清華大學
Tsinghua University

第十六章
主成分分析

主成分分析

- 主成分分析（principal component analysis, PCA）是一种常用的无监督学习方法
- 这一方法利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。
- 主成分的个数通常小于原始变量 的个数，所以主成分分析属于降维方法。
- 主成分分析主要用于发现数据中的基本结构， 即数据中变量之间的关系。

基本想法

- 主成分分析中，首先对给定数据进行规范化，使得数据每一变量的平均值为0，方差为1。
- 之后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。
- 新变量是可能的正交变换中变量的方差的和（信息保存）最大的，方差表示在新变量上信息的大小。
- 可以用主分成近似地表示原始数据，发现数据的基本结构
- 也可以把数据由少数主成分表示，对数据降维

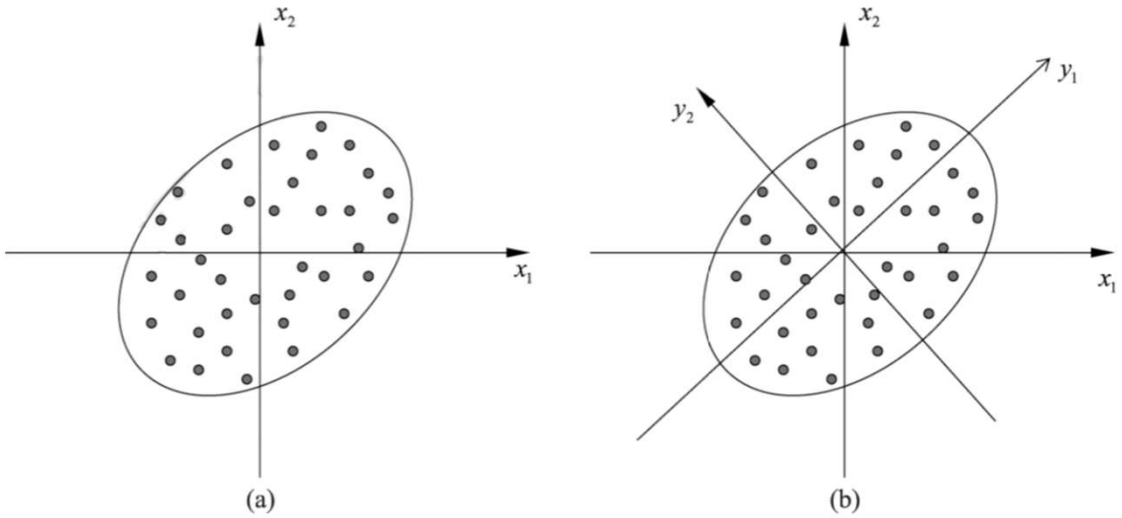
基本想法

- 数据集合中的样本由实数空间（正交坐标系） 中的点表示，空间的一个坐标轴表示一个变量，规范化处理后得到的数据分布在原点附近。
- 对原坐标系中的数据进行主成分分析等价于进行坐标系旋转变换，将数据投影到新坐标系的坐标轴上
- 新坐标系的第一坐标轴、第二坐标轴等分别表示第一主成分、 第二主成分等
- 数据在每一轴上的坐标值的平方表示相应变量的方差
- 这个坐标系是在所有可能的新的坐标系中，坐标轴上的方差的和最大的

例

- 数据由线性相关的两个变量x_1和x_2表示

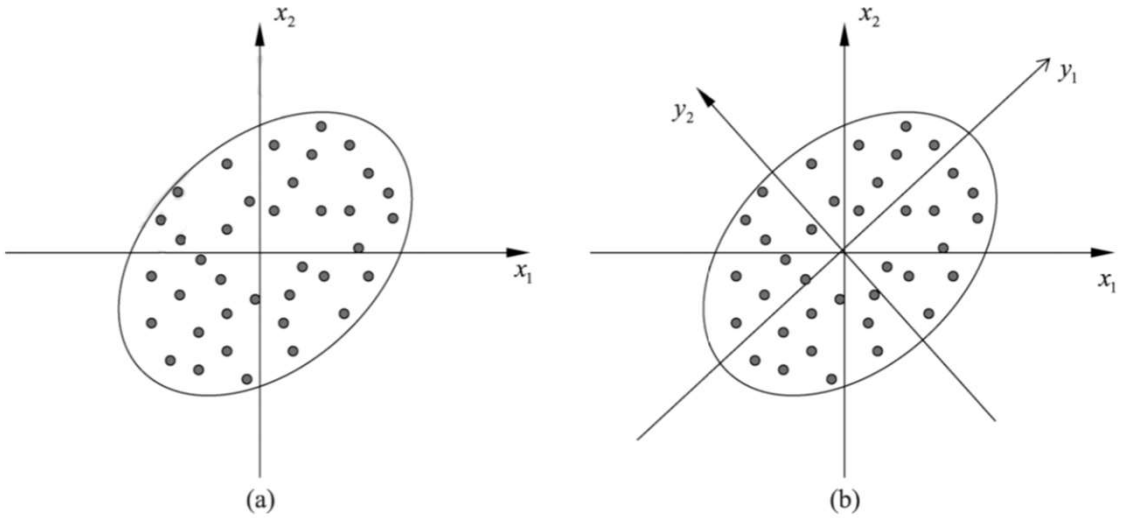
- 主成分分析对数据进行
  正交变换，对原坐标系进行旋转变换，并将数据在新坐标系表示



例

- 主成分分析选择方差最大的方向（第一主成分）作为新坐标系的第一坐标轴，即y_1轴

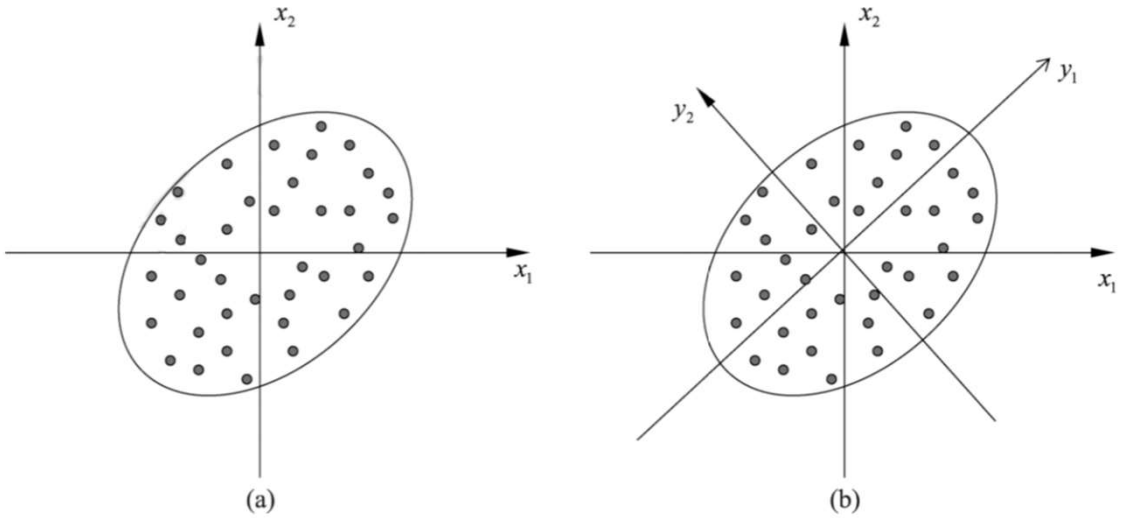
- 之后选择与第一坐标轴正交，且方差次之的方向（第二主成分）作为新坐标系的第二坐标轴，即y_2轴



例

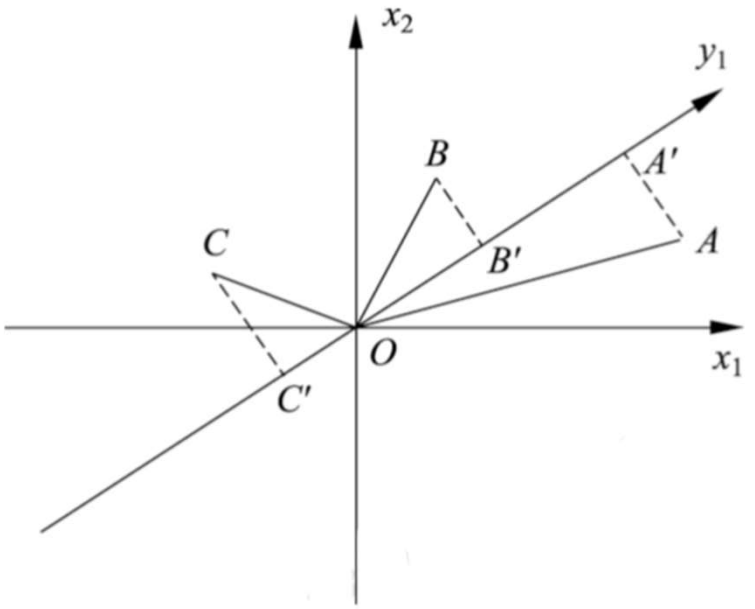
- 在新坐标系里，数据中的变量y_1和y_2是线性无关的，当知道其中一个变量y_1的取值时，对另一个变量y_2的预测是完全随机的，反之亦然

- 如果主成分分析只取第一主成分，即新坐标系的y_1轴，那么等价于将数据投影在椭圆长轴上，用这个主轴表示数据，将二维空间的数据压缩到一维空间中。



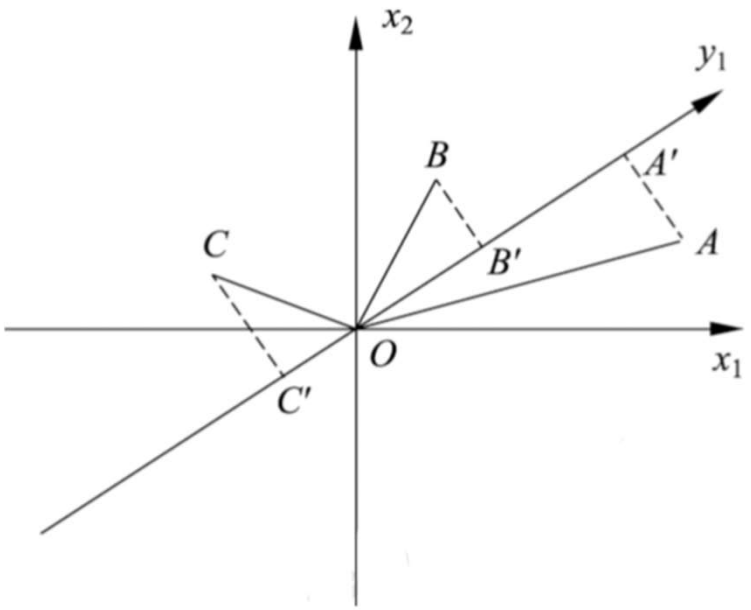
例

- 假设有两个变量x_1和x_2，三个样本点A、B、C，样本分布在由x_1和x_2轴组成的坐标系中
- 对坐标系进行旋转变换，得到新的坐标轴y_1，表示新的变量y_1
- 样本点A、B、C在y_1轴上投影，得到y_1轴的坐标值A'、B'、 C'



例

- 坐标值的平方和$OA'^2 + OB'^2 + OC'^2$表示样本在变量y_1上的方差和
- 主成分分析旨在选取正交变换中方差最大的变量，作为第一主成分，也就是旋转变换中坐标值的平方和最大的轴
- $OA'^2 + OB'^2 + OC'^2$最大等价于样本点到y_1轴的距离的平方和$AA'^2 + BB'^2 + CC'^2$最小
- 主成分分析在旋转变换中选取离样本点的距离平方和最小的轴，作为第一主成分。第二主成分等的选取，在保证与已选坐标轴正交的条件下，类似地进行



主成分分析

- 在数据总体（population)上进行的主成分分析称为总体主成分分析
- 在有限样本上进行的主成分分析称为样本主成分分析
- 总体主成分分析是样本主成分分析的基础

定义和导出

- 假设 $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 是m维随机变量，其均值向量是 $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = E(\boldsymbol{x}) = (\mu_1, \mu_2, \cdots, \mu_m)^{\mathrm{T}}$$

- 协方差矩阵是 Σ

$$\Sigma = \mathrm{cov}(\boldsymbol{x}, \boldsymbol{x}) = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}]$$

- 考虑由m维随机变量x到m维随机变量 $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$ 的线性变换

$$y_i = \alpha_i^{\mathrm{T}} \boldsymbol{x} = \alpha_{1i} x_1 + \alpha_{2i} x_2 + \cdots + \alpha_{mi} x_m$$

- 其中 $\alpha_i^{\mathrm{T}} = (\alpha_{1i}, \alpha_{2i}, \cdots, \alpha_{mi}),\ i = 1, 2, \cdots, m.$

定义和导出

- 由随机变量性质可知

$$E(y_i) = \alpha_i^{\mathrm{T}}\mu, \quad i = 1, 2, \cdots, m$$

$$\mathrm{var}(y_i) = \alpha_i^{\mathrm{T}}\Sigma\alpha_i, \quad i = 1, 2, \cdots, m$$

$$\mathrm{cov}(y_i, y_j) = \alpha_i^{\mathrm{T}}\Sigma\alpha_j, \quad i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, m$$

定义和导出

定义 16.1（总体主成分） 给定一个如式 (16.1) 所示的线性变换，如果它们满足下列条件:

（1）系数向量 α_i^{T} 是单位向量，即 $\alpha_i^{\mathrm{T}}\alpha_i = 1$，$i = 1, 2, \cdots, m$;

（2）变量 y_i 与 y_j 互不相关，即 $\mathrm{cov}(y_i, y_j) = 0 (i \neq j)$;

（3）变量 y_1 是 $\boldsymbol{x}$ 的所有线性变换中方差最大的；y_2 是与 y_1 不相关的 $\boldsymbol{x}$ 的所有线性变换中方差最大的；一般地，y_i 是与 $y_1, y_2, \cdots, y_{i-1}$ $(i = 1, 2, \cdots, m)$ 都不相关的 $\boldsymbol{x}$ 的所有线性变换中方差最大的；这时分别称 $y_1, y_2, \cdots, y_m$ 为 $\boldsymbol{x}$ 的第一主成分、第二主成分、$\cdots$、第 m 主成分。

定义和导出

- 定义中的条件（1）表明线性变换是正交变换，$\alpha_1, \alpha_2, \cdots, \alpha_m$ 是其一组标准正交基

$$\alpha_i^{\mathrm{T}} \alpha_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- 条件（2）（3）给出了一个求主成分的方法：

- 第一步，在x的所有线性变换 $\alpha_1^{\mathrm{T}} \boldsymbol{x} = \sum_{i=1}^{m} \alpha_{i1} x_i$ 中，在 $\alpha_1^{\mathrm{T}} \alpha_1 = 1$ 条件下，求方差最大的，得到x的第一主成分

定义和导出

- 第二步，在与 $\alpha_1^{\mathrm{T}}\boldsymbol{x}$ 不相关的x的所有线性变换 $\alpha_2^{\mathrm{T}}\boldsymbol{x}=\sum_{i=1}^{m}\alpha_{i2}x_i$ 中，在 $\alpha_2^{\mathrm{T}}\alpha_2=1$ 条件下，求方差最大的，得到x的第二主成分

- 第k步，在与 $\alpha_1^{\mathrm{T}}\boldsymbol{x},\alpha_2^{\mathrm{T}}\boldsymbol{x},\cdots,\alpha_{k-1}^{\mathrm{T}}\boldsymbol{x}$ 不相关的x的所有线性变换 $\alpha_k^{\mathrm{T}}\boldsymbol{x}=\sum_{i=1}^{m}\alpha_{ik}x_i$ 中，在 $\alpha_k^{\mathrm{T}}\alpha_k=1$ 条件下，求方差最大的，得到x的第k主成分

主要性质

定理 16.1 设 $\boldsymbol{x}$ 是 m 维随机变量，Σ 是 $\boldsymbol{x}$ 的协方差矩阵，Σ 的特征值分别是 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m \geqslant 0$，特征值对应的单位特征向量分别是 $\alpha_1, \alpha_2, \cdots, \alpha_m$，则 $\boldsymbol{x}$ 的第 k 主成分是

$$y_k = \alpha_k^{\mathrm{T}} \boldsymbol{x} = \alpha_{1k} x_1 + \alpha_{2k} x_2 + \cdots + \alpha_{mk} x_m, \quad k = 1, 2, \cdots, m \tag{16.5}$$

$\boldsymbol{x}$ 的第 k 主成分的方差是

$$\mathrm{var}(y_k) = \alpha_k^{\mathrm{T}} \Sigma \alpha_k = \lambda_k, \quad k = 1, 2, \cdots, m \tag{16.6}$$

即协方差矩阵 Σ 的第 k 个特征值。

主要性质

- 证明：采用拉格朗日乘子法求出主成分

- 首先求x的第一主成分，$y_1 = \alpha_1^{\mathrm{T}} \boldsymbol{x}$，即求系数向量 α_1。由定义16.1知，第一主成分的 α_1 是在 $\alpha_1^{\mathrm{T}}\alpha_1 = 1$ 条件下，x的所有线性变换中使方差 $\mathrm{var}(\alpha_1^{\mathrm{T}}\boldsymbol{x}) = \alpha_1^{\mathrm{T}}\Sigma\alpha_1$ 达到最大的

- 求第一主成分就是求解约束最优化问题：

$$\max_{\alpha_1} \quad \alpha_1^{\mathrm{T}}\Sigma\alpha_1$$

$$\text{s.t.} \quad \alpha_1^{\mathrm{T}}\alpha_1 = 1$$

主要性质

- 定义拉格朗日函数 $\alpha_1^{\mathrm{T}}\Sigma\alpha_1 - \lambda(\alpha_1^{\mathrm{T}}\alpha_1 - 1)$
- 其中 λ 是拉格朗日乘子。将拉格朗日函数对 α_1 求导，并令其为0，得 $\Sigma\alpha_1 - \lambda\alpha_1 = 0$
- 因此，λ 是 Σ 的特征值，α_1 是对应的单位特征向量。于是，目标函数
$$\alpha_1^{\mathrm{T}}\Sigma\alpha_1 = \alpha_1^{\mathrm{T}}\lambda\alpha_1 = \lambda\alpha_1^{\mathrm{T}}\alpha_1 = \lambda$$
- 假设 α_1 是 Σ 的最大特征值 λ_1 对应的单位特征向量，显然 α_1 与 λ_1 是最优化问题的解
- 所以，$\alpha_1^{\mathrm{T}}x$ 构成第一主成分，其方差等于协方差矩阵的最大特征值

$$\mathrm{var}(\alpha_1^{\mathrm{T}}\boldsymbol{x}) = \alpha_1^{\mathrm{T}}\Sigma\alpha_1 = \lambda_1$$

主要性质

- 接着求x的第二主成分 $y_2 = \alpha_2^{\mathrm{T}} \boldsymbol{x}$ 。第二主成分的 α_2 是在 $\alpha_2^{\mathrm{T}} \alpha_2 = 1$ ，且 $\alpha_2^{\mathrm{T}} \boldsymbol{x}$ 与 $\alpha_1^{\mathrm{T}} \boldsymbol{x}$ 不相关的条件下，x的所有线性变换中使方差 $\mathrm{var}(\alpha_2^{\mathrm{T}} \boldsymbol{x}) = \alpha_2^{\mathrm{T}} \Sigma \alpha_2$ 达到最大的
- 求第二主成分需要求解约束最优化问题

$$\begin{aligned} \max_{\alpha_2} \quad & \alpha_2^{\mathrm{T}} \Sigma \alpha_2 \\ \text{s.t.} \quad & \alpha_1^{\mathrm{T}} \Sigma \alpha_2 = 0, \qquad \alpha_2^{\mathrm{T}} \Sigma \alpha_1 = 0 \\ & \alpha_2^{\mathrm{T}} \alpha_2 = 1 \end{aligned}$$

- 注意到 $\alpha_1^{\mathrm{T}} \Sigma \alpha_2 = \alpha_2^{\mathrm{T}} \Sigma \alpha_1 = \alpha_2^{\mathrm{T}} \lambda_1 \alpha_1 = \lambda_1 \alpha_2^{\mathrm{T}} \alpha_1 = \lambda_1 \alpha_1^{\mathrm{T}} \alpha_2$ 以及 $\alpha_1^{\mathrm{T}} \alpha_2 = 0, \qquad \alpha_2^{\mathrm{T}} \alpha_1 = 0$

主要性质

- 定义拉格朗日函数 $\alpha_2^{\mathrm{T}}\Sigma\alpha_2 - \lambda(\alpha_2^{\mathrm{T}}\alpha_2 - 1) - \phi\alpha_2^{\mathrm{T}}\alpha_1$
- 其中 λ, ϕ 是拉格朗日乘子。对 α_2 求导，并令其为0，得

$$2\Sigma\alpha_2 - 2\lambda\alpha_2 - \phi\alpha_1 = 0 \quad 16.10$$

- 将方程左乘以 α_1^{T} 有 $2\alpha_1^{\mathrm{T}}\Sigma\alpha_2 - 2\lambda\alpha_1^{\mathrm{T}}\alpha_2 - \phi\alpha_1^{\mathrm{T}}\alpha_1 = 0$
- 此式前两项为0，且 $\alpha_1^{\mathrm{T}}\alpha_1 = 1,$ ，导出 $\phi = 0$ ，因此式(16.10)成为

$$\Sigma\alpha_2 - \lambda\alpha_2 = 0$$

- 由此，λ 是 Σ 的特征值，α_2 是对应的单位特征向量。于是，目标函数

$$\alpha_2^{\mathrm{T}}\Sigma\alpha_2 = \alpha_2^{\mathrm{T}}\lambda\alpha_2 = \lambda\alpha_2^{\mathrm{T}}\alpha_2 = \lambda$$

主要性质

- 假设 α_2 是 Σ 的第二大特征值 λ_2 对应的单位特征向量，显然 α_2 与 λ_2 是以上最优化问题的解
- 于是 $\alpha_2^{\mathrm{T}}\boldsymbol{x}$ 构成第二主成分，其方差等于协方差矩阵的第二大特征值 $\mathrm{var}(\alpha_2^{\mathrm{T}}\boldsymbol{x}) = \alpha_2^{\mathrm{T}}\Sigma\alpha_2 = \lambda_2$
- 一般地，x的第k主成分是 $\alpha_k^{\mathrm{T}}\boldsymbol{x}$，并且 $\mathrm{var}(\alpha_k^{\mathrm{T}}\boldsymbol{x}) = \lambda_k$，这里 λ_k 是 Σ 的第k个特征值并且 α_k 是对应的单位特征向量。

主要性质

- 按照上述方法求得第一、第二、直到第m主成分，其系数向量 $\alpha_1, \alpha_2, \cdots, \alpha_m$ 分别是 Σ 的第一个、第二个、直到第m个单位特征向量， $\lambda_1, \lambda_2, \cdots, \lambda_m$ 分别是对应的特征值。

- 第k主成分的方差等于 Σ 的第k个特征值，

$$\mathrm{var}(\alpha_k^{\mathrm{T}}\boldsymbol{x}) = \alpha_k^{\mathrm{T}}\Sigma\alpha_k = \lambda_k, \quad k = 1, 2, \cdots, m$$

主要性质

推论 16.1 m 维随机变量 $\boldsymbol{y}=(y_1,y_2,\cdots,y_m)^{\mathrm{T}}$ 的分量依次是 $\boldsymbol{x}$ 的第一主成分到第 m 主成分的充要条件是:

(1) $\boldsymbol{y}=A^{\mathrm{T}}\boldsymbol{x}$, A 为正交矩阵

$$A=\begin{bmatrix}\alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m}\\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m}\\ \vdots & \vdots & & \vdots\\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mm}\end{bmatrix}$$

(2) $\boldsymbol{y}$ 的协方差矩阵为对角矩阵

$$\operatorname{cov}(\boldsymbol{y})=\operatorname{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m)$$

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$$

其中 λ_k 是 Σ 的第 k 个特征值, α_k 是对应的单位特征向量, $k=1,2,\cdots,m$。

主要性质

- 以上证明中，λ_k 是 Σ 第k个特征值，α_k 是对应的特征向量，即

$$\Sigma\alpha_k = \lambda_k\alpha_k, \quad k = 1, 2, \cdots, m$$

- 用矩阵表示为

$$\Sigma A = A\Lambda$$

- 这里 $A = [\alpha_{ij}]_{m\times m}$，$\Lambda$ 是对角矩阵，其第k个对角元素是 λ_k

- 因为A是正交矩阵，即 $A^TA=AA^T=I$，由 $\Sigma A = A\Lambda$ 得

$$A^{\mathrm{T}}\Sigma A = \Lambda \qquad \Sigma = A\Lambda A^{\mathrm{T}}$$

总体主成分的性质

- （1）总体主成分y的协方差矩阵是对角矩阵

$$\mathrm{cov}(\boldsymbol{y}) = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$$

- （2）总体主成分y的方差之和等于随机变量x的方差之和，即

$$\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} \sigma_{ii}$$

- 其中 σ_{ii} 是随机变量 x_i 的方差，即协方差矩阵 Σ 的对角元素。事实上，利用 $\Sigma = A\Lambda A^{\mathrm{T}}$ 及矩阵的迹(trace)的性质，可知

$$\sum_{i=1}^{m} \mathrm{var}(x_i) = \mathrm{tr}(\Sigma^{\mathrm{T}}) = \mathrm{tr}(A\Lambda A^{\mathrm{T}}) = \mathrm{tr}(A^{\mathrm{T}}\Lambda A)$$

$$= \mathrm{tr}(\Lambda) = \sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} \mathrm{var}(y_i)$$

总体主成分的性质

- （3）第k个主成分 y_k 与变量 x_i 的相关系数 $\rho(y_k, x_i)$ 称为因子负荷量（factor loading），它表示第k个主成分 y_k 与变量 x_i 的相关关系。计算公式是

$$\rho(y_k, x_i) = \frac{\sqrt{\lambda_k}\alpha_{ik}}{\sqrt{\sigma_{ii}}}, \quad k, i = 1, 2, \cdots, m \qquad 16.20$$

- 因为

$$\rho(y_k, x_i) = \frac{\mathrm{cov}(y_k, x_i)}{\sqrt{\mathrm{var}(y_k)\mathrm{var}(x_i)}} = \frac{\mathrm{cov}(\alpha_k^{\mathrm{T}}\boldsymbol{x}, e_i^{\mathrm{T}}\boldsymbol{x})}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}}$$

- 其中e_i为基本单位向量，其第i个分量为1，其余为0。再由协方差的性质

$$\mathrm{cov}(\alpha_k^{\mathrm{T}}\boldsymbol{x}, e_i^{\mathrm{T}}\boldsymbol{x}) = \alpha_k^{\mathrm{T}}\Sigma e_i = e_i^{\mathrm{T}}\Sigma\alpha_k = \lambda_k e_i^{\mathrm{T}}\alpha_k = \lambda_k\alpha_{ik}$$

得到式(16.20)

总体主成分的性质

- （4）第k个主成分 y_k 与m个变量的因子负荷量满足

$$\sum_{i=1}^{m}\sigma_{ii}\rho^2(y_k,x_i)=\lambda_k$$

- 由式(16.20)有 $\sum_{i=1}^{m}\sigma_{ii}\rho^2(y_k,x_i)=\sum_{i=1}^{m}\lambda_k\alpha_{ik}^2=\lambda_k\alpha_k^{\mathrm{T}}\alpha_k=\lambda_k$

总体主成分的性质

- （5）m个主成分与第i个变量 x_i 的因子负荷量满足

$$\sum_{k=1}^{m} \rho^2(y_k, x_i) = 1 \qquad 16.22$$

- 由于 $y_1, y_2, \cdots, y_m$ 互不相关

  故 $$\rho^2(x_i, (y_1, y_2, \cdots, y_m)) = \sum_{k=1}^{m} \rho^2(y_k, x_i)$$

- 又因 x_i 可以表为 $y_1, y_2, \cdots, y_m$ 的线性组合，所以 x_i 与 $y_1, y_2, \cdots, y_m$ 的相关系数的平方为1，即 $\rho^2(x_i, (y_1, y_2, \cdots, y_m)) = 1$
- 故得式(16.22)

主成分的个数

- 主成分分析的主要目的是降维，所以一般选择k (k << m）个主成分（线性无关变量）来代替m个原有变量（线性相关变量），使问题得以简化，并能保留原有变量的大部分信息。

定理 16.2 对任意正整数 q，$1 \leqslant q \leqslant m$，考虑正交线性变换

$$\boldsymbol{y} = B^{\mathrm{T}}\boldsymbol{x} \tag{16.23}$$

其中 $\boldsymbol{y}$ 是 q 维向量，B^{T} 是 $q \times m$ 矩阵，令 $\boldsymbol{y}$ 的协方差矩阵为

$$\Sigma_{\boldsymbol{y}} = B^{\mathrm{T}}\Sigma B \tag{16.24}$$

则 $\Sigma_{\boldsymbol{y}}$ 的迹 $\mathrm{tr}(\Sigma_{\boldsymbol{y}})$ 在 $B = A_q$ 时取得最大值，其中矩阵 A_q 由正交矩阵 A 的前 q 列组成。

主成分的个数

- 证明:

- 令 β_k 是B的第k列，由于正交矩阵A的列构成m维空间的基，所以 β_k 可以由A的列表示，即

$$\beta_k = \sum_{j=1}^{m} c_{jk}\alpha_j, \quad k = 1, 2, \cdots, q$$

- 等价地

$$B = AC$$

- 其中C是 m x q 矩阵，其第j行第k列元素为C_{jk}

主成分的个数

- 首先

$$B^{\mathrm{T}}\Sigma B = C^{\mathrm{T}}A^{\mathrm{T}}\Sigma AC = C^{\mathrm{T}}\Lambda C = \sum_{j=1}^{m}\lambda_j c_j c_j^{\mathrm{T}}$$

- 其中 c_j^{T} 是C的第j行。因此

$$\begin{aligned}\mathrm{tr}(B^{\mathrm{T}}\Sigma B) &= \sum_{j=1}^{m}\lambda_j \mathrm{tr}(c_j c_j^{\mathrm{T}}) \\ &= \sum_{j=1}^{m}\lambda_j \mathrm{tr}(c_j^{\mathrm{T}} c_j) \\ &= \sum_{j=1}^{m}\lambda_j c_j^{\mathrm{T}} c_j \\ &= \sum_{j=1}^{m}\sum_{k=1}^{q}\lambda_j c_{jk}^2\end{aligned}$$

主成分的个数

- 其次，由 $B = AC$ 及A的正交性知 $C = A^{\mathrm{T}}B$
- 由于A是正交的，B的列是正交的，所以

$$C^{\mathrm{T}}C = B^{\mathrm{T}}AA^{\mathrm{T}}B = B^{\mathrm{T}}B = I_q$$

- 即C的列也是正交的。于是 $\mathrm{tr}(C^{\mathrm{T}}C) = \mathrm{tr}(I_q)$

$$\sum_{j=1}^{m}\sum_{k=1}^{q} c_{jk}^2 = q$$

- 这样，矩阵C可以认为是某个m阶正交矩阵D的前q列

主成分的个数

- 正交矩阵D的行也正交，所以满足 $d_j^{\mathrm{T}} d_j = 1, \quad j = 1, 2, \cdots, m$
- 其中 d_j^{T} 是D的第j行。由于矩阵D的行包括矩阵C的行的前q个元素，所以 $c_j^{\mathrm{T}} c_j \leqslant 1, \quad j = 1, 2, \cdots, m$
- 即 $\sum_{k=1}^{q} c_{jk}^2 \leqslant 1, \quad j = 1, 2, \cdots, m$
- 因为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_q \geqslant \cdots \geqslant \lambda_m$，能找到 c_{jk} 使得 $\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, \cdots, q \\ 0, & j = q+1, \cdots, m \end{cases}$ 时，$\sum_{j=1}^{m} \left(\sum_{k=1}^{q} c_{jk}^2 \right) \lambda_j$ 最大

主成分的个数

- 而当 $B = A_q$ 时，有 $c_{jk} = \begin{cases} 1, & 1 \leqslant j = k \leqslant q \\ 0, & \text{其他} \end{cases}$ 满足 $\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, \cdots, q \\ 0, & j = q+1, \cdots, m \end{cases}$
- 所以，当 $B = A_q$ 时，$\mathrm{tr}(\Sigma_{\boldsymbol{y}})$ 达到最大值
- 定理16.2表明，当x的线性变换y在 $B = A_q$，其协方差矩阵 $\Sigma_{\boldsymbol{y}}$ 的迹 $\mathrm{tr}(\Sigma_{\boldsymbol{y}})$ 取得最大值
- 这就是说，当取A的前q列取x的前q个主成分时，能够最大限度地保留原有变量方差的信息。

主成分的个数

定理 16.3 考虑正交变换

$$\boldsymbol{y} = B^{\mathrm{T}}\boldsymbol{x}$$

这里 B^{T} 是 $p \times m$ 矩阵，A 和 $\Sigma_{\boldsymbol{y}}$ 的定义与定理 16.2 相同，则 $\mathrm{tr}(\Sigma_{\boldsymbol{y}})$ 在 $B = A_p$ 时取得最小值，其中矩阵 A_p 由 A 的后 p 列组成。

- 当舍弃A的后p列，即舍弃变量x的后p个主成分时，原有变量的方差的信息损失最少。

主成分的个数

定义 16.2 第 k 主成分 y_k 的方差贡献率定义为 y_k 的方差与所有方差之和的比，记作 η_k

$$\eta_k = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i} \tag{16.30}$$

k 个主成分 $y_1, y_2, \cdots, y_k$ 的累计方差贡献率定义为 k 个方差之和与所有方差之和的比

$$\sum_{i=1}^{k} \eta_i = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{16.31}$$

主成分的个数

- 通常取k使得累计方差贡献率达到规定的百分比以上
- 累计方差贡献率反映了主成分保留信息的比例，但它不能反映对某个原有变量x_j保留信息的比例
- 通常利用k个主成分 $y_1, y_2, \cdots, y_k$ 对原有变量x_j的贡献率

主成分的个数

定义 16.3 k 个主成分 $y_1, y_2, \cdots, y_k$ 对原有变量 x_i 的贡献率定义为 x_i 与 $(y_1, y_2, \cdots, y_k)$ 的相关系数的平方，记作 ν_i

$$\nu_i = \rho^2(x_i, (y_1, y_2, \cdots, y_k))$$

计算公式如下:

$$\nu_i = \rho^2(x_i, (y_1, y_2, \cdots, y_k)) = \sum_{j=1}^{k} \rho^2(x_i, y_j) = \sum_{j=1}^{k} \frac{\lambda_j \alpha_{ij}^2}{\sigma_{ii}} \tag{16.32}$$

规范化变量的总体主成分

- 在实际问题中，不同变量可能有不同的量纲，直接求主成分有时会产生不合理的结果。

- 为了消除这个影响，常常对各个随机变量实施规范化，使其均值为0，方差为1.

规范化变量的总体主成分

- 设 $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 为m维随机变量，x_i为第i个随机变量，令

$$x_i^* = \frac{x_i - E(x_i)}{\sqrt{\mathrm{var}(x_i)}}, \quad i = 1, 2, \cdots, m$$

- 其中，$E(x_i)$，$\mathrm{var}(x_i)$ 分别是随机变量x_i的均值和方差，这时 x_i^* 就是x_i的规范化随机变量

- 规范化随机变量的协方差矩阵就是相关矩阵R

规范化变量的总体主成分

- 规范化随机变量的总体主成分有以下性质:

- （1） 规范化变量主成分的协方差矩阵是

$$\Lambda^* = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \cdots, \lambda_m^*)$$

- （2） 协方差矩阵的特征值之和为m

$$\sum_{k=1}^{m} \lambda_k^* = m$$

规范化变量的总体主成分

- （3） 规范化随机变量 x_i^* 与主成分 y_k^* 的相关系数（因子负荷量）为

$$\rho(y_k^*, x_i^*) = \sqrt{\lambda_k^*} e_{ik}^*, \quad k, i = 1, 2, \cdots, m$$

其中 $e_k^* = (e_{1k}^*, e_{2k}^*, \cdots, e_{mk}^*)^{\mathrm{T}}$ 为矩阵R对应于特征值 λ_k^* 的单位特征向量

- （4）所有规范化随机变量 x_i^* 与主成分 y_k^* 的相关系数的平方和等于 λ_k^*

$$\sum_{i=1}^{m} \rho^2(y_k^*, x_i^*) = \sum_{i=1}^{m} \lambda_k^* e_{ik}^{*2} = \lambda_k^*, \quad k = 1, 2, \cdots, m$$

- （5）规范化随机变量 x_i^* 与所有主成分 y_k^* 的相关系数的平方和等于1

$$\sum_{k=1}^{m} \rho^2(y_k^*, x_i^*) = \sum_{k=1}^{m} \lambda_k^* e_{ik}^{*2} = 1, \quad i = 1, 2, \cdots, m$$

样本主成分分析

- 总体主成分分析，是定义在样本总体上的。
- 在实际问题中，需要在观测数据上进行主成分分析，这就是样本主成分分析。
- 样本主成分也和总体主成分具有相同的性质。

样本主成分的定义和性质

- 假设对m维随机变量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$ 进行n次独立观测
- $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 表示观测样本
- $\boldsymbol{x}_j=(x_{1j},x_{2j},\cdots,x_{mj})^{\mathrm{T}}$ 表示第j个观测样本
- x_{ij} 表示第j个观测样本的第i个变量，j=1,2,⋯ ,n
- 观测数据用样本矩阵X表示，记作

$$X=[\ \boldsymbol{x}_1\quad \boldsymbol{x}_2\quad \cdots\quad \boldsymbol{x}_n\ ]=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

样本主成分的定义和性质

- 给定样本矩阵X，可以估计样本均值，以及样本协方差。样本均值向量 $\bar{x}$ 为

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$$

- 样本协方差矩阵S为 $S = [s_{ij}]_{m\times m}$

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j), \quad i,j = 1,2,\cdots,m$$

- 其中，$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik}$ 为第i个变量的样本均值

样本主成分的定义和性质

- 样本相关矩阵R为

$$R = [r_{ij}]_{m\times m}, \quad r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}, \quad i,j = 1,2,\cdots,m$$

- 定义m维向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 到m维向量 $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$ 的线性变换

$$\boldsymbol{y} = A^{\mathrm{T}}\boldsymbol{x}$$

- 其中

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

$$a_i = (a_{1i}, a_{2i}, \cdots, a_{mi})^{\mathrm{T}}, \quad i = 1,2,\cdots,m$$

样本主成分的定义和性质

- 考虑 $\boldsymbol{y} = A^{\mathrm{T}}\boldsymbol{x}$ 的任意一个线性变换

$$\boldsymbol{y}_i = a_i^{\mathrm{T}}\boldsymbol{x} = a_{1i}\boldsymbol{x}_1 + a_{2i}\boldsymbol{x}_2 + \cdots + a_{mi}\boldsymbol{x}_m, \quad i = 1, 2, \cdots, m$$

- 其中y_i是m维向量y的第i个变量，相应于容量为n的样本 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，$y_i$的样本均值 $\bar{y}_i$ 为

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} a_i^{\mathrm{T}}\boldsymbol{x}_j = a_i^{\mathrm{T}}\bar{\boldsymbol{x}}$$

- 其中 $\bar{\boldsymbol{x}}$ 是随机向量x的样本均值 $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$

样本主成分的定义和性质

- y_i的样本方差 $\mathrm{var}(y_i)$ 为

$$\mathrm{var}(y_i) = \frac{1}{n-1}\sum_{j=1}^{n}(a_i^{\mathrm{T}}\boldsymbol{x}_j - a_i^{\mathrm{T}}\bar{\boldsymbol{x}})^2$$

$$= a_i^{\mathrm{T}}\left[\frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^{\mathrm{T}}\right]a_i = a_i^{\mathrm{T}}Sa_i$$

- 对任意两个线性变换 $y_i = \alpha_i^{\mathrm{T}}\boldsymbol{x}, y_k = \alpha_k^{\mathrm{T}}\boldsymbol{x}$，相应于容量为n的样本 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，$y_i$, y_k的样本协方差为

$$\mathrm{cov}(y_i, y_k) = a_i^{\mathrm{T}}Sa_k$$

样本主成分的定义和性质

定义 16.4（样本主成分） 给定样本矩阵 $\boldsymbol{X}$。样本第一主成分 $y_1 = a_1^{\mathrm{T}}\boldsymbol{x}$ 是在 $a_1^{\mathrm{T}}a_1 = 1$ 条件下，使得 $a_1^{\mathrm{T}}\boldsymbol{x}_j\,(j = 1, 2, \cdots, n)$ 的样本方差 $a_1^{\mathrm{T}}Sa_1$ 最大的 $\boldsymbol{x}$ 的线性变换；样本第二主成分 $y_2 = a_2^{\mathrm{T}}\boldsymbol{x}$ 是在 $a_2^{\mathrm{T}}a_2 = 1$ 和 $a_2^{\mathrm{T}}\boldsymbol{x}_j$ 与 $a_1^{\mathrm{T}}\boldsymbol{x}_j\,(j = 1, 2, \cdots, n)$ 的样本协方差 $a_1^{\mathrm{T}}Sa_2 = 0$ 条件下，使得 $a_2^{\mathrm{T}}\boldsymbol{x}_j\,(j = 1, 2, \cdots, n)$ 的样本方差 $a_2^{\mathrm{T}}Sa_2$ 最大的 $\boldsymbol{x}$ 的线性变换；一般地，样本第 i 主成分 $y_i = a_i^{\mathrm{T}}\boldsymbol{x}$ 是在 $a_i^{\mathrm{T}}a_i = 1$ 和 $a_i^{\mathrm{T}}\boldsymbol{x}_j$ 与 $a_k^{\mathrm{T}}\boldsymbol{x}_j\,(k < i,\ j = 1, 2, \cdots, n)$ 的样本协方差 $a_k^{\mathrm{T}}Sa_i = 0$ 条件下，使得 $a_i^{\mathrm{T}}\boldsymbol{x}_j\,(j = 1, 2, \cdots, n)$ 的样本方差 $a_i^{\mathrm{T}}Sa_i$ 最大的 $\boldsymbol{x}$ 的线性变换。

- 样本主成分与总体主成分具有同样的性质
- 总体主成分的定理16.2及定理16.3对样本主成分依然成立

样本主成分的定义和性质

- 在使用样本主成分时，一般假设样本数据是规范化的，即对样本矩阵作如下变换：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_i}{\sqrt{s_{ii}}},\quad i=1,2,\cdots,m;\quad j=1,2,\cdots,n \qquad 16.48$$

- 其中

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij},\quad i=1,2,\cdots,m$$

$$s_{ii}=\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2,\quad i=1,2,\cdots,m$$

样本主成分的定义和性质

- 为了方便，以下将规范化变量 x_{ij}^* 每仍记作 x_{ij}，规范化的样本矩阵仍记作X。这时，样本协方差矩阵S就是样本相关矩阵R

$$R = \frac{1}{n-1} X X^{\mathrm{T}}$$

- 样本协方差矩阵S是总体协方差矩阵 Σ 的无偏估计
- 样本相关矩阵R是总体相关矩阵的无偏估计
- S的特征值和特征向量是 Σ 的特征值和特征向量的极大似然估计。

相关矩阵的特征值分解算法

- 给定样本矩阵X，利用数据的样本协方差矩阵或者样本相关矩阵的特征值分解进行主成分分析。具体步骤如下：

- （1）对观测数据按式(16.48)进行规范化处理，得到规范化数据矩阵，仍以X表示

相关矩阵的特征值分解算法

- （2）依据规范化数据矩阵，计算样本相关矩阵R

$$R = [r_{ij}]_{m\times m} = \frac{1}{n-1}XX^{\mathrm{T}}$$

- 其中

$$r_{ij} = \frac{1}{n-1}\sum_{l=1}^{n} x_{il}x_{lj}, \quad i,j = 1,2,\cdots,m$$

相关矩阵的特征值分解算法

- （3） 求样本相关矩阵R的k个特征值和对应的k个单位特征向量
- 求解R的特征方程 $|R-\lambda I|=0$
- 得R的m个特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$

- 求方差贡献率 $\sum_{i=1}^{k}\eta_i$ 达到预定值的主成分个数k
- 求前k个特征值对应的单位特征向量 $a_i=(a_{1i},a_{2i},\cdots,a_{mi})^{\mathrm{T}}, \quad i=1,2,\cdots,k$

相关矩阵的特征值分解算法

- （4）求k个样本主成分
- 以k个单位特征向量为系数进行线性变换，求出k个样本主成分

$$y_i = a_i^{\mathrm{T}} \boldsymbol{x}, \quad i = 1, 2, \cdots, k$$

- （5）计算k个主成分 y_j 与原变量 x_i 的相关系数 $\rho(x_i, y_j)$，以及k个主成分对原变量 x_i 的贡献率v_i。

相关矩阵的特征值分解算法

- （6）计算n个样本的k个主成分值
- 将规范化样本数据代入k个主成分式 $y_i = a_i^{\mathrm{T}}\boldsymbol{x}, \quad i = 1, 2, \cdots, k$
- 得到n个样本的主成分值

- 第j个 样本 $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$ 的第i主成分值是

$$y_{ij} = (a_{1i}, a_{2i}, \cdots, a_{mi})(x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}} = \sum_{l=1}^{m} a_{li} x_{lj}$$

$$i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, n$$

例

- 假设有n个学生参加四门课程的考试，将学生们的考试成绩看作随机 变量的取值，对考试成绩数据进行标准化处理，得到样本相关矩阵R

课程	语文	外语	数学	物理
语文	1	0.44	0.29	0.33
外语	0.44	1	0.35	0.32
数学	0.29	0.35	1	0.60
物理	0.33	0.32	0.60	1

- 试对数据进行主成分分析

例

- 设变量 x_1, x_2, x_3, x_4 分别表示语文、外语、数学、物理的成绩。 对样本相关矩阵进行特征值分解，得到相关矩阵的特征值，并按大小排序，

$$\lambda_1 = 2.17, \quad \lambda_2 = 0.87, \quad \lambda_3 = 0.57, \quad \lambda_4 = 0.39$$

- 这些特征值就是各主成分的方差贡献率。假设要求主成分的累计方差贡献率大于75%, 那么只需取前两个主成分即可，即k=2，因为

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{4} \lambda_i} = 0.76$$

例

- 求出对应于特征值 λ_1, λ_2 的单位特征向量

项目	x_1	x_2	x_3	x_4	方差贡献率
y_1	0.460	0.476	0.523	0.537	0.543
y_2	0.574	0.486	−0.476	−0.456	0.218

- 由 $y_i = a_i^{\mathrm{T}}\boldsymbol{x}, \quad i = 1, 2, \cdots, k$ 可得第一主成分y_1、第二主成分y_2

$$y_1 = 0.460x_1 + 0.476x_2 + 0.523x_3 + 0.537x_4$$

$$y_2 = 0.574x_1 + 0.486x_2 - 0.476x_3 - 0.456x_4$$

例

- 接下来由特征值和单位特征向量求出第一、第二主成分的因子负荷量，以及第一、 第二主成分对变量 x_i 的贡献率

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	−0.444	−0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

例

- 第一主成分 y_1 对应的因子负荷量 $\rho(y_1, x_i), i = 1, 2, 3, 4,$ 均为正数，表明各门课程成绩提高都可使 y_1 提高
- 也就是说，第一主成分 y_1 反映了学生的整体成绩
- 因子负荷量的数值相近，且 $\rho(y_1, x_4)$ 的数值最大，这 表明物理成绩在整体成绩中占最重要位置

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	−0.444	−0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

例

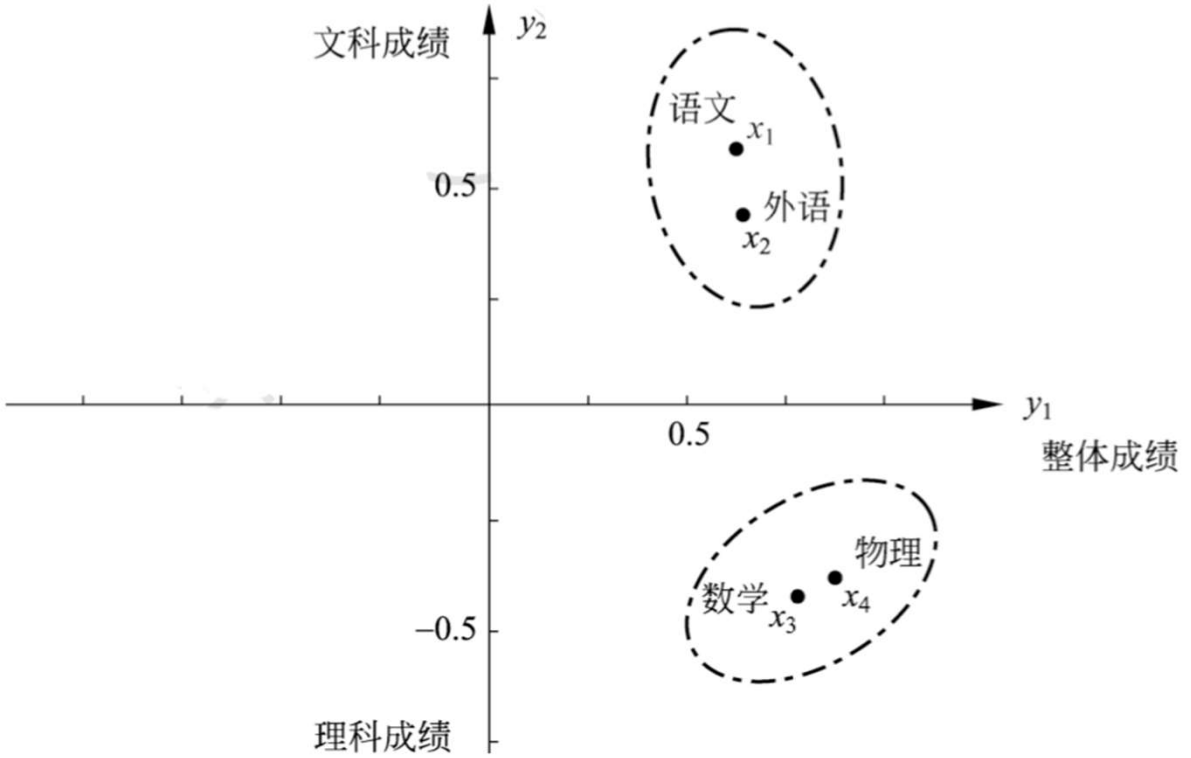
- 第二主成分 y_2 对应的因子负荷量 $\rho(y_2, x_i), i = 1, 2, 3, 4,$ 有正有负
- 正的是语文和外语，负的是数学和物理
- 表明文科成绩提高都可使 y_2 提高，理科成绩提高都可使 y_2 降低
- 也就是说，第二主成分 y_2 反映了学生的文科成绩与理科成绩的关系。

项目	x_1	x_2	x_3	x_4
y_1	0.678	0.701	0.770	0.791
y_2	0.536	0.453	−0.444	−0.425
y_1, y_2 对 x_i 的贡献率	0.747	0.697	0.790	0.806

例

- 将原变量 x_1, x_2, x_3, x_4
  （语文、外语、数学、物理）和
  主成分 y_1, y_2
  （整体成绩、文科对理科成绩）
  的因子负荷量在平面坐标系中
  表示。

- 4个原变量聚成了两类
  因子负荷量相近的语文、外语
  为一类，数学、物理为一类，
  前者反映文科课程成绩，
  后者反映理科课程成绩。



数据矩阵的奇异值分解算法

- 假设有k个主成分，给定样本矩阵x，利用数据矩阵奇异值分解进行主成分分析

- 对于 m x n 实矩阵A，假设其秩为r，0<k<r，则可以将矩阵A进行截断奇异值分解

$$A \approx U_k \Sigma_k V_k^{\mathrm{T}}$$

数据矩阵的奇异值分解算法

- 定义一个新的 n x m 矩阵X’ $X' = \frac{1}{\sqrt{n-1}} X^{\mathrm{T}}$
- X’的每一列均值为0，得

$$X'^{\mathrm{T}} X' = \left(\frac{1}{\sqrt{n-1}} X^{\mathrm{T}}\right)^{\mathrm{T}} \left(\frac{1}{\sqrt{n-1}} X^{\mathrm{T}}\right)$$

$$= \frac{1}{n-1} X X^{\mathbf{T}}$$

- 即$X'^{\mathrm{T}}X'$等于X的协方差矩阵S_X

$$S_X = X'^{\mathbf{T}} X'$$

数据矩阵的奇异值分解算法

- 主成分分析归结于求协方差矩阵S_X的特征值和对应的单位特征向量
- 问题转化为求矩阵 $X'^{T}X'$ 的特征值和对应的单位特征向量

- 假设 X' 的截断奇异值分解为 $X' = U\Sigma V^{T}$，那么V的列向量就是 $S_X = X'^{T}X'$ 的单位特征向量

- 因此，V的列向量就是X的主成分
- 于是，求X主成分可以通过求X'的奇异值分解来实现。

主要成分分析算法

输入：$m \times n$ 样本矩阵 X，其每一行元素的均值为零；

输出：$k \times n$ 样本主成分矩阵 Y。

参数：主成分个数 k

（1）构造新的 $n \times m$ 矩阵

$$X' = \frac{1}{\sqrt{n-1}} X^{\mathrm{T}}$$

X' 每一列的均值为零。

（2）对矩阵 X' 进行截断奇异值分解，得到

$$X' = U \Sigma V^{\mathrm{T}}$$

有 k 个奇异值、奇异向量。矩阵 V 的前 k 列构成 k 个样本主成分。

（3）求 $k \times n$ 样本主成分矩阵

$$Y = V^{\mathrm{T}} X$$