主成分分析

PCA概念

- $\mathbf{x}=(x_1,\cdots,x_p)^T$为随机向量（在实际问题中可以认为是一个观测样本的$p$个不同变量）
- 对x做线性组合有$\mathbf{y}=a\mathbf{x}=\sum_{j=1}^p a_jx_j$
- 不同的a的选取就会有不同的组合，得到不同的$\mathbf{y}_i$，相应的第i组第组合系数为a_i
- 为了尽可能提取$\mathbf{x}$的有关信息，对这个线性组合做出下列约束：
  1. 令$\mathbf{y}_i$的方差在新的组合中最大
  2. $||a_i||_2=1$（这一步操作为称为标准化，因为若不加此限定，可以单纯的将a_i扩大某个倍数以达到更大方差，但这对于解决问题并无实际意义
  3. 对于新的组合$\mathbf{y}_j$，$\mathbf{y}_j\perp\mathbf{y}_i$
- 记$\Sigma=var(\mathbf{x})$，Σ矩阵的特征值为$\lambda_1\cdots\lambda_p$，对应的特征向量为$t_1\cdots,t_p$. 可以证明：t_i即为第i个线性组合的相应系数，$\mathbf{y}_i=t_i^T\mathbf{x}$即为第$i$个主成分
  - 进一步展开上式，$\mathbf{y_i}=t_{1i}x_1+t_{2i}x_2+\cdots+t_{pi}x_p$，这里每一个$t_{ji}$就表示在当前第$i$个主成分中，第$j$个变量$x_{ji}$对于主成分$y_i$的影响重要程度，称这个$t_{ji}$为对应的**载荷**
- 若将这p个主成分写成统一的矩阵形式，则有

$$\mathbf{y}=\mathbf{T}'\mathbf{x}$$

PCA的性质

- p个主成分的协方差矩阵为$diag(\lambda)$，即第i个特征值为第i个主成分的方差，不同的主成分之间彼此独立

$$V(y)=\Lambda$$

- p个主成分的方差之和等于原始p个数据的方差之和

$$\sum V(y_i)=\sum V(x_i)$$

- 进一步由这个信息可以定义**（方差）贡献率**与累积贡献率（及第k个/前k个PCA对总方差的占比）

$$\lambda_i/\sum_{j=1}^p\lambda_j$$

- 还可以定义前k个PCA对于某个变量x_i的变量贡献率，规定其为x_i与$y_1,\cdots,y_m$的复相关系数

$$\rho^2_{i\cdot 1,\cdots,m}=\sum_{i=1}^m\rho^2(x_i,y_k)$$

- 第i个变量x_i与第k个主成分y_k之间的相关系数为该载荷与该主成分方差的乘积

$$Cov(x_i,y_k)=t_{ik}\lambda_k$$

- 相应的，可以进一步求出对应相关系数

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

- 第i个变量的方差等于p个特征值关于对应载荷平方的加权平均，即

$$\sigma_{ii} = t_{i1}^2\lambda_1 + \cdots + t_{pk}^2\lambda_p$$

$$[\ p.f:\ tr(\Lambda) = tr(\mathcal{T}'\Sigma\mathcal{T}) = tr(\Sigma\mathcal{T}\mathcal{T}') = tr(\Sigma)\ ]$$

- 进一步地，由于$\lambda_1 \geq \cdots \geq \lambda_p$，因此若$x_i$的方差较大，则靠前的载荷$t_{i1}$等绝对值较大；反之若绝对值较小，则靠前的载荷较小

PCA的标准化与相关矩阵的PCA

- 与回归分析等不同，是否对原始数据进行标准化将直接影响最后PCA的结果
- 考虑各变量单位不同、方差差异较大等情况，理论而言在进行PCA之前都应当进行标准化，即

$$x_i^* = (x_i - \mu_i)/\sqrt{\sigma_{ii}}$$

- 由PCA的计算性质可知，若未进行标准化，则大量的主成分将倾向于方差变异性较大的变量，而忽略了其他变量本应发挥的作用；尤其是这种变异性不是由数据本身的特征，而是由单位尺度放缩的问题导致时，将对分析的结果造成负面影响
- 对于标准化后的x^*，协方差矩阵退化为原先非标准化x的相关系数矩阵，因此根据相关系数的一些优良性质，上述PCA性质可以得到进一步简化

样本的PCA

- 在实际情况中，我们都是用样本估计总体
- 因此可以用样本协方差矩阵S，相关系数矩阵R对总体PCA进行估计，其中

$$S = \sum(x_i - \bar{x})(x_i - \bar{x})'/(n-1)$$

- 在约束条件中，需要注意这里满足是使得样本的方差最大、协方差为零（而不是总体的方差与协方差）

PCA的理解与补充

- 第一主成分是p维空间中一条最接近n条样本观测的线（此处距离为平方欧式距离）
- 第一与第二主成分共同构成了数据可以达到的最优平面，使得所有数据向这个平面进行投影所得到的方差最大
- 在前述介绍中，认为$\mathbf{x} = (x_1, \cdots, x_p)^T$.对此进行推广，$X = (X_1, \cdots, X_p)$，其中$X$为数据矩阵，其有$p$个数据特征$X_i$（$p$列），对应有$n$条样本观测. 值得说明的是，此推广并未改变其余计算过程
- PCA数量的选取
  - 主成分数量的选取一定程度上是主观的，需要特定情况特定分析
  - 若在前几个主成分中都找不到有价值的模式，那么在更多的主成分中寻找也不大可能有有价值的收获
  - 若前几个主成分是有价值的，则会继续寻找直到找不到更多有价值的模式为止