_k, 即 X 右乘一个列单位正交的矩阵.

定理 2. *设 D 为欧式的距离阵, $X_{n\times p}$ 为其构图, 给定 k $(1 \leq k < p)$, 则一切形如 $\hat{X} = XT_1$(使得 $T = (T_1, T_2)$ 为 p 阶正交矩阵) 的 k 维拟合构图中, 以 $T_1 = V_k$ 时使下式达到最小:*

$$\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2$$

1.2.2 非度量 MDS

- 假设 $m = n(n-1)/2$ 个相异度 (距离) 不能直接测量, 但是我们有其从小到大顺序:

$$\delta_{r_1 s_1} < \delta_{r_2 s_2} < \cdots < \delta_{r_m s_m}$$

- 我们在 R^k 空间寻找一组点, 满足距离

$$d_{r_1 s_1} < d_{r_2 s_2} < \cdots < d_{r_m s_m}$$

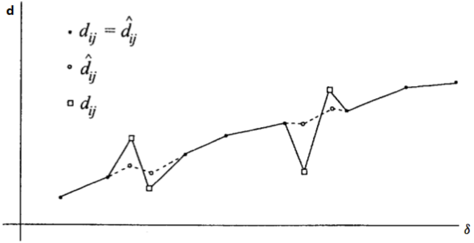
  与原始相异度 (距离) 序相同.

- 给定一组拟合构图, 其两两距离 d_{ij} 对 δ_{ij} 作图未必是单调的. 则记 $\hat{d}_{ij}$ 为使得改图为单调的估计点, 即最小化

$$Stress(k) = \left(\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}\right)^{1/2}$$

满足单调限制条件

$$\hat{d}_{r_1 s_1} \leq \hat{d}_{r_2 s_2} \leq \cdots \leq \hat{d}_{r_m s_m}$$



- $\hat{d}_{ij}$ 并不是距离, 它们仅仅是判断 d_{ij} 单调关系的参考数.

Kruskal **算法** 在给定的 R^k 空间里选择一个拟合构图, 其 $Stress(k)$ 值越小越好:

1. 随机在 R^k 里选择一个拟合构图点
2. 计算给定构图的距离 d_{ij}, 估计 $\hat{d}_{ij}$, 使得单调限制条件成立.
3. 寻找一个新的拟合构图点, 其距离 d_{ij} 使得 $Stress(k)$ 最小
4. 对第 3 步中的距离 d_{ij}, 估计新的满足单调限制条件的 $\hat{d}_{ij}$, 从而得到新的 $Stress(k)$ 值
5. 重复 3-4 直至 $Stress(k)$ 收敛.
6. 作 $Stress(k) \sim k$ 图, 寻找最好的 k.

Kruskal 提出: $Stress(k) \leq 5\%$ 最好, $5\% \sim 10\%$ 次之, 大于 10% 较差.