主成分分析思想原理

目录

方法起源

起源一：寻找重要因素

在若干个相互关联、关系复杂的一组变量 $x_1, x_2, \cdots, x_p$ 中，想找到最为关键的因素，是一个重要的科学问题。

在寻找关键因素过程中，还需要找到能够反映该组变量这个群体的主要特征。

起源二：综合评价要求评价指标线性无关

- 在做综合评价的时候，往往需要将多个评价指标综合成一个指标。综合时除了需要将指标同向，还需要评价指标间线性无关或者不相关。
- 但是很多实际问题中，指标之间是高度关联的，在这种情况下如何进行综合评价？

起源三：建立回归模型的需要

- 在做多元线性回归模型时，理想状态下是需要自变量线性无关的。
- 而且，模型拟合时，还需要样本点的个数 n 与自变量的个数 p 满足一个不等式：
- $$n > 3(p+1)，$$

一旦两个条件有一个满足，回归模型的效果将受影响。

思想原理

原始数据 $x_1, x_2, \cdots, x_p$ 是一组 n 维列向量组

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ & & & \\ x_{n1} & x_{n1} & & x_{np} \end{pmatrix}$$

将其进行标准化处理:

$$x_j^* = \frac{x_{kj} - \bar{x}_j}{s_j}, \quad k = 1,2,\cdots,n; j = 1,2,\cdots,p. \quad s_j$$ 为标准差

$$\begin{cases} Z_1 = u_{11}x_1^* + u_{12}x_2^* + \cdots + u_{1p}x_p^* \\ Z_2 = u_{21}x_1^* + u_{22}x_2^* + \cdots + u_{2p}x_p^* \\ \vdots \\ Z_p = u_{p1}x_1^* + u_{p2}x_2^* + \cdots + u_{pp}x_p^* \end{cases}$$

$$D(Z_i) = \lambda_i, \quad \sum_i \lambda_i = p \quad 且 \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p.$$

对各 u_{ij} 的要求是：

（1）使各个综合指标 Z_i 彼此独立或不相关；

（2）使各个综合指标 Z_i 所反映的各个样品的总信息等于原来 p 个指标 x_j^* 所反映的各个样品的总信息，即

$$\sum D(Z_i) = \sum D(x_j^*) = p$$

主成分分析的系数确定

$$Z=\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}=\begin{pmatrix} u_{11}x_1^*+u_{12}x_2^*+\cdots+u_{1p}x_p^* \\ u_{21}x_1^*+u_{22}x_2^*+\cdots+u_{2p}x_p^* \\ \vdots \\ u_{p1}x_1^*+u_{p2}x_2^*+\cdots+u_{pp}x_p^* \end{pmatrix}=\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{pmatrix}\begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_p^* \end{pmatrix}=UX$$

$$D(Z)=\begin{pmatrix} D(Z_1) & \operatorname{cov}(Z_1,Z_2) & \cdots & \operatorname{cov}(Z_1,Z_p) \\ \operatorname{cov}(Z_2,Z_1) & D(Z_2) & \cdots & \operatorname{cov}(Z_2,Z_p) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}(Z_p,Z_1) & \operatorname{cov}(Z_p,Z_2) & \cdots & D(Z_p) \end{pmatrix}$$

$$=\begin{pmatrix} D(Z_1) & & & \\ & D(Z_2) & & \\ & & & \\ & & & D(Z_p) \end{pmatrix}=\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & & \\ & & & \lambda_p \end{pmatrix}$$

而 $D(UX)=E[UX-E(UX)][UX-E(UX)]'$

$$=E[UX-UE(X)][UX-UE(X)]'$$

$$=E[U(X-EX)][U(X-EX)]'$$

$$=E[U(X-EX)(X-EX)'U']=UD(X)U'=URU',$$

故 $URU'=\Lambda$，这里的 R 为相关系数矩阵.

根据线性代数(特征值与特征向量)中的结论：任何实对称矩阵都能正交相似变换成对角阵，即 $P^{-1}RP=\Lambda$，即 $U=P'=P^{-1}$，P 为正交阵。

主成分分析计算步骤

① 对原始资料矩阵进行标准化处理；

$$x_j^* = \frac{x_{kj} - \bar{x}_j}{s_j}, \quad k = 1,2,\cdots,n;\ j = 1,2,\cdots,p.$$

② 计算相关系数矩阵 R；

③ 计算R 的特征值 λ和单位正交特征向量U'；

$$URU' = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$$

④ 确定主成分个数；

$$\alpha_i = \frac{\lambda_i}{\sum_{i}^{p} \lambda_i}$$

主成分 Z_i 的方差贡献率

$$\eta_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

主成分 Z_1，$Z_2, \cdots$，Z_k 的累积方差贡献率

一般要超过80%

⑤ 建立相应主成分方程；　$Z=U^t X$

应用范畴

$$\begin{cases} Z_1 = u_{11}x_1^* + u_{21}x_2^* + \cdots + u_{p1}x_p^* \\ Z_2 = u_{12}x_1^* + u_{22}x_2^* + \cdots + u_{p2}x_p^* \\ \vdots \\ Z_p = u_{1p}x_1^* + u_{2p}x_2^* + \cdots + u_{pp}x_p^* \end{cases}$$

$$r(Z_i, x_j^*) = \frac{\operatorname{cov}(Z_i, x_j^*)}{\sqrt{DZ_i}\sqrt{Dx_j^*}} = \frac{EZ_i x_j^* - EZ_i Ex_j^*}{\sqrt{\lambda_i}} = \frac{\lambda_i u_{ij}}{\sqrt{\lambda_i}} = \sqrt{\lambda_i} u_{ij}$$

称 $r(Z_i, x_j^*)$ 为 Z_i 在 x_j^* 上的**因子载荷**.

因子载荷的**绝对值和它的符号**可反映主成分与原指标之间相关关系的密切程度

例如，某问题(13个患者反映病情的4个指标)最后得到的主成分表达式为

$$\begin{cases} Z_1 = 0.699964x_1^* + 0.689798x_2^* + 0.087939x_3^* + 0.162777x_4^* \\ Z_2 = 0.095010x_1^* - 0.283647x_2^* + 0.904159x_3^* + 0.304983x_4^* \\ Z_3 = -0.240049x_1^* + 0.058463x_2^* - 0270314x_3^* + 0.930532x_4^* \\ Z_4 = -0.665883x_1^* + 0.663555x_2^* + 0.318895x_3^* - 0.120830x_4^* \end{cases}$$

$$DZ_1 = \lambda_1 = 0.7818, \quad DZ_2 = \lambda_2 = 0.1182$$

可以找出重要因素，认为是 X_1, X_2

- 这个例子还可以对13个患者的病情严重程度排个顺序，怎么做？ 综合评价
- 如果要做四元回归，直接做的效果很差，因为3（4+1）=15，至少需要16个样本点。如果一定要做四元回归，那又怎么做呢？ 主成分回归