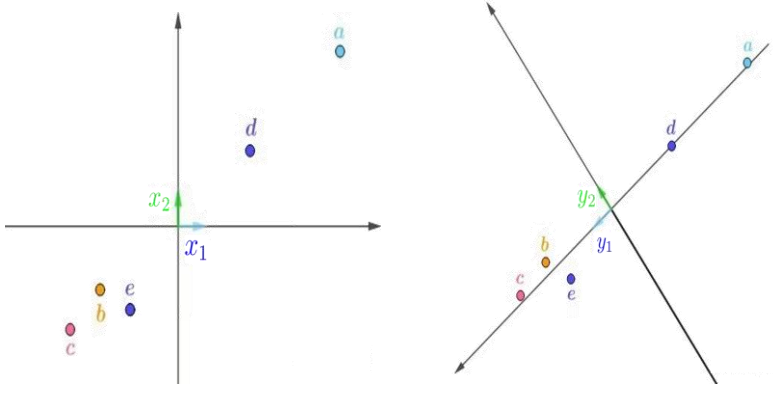
第16章 主成分分析

主成分分析

- 主成分分析(principal component analysis, PCA)
  - 无监督学习方法
- PCA为正交变换
  - 数据由可能存在相关性的变量表示【基(e_1, e_2)，变量为(x_1, x_2)】，经过PCA转换成，由少数线性不相关变量来表示
    - 这些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，PCA属于降维方法
  - PCA用于发现数据表示中的基本结构，即数据表示中变量之间的关系
    - 找出线性意义上独立的变量表示（其实就是基）
- 【注意】
  - 此处的数据矩阵代表数据（非变换）
  - PCA用于分析数据的主要信息分布方向
    - 主要成分，主要信息量



1 总体主成分分析

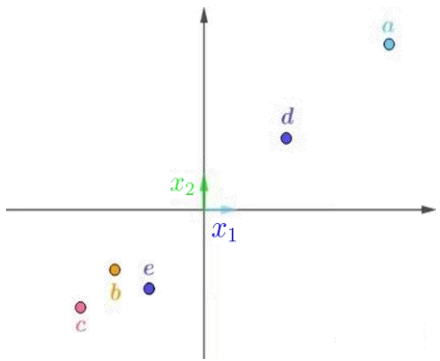
基本思想

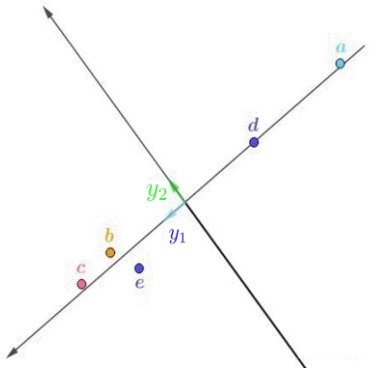
- 对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据
  - 新变量：可能的正交变换中，变量的方差的和（信息量）最大的
    - 方差可用来表示在新变量上信息（量）的大小

- PCA分解
  - 将新变量依次称为第一主成分、第二主成分等
  - **利用主成分近似地表示原始数据**，可理解为发现数据的“**基本结构**”
  - 把数据由少数主成分表示，可理解为对**数据降维**

直观解释 - 线性相关

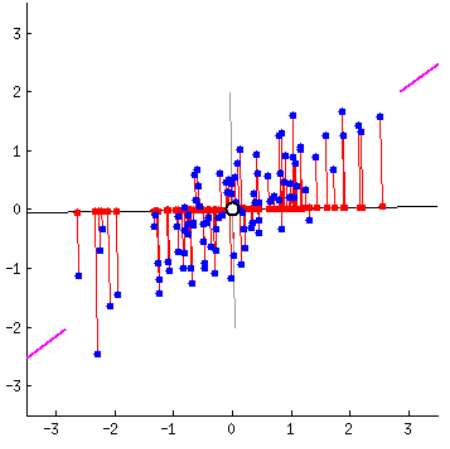
- 考察数据（近似的）线性相关性，并通过变换减少这种相关的冗余性
  - 数据由变量 (x_1, x_2) 表示。变量x_1和x_2是线性相关的。知道其中一个变量 x_1取值时，变量 x_2 的预测不是完全随机的
  - PCA通过正交变换到变量y_1和y_2表示。在新坐标系里，数据中的变量y_1和y_2是线性无关，当知道其中一个变量y_1的取值时，对另一个变量y_2的预测是完全随机的；反之亦然.
- 如数据不严格在一条线上
  - y_1和y_2相关性小（ y_1, y_2可以任意取值），前者x_1, x_2相关性更大（ x_1, x_2近似满足线性关系，和直线y_1的误差比较小）
  - 用一个分量表示时， y_1的误差别x_1/x_2小！





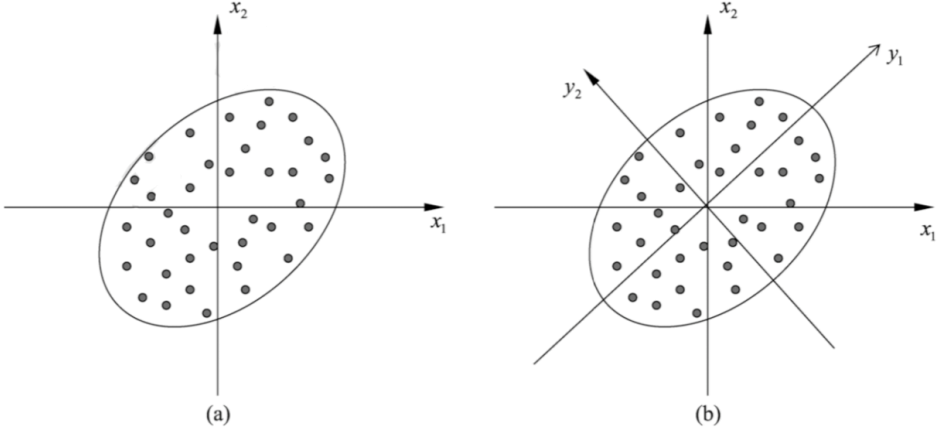
红酒列表的PCA特性

- 寻找一些在所有红酒中很不相同的属性（特性）。PCA寻找能尽可能体现红酒差异的属性。
  - 红点如何“散布”（称之为“方差”）的；它们何时最大化？
- 寻找一些属性，这些属性允许预测或者说“重建”原本的红酒特性。PCA寻找能够尽可能好地重建原本特性的属性。
  - 基于新特性（红点的位置）重建原本的两个特性（蓝点的位置），连接红线的长度重建误差
  - 红线的长度总长度何时最小化？



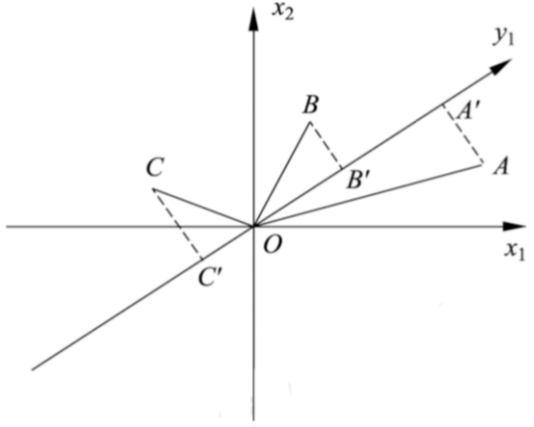
直观解释 – 主成分

- 首先，PCA选择方差和最大的方向 (第一主成分)作为新坐标系的第一坐标轴，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
- 在新坐标系里，数据中的y_1分量和y_2分量线性无关
- 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。



直观解释 – 方差和最大

- 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。样本点A, B, C在y_1轴上投影为A', B', C'。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
- PCA旨在选取正交变换中方差最大的变量作为第一主成分
  - 根据勾股定理，坐标值的平方和$OA'^2 + OB'^2 + OC'^2$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【方差和最大等价于投影误差和最小】
  - 等价地，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



PCA分解

要点：变换后的数据方差是数据协方差矩阵的特征值

主成分分析类别

- 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

- 【理论分析和样本数据分析】

均值与方差

- 均值（mean）

  标量： $\mu = E(X)$

  向量：假设$x = (x_1, x_2, \cdots, x_m)^T$是$m$维随机变量，均值向量

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

- 方差（variance），偏离均值的程度

  标量：$var(X) = E\left[\left(X - E(X)\right)^2\right] = \mathrm{E}[X^2] - E[X]^2$

  向量：协方差矩阵(Covariance)，$m \times m$对称矩阵

  $$cov(X, X) = E[(X - E[X])(X - E[X])^{\mathrm{T}}]$$

  【注】

  - 协方差矩阵是对标量随机变量方差的一般化推广。
  - 考察随机变量的不同分量/属性/变量之间的（协同）变化规律（偏离均值的变化），即描述这些分量是同时变大变小（相关，包括正相关负相关）或者没有关系（不相关）
  - Σ为对称矩阵，可以对角化

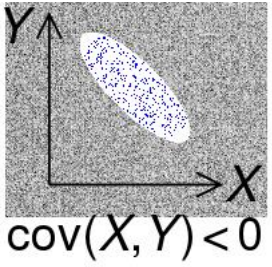
协方差矩阵

协方差（Covariance）：衡量两个随机变量的偏离均值的联合变化情况（是否相同偏离的性质）。X与Y之间的协方差$\mathrm{cov}(X,Y)$

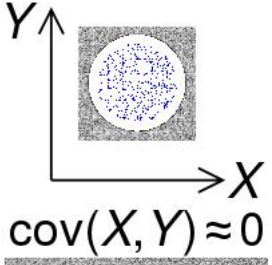
$$\begin{aligned}\mathrm{cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY - E[X]Y - E[Y]X + E[X]E[Y]] \\ &= E[XY] - 2E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

- 直观上，协方差表示两个随机变量不同分量/属性之间，偏离均值的变化规律
  - 如两个分量的偏离变化趋势一致，即，如一变量大于期望时另一变量也大于期望，那么协方差分量为正，反之亦然
  - 如X与Y统计独立，$\mathrm{E}[XY] = E[X]E[Y]$，那么二者之间的协方差为0。反过来并不成立
- 协方差为0的两个随机变量称为是不相关的

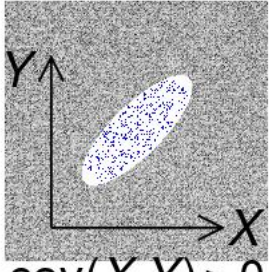
协方差矩阵



$\mathrm{cov}(X,Y)<0$



$\mathrm{cov}(X,Y)\approx 0$



$\mathrm{cov}(X,Y)>0$

线性变换后的均值和方差

由m维随机变量x到m维随机变量$y=(y_1,y_2,\cdots,y_m)^T$的线性变换

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

y为新坐标系下表达，y_i为分量/成分，$\alpha_i^T=(\alpha_{1i},\alpha_{2i},\cdots,\alpha_{mi}),i=1,2,\cdots,m$，于是

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

【注】$E(y_i)=E(\alpha_i^T x)=\alpha_i^T E(x)=\alpha_i^T\mu$

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

【注】$var(y_i)=E\left[(y_i-E(y_i))(y_i-E(y_i))^T\right]=E\left[(\alpha_i^T x-\alpha_i^T\mu)(\alpha_i^T x-\alpha_i^T\mu)^T\right]=\alpha_i^T E[(x-\mu)(x-\mu)^T]\alpha_i=\alpha_i^T\Sigma\alpha_i$，$var(y)=\mathrm{M}=\mathrm{A}^T\Sigma\mathrm{A}$，A为正交变换矩阵（$\mathrm{A}^T=\mathrm{A}^{-1}$）。因此，变换前后的M和$\Sigma$相似（$\mathrm{M}=\mathrm{A}^{-1}\Sigma\mathrm{A}$）

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

【注】$cov(y_i,y_j)=E\left[(y_i-E(y_i))\left(y_j-E(y_j)\right)^T\right]=E\left[(\alpha_i^T x-\alpha_i^T\mu)(\alpha_j^T x-\alpha_j^T\mu)^T\right]=\alpha_i^T E[(x-\mu)(x-\mu)^T]\alpha_j=\alpha_i^T\Sigma\alpha_j$

总体主成分

给定线性变换 $y_i = \alpha_i^{\mathrm{T}} x = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \cdots + \alpha_{mi}x_m$，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$cov(y_i, y_j) = 0(i \neq j)$

3) 变量y_1是x的所有线性变换中方差最大； y_2是与y_1不相关的，x的所有线性变换中方差最大…

这时分别称$y_1, y_2, \cdots, y_m$为x的第一主成分、…、第m主成分

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：

- 在x的所有线性变换$\alpha_1^{\mathrm{T}}x = \sum_{i=1}^{m}\alpha_{i1}x_i$ 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
- 在x的所有与$\alpha_1^{\mathrm{T}}x$ 不相关的线性变换$\alpha_2^{\mathrm{T}}x = \sum_{i=1}^{m}\alpha_{i2}x_i$ 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

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

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

$\boldsymbol{x}$的第k主成分的方差是

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

即协方差矩阵Σ的第k个特征值

主要性质

$$y_k = \alpha_k^{\mathrm{T}}\boldsymbol{x} = \alpha_{1k}x_1 + \alpha_{2k}x_2 + \cdots + \alpha_{mk}x_m, k = 1,2,\cdots,m$$
$$\mathrm{var}(y_k) = \alpha_k^{\mathrm{T}}\Sigma\alpha_k = \lambda_k, k = 1,2,\cdots,m$$

- PCA寻找最大方差方向
  - 线性变换下的方差，为Σ相关函数（ $\mathrm{var}(\alpha_1^{\mathrm{T}}\boldsymbol{x}) = \alpha_1^{\mathrm{T}}\Sigma\alpha_1$ ）（也意味着找不相关方向分量）。Σ蕴含了数据不同分量之间的相关性信息。
  - Σ如果代表变换，其特征向量（可能）包含了变换的本质特征（因为变换描述了数据的分布信息）。
- 数据矩阵不包含变换信息，因此其特征值等不包含数据本身的信息。
- 正交变换的协方差矩阵彼此相似（ $var(y) = \mathrm{M} = \mathrm{A}^T\Sigma\mathrm{A}$ ），特征值和对应特征向量相等
  - 那么，存在正交矩阵A，使得变换后的协方差矩阵M就是原来坐标系协方差矩阵Σ的对角化矩阵（对角线由特征值构成），此时，A就是主成分分析的变换矩阵。

证明

在所有线性变换$\alpha_1^{\mathrm{T}}x=\sum_{i=1}^{m}\alpha_{i1}x_i$中，$\alpha_1^{\mathrm{T}}\alpha_1=1$条件下，使方差达到最大

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

求第一主成分为求解约束最优化问题

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

定义拉格朗日函数：$\alpha_1^{\mathrm{T}}\Sigma\alpha_1-\lambda\left(\alpha_1^{\mathrm{T}}\alpha_1-1\right)$，$\lambda$是拉格朗日乘子。对$\alpha_1$求导，并令其为0

$$\Sigma\alpha_1-\lambda\alpha_1=0$$
$$\Sigma\alpha_1=\lambda\alpha_1$$

即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$\alpha_1$是$\Sigma$的最大特征值$\lambda_1$对应的单位特征向量，显然$\alpha_1$与$\lambda_1$是最优化问题的解

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

第二主成分

第二主成分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【$\mathrm{cov}\left(\alpha_1^{\mathrm{T}}\boldsymbol{x}, \alpha_2^{\mathrm{T}}\boldsymbol{x}\right) = 0$】，$\boldsymbol{x}$的所有线性变换中使方差最大

$$\mathrm{var}\left(\alpha_2^{\mathrm{T}}\boldsymbol{x}\right) = \alpha_2^{\mathrm{T}}\Sigma\alpha_2$$

求第二主成分为求解约束最优化问题

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

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

定义拉格朗日函数$\alpha_2^{\mathrm{T}}\Sigma\alpha_2 - \lambda\left(\alpha_2^{\mathrm{T}}\alpha_2 - 1\right) - \phi\alpha_2^{\mathrm{T}}\alpha_1$，其中$\lambda, \phi$是拉格朗日乘子。对$\alpha_2$求导，令为0

$$2\Sigma\alpha_2 - 2\lambda\alpha_2 - \phi\alpha_1 = 0 \Rightarrow \Sigma\alpha_2 - \lambda\alpha_2 = 0$$

【将方程左乘以α_1^{T}，$2\alpha_1^{\mathrm{T}}\Sigma\alpha_2 - 2\lambda\alpha_1^{\mathrm{T}}\alpha_2 - \phi\alpha_1^{\mathrm{T}}\alpha_1 = 0$，前两项为0, 且$\alpha_1^{\mathrm{T}}\alpha_1 = 1$, 导出$\phi = 0$】

设α_2是Σ的第二大特征值λ_2对应的单位特征向量，α_2与λ_2是以上最优化问题的解。推广到第m主成分

性质

推论 16.1

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

(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

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

推论 16.1

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, $k = 1,2,\cdots,m$ 。即

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

用矩阵表示即为$\Sigma A = A\Lambda$，这里 $A = \left[\alpha_{ij}\right]_{m\times m}$, Λ 是对角矩阵，其第 k 个对角元素是 λ_k。

$$A = [\alpha_1, \alpha_2, \ldots \alpha_m]$$

$$\Sigma A = \Sigma[\alpha_1, \alpha_2, \ldots \alpha_m] = [\lambda_1\alpha_1, \lambda_2\alpha_2, \ldots \lambda_m\alpha_m] = [\alpha_1, \alpha_2, \ldots \alpha_m]\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} = A\Lambda$$

因为 A 是正交矩阵，即 $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I$, 于是，上式两端左乘/右乘$A^{\mathrm{T}}$，可推出

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

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

描述了协方差矩阵对角化属性

总体主成分的性质

1)总体主成分 $\boldsymbol{y}$ 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 $\boldsymbol{y}$ 的方差之和等于随机变量 $\boldsymbol{x}$ 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。

【注】事实上，利用式 $\Sigma = A\Lambda A^{\mathrm{T}}$及矩阵的迹 (trace) 的性质，可知

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

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

此处主要利用了，正交矩阵的性质，变换保持长度

总体主成分的性质

3)第 k 个主成分 y_k 与变量 x_i【 $x_i = e_i^{\mathrm{T}}\boldsymbol{x}$，原始坐标系下的变量】 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

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

【注】因为

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

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

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

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

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

由属性3)有

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

5)m 个主成分与第 i 个变量 x_i 的因子负荷量满足

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

【由于 $y_1, y_2, \cdots, y_m$ 互不相关，故

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

又x_i可以表为$y_1, y_2, \cdots, y_m$的线性组合, 故x_i与$y_1, y_2, \cdots, y_m$的相关系数的平方为1, 即$\rho^2\left(x_i, (y_1, y_2, \cdots, y_m)\right) = 1$】

主成分的个数

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

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

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

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

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

【注】当x的线性变换$\boldsymbol{y}$在$B = A_q$时，其协方差矩阵Σ_y的迹$\mathrm{tr}(\Sigma_y)$取得最大值，即：当取A的前q列取$\boldsymbol{x}$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。

【定理16.3】 考虑正交变换

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

其中B^{T}是 $p \times m$ 矩阵, A和$\Sigma_{\boldsymbol{y}}$的定义见定理16.2，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。

【注】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η_k定义为y_k的方差与所有方差之和的比

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

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

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

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_i保留信息的比例

【定义16.3】 主成分对原变量x_i的贡献率v_i定义为x_i与$(y_1, \cdots, y_k)$的相关系数的平方

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

x_i对$y_1, y_2, \cdots, y_k$都有贡献

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，直接求主成分可能产生不合理的结果。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(归一化)】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，$i=1,2,\cdots,m$，令

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

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

1)规范化变量主成分的协方差矩阵是

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

其中$\lambda_1^* \geqslant \lambda_2^* \geqslant \cdots \geqslant \lambda_m^* \geqslant 0$为相关矩阵$R$的特征值。

2)协方差矩阵的特征值之和为m

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

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$\lambda_k^*$的单位特征向量。

规范化变量的总体主成分

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^*

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

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

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

自相关矩阵

自相关矩阵、自协方差矩阵、互相关矩阵以及互协方差矩阵都是描述随机变量分布情况的工具

列向量$x \in C^{m\times 1}$和$y \in C^{n\times 1}$的外积(得到一个$m \times n$的矩阵):

$$x^{\circ}y = xy^{T}$$

自相关矩阵定义R_x为随机向量x与自身的外积的数学期望:

$$R_x \overset{\text{def}}{\Rightarrow} E\{x(\omega)x^{T}(\omega)\} = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix}$$

其中，r_{ii}是随机变量$x_i(\omega)$的自相关系数，下标$i = 1, \ldots, m$，$r_{ii} = E\{|x_i(\omega)|^2\}$

r_{ij}是$x_i(\omega)$和$x_j(\omega)$的互相关系数，$r_{ij} = E\{x_i(\omega)x_j(\omega)^T\}$

2 样本主成分分析

样本主成分分析

- 总体主成分分析，定义在样本总体上
- 实际问题中，需要在观测数据上进行主成分分析，样本主成分分析
- 样本主成分和总体主成分具有相同的性质

样本主成分的定义 - x

m维随机变量$x=(x_1,\cdots,x_m)^{\mathrm{T}}$进行$n$次独立观测，$\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$观测样本，其中$\boldsymbol{x}_j=(x_{1j},\cdots,x_{mj})^{\mathrm{T}}$为第$j$个观测样本，$x_{ij}$为第$j$个观测样本的第$i$个变量/分量，$j=1,\cdots,n$

- 样本矩阵：$X=[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n]=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$
- 样本均值向量：$\bar{x}=\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$，样本协方差矩阵：$S=[s_{ij}]_{m\times m}$
  - 其中$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$，$i,j=1,2,\cdots,m$
  - 其中$\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ik}$为第$i$个变量/分量的样本均值
- 样本相关矩阵：$R=[r_{ij}]_{m\times m}$，$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$，$i,j=1,2,\cdots,m$

样本主成分的定义 – y

定义m维向量$x = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$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$A = [a_1 \quad a_2 \quad \cdots \quad a_m] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$,

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

考虑任意一个线性变换

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

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$\bar{y}_i$

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

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

样本主成分的定义 – y

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$\bar{y}_i$

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

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

y_i的样本方差$\mathrm{var}(y_i)$为

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

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

【定义16.4】(样本主成分)给定样本矩阵$\boldsymbol{X}$

样本第一主成分$y_1 = a_1^{\mathrm{T}}\boldsymbol{x}$在$a_1^{\mathrm{T}}a_1 = 1$时，使得$a_1^{\mathrm{T}}\boldsymbol{x}_j(j = 1,\cdots,n)$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$a_1^{\mathrm{T}}\boldsymbol{x}_j(j = 1,\cdots,n)$的样本协方差$a_1^{\mathrm{T}}Sa_2 = 0$条件下，使得$a_2^{\mathrm{T}}\boldsymbol{x}_j(j = 1,\cdots,n)$的样本方差$a_2^{\mathrm{T}}Sa_2$最大的$x$的线性变换;

一般地，样本第i主成分$y_i = a_i^{\mathrm{T}}x$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。

在使用样本主成分时，一般对样本矩阵作规范化变换:

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

将规范化变量x_{ij}^*仍记作x_{ij},规范化的样本矩阵仍记作X。样本协方差矩阵S就是样本相关矩阵

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

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

- 主成分分析
  - 传统，通过数据的协方差矩阵或相关矩阵的特征值分解进行
  - 现在常用方法，通过数据矩阵的奇异值分解进行

- 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

相关矩阵的特征值分解算法

- 1)对观测数据进行规范化处理，得到规范化数据矩阵，仍以 X 表示。
- 2)依据规范化数据矩阵，计算样本相关矩阵

$$R = [r_{ij}]_{m\times m} = \frac{1}{n-1}XX^{\mathrm{T}}，\text{其中}r_{ij} = \frac{1}{n-1}\sum_{l=1}^{n} x_{il}x_{lj}\ ,\ i,j = 1,2,\cdots,m$$

- 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

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

- 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，求出样本主成分

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

相关矩阵的特征值分解算法

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

数据矩阵的奇异值分解算法

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

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

式中U_k是$m \times k$矩阵，V_k是$n \times k$矩阵，Σ_k是k阶对角矩阵；A的完全奇异值分解的矩阵U, V, Σ。U_k, V_k分别由取U, V的前k列，Σ_k取Σ的前k个对角线元素得到。

定义一个新的 $n \times m$ 矩阵 $X' = \frac{1}{\sqrt{n-1}} X^{\mathrm{T}}$，$X'$ 的每一列均值为零。则

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

即 $X'^{\mathrm{T}} X'$ 等于 X 的协方差矩阵$S_X = X'^{\mathbf{T}} X'$

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，所以问题转化为求矩阵 $X'^{\mathrm{T}} X'$ 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 $X' = U \Sigma V^{\mathrm{T}}$, 那么$V$的列向量就是$S_X = X'^{\mathrm{T}} X'$的单位特征向量。因此，$V$的列向量就是$X$的主成分。

主成分分析算法

- 【算法 16.1(主成分分析算法)】
- 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
- $X' = U\Sigma V^{\mathrm{T}}$