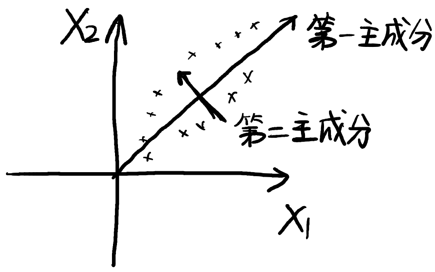
主成分分析(Principal Component Analysis)

目的：降维

1. 总体主成分分析

PCA利用正交变换把相关变量表示的观测数据转换为几个由线性无关变量表示的数据



问题：如何利用一个超平面，对样本恰当表达？

最近重构性：样本点到超平面的距离足够近

最大可分性：样本点在超平面的投影尽可能分开

在总体上进行的主成分分析称为总体主成分分析

1.1 定义和导出

设 $X=(x_1,x_2,\cdots,x_n)^T$ 是n维随机变量，$E(X)=\mu$，$Cov(X)=\Sigma\in R^{m\times m}$

考虑X到随机变量 $Y_i=\alpha_i^T X$，

$E(Y_i)=\alpha_i^T\mu$，$Var(Y_i)=\alpha_i^T\Sigma\alpha_i$，$cov(Y_i,Y_j)=\alpha_i^T\Sigma\alpha_j$

定义(总体主成分) 设 $X=(x_1,x_2,\cdots,x_n)^T$，$E(X)=\mu$，$Cov(X)=\Sigma$

称 $Y_i=\alpha_i^T X$ 是X的第i个主成分，如果：

(1) $\alpha_i^T\alpha_i=1 \quad (i=1,2,\cdots,m)$

(2) 变量 Y_i 与 Y_j 互不相关，$cov(Y_i,Y_j)=0 \ (i\neq j)$ (信息不重合)

(3) $Var(Y_i)=\max\limits_{\substack{\alpha^T\alpha=1 \\ \alpha^T\Sigma\alpha_j=0}} Var(\alpha^T X)$

1.2 主要性质

设Σ的特征值是$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，对应单位特征向量为$\alpha_1, \alpha_2, \cdots, \alpha_m$，

X的第K个主成分$Y_k = \alpha_k^T X$，$Var(Y_k) = \alpha_k^T \Sigma \alpha_k$

以求第一个主成分为例，相当于求解以下最优化问题

$$\max_{\alpha_i} \alpha_i^T \Sigma \alpha_i$$
$$s.t. \ \alpha_i^T \alpha_i = 1$$

定义拉格朗日函数 $\alpha_i^T \Sigma \alpha_i - \lambda(\alpha_i^T \alpha_i - 1) = L(\alpha_i)$

则 $L'(\alpha_i) = 2\{\Sigma \alpha_i - \lambda \alpha_i\} = 0 \qquad \Sigma \alpha_i = \lambda \alpha_i$

α_i是Σ的最大特征值对应的特征向量，$\alpha_i^T X$构成第一主成分

性质：(1) $cov(Y) = diag\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$（对角矩阵）

(2) $\sum_i \lambda_i = \sum_i \sigma_{ii}$（求和）

(3) 因子负荷量 $\rho(Y_k, X_i) = cor(Y_k, X_i) = \frac{\sqrt{\lambda_k}\alpha_{ki}}{\sqrt{\sigma_{ii}}}$，其中$\alpha_k = (\alpha_{k1}, \cdots, \alpha_{km})^T$

(4) $\sum_{i=1}^m \sigma_{ii} \rho(Y_k, X_i)^2 = \lambda_k$

(5) $\sum_k \rho^2(Y_k, X_i) = 1$

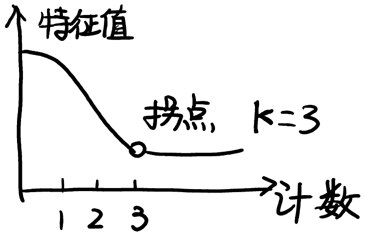
前K个主成分对变量X_i的贡献率为 $\nu_i = \sum_{j=1}^{k} \rho^2(Y_j, X_i)$

1.3 主成分的个数

前K个主成分的方差贡献率，定义为前K个主成分的方差和总方差的比值

$\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^n \lambda_j}$，通常选K使方差贡献率比较高，例如70%、80%以上

(1) 崖底碎石图 (scree plot)



选择拐点前一个点，选 $k=2$

(2) 累计方差贡献率

要求到达 70%~80% 以上

(3) Kaiser 准则

假设 $\sigma_{ii}=1$，选大于 1 的特征值的数目，$\sum_{k=1}^{m}\lambda_k=m$

2. 样本 PCA

观测数据用矩阵 $X=(x_1,x_2,\cdots,x_n)^T\in R^{n\times m}$ 表示，

样本协方差矩阵 $S=\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T$，均值 $\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j$

相关矩阵：$R=\mathrm{diag}(S)^{-1/2}S\,\mathrm{diag}(S)^{1/2}$

定义 2：(样本主成分) 通过样本协方差定义样本标准化 $x_{ij}^{*}=\frac{x_{ij}-\bar{x}_i}{\sqrt{s_{ii}}}$

第 k 个主成分

(1) 对 R 进行特征值分解，α_k 是第 k 个特征向量

(2) $y_k=X\alpha_k\in R^n$ 代表样本主成分 $Y\in R^{n\times k}$ 可以作为其它机器学习算法的输入，例如聚类分析

探索性因子分析

一个变量的变异性可以归结为公共因子+特殊因子，目的是寻找少量的公共因子解释一组输入变量

因子模型 设 $X=(X_1, X_2, \cdots, X_p)^T \in R^p$

$F_1, \cdots, F_q$ 公因子

$$X_1-\mu_1 = a_{11}F_1+a_{12}F_2+\cdots+a_{1q}F_q+\varepsilon_1$$

$$\vdots$$

$$X_p-\mu_p = \underbrace{a_{p1}F_1+a_{p2}F_2+\cdots+a_{pq}F_q}_{\text{公因子}}+\underbrace{\varepsilon_q}_{\text{特殊因子}}$$

写成矩阵形式 $X-\mu = AF+E$

$A\in R^{p\times q}$ 载荷矩阵

因子分析

$X\in R^p \quad F\in R^q \quad (q<<p)$

$X-\mu = \underset{\text{载荷矩阵}}{\underline{A}}F+\varepsilon \rightarrow$ 特殊因子 $\quad A=(a_{ij})_{p\times q}$

正交因子：

$E(F)=0, \quad Var(F)=1 \quad cov(F_i, F_j)=0$

$E(\varepsilon_k)=0, \quad Var(\varepsilon_k)=\sigma_k^2, \quad cov(\varepsilon_k, \varepsilon_m)=0$

$cov(F_i, \varepsilon_k)=0$

$$Var(X_k)= \underbrace{a_{k1}^2+a_{k2}^2+\cdots+a_{kp}^2}_{\text{共性方差}}+\underbrace{\sigma_k^2}_{\text{特殊方差}}$$

$Cov(X_k, X_m)=a_{k1}a_{m1}+a_{k2}a_{m2}+\cdots+a_{kp}a_{mp}$ （只与A有关）

公共因子的解释

主要通过载荷矩阵的绝对值较大的系数来解释

① 载荷矩阵系数正负无意义

② 正负对比有意义

模型估计

$cov(X)=\Sigma = AA^T+\Phi \qquad \Phi = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$

$A\in R^{p\times q}$ 对任意一个Q为正交矩阵

$A^*=AQ \quad A^*A^{*T}=AQQ^TA^T=AA^T$

则 $F^*=QF$，$A^*F^*=AF$

① 因子载荷矩阵有无穷多个解

② 首先得到一个载荷矩阵估计的初值，再进行旋转以得到更好的解释

模型估计方法

1. 主成分法

令 $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ 表示 Σ 的特征值，则 $\hat{\Sigma}$ 的特征值分解

$$\hat{\Sigma} = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_p v_p v_p^T = V \Lambda V^T$$

$$\hat{\Sigma} = \frac{1}{N-1} \sum_i (X_i - \bar{X})(X_i - \bar{X})^T$$

（样本协方差）

载荷矩阵 $\hat{A}$：第 i 列为 $\sqrt{\lambda_i} v_i$，$A = (\sqrt{\lambda_1} v_1, \sqrt{\lambda_2} v_2, \cdots, \sqrt{\lambda_q} v_q)$

$$\tilde{A}\tilde{A}^T = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_q v_q v_q^T \qquad \Sigma = AA^T + \Phi$$

对角矩阵

$$\hat{\sigma}_k^2 = \hat{\Sigma}_{kk} - \sum_{i=1}^q \tilde{a}_{ki}^2$$

2. 极大似然估计法

假定 $F_1, F_2, \cdots, F_q$ 服从多元正态分布 $F \sim N(\mu, \Sigma)$

由于载荷矩阵的不唯一性，附上条件 $A^T \Phi^{-1} A$ 为对角矩阵

在此约束下求解 A 和 Φ 的极大似然估计

因子旋转

得到载荷矩阵的初值估计了之后，可通过因子旋转使其解释性提高

①对于任意因子，只有少数变量在该因子上的载荷绝对值较大，其他 ≈ 0.

②对于任意输入变量，只在少数因子上载荷绝对值较大，其它 ≈ 0

③任意两个因子的载荷呈现不同模式

1. 正交旋转：采用正交矩阵对因子进行旋转，保持了因子之间的正交性

2. 斜交旋转：采用非正交矩阵对因子进行旋转，可以更好的简化载荷矩阵
   但旋转后因子存在相关性

最大方差旋转 (variance rotation)

应用最广泛的因子旋转法

(1) 它是一种正交旋转

(2) $\sum_{k=1}^{p} \sum_{i=1}^{q} \left(a_{ki}^2 - \frac{1}{pq} \sum_{k=1}^{p} \sum_{i=1}^{q} a_{ki}^2\right)^2$ 使载荷平方方差最大化

因子数目 q 的选择

① Kaiser 准则 $\sum_{k=1}^p a_{ki}^2 / \sum_{k=1}^p \sigma_{kk}^2 > \frac{1}{p}$

②崖底碎石图 (scree plot)：选择拐点前一点

③如果载荷矩阵由最大似然估计而得，可以采用假设检验

因子得分

对公共因子 $F=(F_1,F_2,\cdots,F_q)$ 的估计值为因子得分

$X_1,X_2,\cdots,X_N,\ X_i\in R^p,$

采用最小二乘法得到 $\hat{F}_i$

假设收集到样本矩阵 $X\in R^{N\times p}$ $F=(F_1,F_2,\cdots,F_N)^T\in R^{N\times q}$

$$X^T=AF^T+E^T$$

最小二乘目标 $\sum_{i=1}^{N}\|X_i-AF_i\|^2=\|X^T-AF^T\|_F^2$ $\|M\|_F=\sqrt{tr(M^TM)}$

$$\hat{F}_i=(A^TA)^{-1}(A^TX_i)$$ （第i个因子得分）

$$\hat{F}^T=(A^TA)^{-1}(A^TX^T)$$

与主成分分析的区别：

①因子分析有"模型"

②因子分析解释公共变异性，主成分分析解释总变异。