1 Chapter 15：奇异值分解

15.2

试求矩阵

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

的奇异值分解并写出其外积展开式

求对称矩阵

$$A^T A = \begin{bmatrix} 5 & 11 \\ 11 & 25 \end{bmatrix}$$

其特征值和特征向量（顺便单位化）为 $\lambda_1 = 15 + \sqrt{221}, \lambda_2 = 15 - \sqrt{221}$

后面很难算，于是用 Python 计算（直接调库： `U, S, V = numpy.linalg.svd(A.T)` ）

```python
U = [[-0.40455358  0.9145143 ]
     [-0.9145143  -0.40455358]]
S = [5.4649857  0.36596619]
V = [[-0.81741556 -0.57604844  0.          0.        ]
     [ 0.57604844 -0.81741556  0.          0.        ]
     [ 0.          0.          1.          0.        ]
     [ 0.          0.          0.          1.        ]]
```

其外积展开式为 $A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T =$

```python
  [[ 1.80720735  1.27357371 -0.         -0.        ]
   [ 4.08528566  2.87897923 -0.         -0.        ]]
+ [[ 0.19279265 -0.27357371  0.          0.        ]
   [-0.08528566  0.12102077 -0.         -0.        ]]
```

15.4

证明任何一个秩为 1 的矩阵可写成两个向量的外积形式，并给出实例

Proof

由奇异值分解， $A = U\Sigma V^T$ ，其中 Σ 为对角矩阵且只有第一个元素不为 0 ，由外积展开式 $A = \sigma_1 u_1 v_1^T$ ，这就是所求的两个向量的外积形式。举例

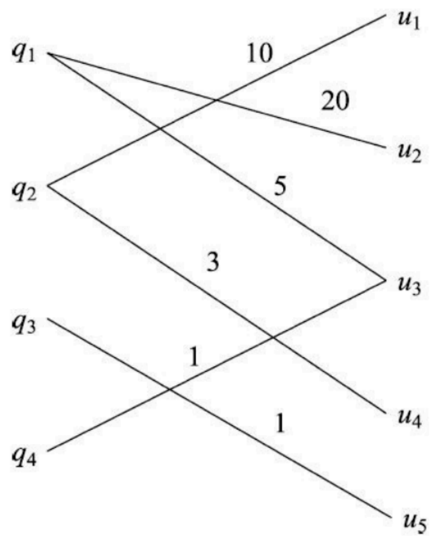
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \lambda_1 = 4, \lambda_2 = 0, \quad A = 2 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

□

15.5

搜索中的点击数据记录用户搜索时提交的查询语句，点击的网页 URL，以及点击的次数，构成一个二部图，其中一个结点集合 $\{q_i\}$ 表示查询，另一个结点集合 $\{u_j\}$ 表示 URL，边表示点击关系，边上的权重表示点击次数。

图 15.2 是一个简化的点击数据例。点击数据可以由矩阵表示，试对该矩阵进行奇异值分解，并解释得到的三个矩阵所表示的内容。



由题意，对应矩阵为

$$A = \begin{bmatrix} 0 & 20 & 5 & 0 & 0 \\ 10 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 100 & 0 & 0 & 30 & 0 \\ 0 & 400 & 100 & 0 & 0 \\ 0 & 100 & 26 & 0 & 0 \\ 30 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

解得特征值为

$$\lambda_1 = 213 + \sqrt{44969}, \lambda_2 = 109, \lambda_3 = 1, \lambda_4 = 213 - \sqrt{44969}, \lambda_5 = 0$$

后面很难算，于是用 Python 计算

```python
U = [[-0.99993  0.       -0.       -0.01179]
     [ 0.        1.        0.        0.     ]
     [-0.        0.       -1.        0.     ]
     [-0.01179  0.        0.        0.99993]]
S = [20.61696 10.44031  1.        0.97008]
V = [[ 0.        0.95783  0.       -0.        0.28735]
     [-0.97001 -0.        0.       -0.24307 -0.     ]
     [-0.24307  0.        0.        0.97001  0.     ]
     [ 0.        0.28735  0.        0.       -0.95783]
     [ 0.        0.       -1.        0.        0.     ]]
```

将 V 的每一列看作是一个 URL，因为第五个奇异值为 0，根据外积展开式，去掉第五列。每列的各元素表示该维度的特征对当前 URL 的重要性：如第一列的表示 URL1 的第二个维度的特征较为显著，第二列表示 URL2 的第一个维度的特征比较显著。

将 U 的每一列看作是一个查询，每列的各元素值表示该维度的特征对当前查询的重要性，如第一个查询倾向于第一个维度的特征比较显著的 URL，结合 V 矩阵可以知道，URL2 的第一个维度的特征比较显著，则第一个查询倾向于 URL2。

通过奇异值分解将 Query 和 URL 之间的维度特征提取了出来。

2 Chapter 16：主成分分析

16.1

对以下样本数据进行主成分分析：

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

样本均值和协方差矩阵为

$$\overline{x} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}, S = \begin{bmatrix} \frac{16}{5} & \frac{17}{5} \\ \frac{17}{5} & \frac{16}{5} \end{bmatrix}$$

矩阵规范化为

$$X' = \begin{bmatrix} -\frac{\sqrt{5}}{2} & -\frac{\sqrt{5}}{4} & -\frac{\sqrt{5}}{4} & 0 & \frac{\sqrt{5}}{4} & \frac{3\sqrt{5}}{4} \\ -\frac{3}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

于是样本相关矩阵为

$$R = \begin{bmatrix} 1 & \frac{17\sqrt{5}}{40} \\ \frac{17\sqrt{5}}{40} & 1 \end{bmatrix}$$

求解特征值 $\lambda_1 = 1 + \frac{17\sqrt{5}}{40}, \lambda_2 = 1 - \frac{17\sqrt{5}}{40}$，单位特征向量 $a_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, a_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

$\frac{\lambda_1}{\lambda_1+\lambda_2} \approx 0.9752$ 已经足够大，故只取第一主成分

$$y_1 = a_1^T \boldsymbol{x} = \frac{\sqrt{2}}{2} x_1 + \frac{\sqrt{2}}{2} x_2$$

一共 $k(=1)$ 个主成分 y_1 对原有变量（规范化后）的贡献率为

$$\nu_1 = \rho^2(x_1, y_1) = \lambda_1 a_{11}^2 = 0.9752$$
$$\nu_2 = \rho^2(x_2, y_1) = \lambda_1 a_{21}^2 = 0.9752$$

将 $N = 6$ 个样本代入对应的主成分式，得到主成分矩阵

$$Y = a_1^T X = \begin{bmatrix} -\frac{\sqrt{10}+3\sqrt{2}}{4} & -\frac{\sqrt{10}+2\sqrt{2}}{8} & -\frac{\sqrt{10}}{8} & 0 & \frac{\sqrt{10}+2\sqrt{2}}{8} & \frac{3\sqrt{10}+6\sqrt{2}}{8} \end{bmatrix}$$

16.2

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。

Proof

设 $x_1, x_2, ..., x_n$ 是从总体 X 中得到的随机样本，记 X 的期望为 $E(X) = E(x_i) = \mu$，协方差矩阵为 $\Sigma = \mathrm{Cov}(X, X) = \mathrm{Cov}(x_i, x_i)$

考虑样本独立性，有 $\mathrm{Cov}(x_i, x_j) = 0, i \neq j \in \{1, 2, ..., n\}$，则样本均值向量的期望和协方差满足:

$$
\begin{aligned}
E(\overline{x}) &= \frac{1}{n} E\left(\sum_{i=0}^{n} x_i\right) = \frac{n\mu}{n} = \mu, \\
\mathrm{Cov}(\overline{x}, \overline{x}) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=0}^{n} x_i, \frac{1}{n}\sum_{j=0}^{n} x_j\right) \\
&= \frac{1}{n^2}\sum_{i=0}^{n}\sum_{j=0}^{n} \mathrm{Cov}(x_i, x_j) \\
&= \frac{1}{n^2}\sum_{i=0}^{n} \mathrm{Cov}(x_i, x_i) \\
&= \frac{1}{n}\Sigma
\end{aligned}
$$

于是有

$$
\begin{aligned}
E(S) &= E\left[\frac{1}{n-1}\sum_{i=0}^{n}(x_i - \overline{x})(x_i - \overline{x})^T\right] \\
&= E\left[\frac{1}{n-1}\sum_{i=0}^{n}\left(((x_i - \mu) - (\overline{x} - \mu))((x_i - \mu) - (\overline{x} - \mu))^T\right)\right] \\
&= \frac{1}{n-1}\sum_{i=0}^{n} E\left[(x_i - \mu)(x_i - \mu)^T\right] - \frac{n}{n-1} E\left[\overline{x}\overline{x}^T - \overline{x}\mu^T - \mu\overline{x}_T + \mu\mu^T\right] \\
&= \frac{1}{n-1}\sum_{i=0}^{n} \mathrm{Cov}(x_i, x_i) - \frac{n}{n-1} E\left[(\overline{x} - \mu)(\overline{x} - \mu)^T\right] \\
&= \frac{n}{n-1}\Sigma - \frac{n}{n-1}\mathrm{Cov}(\overline{x}, \overline{x}) \\
&= \frac{n}{n-1}\Sigma - \frac{1}{n-1}\Sigma \\
&= \Sigma
\end{aligned}
$$

□

16.3

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：

$$\min_{L} \|X - L\|_F$$
$$s.t. \operatorname{rank}(L) \leq k$$

这里 F 是弗罗贝尼乌斯范数， k 是主成分个数。试问为什么？

Proof

记 $X' = \frac{1}{\sqrt{n-1}} X^T$ ，有

$$X'^T X' = \frac{1}{n-1} X X^T = R$$

即主成分分析对 R 求特征值和特征向量，等价于对 X' 求奇异值分解。

回忆奇异值分解相关定理，矩阵 A 的秩为 k 的截断奇异值分解为弗罗贝尼乌斯范数在所有秩不超过 k 的矩阵中的最优近似。即 $A \approx A_k = U_k \Sigma_k V_k$ 满足

$$\|A - A_k\| = \min_{S} \|A - S\|_F \quad s.t. \operatorname{rank}(S) \leq k$$

将式中的 A 代换为 X' ，再代换为 X ，即可得到：
主成分分析等价于 $\frac{1}{\sqrt{n-1}} X^T \approx \frac{1}{\sqrt{n-1}} X_k^T = U_k \Sigma_k V_k$ 满足

$$\left\| \frac{1}{\sqrt{n-1}} X^T - \frac{1}{\sqrt{n-1}} X_k^T \right\| = \min_{S} \left\| \frac{1}{\sqrt{n-1}} X^T - S \right\|_F \quad s.t. \operatorname{rank}(S) \leq k$$

这个最优化问题跟题设显然是一样的，只相差一个转置与系数 $\frac{1}{\sqrt{n-1}}$ （注意到 L 转置和乘系数不改变秩）。 □