

# 奇异值分解

颜文斌  
清华大学

# 内容提要

- 奇异值分解
- 主成分分析
- 奇异值分解和几何

# 矩阵对角化

- 矩阵对角化有很多应用：简化计算、解方程.....
- 不是所有矩阵都可以对角化的
  - 可对角化矩阵例：对称矩阵
- 特征值和特征向量只使用于方阵
- 对于一般的 $m \times n$ 矩阵 $A$ ，有没有类似的操作？
  - 回忆方程 $Ax = b$ 不一定有解，但是 $A^T Ax = A^T b$ 有解
  - 考虑方阵 $A^T A$ 和 $AA^T$ ，他们都是**半正定**矩阵，所以可以对角化而且特征值大于等于0

# 对角化 $A^T A$ 和 $AA^T$

- $m \times n$  矩阵  $A$ ,  $A^T A$  和  $AA^T$  都是**半正定**矩阵,
  - 证明:  $\forall \mathbf{x}, \mathbf{x}^T A^T A \mathbf{x} = \|A \mathbf{x}\|^2 \geq 0$ , 所以  $A^T A$  是半正定的。  $AA^T$  同理
- $A^T A$  和  $AA^T$  都可以对角化
  - $A^T A = V \Lambda_1 V^T$ ,  $AA^T = U \Lambda_2 U^T$
  - $V^T A^T A V = (AV)^T A(V) = \Lambda_1$
  - $U^T A A^T U = (U^T) A (A^T U) = \Lambda_2$
- 猜测: 找到正交矩阵  $U$  和  $V$  使得  $m \times n$  矩阵  $U^T A V$  可以写成  $\Sigma$ ? 其中  $\Sigma$  是某种意义上的“对角”矩阵

# 奇异值 (singular value)

- $m \times n$  的实矩阵  $A$ ,  $A^T A$  是一个  $n \times n$  的对称矩阵,  $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n\}$  是由  $A^T A$  的特征向量构成的  $\mathbb{R}^n$  中的正交归一基, 对应的实特征值为  $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ , 假设  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$
- $A^T A$  的所有特征值都非负:  $\|A\mathbf{q}_i\|^2 = \mathbf{q}_i^T A^T A \mathbf{q}_i = \lambda_i \geq 0$
- 矩阵  $A$  的奇异值定义为  $A^T A$  的特征值的平方根:  $\sigma_i = \sqrt{\lambda_i}$ 
  - $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$
  - $\sigma_i = \|A\mathbf{q}_i\|$

# 例

- 求以下矩阵的奇异值

$$A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} \quad A^T A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix} \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}$$

- 特征方程：
$$\det \begin{bmatrix} 80 - \lambda & 100 & 40 \\ 100 & 170 - \lambda & 140 \\ 40 & 140 & 200 - \lambda \end{bmatrix}$$
$$= -\lambda(\lambda - 90)(\lambda - 360)$$

- 特征值：360, 90, 0。奇异值： $6\sqrt{10}$ ,  $3\sqrt{10}$ , 0

# $A$ 的秩和 $A^T A$ 的秩

- 定理：  $m \times n$  矩阵  $A$ ,  $\text{rank}(A^T A) = \text{rank}(A)$
- 证明：
  - $\text{rank}(A^T A) = n - \dim N(A^T A)$ ,  $\text{rank}(A) = n - \dim N(A)$ , 所以我们只需证明  $\dim N(A^T A) = \dim N(A)$
  - 如果  $\mathbf{x} \in N(A)$ , 则  $A\mathbf{x} = 0$ , 等号两边同时左乘  $A^T$  得到  $A^T A\mathbf{x} = 0$ , 所以  $\mathbf{x} \in N(A^T A)$
  - 如果  $\mathbf{x} \in N(A^T A)$ , 则  $A^T A\mathbf{x} = 0$ , 等号两边同时左乘  $\mathbf{x}^T$  得到  $\mathbf{x}^T A^T A\mathbf{x} = 0$ , 所以  $\|A\mathbf{x}\|^2 = 0$ , 所以  $A\mathbf{x} = 0$ , 所以  $\mathbf{x} \in N(A)$
  - $N(A)$  和  $N(A^T A)$  存在一一映射, 所以  $\dim N(A^T A) = \dim N(A)$ ,
- 推论： $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(AA^T) = \text{rank}(A^T A)$

# 非零奇异值的数量

- 定理： $m \times n$ 的实矩阵 $A$ 的非零奇异值 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ 的个数 $r$ 等于 $A$ 的秩 $\text{rank}(A)$
- 证明：
  - 设 $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ 为 $\mathbb{R}^n$ 中可以把 $A^T A$ 对角化的正交归一基， $\{\lambda_1, \cdots, \lambda_n\}$ 为对应的特征值，则 $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_n\}$ 是一个**正交向量集合**，也就是说 $\forall i \neq j$ ,  
 $(A\mathbf{v}_i)^T (A\mathbf{v}_j) = \mathbf{v}_i^T A^T A \mathbf{v}_j = \mathbf{v}_i^T (\lambda_j \mathbf{v}_j) = 0$
  - 假设 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ 是所有正特征值， $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_n = 0$ 是所有0特征值，则 $\{A\mathbf{v}_{r+1}, \cdots, A\mathbf{v}_n\}$ 都是零向量
  - 因为 $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ 为 $\mathbb{R}^n$ 中可以把 $A^T A$ 对角化的正交归一基，所以对于任意的 $\mathbf{x} \in \mathbb{R}^n$ ， $\mathbf{x}$ 可以写成 $\mathbf{x} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n$ 。



# 非零奇异值的数量（续）

- 定理： $m \times n$ 的实矩阵 $A$ 的非零奇异值 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ 的个数 $r$ 等于 $A$ 的秩 $\text{rank}(A)$
- 证明（续）：
  - 因为 $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ 为 $\mathbb{R}^n$ 中可以把 $A^T A$ 对角化的正交归一基，所以对于任意的 $\mathbf{x} \in \mathbb{R}^n$ ， $\mathbf{x}$ 可以写成 $\mathbf{x} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n$ 。
  - 我们需要证明 $A$ 的秩为 $r$ ，只需要证明列空间 $C(A)$ 的维度为 $r$ ，需要找到列空间的一组基
  - 因为 $C(A)$ 是由 $A$ 所有列的线性组合得到的，所以 $\forall \mathbf{y} \in C(A)$ ， $\mathbf{y} = A\mathbf{x} = c_1 A\mathbf{v}_1 + \cdots + c_n A\mathbf{v}_n = c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r$ ，也就是说 $C(A)$ 中的任何向量都可以写成 $\{A\mathbf{v}_1, \dots, A\mathbf{v}_r\}$ 的线性组合
  - 所以 $\{A\mathbf{v}_1, \dots, A\mathbf{v}_r\}$ 是 $C(A)$ 的一组正交基，所以 $r = \text{rank}(A)$

# 奇异值分解：陈述

- 广义“对角”矩阵： $m \times n$ 矩阵 $\Sigma$

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

- $D$ 是一个 $r \times r$ 的对角矩阵， $\Sigma$ 所有大于 $r$ 的行和列都是0

- **定理（奇异值分解）**： $m \times n$ 矩阵 $A$ 的秩为 $r$ 。则存在一个形状如上的 $m \times n$ 矩阵 $\Sigma$ 且 $D$ 的对角元是 $A$ 的前 $r$ 个（非零）的奇异值， $m \times m$ 的正交矩阵 $U$ 和 $n \times n$ 的正交矩阵 $V$ ，而且以上矩阵满足关系

$$A = U \Sigma V^T$$

- 可以用两个正交矩阵把任意矩阵 $A$ 变成简单形式
- 考虑 $\text{rank}(A)=1$ ，矩阵可以写成两个向量的乘积（压缩）

# 奇异值分解：证明

- 证明：直接构造相应的矩阵
  - $A^T A$ 是对称矩阵，由谱定理可知存在一组 $\mathbb{R}^n$ 中的正交归一基，将 $A^T A$ 对角化
  - 设 $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ 为 $\mathbb{R}^n$ 中可以把 $A^T A$ 对角化的正交归一基， $\{\lambda_1, \dots, \lambda_n\}$ 为对应的特征值。由之前定理 $\{\lambda_1, \dots, \lambda_r\}$ 是正特征值， $\{\lambda_{r+1}, \dots, \lambda_n\}$ 是零特征值
  - 因为 $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ 之间是正交的，则 $\forall i \neq j$ ,  $(A\mathbf{v}_i)^T(A\mathbf{v}_j) = \mathbf{v}_i^T A^T A \mathbf{v}_j = \mathbf{v}_i^T(\lambda_j \mathbf{v}_j) = 0$ ，所以 $\{A\mathbf{v}_1, \dots, A\mathbf{v}_r\}$ 之间也是正交的， $\{A\mathbf{v}_{r+1}, \dots, A\mathbf{v}_n\}$ 是零向量
  - 令 $u_i = \frac{1}{\|A\mathbf{v}_i\|} A\mathbf{v}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$ ,  $i = 1, \dots, r$ ，则 $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 是 $C(A)$ 的一组正交归一基。 $\sigma_i = \sqrt{\lambda_i}$

# 奇异值分解：证明（续）

- 证明：直接构造相应的矩阵

- 令  $\mathbf{u}_i = \frac{1}{\|A\mathbf{v}_i\|} A\mathbf{v}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$ ,  $i = 1, \dots, r$ , 则  $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$  是  $C(A)$  的一组正交归一基。  $\sigma_i = \sqrt{\lambda_i}$
- 再设  $\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\}$  是  $N(A^T)$  中的一组正交归一基。因为  $N(A^T)$  是  $C(A)$  的正交补, 则  $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$  是  $\mathbb{R}^m$  的一组正交归一基
- 设矩阵  $U = (\mathbf{u}_1, \dots, \mathbf{u}_m)$  和  $V = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ ,  $U$  和  $V$  都是正交的
- $AV = (A\mathbf{v}_1, \dots, A\mathbf{v}_n) = (\sigma_1 \mathbf{u}_1, \dots, \sigma_r \mathbf{u}_r, 0, \dots, 0) = U\Sigma$
- 所以  $A = U\Sigma V^T$ , 我们得到了  $A$  的奇异值分解  $A = U\Sigma V^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$

# 奇异值分解：例

- 矩阵  $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$  的奇异值分解
- $A^T A = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}$
- $A^T A$  的特征值：360, 90, 0。  $A$  的奇异值： $6\sqrt{10}$ ,  $3\sqrt{10}$ , 0
- $V = \begin{bmatrix} 1/3 & -2/3 & 2/3 \\ 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \end{bmatrix}$

# 奇异值分解——例

- 矩阵  $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$  的奇异值分解
- $A$  的奇异值：  $6\sqrt{10}$ ,  $3\sqrt{10}$ ,  $0$
- $V = \begin{bmatrix} 1/3 & -2/3 & 2/3 \\ 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \end{bmatrix}$ ,  $AV = \begin{bmatrix} 4 & 11 \\ 8 & 7 \end{bmatrix}$
- $\mathbf{u}_1 = \frac{1}{\sigma_1} A \mathbf{v}_1 = \begin{bmatrix} 3/\sqrt{10} \\ 1/\sqrt{10} \end{bmatrix}$ ,  $\mathbf{u}_2 = \frac{1}{\sigma_2} A \mathbf{v}_2 = \begin{bmatrix} 1/\sqrt{10} \\ -3/\sqrt{10} \end{bmatrix}$
- $U = (\mathbf{u}_1 \quad \mathbf{u}_2)$

# 奇异值分解——例

• 矩阵  $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$  的奇异值分解  $A = U\Sigma V^T$

•  $A$  的奇异值：  $6\sqrt{10}$ ,  $3\sqrt{10}$ ,  $0$ ,  $\Sigma = \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix}$

•  $V = \begin{bmatrix} 1/3 & -2/3 & 2/3 \\ 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \end{bmatrix}$ ,  $U = \begin{bmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{bmatrix}$

•  $A = \begin{bmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ -2/3 & -1/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \end{bmatrix}$

# $A$ 和 $A^T$ 的奇异值分解

- 练习：矩阵 $A^T = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix}$ 的奇异值分解
- 思考：已知 $A$ 的奇异值分解，求矩阵 $A^T$ 的奇异值分解？
- 结论：若 $A$ 的奇异值分解为 $U\Sigma V^T$ ， $A^T$ 的奇异值分解为 $A^T = V\Sigma^T U^T$
- 推论： $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ 是矩阵 $AA^T$ 的特征向量
- 证明：
  - $AA^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T$ ，所以 $AA^T U = U\Sigma\Sigma^T$
- $V$ 和 $U$ 分别是将 $A^T A$ 和 $AA^T$ 对角化的正交矩阵



# $A$ 和 $A^T$ 的奇异值分解

- **定理：**  $A^T A$ 和 $AA^T$ 的非零特征值都相同
- **证明：**可以直接用 $A$ 的奇异值分解证明，我们这里直接证明
  - 假设 $x_i$ 是 $A^T A$ 的特征值为 $\lambda_i \neq 0$ 的特征向量， $(\lambda_i I - A^T A)x_i = 0$
  - 等号两端同时左乘： $0 = A(\lambda_i I - A^T A)x_i = (\lambda_i A - AA^T A)x_i = (\lambda_i I - AA^T)(Ax_i)$ 。又因为 $x^T A^T A x = \lambda_i x^T x > 0$ ，所以 $Ax_i \neq 0$ 。所以 $Ax_i$ 是 $AA^T$ 的特征值为 $\lambda_i$ 的特征向量
  - 同理，如果 $x_i$ 是 $AA^T$ 的特征值为 $\lambda_i \neq 0$ 的特征向量， $A^T x_i$ 是 $A^T A$ 的特征值为 $\lambda_i$ 的特征向量
  - 这样，我们就有 $A^T A$ 和 $AA^T$ 非零特征值和非零特征向量的一一对应

# 应用：四个子空间的正交归一基

- $A = U\Sigma V^T$ 
  - $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_r\}$  是  $C(A^T)$  的正交归一基,  $V_r = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_r)$
  - $\{\boldsymbol{v}_{r+1}, \dots, \boldsymbol{v}_n\}$  是  $N(A)$  的正交归一基,  $V_{n-r} = (\boldsymbol{v}_{r+1}, \dots, \boldsymbol{v}_n)$
  - $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_r\}$  是  $C(A)$  的正交归一基,  $U_r = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_r)$
  - $\{\boldsymbol{u}_{r+1}, \dots, \boldsymbol{u}_m\}$  是  $N(A^T)$  的正交归一基,  $U_{n-r} = (\boldsymbol{u}_{r+1}, \dots, \boldsymbol{u}_m)$
- 证明？

# 应用：数据压缩

- 假设 $\text{rank}(A) < \min(m, n)$ , 则

$$A = (U_r, U_{m-r}) \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} = U_r D V_r^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

- 可以用 $U_r, D, V_r$ 这三个矩阵的 $r(m+1+n)$ 个分量完全决定 $A$  ( $mn$ )
- **图像压缩**
  - 先考虑黑白图片, 可以用一个 $m \times n$ 的矩阵描述, 每个元素是该像素的灰度 (0-255之间的整数, 0是黑, 255是白)
  - 如果 $r(m+1+n) < mn$ , 我们可以只储存或者传输 $U_r, D, V_r$  (无损)。例如矩阵秩为1的时候我们只需要储存一个行向量和一个列向量
  - 甚至可以把很小的奇异值当成零忽略, 进一步压缩图片 (有损)

# 数据压缩的误差估计

- $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ , 我们需要估算取前k个奇异值得到的矩阵和真实矩阵之间的误差

- 设  $A(k) = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ , 则  $\delta A = A - A(k) = \sum_{l=k+1}^r \sigma_l \mathbf{u}_l \mathbf{v}_l^T$

- $\delta A$ 的ij分量的绝对值:

$$|\delta A_{ij}| = \left| \sum_{l=k+1}^r \sigma_l (\mathbf{u}_l)_i (\mathbf{v}_l^T)_j \right| \leq \sum_{l=k+1}^r \sigma_l |(\mathbf{u}_l)_i| |(\mathbf{v}_l^T)_j|$$

- $\mathbf{u}_l$ ,  $\mathbf{v}_l$ 的长度都是1, 所以  $|(\mathbf{u}_l)_i| \leq 1$ ,  $|(\mathbf{v}_l^T)_j| \leq 1$

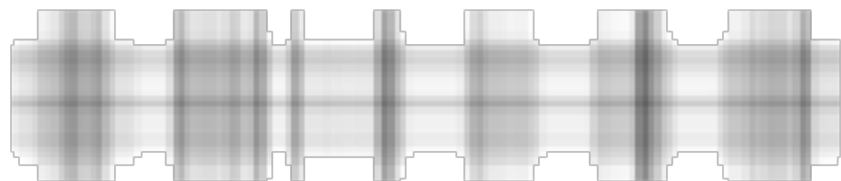
- 所以  $|\delta A_{ij}| \leq \sum_{l=k+1}^r \sigma_l$

- 结论: 误差由忽略的奇异值控制, 取得奇异值越多误差越小

# 图像压缩例

- 一个 $128 \times 623 = 79744$ 像素的黑白图片
  - 秩为24。奇异值分解后有效信息 $24 \times (128 + 623) = 18024$
  - 取前 $k$ 个奇异值的近似

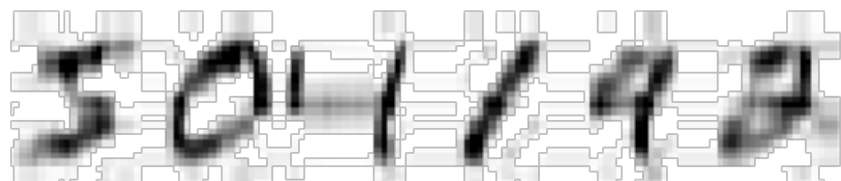
504192



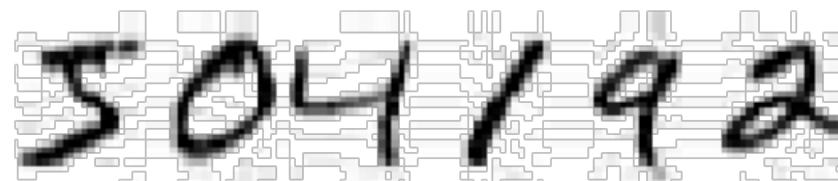
k=1



k=3



k=5



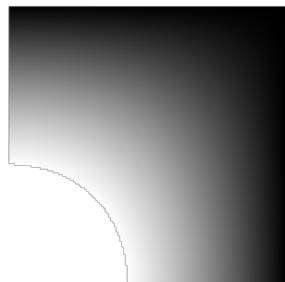
k=10

# 图像压缩例

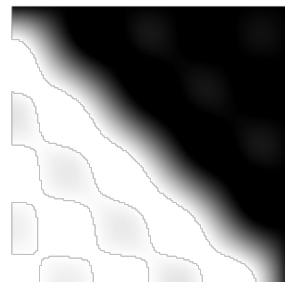
- 考虑一个 $300 \times 300$ 的下三角矩阵，对角线和下面的元素全是255



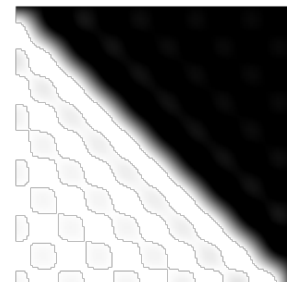
原图：300x300



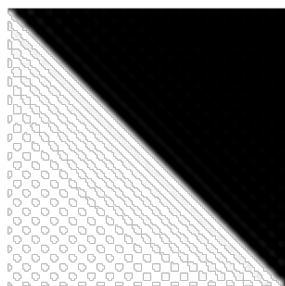
前1个奇异值 ~ 600



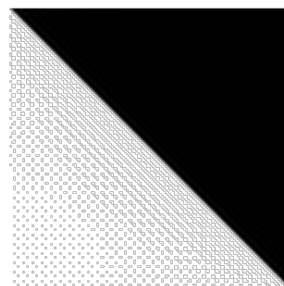
前5个 ~ 3000



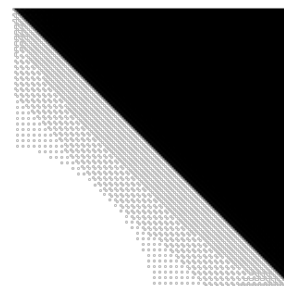
前10个 ~ 6000



前30个 ~ 18000



前60个 ~ 36000



前100个 ~ 60000

# 小结

1. The SVD factors  $A$  into  $U\Sigma V^T$ , with  $r$  singular values  $\sigma_1 \geq \dots \geq \sigma_r > 0$ .
2. The numbers  $\sigma_1^2, \dots, \sigma_r^2$  are the nonzero eigenvalues of  $AA^T$  and  $A^T A$ .
3. The orthonormal columns of  $U$  and  $V$  are eigenvectors of  $AA^T$  and  $A^T A$ .
4. Those columns hold orthonormal bases for the four fundamental subspaces of  $A$ .
5. Those bases diagonalize the matrix:  $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$  for  $i \leq r$ . This is  $\mathbf{A}\mathbf{V} = \mathbf{U}\Sigma$ .
6.  $A = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \dots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$  and  $\sigma_1$  is the maximum of the ratio  $\|A\mathbf{x}\| / \|\mathbf{x}\|$ .

# 内容提要

- 奇异值分解
- 主成分分析
- 奇异值分解和几何



# 统计知识

- 假设一组数据来源于n个样本 $\{\mu_1, \mu_2, \dots, \mu_n\}$
- 例：所有同学的期中考试成绩
- 平均值 (mean) :  $\bar{\mu} = \frac{\sum_i \mu_i}{n}$
- 标准差 (standard deviation) :  $\sigma = \sqrt{\frac{\sum_{i=1}^n (\mu_i - \bar{\mu})^2}{n-1}}$ 
  - n-1个自由度，因为平均值也是一个自由度
  - 数据的分散程度，标准差越大，数据越分散：

# 统计知识

- 假设n个样本，每个样本i我们得到两个数据 $\mu_i$ 和 $\rho_i$
- 例：所有同学的期中考试成绩和平时作业成绩
- **协方差**：
$$\text{cov}(\mu, \rho) = \frac{\sum_{i=1}^n (\mu_i - \bar{\mu})(\rho_i - \bar{\rho})}{n-1}$$
  - 描述了 $\mu$ 和 $\rho$ 之间的相关性
  - $\text{cov}(\mu, \rho) > 0$ 正相关， $\text{cov}(\mu, \rho) < 0$ 负相关

# 协方差矩阵

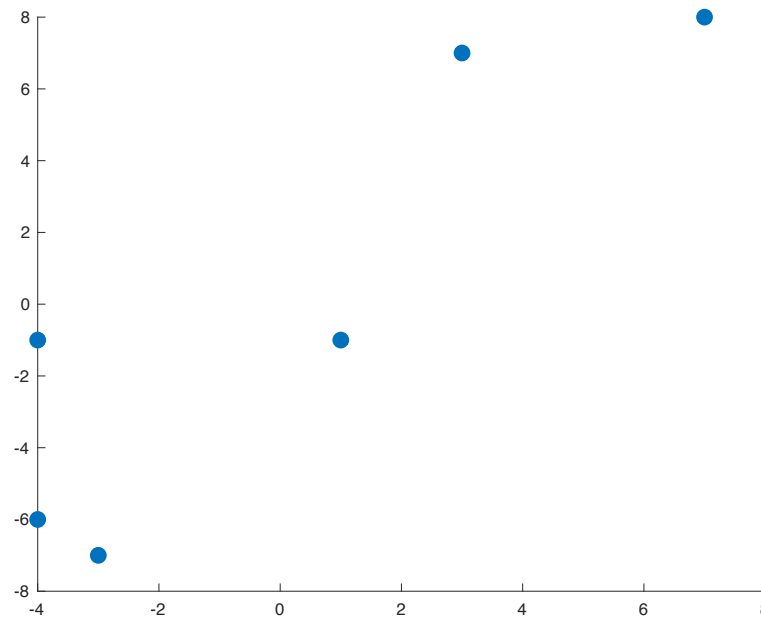
- 将数据存在一个 $m \times n$ 的矩阵 $A_0$ 中，每一行对应一种数据，每一列代表一个样本
- 矩阵 $A$ :
  - 由 $A_0$ 的每一个元素减去它所在行的平均值得到 $A_{ij} = (A_0)_{ij} - \frac{\sum_{k=1}^n (A_0)_{ik}}{n}$
  - $A$ 中每一行的数据都是以0为中心分布
- 协方差矩阵 (**covariance matrix**) :  $S = \frac{AA^T}{n-1}$ 
  - 样本**方差** :  $S_{ii} = \sigma_i^2$ ，第 $i$ 种数据的**标准差**平方。  $S_{ij}$  : 第 $i$ 种和第 $j$ 种数据的协方差
  - 总方差 (**total variance**) :  $\text{tr}S = \sum_i S_{ii} = \sum_i \sigma_i^2$

例：

- 6个同学的数学和历史成绩（已经减去平均值）

$$A = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{bmatrix}$$

$$S = \frac{AA^T}{5} = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix}$$



# 主成分分析 (PCA)

- 一般来说数据i和数据j可能会有相关，也就是说它们之间的协方差 $S_{ij}$ 不等于0
- 主成分分析：找到原有数据的一系列线性组合作为新的数据，新数据之间的协方差为0
  - 奇异值分解： $A = U\Sigma V^T$
- 定义新的数据矩阵： $B = U^T A = \Sigma V^T$ ， $B$ 的协方差矩阵为 $\frac{\Sigma V^T V \Sigma}{n-1} = \frac{\Sigma^2}{n-1}$ ，因为 $\Sigma^2$ 是对角矩阵， $B$ 的数据之间协方差为0
  - 总方差不变： $\text{tr} \frac{BB^T}{n-1} = \text{tr} \frac{U^T A A^T U}{n-1} = \frac{\text{tr} U^T A A^T U}{n-1} = \frac{\text{tr} A A^T U U^T}{n-1} = \frac{\text{tr} A A^T}{n-1}$

# 主成分分析

- 原数据矩阵： $A = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ ，第*i*列向量 $\mathbf{a}_i$ 对应样本*i*的数据
- 新数据矩阵： $B = (\mathbf{b}_1, \dots, \mathbf{b}_n) = U^T A = (U^T \mathbf{a}_1, \dots, U^T \mathbf{a}_n)$ 
  - $B$ 的第*i*列向量 $\mathbf{b}_i$ 对应样本*i*的数据，这些数据由 $\mathbf{a}_i$ 的分量决定： $\mathbf{b}_i = U^T \mathbf{a}_i$
  - 因为 $U$ 是正交矩阵 $UU^T = U^T U = I$ ， $\mathbf{a}_i = U \mathbf{b}_i$ 。
- $B$ 的协方差矩阵 $\frac{BB^T}{n-1} = \frac{\Sigma V^T V \Sigma}{n-1} = \frac{\Sigma^2}{n-1}$ 是对角矩阵
  - 非对角元为0，新数据之间互不相关
  - 新数据的方差就是 $A$ 的奇异值平方。
  - 原方差 $\frac{(AA^T)_{ii}}{n-1} = \frac{(UBB^T U^T)_{ii}}{n-1} = \frac{\sum_{k=1}^r \sigma_k^2 U_{ik}^2}{n-1}$

# 主成分分析

- 原数据矩阵：  $A = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ ，第*i*列向量 $\mathbf{a}_i$ 对应样本*i*的数据
- 新数据矩阵：  $B = (\mathbf{b}_1, \dots, \mathbf{b}_n) = U^T A = (U^T \mathbf{a}_1, \dots, U^T \mathbf{a}_n)$
- $B$ 的协方差矩阵  $\frac{BB^T}{n-1} = \frac{\Sigma V^T V \Sigma}{n-1} = \frac{\Sigma^2}{n-1}$  是对角矩阵
  - 新数据的方差就是 $A$ 的奇异值平方。
  - $A$ 的非零奇异值的数量是 $A$ 的秩 $r$ ， $r+1$ 到 $m$ 的新数据的方差是0
  - 所有的数据都在 $\mathbb{R}^m$ 的 $m-r$ 个平面 $\sum_{j=1}^m U_{ji} x_j = 0, i = r+1, \dots, m$ 的交集上
  - 所有数据点分布在一个 $r$ 维的空间中，这个空间由 $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 张成（ $C(A)$ 的正交归一基）
  - 如果第*i*个奇异值很接近0，说明数据很靠近平面 $\sum_{j=1}^m U_{ji} x_j = 0$ 。

# 主成分： $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$

- 原数据矩阵： $A = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ ，第 $i$ 列向量 $\mathbf{a}_i$ 对应样本 $i$ 的数据
- $A$ 的奇异值分解： $A = U\Sigma V^T$
- 新数据矩阵： $B = (\mathbf{b}_1, \dots, \mathbf{b}_n) = U^T A = (U^T \mathbf{a}_1, \dots, U^T \mathbf{a}_n)$
- 所有数据点分布在一个 $r$ 维的空间中，这个空间由 $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 张成（ $C(A)$ 的正交归一基）
  - $\mathbf{u}_1$ 是所有数据变化最大的方向（对应的方差最大）， $\mathbf{u}_2$ 次之。。。
  - $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 被称作主成分（principle component）
  - 主成分是描述整组数据最重要的线性组合，而且互相独立
  - 由于 $r$ 小于等于 $m$ ，所以虽然每个样本测了 $m$ 个数据，里面只有 $r$ 个是独立的

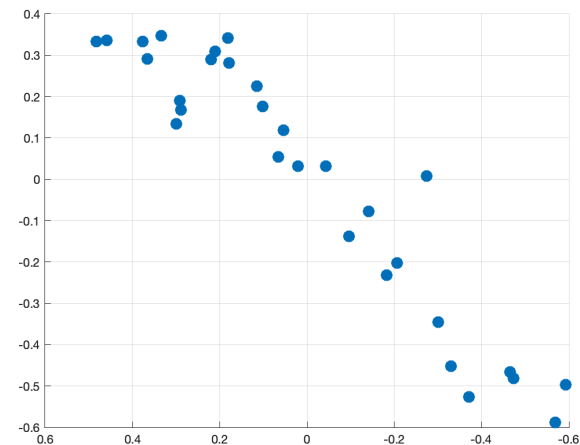
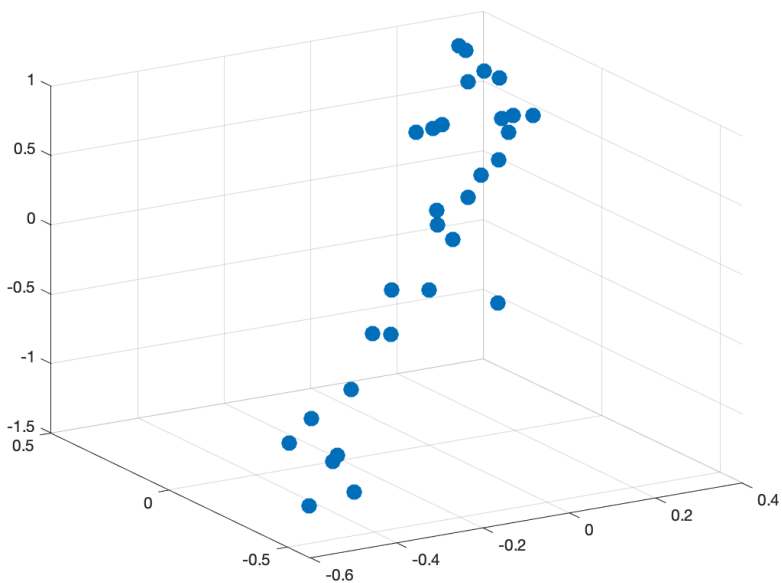


# $\{\sigma_1 \mathbf{v}_1, \dots, \sigma_r \mathbf{v}_r\}$ 的意义

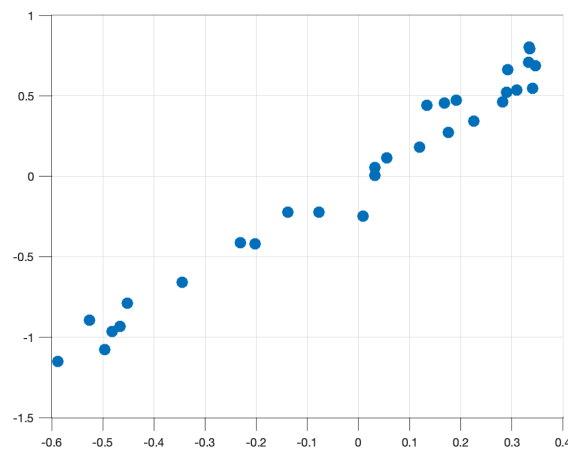
- 原数据矩阵： $A = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ ，第*i*列向量 $\mathbf{a}_i$ 对应样本*i*的数据
- $A$ 的奇异值分解： $A = U\Sigma V^T = \sum_{k=1}^r \sigma_k \mathbf{u}_k \mathbf{v}_k^T$ 
  - 主成分 $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 是描述整组数据最重要的线性组合，而且互相独立
- $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ 是什么？
  - $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ 都是*n*维向量，每个分量对应一个样本
  - 第一主成分的数值： $\mathbf{u}_1^T A = \mathbf{u}_1^T (\sum_{k=1}^r \sigma_k \mathbf{u}_k \mathbf{v}_k^T) = \sigma_1 \mathbf{v}_1^T$
  - $\sigma_1 \mathbf{v}_1$ 的第*i*个分量是第*i*个样本的第一主成分的值，同理 $\sigma_j \mathbf{v}_j$ 的第*i*个分量是第*i*个样本的第*j*个主成分的值
  - $\mathbf{v}_j$ 是单位向量，所以每个分量的绝对值小于等于1，数据的分散程度取决于 $\sigma_j$

# 例

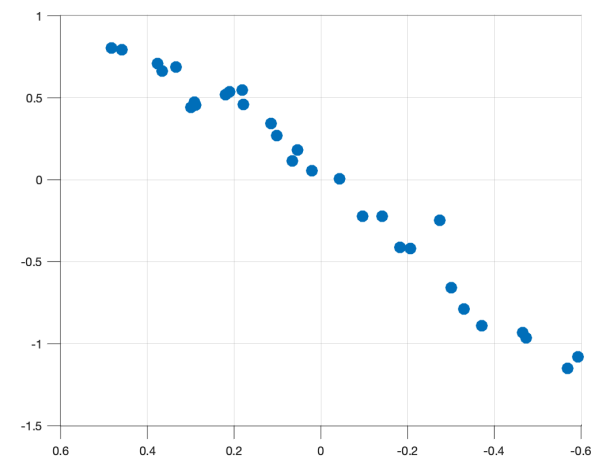
- 30个样本，每个样本测三个数据 $x_1, x_2, x_3$



$$x_1 - x_2$$



$$x_1 - x_3$$



$$x_2 - x_3$$

## 例（续）

- 协方差矩阵： $\frac{AA^T}{29} = \begin{bmatrix} 0.0987 & 0.0941 & 0.1940 \\ 0.0941 & 0.1004 & 0.1951 \\ 0.1940 & 0.1951 & 0.3908 \end{bmatrix}$

- $A$ 的奇异值分解： $A = U\Sigma V^T$

- $U = \begin{bmatrix} -0.4057 & 0.6838 & 0.6065 \\ -0.4084 & -0.7292 & 0.5492 \\ -0.8177 & 0.0249 & -0.5751 \end{bmatrix}$

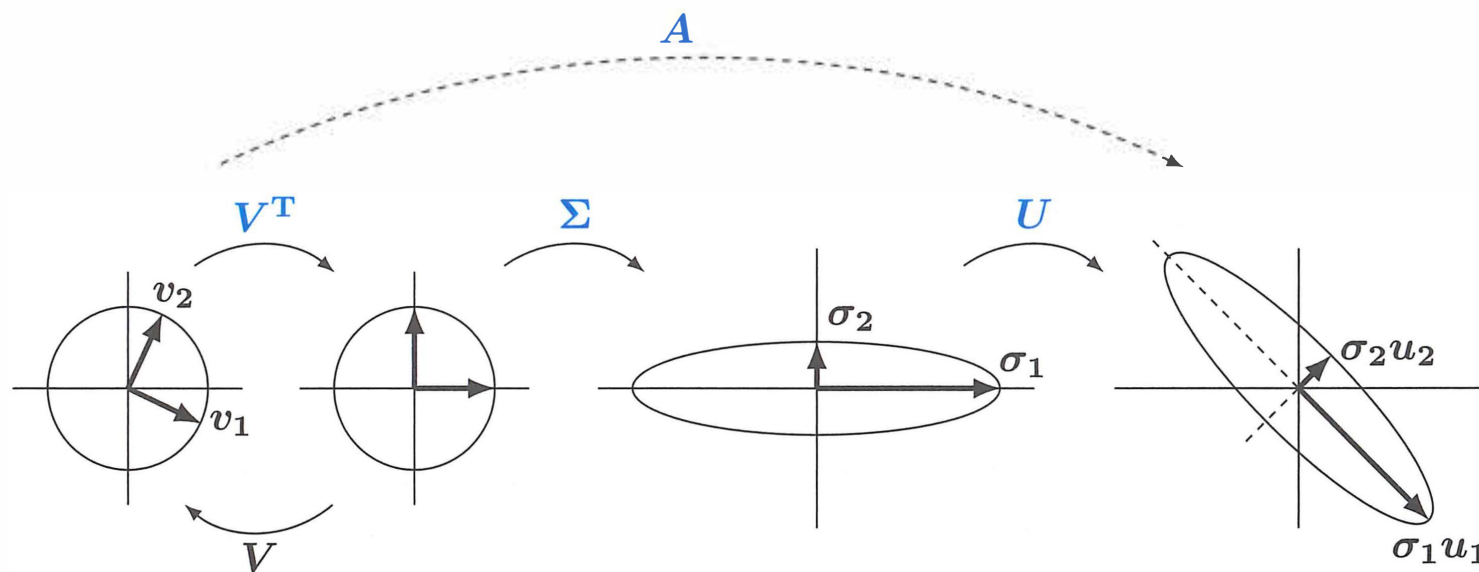
- $D = \begin{bmatrix} 4.1171 & 0 & 0 \\ 0 & 0.3967 & 0 \\ 0 & 0 & 0.0252 \end{bmatrix}$

# 内容提要

- 奇异值分解
- 主成分分析
- 奇异值分解和几何

# 奇异值分解的几何图像

- 考虑2x2的情况  $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{bmatrix} = U \Sigma V^T$ .



- 一般情况：（转动+反射）x（拉伸）x（转动+反射）

# 矩阵的模

- 我们用内积定义了向量的模（长度）： $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$
- 矩阵的模： $\|A\| = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_1$
- 证明：
  - $\|A\mathbf{x}\|^2 = \mathbf{x}^T A^T A \mathbf{x} = \mathbf{x}^T V \Sigma^T \Sigma V^T \mathbf{x} = \sum_{k=1}^r \mathbf{x}^T \mathbf{v}_k \sigma_k^2 \mathbf{v}_k^T \mathbf{x}$
  - $\sum_{k=1}^r \mathbf{x}^T \mathbf{v}_k \sigma_k^2 \mathbf{v}_k^T \mathbf{x} = \sigma_1^2 \sum_{k=1}^n \mathbf{x}^T \mathbf{v}_k \mathbf{v}_k^T \mathbf{x} - \sum_{k=2}^r (\sigma_1^2 - \sigma_k^2) \mathbf{x}^T \mathbf{v}_k \mathbf{v}_k^T \mathbf{x}$
  - $\sigma_1^2 \sum_{k=1}^n \mathbf{x}^T \mathbf{v}_k \mathbf{v}_k^T \mathbf{x} = \sigma_1^2 \mathbf{x}^T V V^T \mathbf{x} = \sigma_1^2 \mathbf{x}^T \mathbf{x} = \sigma_1^2 \|\mathbf{x}\|^2$
  - $\|A\mathbf{x}\|^2 = \sigma_1^2 \sum_{k=1}^n \mathbf{x}^T \mathbf{v}_k \mathbf{v}_k^T \mathbf{x} \leq \sigma_1^2 \|\mathbf{x}\|^2$
  - 所以  $\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\sigma_1 \|\mathbf{x}\|}{\|\mathbf{x}\|} = \sigma_1$

# 矩阵的模的性质

- 由矩阵模的定义可知： $\forall \mathbf{x} \neq 0, \|\mathbf{Ax}\| \leq \|A\| \|\mathbf{x}\|$
- **三角不等式**： $\|A + B\| \leq \|A\| + \|B\|$
- 证明：利用向量模的三角不等式
  - $\|(A + B)\mathbf{x}\| = \|\mathbf{Ax} + \mathbf{Bx}\| \leq \|\mathbf{Ax}\| + \|\mathbf{Bx}\| \leq \|A\| \|\mathbf{x}\| + \|B\| \|\mathbf{x}\|$
  - $\frac{\|(A+B)\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| + \|B\|$
  - 所以  $\max_{\mathbf{x} \neq 0} \frac{\|(A+B)\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| + \|B\|$
  - 也就是说： $\|A + B\| \leq \|A\| + \|B\|$
- **乘积不等式**： $\|AB\| \leq \|A\| \|B\|$

# 矩阵模的应用

- 两个向量差的模代表了两个向量端点的距离
- 类似：两个矩阵差的模量度了两个矩阵之间的“差距”
- **Eckart-Young-Mirsky定理**：同矩阵 $A$ 最接近的秩为 $k$ 的矩阵是 $A_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T$
- 证明：
  - 我们需要证明对于任意秩为 $k$ 的矩阵 $B$ ，都有 $\|A - B\| \geq \|A - A_k\| = \sigma_{k+1}$



# 矩阵模的应用

- **Eckart-Young-Mirsky定理**：同矩阵 $A$ 最接近的秩为 $k$ 的矩阵是 $A_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T$
- 证明：
  - 我们需要证明对于任意秩为 $k$ 的矩阵 $B$ ，都有 $\|A - B\| \geq \|A - A_k\| = \sigma_{k+1}$
  - $w = c_1 \mathbf{v}_1 + \cdots + c_{k+1} \mathbf{v}_{k+1}$ ,  $Bw = c_1 B\mathbf{v}_1 + \cdots + c_{k+1} B\mathbf{v}_{k+1}$
  - 因为的秩为 $k$ ，所以 $k+1$ 个向量 $B\mathbf{v}_1, \dots, B\mathbf{v}_{k+1}$ 线性相关，所以必然存在非零的 $\{c_i\}$ 使的 $Bw = 0$ 。我们可以再假设 $\|w\| = 1$
  - $\|A - B\|^2 \geq \|(A - B)w\|^2 = \|Aw\|^2 = \sigma_1^2 c_1^2 + \cdots + \sigma_{k+1}^2 c_{k+1}^2 \geq \sigma_{k+1}^2 (c_1^2 + \cdots + c_{k+1}^2) = \sigma_{k+1}^2$

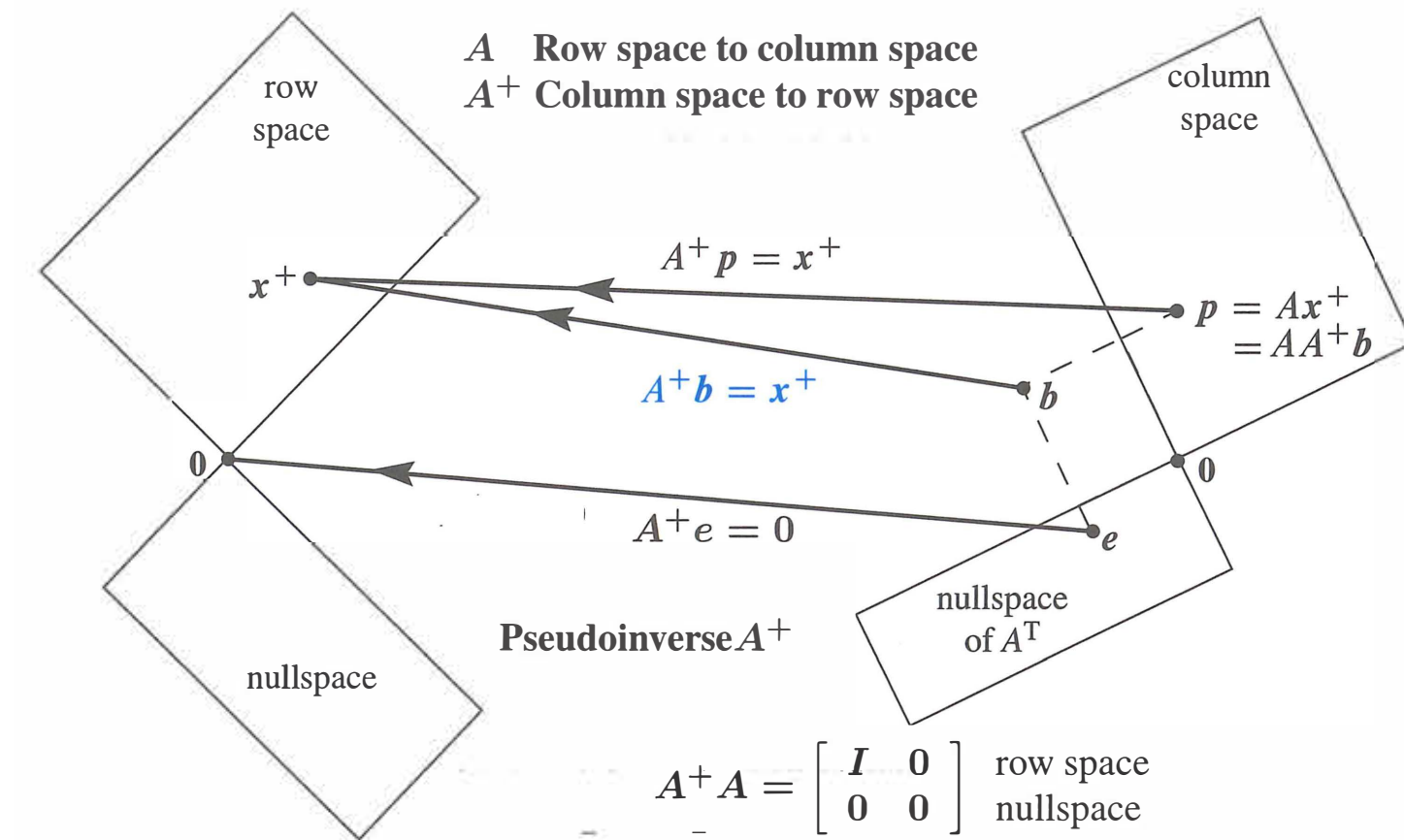
# 极分解 $A = QS$

- 非零复数： $x + iy = re^{i\theta}$ ， $r > 0$ 是一个1x1的正定矩阵， $e^{i\theta}$ 是1x1的幺正矩阵（正交矩阵在复数域上的推广）
- 实方阵的极分解： $A = U\Sigma V^T = (UV^T)(V\Sigma V^T) = QS$ 
  - $Q = UV^T$ 是一个正交矩阵
  - $S$ 是一个半正定矩阵（因为 $\Sigma$ 可能有0特征值）
- 如果 $A$ 可逆，则 $Q$ 和 $S$ 都可逆，这时候 $S$ 是一个正定矩阵

# 伪逆 (pseudoinverse)

- $m \times n$  矩阵  $A = U\Sigma V^T$
- 伪逆 :  $A^+ = V\Sigma^+ U^T = \sum_{k=1}^r \mathbf{v}_k \sigma_k^{-1} \mathbf{u}_k^T$
- 其中  $\Sigma^+ = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}$  是一个  $n \times m$  的矩阵
- $A^+ A = V\Sigma^+ U^T U \Sigma V^T = \begin{bmatrix} I_{r \times r} & 0 \\ 0 & 0 \end{bmatrix}$  是一个  $n \times n$  的矩阵
- $AA^+ = U\Sigma V^T V \Sigma^+ U^T = \begin{bmatrix} I_{r \times r} & 0 \\ 0 & 0 \end{bmatrix}$  是一个  $m \times m$  的矩阵

# 伪逆



# 伪逆的应用

- 最小二乘法： $A^T A \mathbf{x} = A^T \mathbf{b}$
- 解： $\mathbf{x}^+ = A^+ \mathbf{b} = V \Sigma^+ U^T \mathbf{b}$
- 证明：
  - $A^T A \mathbf{x}^+ = V \Sigma^T U^T U \Sigma V^T V \Sigma^+ U^T \mathbf{b} = V \Sigma^T U^T \mathbf{b} = A^T \mathbf{b}$