HOMEWORK 9

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

证明以下性质:

$$\sum_k \rho^2(Y_k,X_i)=1$$

其中，$\rho(Y_k,X_i)=\frac{\sqrt{\lambda_k}\alpha_{ki}}{\sqrt{\sigma_{ii}}}$, $\sigma_{ii}=\operatorname{var}(X_i)$, $\alpha_{ki}=e_i^\top\alpha_k$, e_i 为基本单位向量, 其第 i 个变量为 1，其余为 0。

1.2 对以下样本数据进行主成分分析:

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

1.3 证明样本协方差矩阵S 是总体协方差矩阵方差Σ 的无偏估计。

1.4 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:

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

这里 F 是弗罗贝尼乌斯范数, k 是主成分个数。试问为什么?

以上证明题请以 PDF 格式提交。

2 数据分析及算法实现。

数据集介绍: NBA 数据集，具体数据描述见 WORD 文档。

完成任务文档 (1)–(4) 题。

3 编程练习:

请写一个 PCA 算法，要求：

(1) 函数命名: PCA

(2) 输入参数：

Dat: 样本数据集（维度 $n \times m$，样本量 n，变量 m）

max.k: 主成分的最大个数（取值为正整数）

(3) 输出: (a) 前 max.k 个主成分的方差 (一个长度是 max.k 的向量) (b) 主成分系数矩阵（$m\times$max.k 的系数矩阵）

测试：对“NBA 数据”进行测试，设置 max.k = 10，打印前 max.k 个主成分的方差，绘制碎石图 screeplot。

注意：要求代码简洁、高效、可读性强；结果正确无误。提交 HTML 格式的代码文件。

提交时间：12 月 28 日，晚 20:00 之前。请预留一定的时间，迟交作业扣 3 分，作业抄袭 0 分。