

统计学习：第十六章

1 对以下样本数据进行主成分分析：

$$X = \begin{bmatrix} 2 & 3 & 3 & 4 & 5 & 7 \\ 2 & 4 & 5 & 5 & 6 & 8 \end{bmatrix},$$

答：

$$V = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix},$$

2 证明样本协方差矩阵 S 是总体协方差矩阵方差 E 的无偏估计

答：

$$\begin{aligned} E[s_{ij}] &= E \left[\frac{1}{n-1} \sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \right] \\ &= \frac{1}{n-1} E \left[\sum_{k=1}^n (x_{ik} - \frac{1}{n} \sum_{l=1}^n x_{il})(x_{jk} - \frac{1}{n} \sum_{l=1}^n x_{jl}) \right] \\ &= \frac{1}{n-1} \sum_{k=1}^n E \left[(x_{ik} - \frac{1}{n} \sum_{l=1}^n x_{il})(x_{jk} - \frac{1}{n} \sum_{l=1}^n x_{jl}) \right] \\ &= \frac{1}{n-1} \sum_{k=1}^n E \left[(x_{ik} - \mu_i - \frac{1}{n} \sum_{l=1}^n (x_{il} - \mu_i))(x_{jk} - \mu_j - \frac{1}{n} \sum_{l=1}^n (x_{jl} - \mu_j)) \right] \\ &= \frac{n}{n-1} (cov(i, j) - \frac{2}{n} cov(i, j) + \frac{n}{n^2} cov(i, j)) \\ &= cov(i, j) \end{aligned}$$

3 设 X 为数据规范化样本矩阵，则主成分等价于求解以下最优化问题

$$\begin{aligned} \min_L \quad & \|X - L\|_F \\ \text{s.t.} \quad & \text{rank}(L) \leq k \end{aligned}$$

这里 F 是弗罗贝尼乌斯范数， k 是主成分个数。试问为什么？答：有上一章知奇异值分解 $U\Sigma_k V^T$ 为一个最优解，而奇异值分解等价于求 X 的主成分。