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**。并且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。在数据总体上进行的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。

基础知识

- 正交矩阵:

  如果n阶矩阵A满足$A^TA=E$即$A^{-1}=A^T$，那么称A为正交矩阵。方阵A为正交矩阵的充分必要条件是A的列向量都是单位向量且两两正交。

- 方差:

  设x是一个随机变量，若$E\{[x-E(x)]^2\}$存在，则称其为x的方差，记为$Var(x)$或$D(x)$。

- 协方差:

  称$E\{[x-E(x)][y-E(y)]\}$为随机变量x,y的协方差，记为$Cov(x,y)$。

- 相关系数:

  $\rho_{xy}=\dfrac{Cov(x,y)}{\sqrt{D(x)}\sqrt{D(y)}}$称为随机变量$x,y$的相关系数。

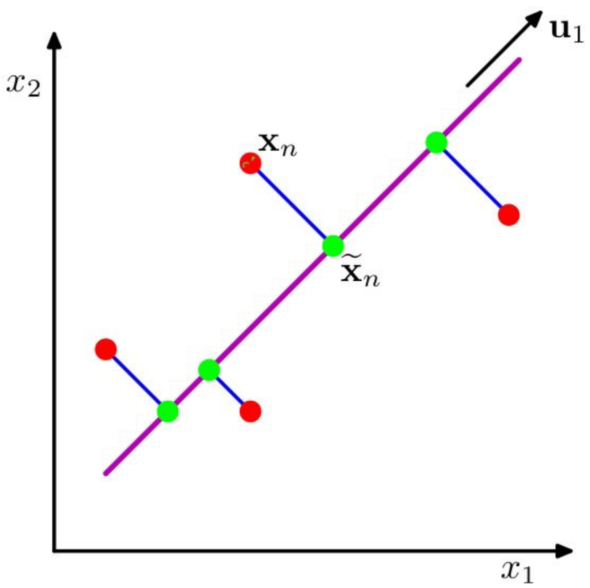
- 协方差矩阵:

- 向量内积的几何意义:

  向量a和b的内积可以表示为$a\cdot b=|a|\times|b|cos\theta$，几何意义就是一个向量在另一个向量上的投影与这个向量模长的积，也就是同方向的积。特别地，如果一个向量a是某个坐标轴的单位坐标向量，那么向量的内积自然就是$|b|cos\theta$，也就是向量b在此坐标轴上的坐标值。因此，如果想要将一个向量变换到新的坐标系，那么只要对新坐标系向量进行内积运算即可。

总体主成分分析

基本想法



定义和导出

假设$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^T$是$m$维随机变量，意即每一维都是单独的随机变量，但是相关性未知。设其均值向量是$\boldsymbol{\mu}=E(\boldsymbol{x})=(\mu_1.\mu_2,\cdots,\mu_m)^T$，**协方差矩阵**是$\boldsymbol{\Sigma}=Cov(\boldsymbol{x},\boldsymbol{x})=E\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\}$。

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_m-\mu_m \end{bmatrix} \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_m-\mu_m \end{bmatrix} \\
&= \begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_m-\mu_m) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \cdots & (x_2-\mu_2)(x_m-\mu_m) \\ \vdots & \vdots & \vdots & \vdots \\ (x_m-\mu_m)(x_1-\mu_1) & (x_m-\mu_m)(x_2-\mu_2) & \cdots & (x_m-\mu_m)^2 \end{bmatrix}
\end{aligned}
$$

以上可以看出协方差矩阵$\boldsymbol{\Sigma}$是**对称矩阵。**

考虑由m维随机变量$\boldsymbol{x}$到m维随机变量$\boldsymbol{y}=(y_1,y_2,\cdots,y_3)$的线性变换$\boldsymbol{\alpha}_i^T$：

$\boldsymbol{\alpha}_i^T$本质上是个实数向量，不是随机变量。

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

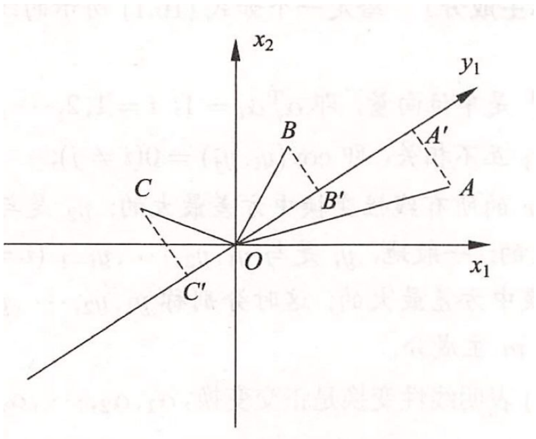
由随机变量的性质可知：

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

$$
\begin{aligned}
Var(y_i) &= E\{[y_i-E(y_i)]^2\} \\
&= E\{(\alpha_i^T x-\alpha_i^T\mu)^2\} \\
&= E\{[\alpha_i^T(x-\mu)][\alpha_i^T(x-\mu)]\} \\
&= E\{[\alpha_i^T(x-\mu)(x-\mu)^T\alpha_i]\} \\
&= \alpha_i^T E\{[(x-\mu)(x-\mu)^T]\}\alpha_i \\
&= \alpha_i^T\Sigma\alpha_i
\end{aligned}
$$

$$
\begin{aligned}
Cov(y_i,y_j) &= E\{(y_i-E(y_i))((y_j-E(y_j))\} \\
&= E\{(\alpha_i^T x-\alpha_i^T\mu)(\alpha_j^T x-\alpha_j^T\mu)\} \\
&= E\{(\alpha_i^T xx^T\alpha_j-2\alpha_i^T\mu x\alpha_j+\alpha_i^T\mu\mu^T\alpha_j)\} \\
&= E\{\alpha_i^T((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T)\alpha_j\} \\
&= \alpha_i^T\Sigma\alpha_j
\end{aligned}
$$

值得注意的是：以上三个的最终结果均是一个数。



主成分分析对数据进行正交变换，具体地，对原坐标系进行旋转变换，并将数据在新坐标系表示。比如原数据由变量(x_1, x_2)表征，那么变换后在新坐标系里就由变量(y_1, y_2)表征。主成分分析选择方差最大的方向作为新坐标系的第一坐标轴y_1轴，然后选择与第一坐标轴正交，且方差最大的方向作为新坐标系的第二坐标轴y_2轴。

首先，我们来求x的第一主成分$y_1 = \alpha_1^T x$，即求系数向量α_1，这个α_1实际上就是变换后的坐标轴。因为α_1是单位向量，所以由内积的几何意义得知，y_1就是x变换后在α_1上的坐标。而选取方差最大的变量，也就是旋转变换后坐标值的平方和最大的轴。

定理：设$\boldsymbol{x}$是m维随机变量，$\boldsymbol{\Sigma}$是$\boldsymbol{x}$的**协方差矩阵**，$\boldsymbol{\Sigma}$的**特征值**分别为$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_m \geqslant 0$，特征值对应的**单位特征向量**分别为$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m$，则$\boldsymbol{x}$的第$k$个主成分是

$$y_k = \boldsymbol{\alpha}_k^T \boldsymbol{x} = \alpha_{1k}^T x_1 + \alpha_{2k}^T x_2 + \cdots + \alpha_{mk}^T x_m$$

$\boldsymbol{x}$的第k个主成分的**方差**为

$$Var(y_k) = \boldsymbol{\alpha}_k^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_k = \lambda_k$$

推论：m维随机变量$\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^T$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：

- $\boldsymbol{y} = \boldsymbol{A}^T \boldsymbol{x}$，$\boldsymbol{A}$为正交矩阵。
- $\boldsymbol{y}$的协方差矩阵为对角矩阵即$Cov(\boldsymbol{y}) = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$，其中$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$。$\lambda_k$是$\boldsymbol{x}$的协方差矩阵的第$k$个特征值，$\boldsymbol{\alpha}_k$是对应的单位特征向量。

总体主成分的性质

- 总体主成分$\boldsymbol{y}$的协方差矩阵是对角矩阵，对角线上的元素即是每个分量的方差。
- 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$\sigma_{ii}$是随机变量$x_i$的方差，即协方差矩阵$\boldsymbol{\Sigma}$的对角元素。实际上是说，$\boldsymbol{y}$的协方差矩阵的迹等于$\boldsymbol{x}$的协方差矩阵的迹。
- 第k个主成分与变量x_i的相关系数$\rho(y_k, x_i)$称为因子负荷量。计算公式为

$$\rho(y_k, x_i) = \frac{Cov(y_k, x_i)}{\sqrt{Var(y_k)Var(x_i)}} = \frac{\sqrt{\lambda_k}\alpha_{ik}}{\sqrt{\sigma_{ii}}}$$

- 第k个主成分y_k与m个变量的因子负荷量满足$\sum_{i=1}^{m} \alpha_{ii} \rho^2(y_k, x_i) = \lambda_k$。
- m个主成分与第i个变量x_i的因子负荷量满足$\sum_{i=1}^{m} \rho^2(y_k, x_i) = 1$。

样本主成分分析

上述的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。

定义和性质

假设对m维随机变量$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)^T$进行$n$次独立观测，$\boldsymbol{x}_1,\boldsymbol{x}_2\cdots,\boldsymbol{x}_n$表示观测样本，其中$x_j=(x_{1j},x_{2j},\cdots,x_{mj})^T$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，则$n$次观测数据可以用样本矩阵$\boldsymbol{X}$表示，记作：

$$\boldsymbol{X}=[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n]=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{31} & x_{32} & \cdots & x_{mn} \end{bmatrix}$$

给定了样本我们就可以估计样本均值$\bar{\boldsymbol{x}}$和样本协方差$\boldsymbol{S}$：

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

$$\boldsymbol{S}=[s_{ij}]_{m\times m}$$

$$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$$

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

其中

$$A=[\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_m]=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

考虑$\boldsymbol{y}$中的任意的一个线性变换

$$y_i=\boldsymbol{a}_i^T\boldsymbol{x}=a_{1i}^Tx_1+a_{2i}^Tx_2+\cdots+a_{mi}^Tx_m$$

其中y_i是m维向量$\boldsymbol{y}$的第i个随机变量，均值为$\bar{y_i}=\dfrac{1}{n}\sum\limits_{j=1}^{n}a_i^Tx_j=a_i^T\bar{x}$。$y_i$的样本方差

$$\begin{aligned} Var(y_i)&=\frac{1}{n-1}\sum_{j=1}^{n}(a_i^Tx_j-a_i^T\bar{x})^2 \\ &=a_i^T[\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T]a_i=a_i^TSa_i \end{aligned}$$

对任意两个线性变换$y_i=a_i^Tx, y_k=a_k^Tx$，样本协方差为$Cov(y_i,y_k)=a_i^TSa_k$。

对于一组给定的数据点$\{v_1, v_2, \cdots, v_n\}$，中心化后的表示为$\{x_1, x_2, \cdots, x_n\} = \{v_1 - \mu, v_2 - \mu, \cdots, v_n - \mu\}$，其中$\mu = \frac{1}{n}\sum_{i=1}^{n} v_i$。由向量内积的几何意义得出，我们要找一个投影方向$a$使得$x_1, x_2, \cdots, x_n$在$a$上的投影坐标方差尽可能大。由于我们提前中心化的原因，投影之后均值为$0, (\mu' = \frac{1}{n}\sum_{i=1}^{n} x_i^T a = (\frac{1}{n}\sum_{i=1}^{n} x_i^T)a = 0)$，这正是提前中心化的意义。

因此投影后的方差可以表示为

$$
\begin{aligned}
D(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i^T a)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(a^T x_i x_i^T a) \\
&= a^T \sum_{i=1}^{n}(\frac{1}{n} x_i x_i^T) a
\end{aligned}
$$