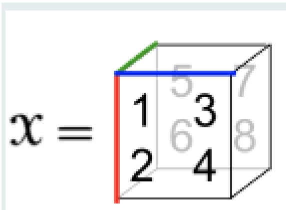
矩阵分析上篇

一：线性代数基础

1.1 tensor

- tensor是一个多维的矩阵



- 将tensor转换为一个矩阵， （分别按照三个坐标轴展开，即图中红，绿，蓝三条轴）

$$\mathbf{X}_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$$

$$\mathbf{X}_{(2)} = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}$$

$$\mathbf{X}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

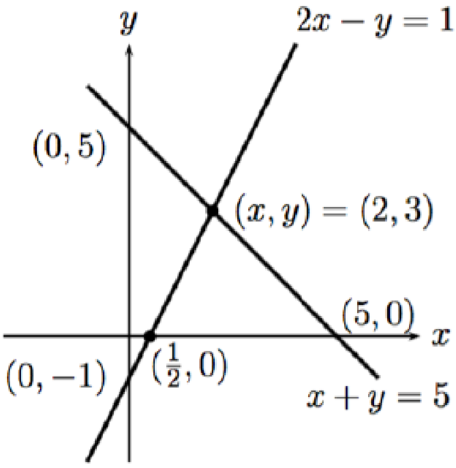
1.2 $\mathbf{Ax} = \mathbf{b}$的行视图

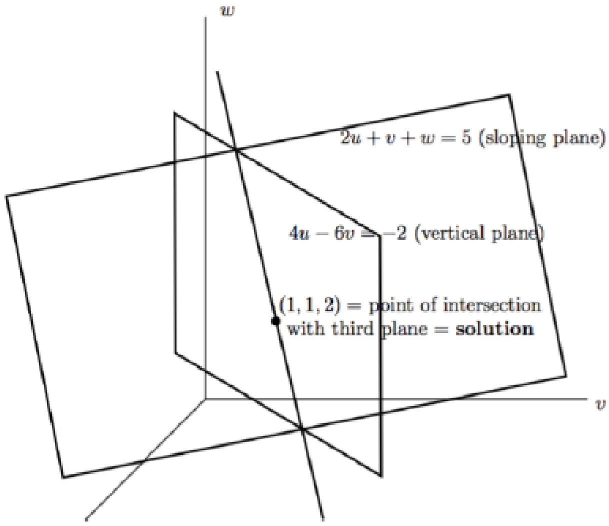
- 假设$A \in \mathbb{R}^{m \times n}$

$$\underbrace{\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}}_{A \in \mathbb{R}^{2\times 2}} \underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{x \in \mathbb{R}^2} = \underbrace{\begin{bmatrix} 1 \\ 5 \end{bmatrix}}_{b \in \mathbb{R}^2}$$

$$\underbrace{\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}}_{A \in \mathbb{R}^{3\times 3}} \underbrace{\begin{bmatrix} u \\ v \\ w \end{bmatrix}}_{x \in \mathbb{R}^3} = \underbrace{\begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix}}_{b \in \mathbb{R}^3}$$

- 行视图：可以理解为凸优化中的超平面，每一行代表一个超平面（二维中是直线，三维中是平面，高维中是超平面）

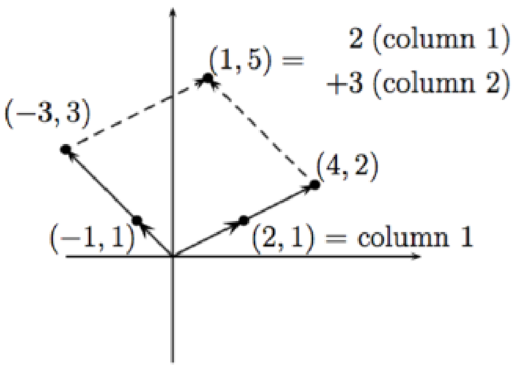


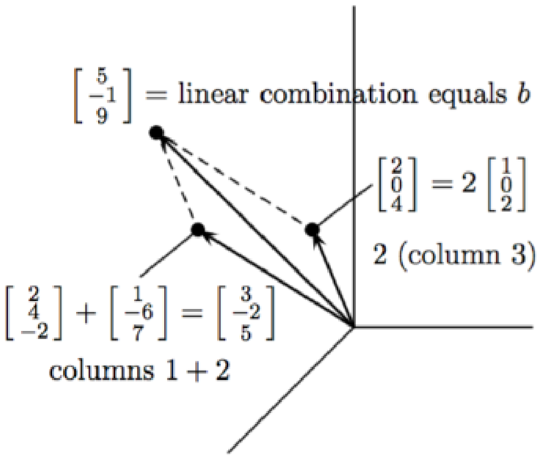


1.3 $\mathbf{Ax} = \mathbf{b}$的列视图

- 列视图可以理解为矩阵列的线性组合。解方程相当于解线性组合前的系数。

$$x\begin{bmatrix}2\\1\end{bmatrix} + y\begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}1\\5\end{bmatrix}$$





1.4 线性相关和线性无关

- 线性相关：矢量集合$[\mathbf{a}_1,\ldots,\mathbf{a}_n]$是线性相关的，如果$\sum_{k=1}^{n} c_k a_k = 0$，当且仅当$c_1, c_2, \ldots, c_n \neq 0$,即至少有一个向量可以由其它向量线性导出(如下式)，换句话说，就是不能是所有的系数同时为0.

$$\mathbf{a}_l = -\frac{1}{c_l}\sum_{k=1,k\neq 1}^{n} c_k a_k$$

- 线性无关：矢量集合$[\mathbf{a}_1,...,\mathbf{a}_n]$是线性无关的，如果$\sum_{k=1}^n c_k a_k = 0$，当且仅当$c_1, c_2, ..., c_n = 0$
- 定义$\mathbf{A} = [\mathbf{a}_1,...,\mathbf{a}_n]$，则$\mathbf{A}\mathbf{x} = 0$只有$\mathbf{x} = \mathbf{0}$，没有其它的线性组合能产生$\mathbf{0}$，此时$\mathbf{A}$可逆。此时A中的向量线性无关。

1.5 Span、基和子空间

- span（子空间）：

$$\text{span}[\mathbf{a}_1,...,\mathbf{a}_n] = \left\{\mathbf{y} \in \mathbb{R}^m | \mathbf{y} = \sum_{k=1}^n c_k a_k\right\} = S$$

- 其实span是向量集合$[\mathbf{a}_1,...,\mathbf{a}_n]$所有的线性组合。此时如果$[\mathbf{a}_1,...,\mathbf{a}_n]$是线性无关的，那么$[\mathbf{a}_1,...,\mathbf{a}_n]$是S的一组基。
- 正交基是指$\mathbf{a_i^T a_j} = 0$，即基中的向量不仅线性无关，而且两两正交。
- S可以有不同的一组基，但是基里向量的个数是相同的，被称为S的维数。等于rank(A)。
- 一个子空间用一组基就可以表示了！
- 对于基的理解：恰到好处。基中向量的个数是子空间的维数，也是最大线性无关向量的个数。子空间中的任何一个向量都可以用一组基的线性组合来表示。

二：线性代数精华

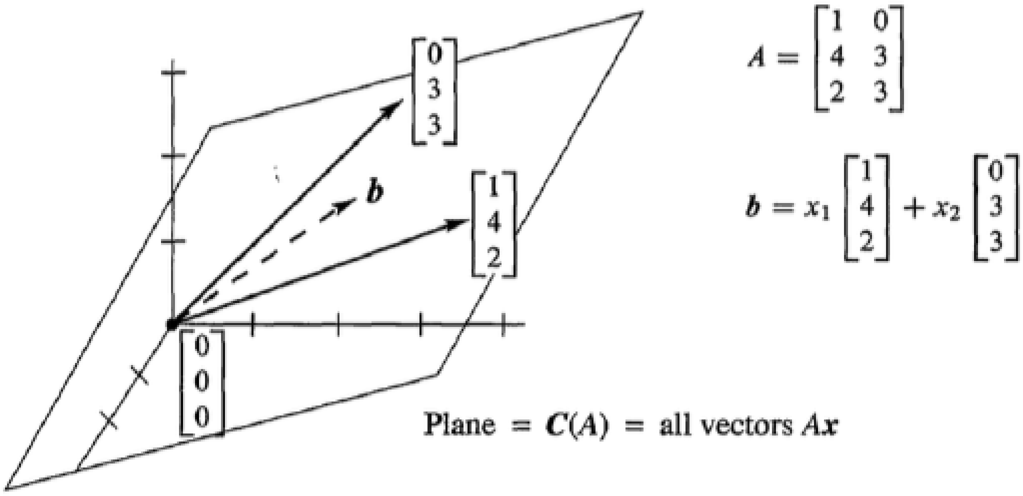
四个基本的子空间，包括列空间，零空间，行空间，左零空间。

2.1 列空间

$\mathbf{A}$是一个$m*n$的矩阵，$\mathbf{C(A)}$是$\mathbf{R^m}(not\ \mathbf{R^n})$的子空间。$\mathbf{C(A)}$包含所有列的线性组合，即$\mathbf{C(A)} = \{\mathbf{y = Ax, x} \in \mathbf{R^n}\}$.

$$A = \begin{bmatrix} 1 & 0 \\ 4 & 3 \\ 2 & 3 \end{bmatrix}, C(A) = \text{span}\left[\begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix}\right], \text{构成了一个}R^3\text{的子空间。}$$

其实，该列空间$\mathbf{C(A)}$是全空间的一个平面而已，是全空间中的一部分。列的线性组合无法组合出平面外的任何向量。



2.2 零空间

$\mathbf{A}$是一个$m*n$的矩阵，$\mathbf{N(A)}$是$\mathbf{R^n}(not\ \mathbf{R^m})$的子空间。

定义：$\mathbf{N(A)}$是包含$\mathbf{Ax}=\mathbf{0}$的所有解的集合。注意：$\mathbf{Ax}=\mathbf{b}$的解并不形成一个子空间（因为不包含0向量）。

例子：

$$A=\begin{bmatrix} 1 & 2 & 2 & 4 \\ 3 & 8 & 6 & 16 \end{bmatrix} \to U=\begin{bmatrix} 1 & 2 & 2 & 4 \\ 0 & 2 & 0 & 4 \end{bmatrix}$$

即$\mathbf{Ux}=\mathbf{0}$，于是有$S_1=\begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \end{bmatrix} S_2=\begin{bmatrix} 0 \\ -2 \\ 0 \\ 1 \end{bmatrix}$

因此$N(A)=C\left(\begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \end{bmatrix},\begin{bmatrix} 0 \\ -2 \\ 0 \\ 1 \end{bmatrix}\right)$，是$\mathbf{R^4}$的子空间。

2.3 行空间

行空间（row space）:$\mathbf{C(A^T)}$是$\mathbf{R^n}$的子空间

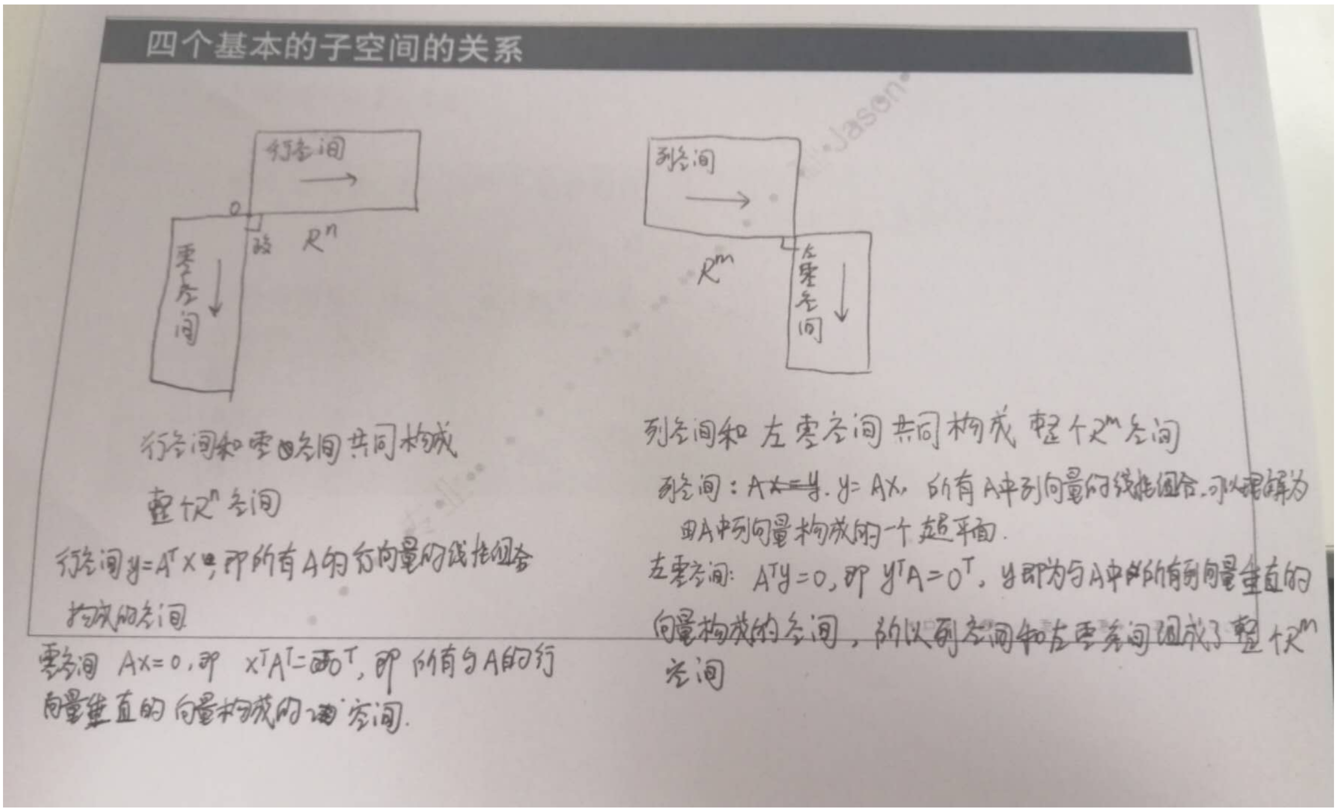
定义：包含所有行的线性组合。

$$A=\begin{bmatrix}1&2&2&4\\3&8&6&16\end{bmatrix}, \text{则} C\left(A^{\top}\right)=C\left(\begin{bmatrix}1\\2\\2\\4\end{bmatrix},\begin{bmatrix}3\\8\\6\\16\end{bmatrix}\right)$$

2.4 左零空间

$\mathrm{N}\left(\mathrm{A}^{\top}\right)=\left\{\mathrm{A}^{\top}\mathrm{y}=0\right\}$的解的集合。是$\mathbf{R}^{\mathbf{m}}$的子空间。

2.5 四个基本的子空间的关系



2.6 利用子空间重新看待方程组的解

$$\begin{cases}\text{无解} \quad b\notin C(A)\\ \text{有解} \quad b\in C(A)\begin{cases}N(A)\text{维数为}0\text{，则有唯一解。说明此时没有通解，只有特解。}\\ N(A)\text{维数大于}0\text{，则有无穷解。说明此时有通解。}\end{cases}\end{cases}$$

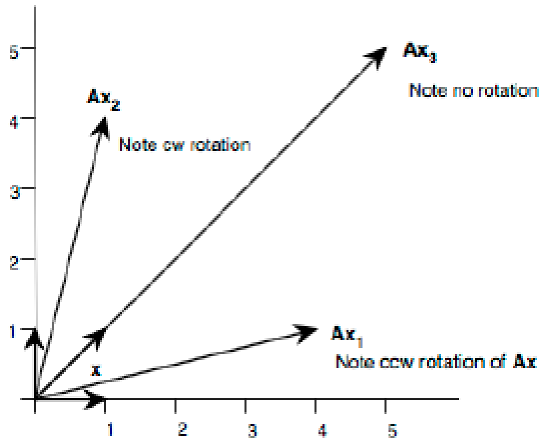
- 如果有解，解的形式为：$\mathbf{x}=\mathbf{p}+\mathbf{v}$，其中$\mathbf{p}$为特解（$\mathbf{Ap}=\mathbf{b}$），$\mathbf{v}$是通解($\mathbf{Av}=\mathbf{0}$)。

三：特征分解

3.1 方阵的特征值与特征向量

给定一个矩阵$A=\begin{bmatrix}4&1\\1&4\end{bmatrix}$，对于$\mathbf{X}_1=\begin{bmatrix}1\\0\end{bmatrix}$，则有$\mathbf{Ax_1}=\begin{bmatrix}4\\1\end{bmatrix}$，对于$\mathbf{X}_3=\begin{bmatrix}1\\1\end{bmatrix}$，则有$\mathbf{Ax_3}=5\begin{bmatrix}1\\1\end{bmatrix}$。

特征值可以理解为对一个向量的伸缩程度。



3.2 特征分解的一般性质

- $\mathbf{Ax}=\lambda\mathbf{x}$，其中$\lambda$为特征值，$\mathbf{x}$为特征向量。
- 求取特征值的过程：$\mathbf{(A-\lambda I)x=0}$,$det(\mathbf{A}-\lambda\mathbf{I})=0$.即保证其不可逆（行列式的值为0）
- 对于$\mathbf{Ax_i}=\lambda\mathbf{x_i}$,如果所有的特征值都不相同，则相应的所有的特征向量线性无关。此时A可以被对角化为：

$$\mathbf{A}=\mathbf{V\Lambda V^{-1}}$$

  其中：$\mathrm{V}=[\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}]$，$\quad\Lambda=\mathrm{Diag}(\lambda_1,\ldots,\lambda_\mathrm{n})$,$\mathbf{x_i}$是特征向量，$\lambda$为特征值。

- 注意：并不是所有的方阵都可以被对角化。当特征值中有重根时，便不一定可以被对角化。

3.3 对称矩阵的特征分解

- 一个对称矩阵，无论其特征值相同或者不同，则其相应的所有的特征向量正交（$UU^\top=U^\top U=I$)
- 正交意味着：任意两个特征向量垂直，每个特征向量的模长为1。

  此时进行特征分解：

$$
\begin{aligned}
\mathsf{A} &= \mathsf{U}\Lambda\mathsf{U}^{\top} \\
&= [\mathbf{u}_1, \ldots, \mathbf{u}_n] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^{\top} \\ \vdots \\ \mathbf{u}_n^{\top} \end{bmatrix} \\
&= \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\top}
\end{aligned}
$$

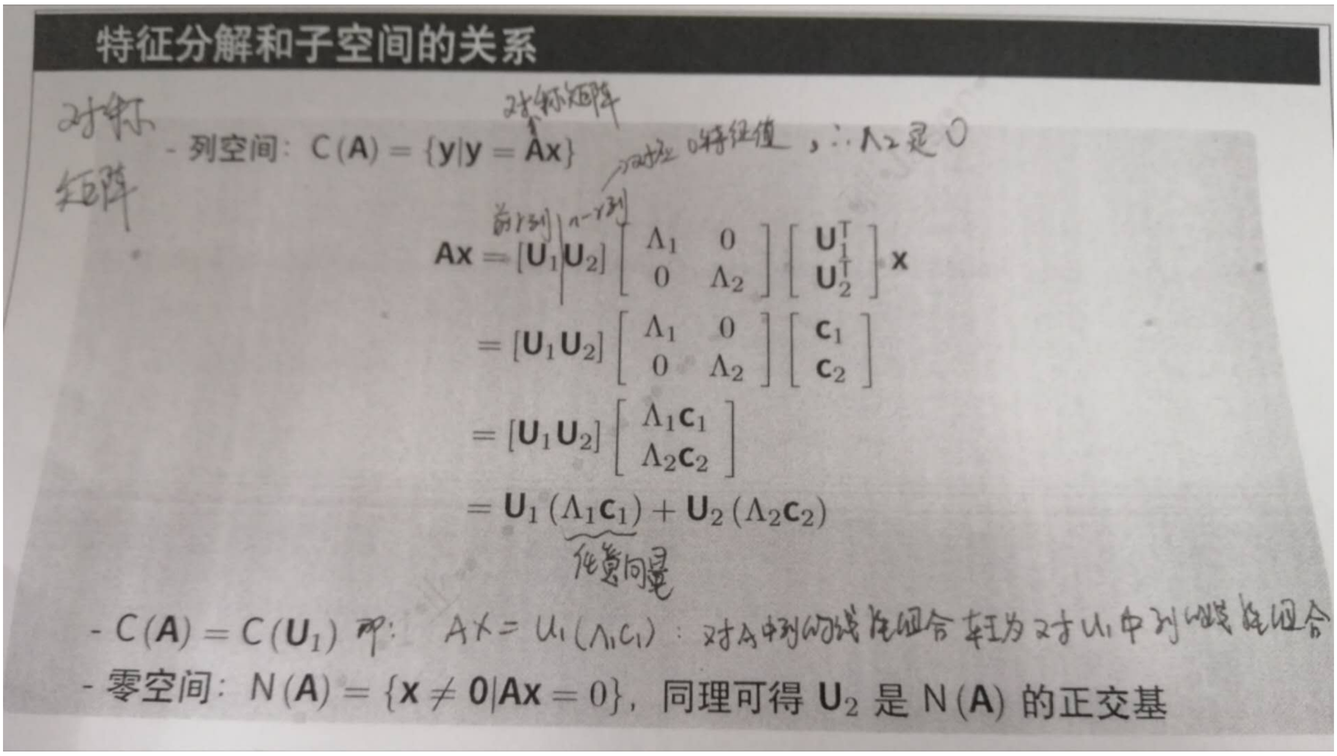
- 对称矩阵的特征值是实数
- 如果$\mathbf{A} \in \mathbf{R}^{\mathbf{n}\times\mathbf{n}}$,是一个对称矩阵且 $rankr \le n$,则有:

$$\underbrace{|\lambda_1| \ge |\lambda_2| \ge \ldots \ge |\lambda_r|}_{r} > \underbrace{\lambda_{r+1} = \ldots \lambda_n}_{n-r} = 0$$

$$\operatorname{Rank}\left(\mathbf{A}^{\top}\mathbf{A}\right) = \operatorname{Rank}\left(\mathbf{A}\mathbf{A}^{\top}\right) = \operatorname{Rank}(\mathbf{A}) = \operatorname{Rank}(\Lambda)$$

- 特征值在某些程度上可以反应能量的大小，在某些时候，后面较小的特征值可以删除。

3.4 特征分解和子空间的关系



特征分解和子空间的关系

对称矩阵

- 列空间：$C(\mathbf{A}) = \{\mathbf{y}|\mathbf{y} = \mathbf{A}\mathbf{x}\}$ 对称矩阵

对应0特征值，∴Λ2是0

前r列 | n−r列

$$
\begin{aligned}
\mathbf{A}\mathbf{x} &= [\mathbf{U}_1 \mathbf{U}_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^{\top} \\ \mathbf{U}_2^{\top} \end{bmatrix} \mathbf{x} \\
&= [\mathbf{U}_1 \mathbf{U}_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} \\
&= [\mathbf{U}_1 \mathbf{U}_2] \begin{bmatrix} \Lambda_1 \mathbf{c}_1 \\ \Lambda_2 \mathbf{c}_2 \end{bmatrix} \\
&= \mathbf{U}_1 \underbrace{(\Lambda_1 \mathbf{c}_1)}_{\text{任意向量}} + \mathbf{U}_2 (\Lambda_2 \mathbf{c}_2)
\end{aligned}
$$

- $C(\mathbf{A}) = C(\mathbf{U}_1)$ 即：$Ax = U_1(\Lambda_1 C_1)$：对A中列的线性组合转为对U_1中列的线性组合
- 零空间：$N(\mathbf{A}) = \{\mathbf{x} \ne \mathbf{0} | \mathbf{A}\mathbf{x} = 0\}$，同理可得 $\mathbf{U}_2$ 是 $N(\mathbf{A})$ 的正交基

四：PCA

4.1 优化问题

- 约束问题

$$\max\ \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} \qquad \text{subject to } ||\mathbf{x}||_2^2 = 1$$

- 拉格朗日函数

$$\mathsf{L}(\mathbf{x}, \lambda) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - \lambda\mathbf{x}^{\mathsf{T}}\mathbf{x}$$

- 计算梯度

$$\nabla\mathsf{L}(\mathbf{x}, \lambda) = 2\mathbf{A}\mathbf{x} - 2\lambda\mathbf{x} = \mathbf{0}$$

- 可得

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

则：$\mathbf{x^T A x} = \lambda\mathbf{x^T x} = \lambda||\mathbf{x}||_\mathbf{2}^\mathbf{2} = \lambda$,因此，原问题变为求A的最大特征值。

4.2 正交变换

- 维数灾难
- 通过数学变换将原始高维属性空间转变成一低维空间
- 给定 d 维空间的样本 $\mathsf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\mathsf{N}] \in \mathbb{R}^{\mathsf{d}\times\mathsf{N}}$，变换之后得到 $\mathsf{d}' \leq \mathsf{d}$ 维空间中的样本

$$\mathbf{Z} = \mathsf{W}^{\mathsf{T}}\mathsf{X}$$

其中 $\mathsf{W} \in \mathbb{R}^{\mathsf{d}\times\mathsf{d}'}$ 是变换矩阵，$\mathbf{Z} \in \mathbb{R}^{\mathsf{d}'\times\mathsf{N}}$ 是样本在新空间中的表达.

- 投影解释：$\mathsf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{\mathsf{d}'}]$：$\mathsf{d}'$ 个 d 维向量，$\mathbf{z}_\mathsf{n} = \mathsf{W}^{\mathsf{T}}\mathbf{x}_\mathsf{n}$

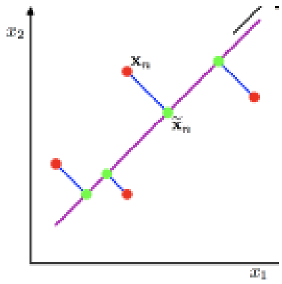
- 降维：维度降低，但是样本数量不变

4.3 PCA

- 假定 $d' = 1$，$\mathbf{w}_1^T\mathbf{w}_1 = 1$．每个数据点投影值为 $\mathbf{w}_1^T\mathbf{x}_n$，样本均值为 $\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$，投影数据均值是．投影数据的方差为 $\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{w}_1^T\mathbf{x}_n - \mathbf{w}_1^T\overline{\mathbf{x}}\right\}^2 = \mathbf{w}_1^T\mathbf{C}_X\mathbf{w}_1$
- 其中协方差矩阵定义为 $\mathbf{C}_X = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T$
优化目标为

$$\max\ \mathbf{w}_1^T\mathbf{C}_X\mathbf{w}_1 \qquad \text{subject to } ||\mathbf{w}_1||_2^2 = 1$$

- 特征向量被称为第一主成分．$d' > 1$ 数学归纳法类推．选取准则：$\frac{\sum_{i=1}^{d'}\lambda_i}{\sum_{i=1}^{d}\lambda_i} \geq t$，例如 $t = 95\%$



- 选取原则可以这样理解：前d'个特征值的能量占到所有特征值能量的95%。
- 转换矩阵的求解方法：
    1. 求原矩阵的协方差矩阵
    2. 对协方差矩阵进行特征分解，
    3. 对特征值进行排序，
    4. 找出较大的特征值对应的特征向量，即可组成转换矩阵
- 用方差来衡量点的分散程度。希望投影后的值越分散越好。如果所有数据都投影到某个集中的未知上，则数据信息丢失过多。希望方差尽可能大是为了尽可能地是数据区分开，从而尽可能多地保留数据地原始信息。所以数据越分散越好。

4.4 PCA举例

$\mathbf{X} = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}$，$\mathbf{C}_X = \begin{bmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{bmatrix}$

- 计算 $\mathbf{C}_X$ 特征值为：$\lambda_1 = 2,\ \lambda_2 = 2/5$，特征值特征向量为 $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$，$\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

- 降维：$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}\mathbf{X} = \begin{bmatrix} -\frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$，此时可验证 $\mathbf{C}_Y = 2 = \lambda_1$

- 注意有5个样本，每个样本是2维。