主成分分析 （Principal Component Analysis）

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。

主成分分析是一种降维方法。

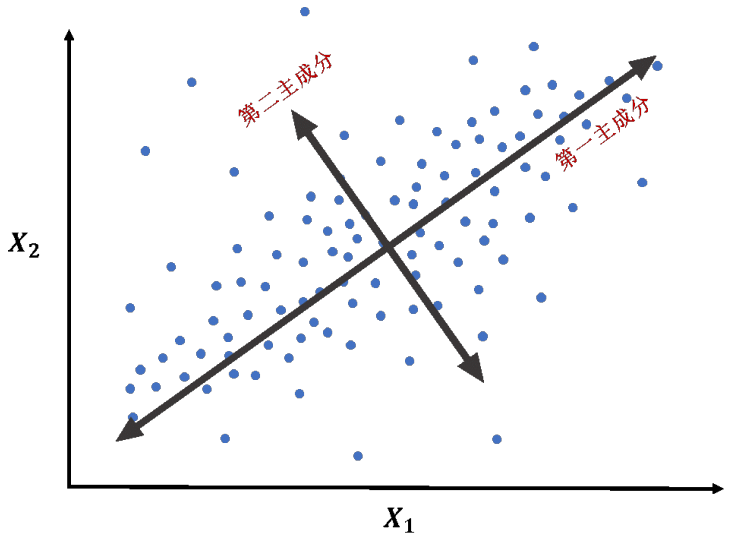

Figure 1: 主成分分析

理解：如何利用一个超平面，对样本恰当的表达？
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（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。

1.1. 定义和导出

设$X = (X_1, \cdots, X_m)^\top$ 是m维随机变量，均值为 $E(X) \overset{\text{def}}{=} \mu$, 协方差矩阵为 $\text{cov}(X) \overset{\text{def}}{=} \Sigma$.

考虑m维随机变量X到m维随机变量Y的线性变换： $Y_i = \alpha_i^\top X$, 其中 $\alpha_i = (\alpha_{1i}, \cdots, \alpha_{mi})^\top$. 由此可知 $E(Y_i) = \alpha_i^\top \mu$, $\text{var}(Y_i) = \alpha_i^\top \Sigma a_i$ $\text{cov}(Y_i, Y_j) = \alpha_i^\top \Sigma \alpha_j$.

定义1（总体主成分）

设$X = (X_1, \cdots, X_m)^\top$是m维随机向量，均值$E(X) = \mu$，协方差阵$D(X) = \Sigma$，称$Y_i = \alpha_i^\top X$为X的第i主成分（i=1,2,...,m）如果：

（1）$\alpha_i^\top \alpha_i = 1$ $(i = 1, 2, ..., m)$；

（2）变量Y_i与Y_j互不相关，即$\text{cov}(Y_i, Y_j) = 0$ $(i \neq j)$；

（3）$Var(Y_i) = \max\limits_{\alpha^\top \alpha = 1, \alpha^\top \Sigma \alpha_j = 0 \ (j=1,...,i-1)} Var(\alpha^\top X)$.

1.2. 主要性质

设Σ的特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ 特征值对应的单位特征向量为 $\alpha_1, \cdots, \alpha_m$, 则x的第k个主成分是 $Y_k = \alpha_k^\top X$, 方差为 $\text{var}(Y_k) = \alpha_k^\top \Sigma \alpha_k$.

以第一主成分为例: 求第一主成分等价于求解以下最优化问题：

$$\max_{\alpha_1} \alpha_1^\top \Sigma \alpha_1 \tag{1.1}$$

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

定义拉格朗日函数

$$\alpha_1^\top \Sigma \alpha_1 - \lambda \left(\alpha_1^\top \alpha_1 - 1 \right)$$

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

$$\alpha_1^\top \Sigma \alpha_1 = \alpha_1^\top \lambda \alpha_1 = \lambda \alpha_1^\top \alpha_1 = \lambda$$

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$\alpha_1^\top X$构成第一主成分，其方差等于协方差矩阵的最大特征值

$$\text{var}\left(\alpha_1^\top X\right) = \alpha_1^\top \Sigma \alpha_1 = \lambda_1 \tag{1.2}$$

性质：

(1) $\text{cov}(Y) = \text{diag}(\lambda_1, \cdots, \lambda_m)$;

(2) $\sum_i \lambda_i = \sum \sigma_{ii}$ $(\sigma_{ii} = \text{var}(X_i))$.

(3) $\rho(Y_k, X_i) = \text{cor}(Y_k, X_i)$ 称为因子负荷量（factor loading）。

$$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}\alpha_{ki}}{\sqrt{\sigma_{ii}}} \tag{1.3}$$

通过因子负荷量的高低可以对主成分进行解读

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

(5) $\sum_k \rho^2(Y_k, X_i) = 1$ (作业：证明这个性质)

前k个主成分对变量X_i的贡献率定义为 $\nu_i \overset{\text{def}}{=} \sum_{j=1}^k \rho^2(Y_j, X_i)$.

1.3. 主成分的个数

第k个主成分的方差贡献率定义为 Y_k的方差与总方差的比值 $(\sum_{j=1}^k \lambda_j)/(\sum_{j=1}^m \lambda_j)$

通常选取k使得方差贡献率比较高，例如70% 80%以上。

在实际数据分析中，我们常采用一些准则选取最佳的主成分数目，常用的方式有：

（1）崖底碎石图（Scree Plot）

根据每个主成分对应的特征值绘制点线图，当主成分数目从k增加到k+1时，如果特征值出现较剧烈的下降，即第k+1个点为拐点，表示相比较于前k个主成分所提取的变量信息，第k+1个主成分只包含了非常少量的变量信息，因此提取k个主成分

较为合适。如图2所示，当拐点出现在k=3处，应当选择2个主成分。

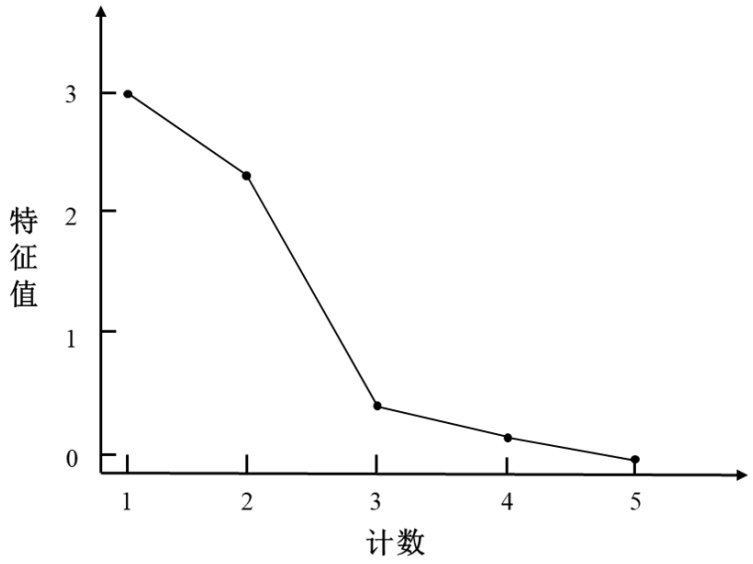

Figure 2: 崖底碎石图示例

（2）累计方差贡献率

这个方法的思想和崖底碎石图相似，都是通过特征值的相对大小来选择最佳的主成分数目。每个主成分的方差贡献率为$\lambda_i / \sum_{i=1}^{m} \lambda_i$，因此前k个主成分的累计方差贡献率为$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{m} \lambda_i$，选择最佳主成分数目，使得前k个主成分的累计方差解释比例达到一定数值（通常为70%至80%）。

（3）Kaiser准则

根据前文的推导，主成分的表达式是求解观测变量的协方差阵的特征值和特征向量得到的。由于m个原变量所提供的总信息（变量总方差）的绝大部分只需要用前k个主成分来代替，则存在：

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

在对变量进行标准化后，有$\mathrm{tr}(\Sigma) = \Sigma_{i=1}^{m} \sigma_{ii} = m$，则平均的原始变量信息为1。Kaiser准则的想法非常简单：主成分至少要能解释一个原始变量的信息，因此在这个准则下我们只需要选择与大于1的特征值数目即可。

2. 样本主成分分析

在实际数据分析中，我们无法取得总体，但可以在样本观测数据上进行主成分分析。

观测数据用样本矩阵$\mathbf{X} = (x_1, \cdots, x_n)^\top \in \mathbb{R}^{n\times m}$ 表示. 样本协方差矩阵可以表示为： $S = \sum_j (x_j - \overline{x})(x_j - \overline{x})^\top/(n-1)$， 其中 $\overline{x} = n^{-1}\sum_j x$. 样本的相关矩阵为 $R = \mathrm{diag}(S)^{-1/2} S \mathrm{diag}(S)^{-1/2}$

定义2:（样本主成分）

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

在实际中，需要对样本矩阵做标准化处理：

$$x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{\sqrt{s_{jj}}}. \tag{2.1}$$

其中 $\overline{x}_j$ 是第j列的平均值。

求解k个样本主成分：

(1) 对 R进行特征值分解，设 a_i 表示第i个特征向量（按照特征值降序）

(2) $y_i = \mathbf{X} a_i$ 代表样本主成分（也称为主成分得分）

注*：这里计算主成分得分用到的$\mathbf{X}$是标准化后的样本矩阵$\mathbf{X}$。

主成分分析可以作为对其他机器学习方法的输入，例如可以利用前k个样本主成分进行聚类分析。