

海量高维数据与维度约减

七月在线 龙老师
2016年7月23日



主要内容

- 为什么要数据降维
- 为什么能数据降维
- SVD
 - 基本概念与性质
 - 怎么用SVD降维
 - 实际案例
- CUR

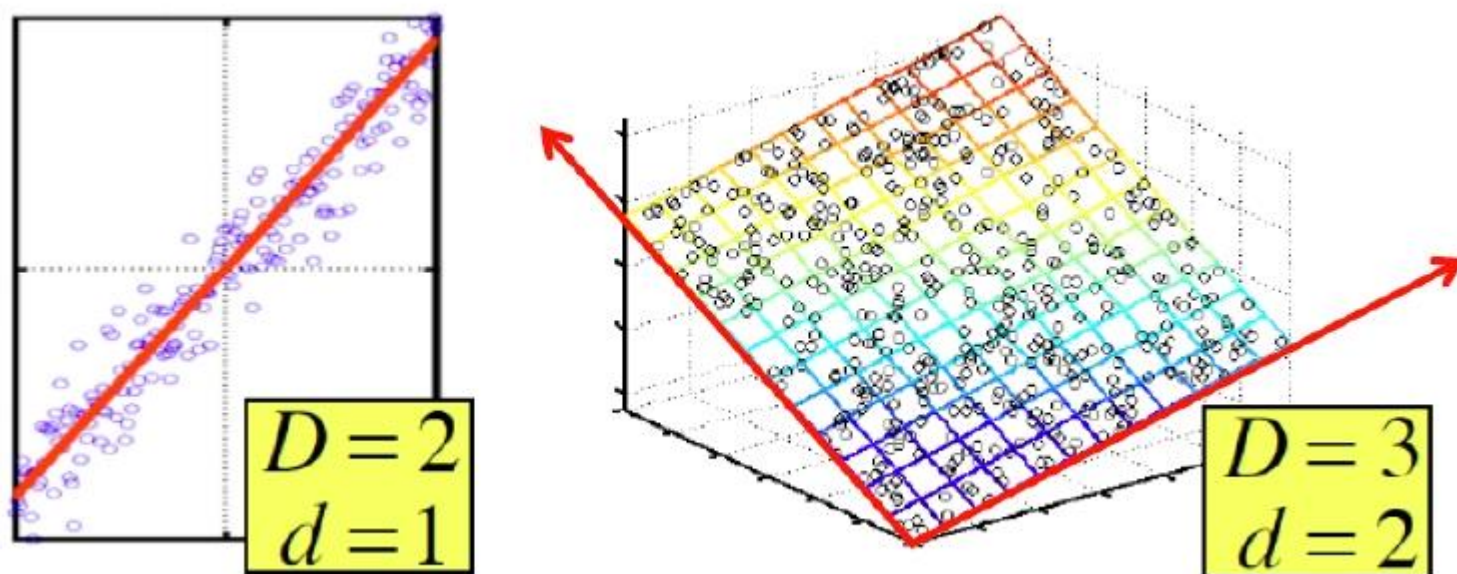


为什么要降维

- 海量数据太大，不得不降维。
- 可以让你使用简单的模型运算的更快，更容易理解，更容易维护。
- 优质的降维数据可以在使用不是最优的模型参数的情况下得到不错的预测结果，这样你就不必费力去选择最适合的模型和最优的参数了。



为什么能降维



- 假设：数据实际上是存在或者靠近一个低维子空间中。
- 子空间的坐标轴能够最有效地表达这个数据

回忆：矩阵的秩

- 矩阵的秩：矩阵的线性独立的列（行）的个数。

$$\text{矩阵 } \mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \text{ 的秩 } r=2$$

- 为什么关心矩阵的秩？

- 我们可以把A用两个新的基向量表示：

- $[1 \ 2 \ 1] \ [-2 \ -3 \ 1]$

- 则相应的坐标就变为： $[1 \ 0] \ [0 \ 1] \ [1 \ 1]$

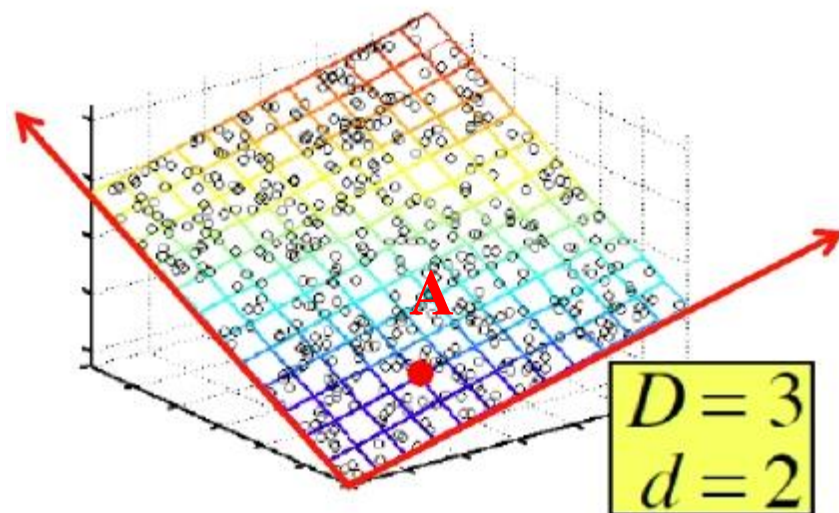


秩其实就是最小的维度

□ 观察三维视图：

■ 把矩阵的每一行作为一点在三维空间的坐标

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \begin{matrix} A \\ B \\ C \end{matrix}$$



秩其实就是最小的维度

□ 我们可以重新采用一个坐标系，这个坐标系是以原矩阵的前两行作为坐标基底的。

■ 老坐标基底: $[1\ 0\ 0]\ [0\ 1\ 0]\ [0\ 0\ 1]$

■ 新坐标基底: $[1\ 2\ 1]\ [-2\ -3\ 1]$

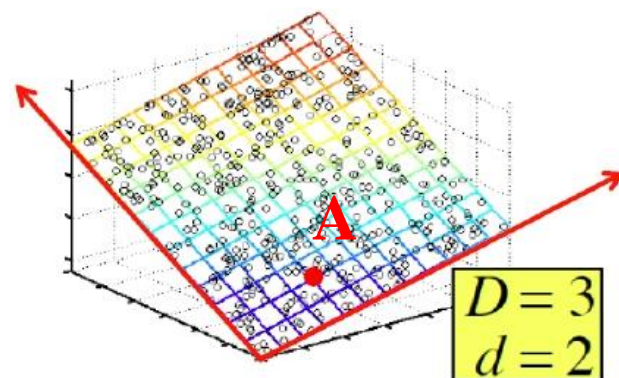
□ 则ABC三点的新坐标分别为:

■ A: $[1\ 0]$,

■ B: $[0\ 1]$,

■ C: $[1\ 1]$

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \end{matrix}$$

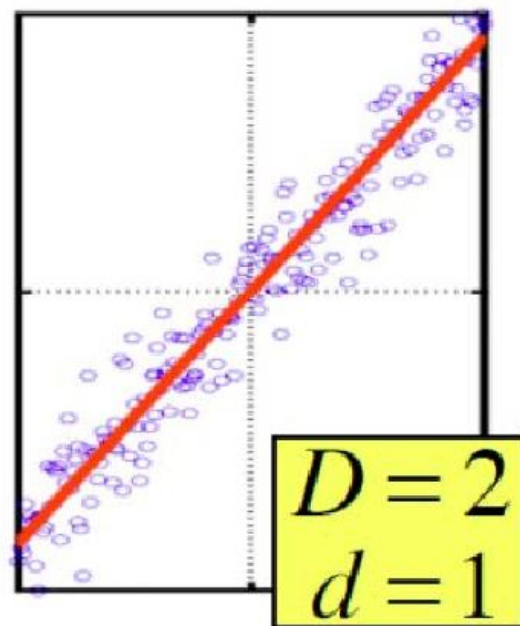


□ 注意：我们已经把ABC三点降维了



降维的关键

- 降维的关键就是找到能够表达数据的最少维度，用最少的坐标轴表示数据。
- 右图的点在二维空间中
- 但大量聚集在红线附近
- 所以就可以用红线所代表的一维坐标来表示。
- 当然这样做有一点误差



SVD

$$\mathbf{A}_{[m \times n]} = \mathbf{U}_{[m \times r]} \Sigma_{[r \times r]} (\mathbf{V}_{[n \times r]})^T$$

A: 输入矩阵

$m \times n$ (e.g., m 篇文章, n 个词语)

U: 左奇异向量矩阵

$m \times r$ (m 篇文章, r 个主题)

Σ : 奇异值矩阵

$r \times r$ 对角阵 (每个主题的重要性) (r : 矩阵A的秩)

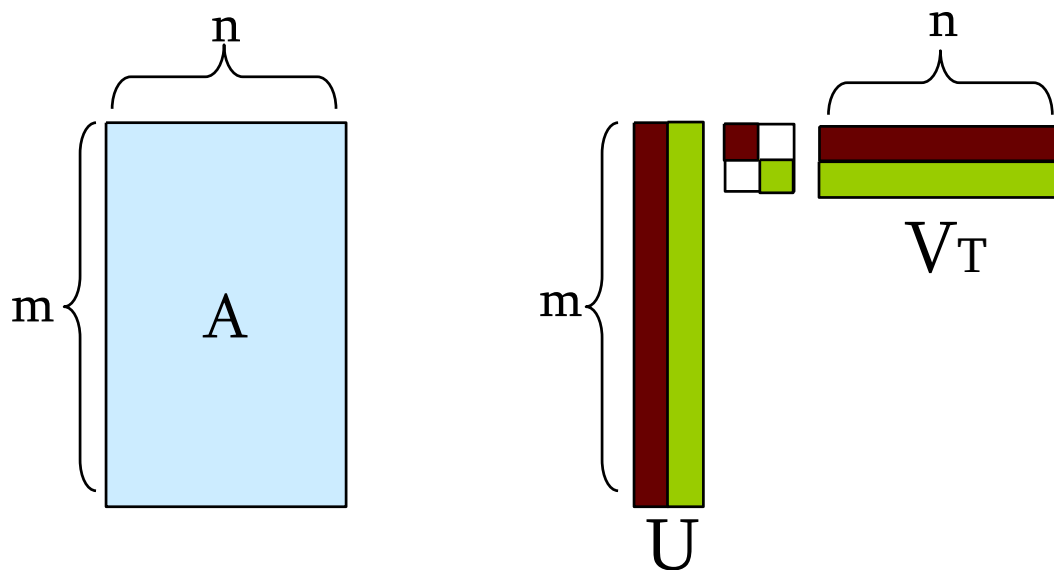
V: 右奇异向量矩阵

$n \times r$ (n 个词语, r 个主题)



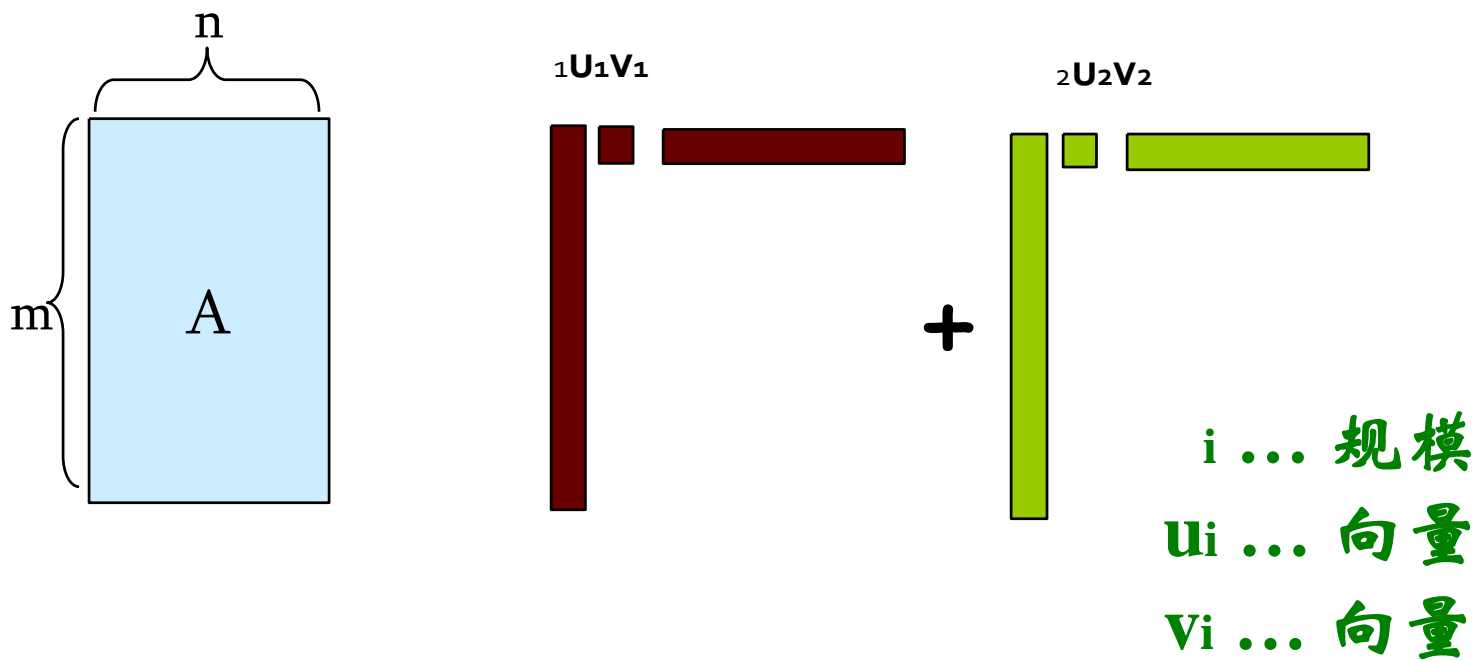
奇异值分解

$$A \approx U \Sigma V^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^T$$



奇异值分解

$$A \approx U \Sigma V^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^T$$

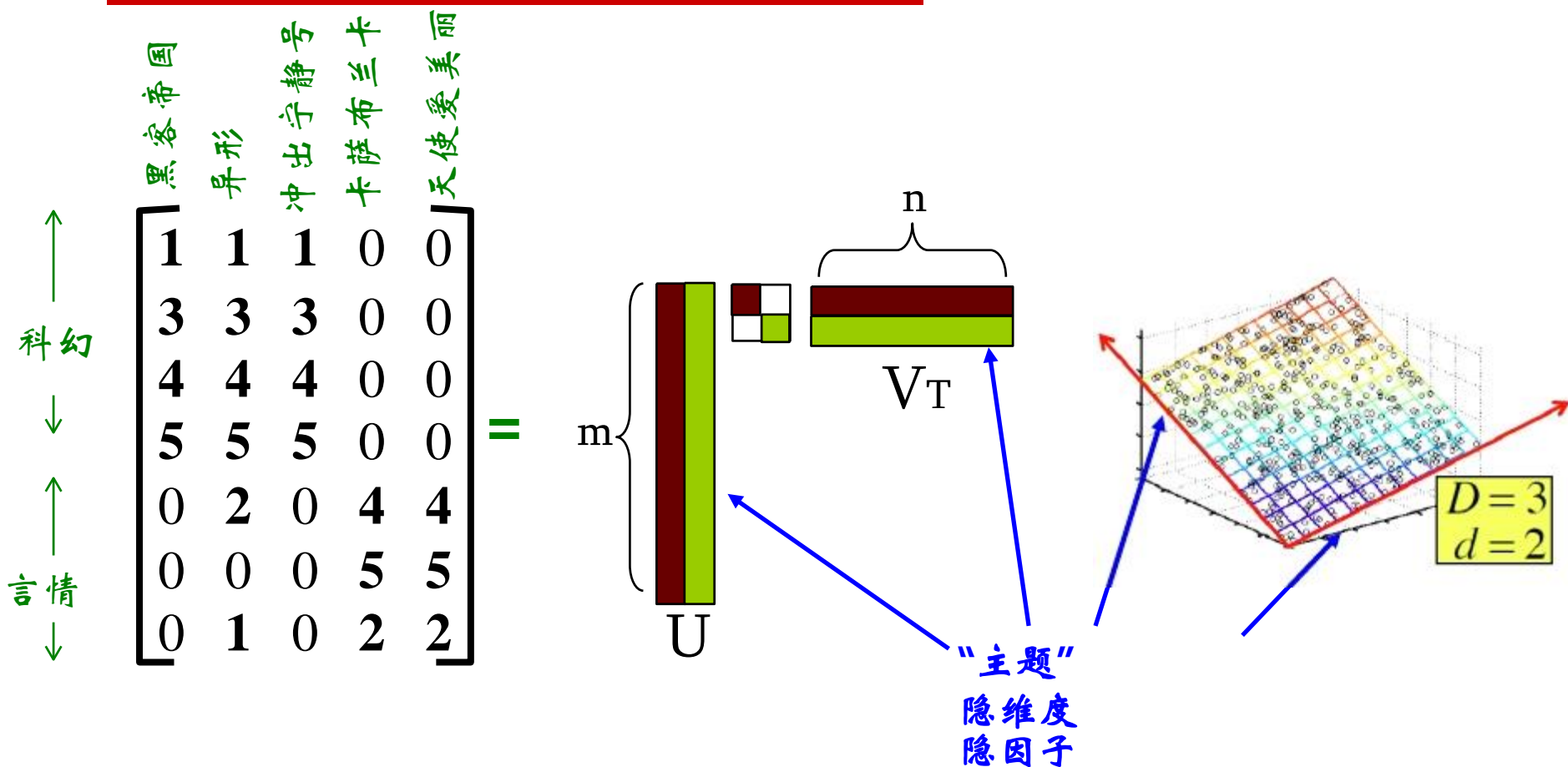


SVD的性质

- 对于一个实数矩阵，总能够拆解成三个矩阵相乘 $A = U \cdot \Sigma \cdot V_T$
- 这三个矩阵满足如下性质：
 - U 、 Σ 、 V 是唯一的
 - U 、 V 的列是单位标准正交基
 - Σ 是对角阵，对角上每一个值是奇异值，是正数，并且按降序排序。



SVD案例：用户看电影



SVD案例：用户看电影

↑ 科幻
↓
↑ 言情
↓

$$\begin{bmatrix}
 \text{黑} & \text{容} & \text{帝} & \text{国} \\
 \text{异} & \text{形} & & \\
 \text{冲} & \text{出} & \text{宁} & \text{静} \\
 \text{卡} & \text{萨} & \text{布} & \text{兰} \\
 \text{天} & \text{使} & \text{爱} & \text{美} \\
 \text{丽} & & &
 \end{bmatrix}
 \begin{bmatrix}
 1 & 1 & 1 & 0 & 0 \\
 3 & 3 & 3 & 0 & 0 \\
 4 & 4 & 4 & 0 & 0 \\
 5 & 5 & 5 & 0 & 0 \\
 0 & 2 & 0 & 4 & 4 \\
 0 & 0 & 0 & 5 & 5 \\
 0 & 1 & 0 & 2 & 2
 \end{bmatrix}
 =
 \begin{bmatrix}
 0.13 & 0.02 & -0.01 \\
 0.41 & 0.07 & -0.03 \\
 0.55 & 0.09 & -0.04 \\
 0.68 & 0.11 & -0.05 \\
 0.15 & -0.59 & 0.65 \\
 0.07 & -0.73 & -0.67 \\
 0.07 & -0.29 & 0.32
 \end{bmatrix}
 \times
 \begin{bmatrix}
 12.4 & 0 & 0 \\
 0 & 9.5 & 0 \\
 0 & 0 & 1.3
 \end{bmatrix}
 \times
 \begin{bmatrix}
 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\
 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
 0.40 & -0.80 & 0.40 & 0.09 & 0.09
 \end{bmatrix}$$



SVD案例：用户看电影

↑ 科幻
↓
↑ 言情
↓

| 黑 容 帝 国 | 异形 | 号 冲 出 宁 静 | 卡 萨 布 兰 | 卡 萨 布 兰 | 丽 美 爱 天 使 |
|------------------|----|-----------------------|------------------|------------------|-----------------------|
| 1 | 1 | 1 | 0 | 0 | |
| 3 | 3 | 3 | 0 | 0 | |
| 4 | 4 | 4 | 0 | 0 | |
| 5 | 5 | 5 | 0 | 0 | |
| 0 | 2 | 0 | 4 | 4 | |
| 0 | 0 | 0 | 5 | 5 | |
| 0 | 1 | 0 | 2 | 2 | |

科幻-主题 言情-主题

$$\begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



SVD案例：用户看电影

↑ 科幻
↓
↑ 言情
↓

| 黑 帝 国 | 星 形 | 号 冲 出 静 夜 | 卡 萨 布 兰 | 丽 美 爱 天 使 |
|-------------|--------|-----------------------|------------------|-----------------------|
| 1 | 1 | 1 | 0 | 0 |
| 3 | 3 | 3 | 0 | 0 |
| 4 | 4 | 4 | 0 | 0 |
| 5 | 5 | 5 | 0 | 0 |
| 0 | 2 | 0 | 4 | 4 |
| 0 | 0 | 0 | 5 | 5 |
| 0 | 1 | 0 | 2 | 2 |

科幻-主题 言情-主题

$$= \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

U 是“用户-主题”相似矩阵



SVD案例：用户看电影

↑ 科幻
↓
↑ 言情
↓

$$\begin{bmatrix}
 \text{黑} & \text{容} & \text{帝} & \text{国} \\
 \text{异形} & & & \\
 \text{冲出} & \text{宁静} & \text{号} & \\
 \text{卡} & \text{萨} & \text{布} & \text{兰} \\
 \text{天使} & \text{爱} & \text{美} & \text{丽}
 \end{bmatrix}
 =
 \begin{bmatrix}
 \text{科幻-主题} & & & \\
 0.13 & 0.02 & -0.01 & \\
 0.41 & 0.07 & -0.03 & \\
 0.55 & 0.09 & -0.04 & \\
 0.68 & 0.11 & -0.05 & \\
 0.15 & -0.59 & 0.65 & \\
 0.07 & -0.73 & -0.67 & \\
 0.07 & -0.29 & 0.32 &
 \end{bmatrix}
 \times
 \begin{bmatrix}
 \text{科幻-主题的“强度”} & & & \\
 12.4 & 0 & 0 & \\
 0 & 9.5 & 0 & \\
 0 & 0 & 1.3 &
 \end{bmatrix}
 \times
 \begin{bmatrix}
 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\
 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
 0.40 & -0.80 & 0.40 & 0.09 & 0.09
 \end{bmatrix}$$



SVD案例：用户看电影

↑ 科幻
↓
↑ 言情
↓

| 黑 帝 国 | 异 形 | 号 冲 出 静 夜 | 卡 萨 布 兰 | 丽 美 爱 天 使 |
|-------------|--------|-----------------------|------------------|-----------------------|
| 1 | 1 | 1 | 0 | 0 |
| 3 | 3 | 3 | 0 | 0 |
| 4 | 4 | 4 | 0 | 0 |
| 5 | 5 | 5 | 0 | 0 |
| 0 | 2 | 0 | 4 | 4 |
| 0 | 0 | 0 | 5 | 5 |
| 0 | 1 | 0 | 2 | 2 |

科幻-主题

| | | |
|------|-------|-------|
| 0.13 | 0.02 | -0.01 |
| 0.41 | 0.07 | -0.03 |
| 0.55 | 0.09 | -0.04 |
| 0.68 | 0.11 | -0.05 |
| 0.15 | -0.59 | 0.65 |
| 0.07 | -0.73 | -0.67 |
| 0.07 | -0.29 | 0.32 |

V 是“电影-主题”相似矩阵

| | | |
|------|-----|-----|
| 12.4 | 0 | 0 |
| 0 | 9.5 | 0 |
| 0 | 0 | 1.3 |

科幻-主题

| | | | | |
|------|-------|------|-------|-------|
| 0.56 | 0.59 | 0.56 | 0.09 | 0.09 |
| 0.12 | -0.02 | 0.12 | -0.69 | -0.69 |
| 0.40 | -0.80 | 0.40 | 0.09 | 0.09 |



SVD 的深入理解

□ “电影”、“用户”和“主题”

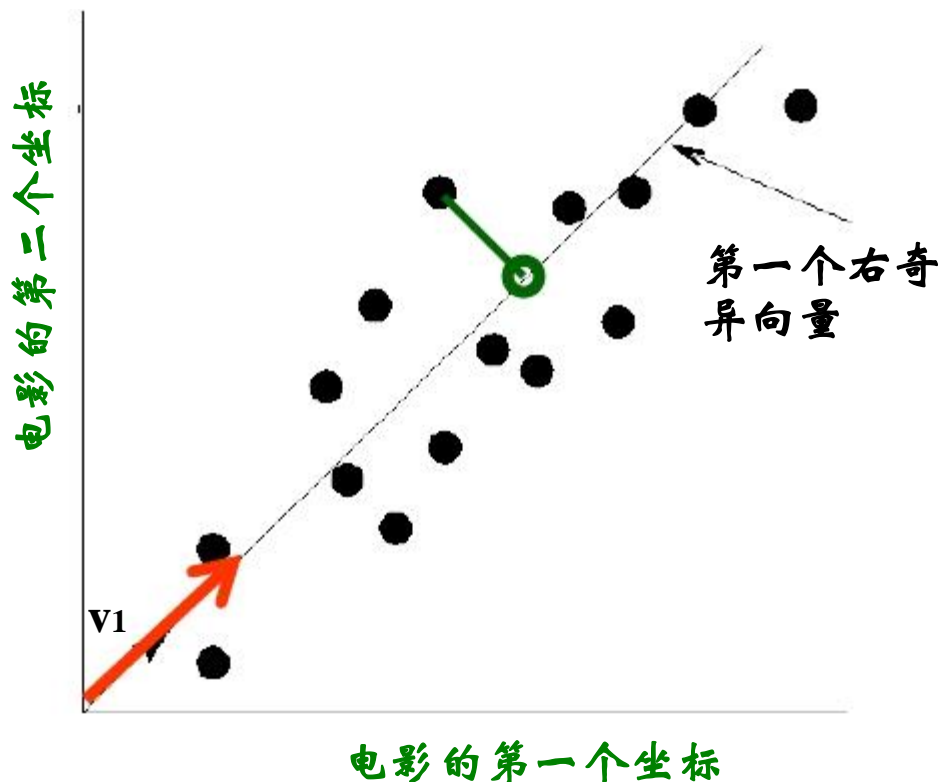
- U: “用户-主题”相似矩阵
- V: “电影-主题”相似矩阵
- Σ : 其对角元素是每一个主题的“强度”



SVD进行降维

□ SVD能够给出“最好”的结果

■ 所谓“最好”就是使得平方误差（投影）最小



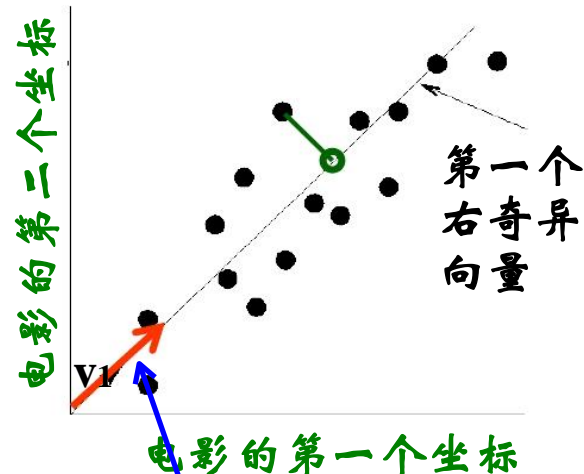
SVD进行降维

- $A = U \cdot \Sigma \cdot V_T$
- U: “用户-主题”相似矩阵
- V: “电影-主题”相似矩阵

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times$$

$$\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times$$

$$\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



SVD进行降维

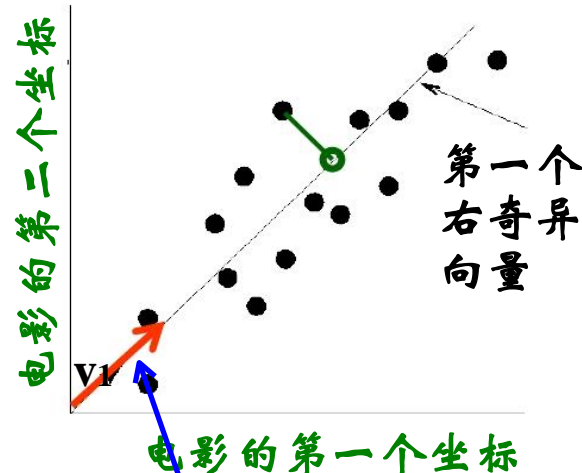
$$A = U \cdot \Sigma \cdot V_T$$

V1坐标轴的方差
（“扩散”）

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times$$

$$\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times$$

$$\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



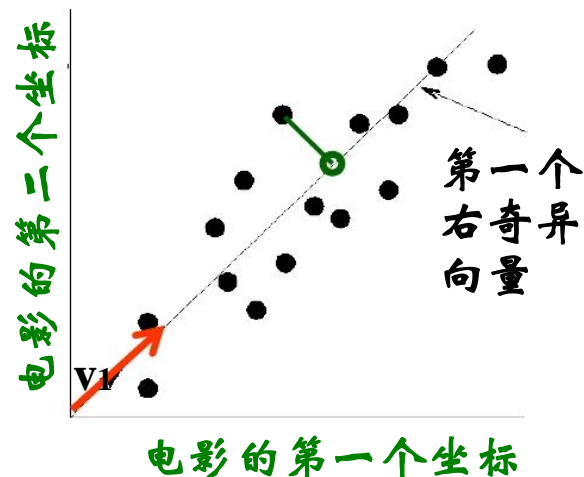
SVD进行降维

$$A = U \cdot \Sigma \cdot V_T$$

用户坐标在“科幻-主题”坐标轴上的投影 $(U\Sigma)^T$:

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 3 | 3 | 3 | 0 | 0 |
| 4 | 4 | 4 | 0 | 0 |
| 5 | 5 | 5 | 0 | 0 |
| 0 | 2 | 0 | 4 | 4 |
| 0 | 0 | 0 | 5 | 5 |
| 0 | 1 | 0 | 2 | 2 |

| | | |
|------|-------|-------|
| 0.13 | 0.02 | -0.01 |
| 0.41 | 0.07 | -0.03 |
| 0.55 | 0.09 | -0.04 |
| 0.68 | 0.11 | -0.05 |
| 0.15 | -0.59 | 0.65 |
| 0.07 | -0.73 | -0.67 |
| 0.07 | -0.29 | 0.32 |



SVD进行降维

□ Q:SVD怎么降维？能够精确到什么程度？

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & \cancel{1.3} \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & \cancel{1.3} \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$



SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Note: In the original image, the third column of the second matrix and the third row of the third matrix are crossed out with red lines, indicating they are the smallest singular values being set to zero for dimensionality reduction.



SVD进行降维

- Q:SVD怎么降维？能够精确到什么程度？
- A:把最小的奇异值设为0

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \end{bmatrix}$$



SVD进行降维

□ Q:SVD怎么降维？能够精确到什么程度？

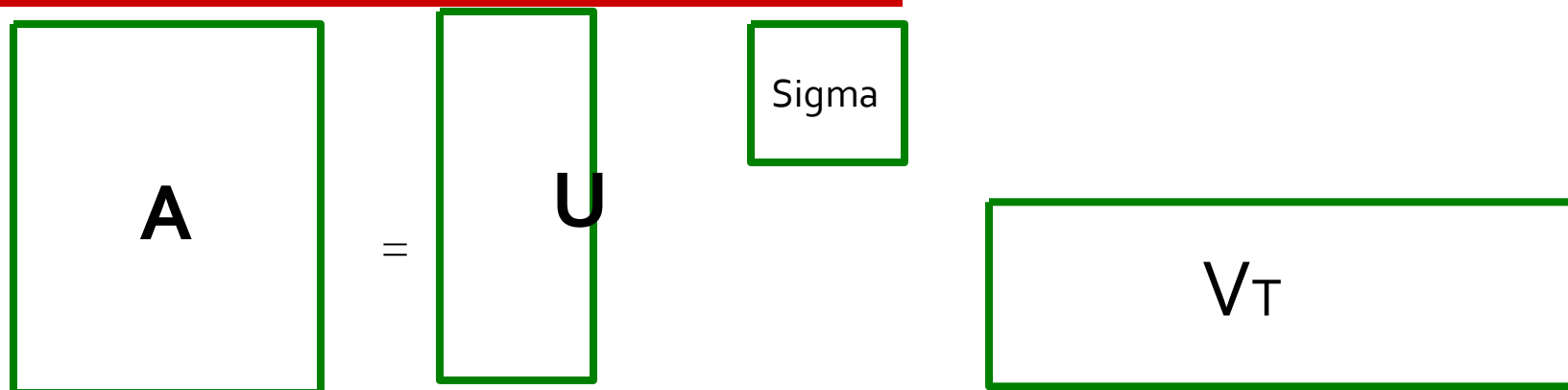
□ A:把最小的奇异值设为0

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.92 & 0.95 & 0.92 & 0.01 & 0.01 \\ 2.91 & 3.01 & 2.91 & -0.01 & -0.01 \\ 3.90 & 4.04 & 3.90 & 0.01 & 0.01 \\ 4.82 & 5.00 & 4.82 & 0.03 & 0.03 \\ 0.70 & 0.53 & 0.70 & 4.11 & 4.11 \\ -0.69 & 1.34 & -0.69 & 4.78 & 4.78 \\ 0.32 & 0.23 & 0.32 & 2.01 & 2.01 \end{bmatrix}$$

弗罗宾尼斯范数 $\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2}$ $\|A-B\|_F = \sqrt{\sum_{ij} (A_{ij}-B_{ij})^2}$ 很小

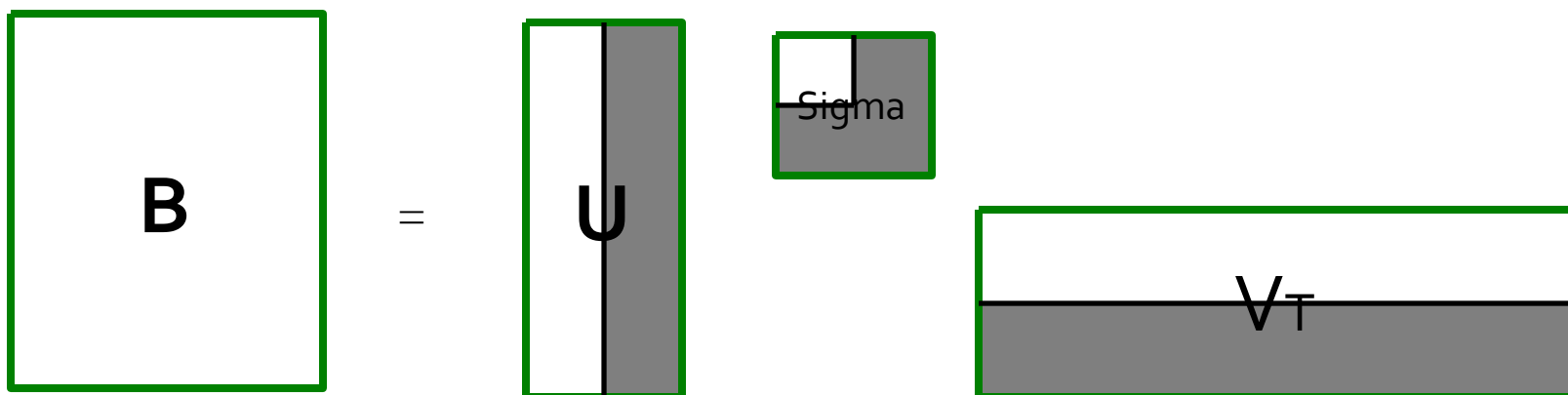


SVD: 最好的低秩近似



B 是 A 最好的近似:

$$\|A-B\|_F = \sqrt{\sum_{ij} (A_{ij}-B_{ij})^2} \text{ 很小}$$



SVD:最好的低秩近似

□ 理论证明:

- $A=U\Sigma V^T$ 和 $B=USV^T$, 其中S是 $r*r$ 的对角矩阵,
- $s_i=\sigma_i (i=1\dots k)$, 剩下的 $s_i=0$
- 于是B是 $\text{rank}(B)=k$ 的情况下对A的最好的近似。

□ 所谓“最好”是指:

- B是 $\|A-B\|_F$ 在 $\text{rank}(B)=k$ 情况下取最小值的解

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}_{m \times n} = \begin{pmatrix} u_{11} & \dots & & \\ \vdots & \ddots & & \\ u_{m1} & & & \end{pmatrix}_{m \times r} \begin{pmatrix} \sigma_{11} & 0 & \dots \\ 0 & \ddots & \\ \vdots & & \end{pmatrix}_{r \times r} \begin{pmatrix} v_{11} & \dots & v_{1n} \\ \vdots & \ddots & \\ & & \end{pmatrix}_{r \times n}$$

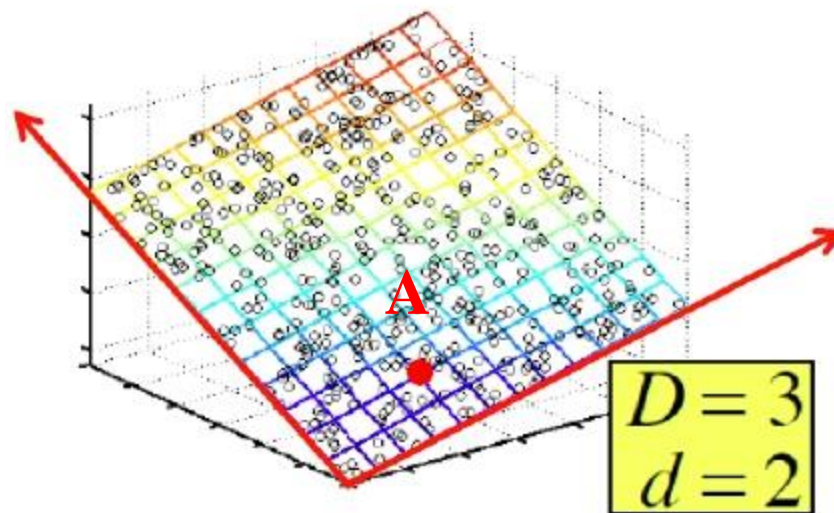


保留多少奇异值

□ 80-90% 的能量 = $\sum_i \sigma_i^2$

■ ——一般经验值

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \end{matrix}$$



SVD: 计算复杂度

□ 完全计算复杂度:

- $O(nm^2)$ 或 $O(n^2m)$, 取最小的。

□ 但是我们可以减少计算量:

- 如果只需要奇异值
- 或者只需要前k个奇异向量
- 或者矩阵是稀疏矩阵

□ 一般用开源线性运算算法包:

- LINPACK, SciPy, Matlab, SPlus, Mathematica



案例：给用户推荐电影

□ Q: 找到潜在喜欢“黑客帝国”电影的用户

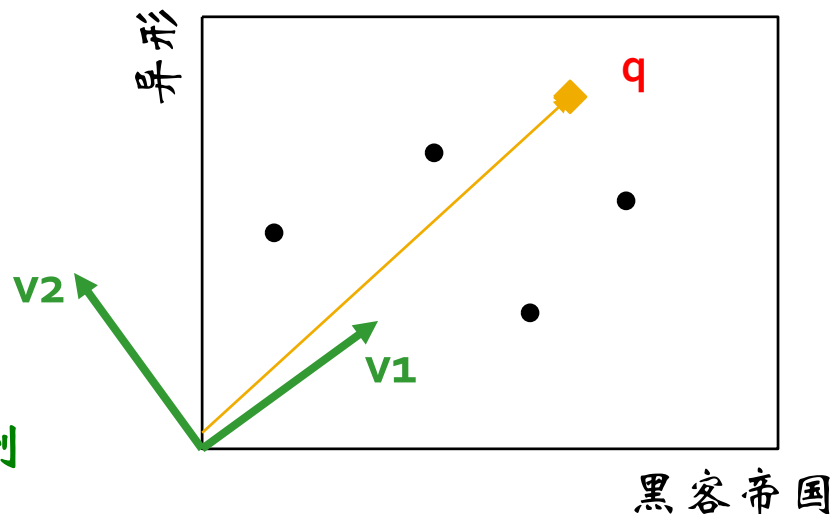
□ A: 将用户的搜索统计映射到“主题空间”

■ —— 怎么做？

黑客帝国 异形 冲出宁静号 卡萨布兰卡 天使爱美丽

$$q = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

与主题向量的基底做内积，投影到“主题空间”中。



案例：给用户推荐电影

□ Q: 找到潜在喜欢“黑客帝国”电影的用户

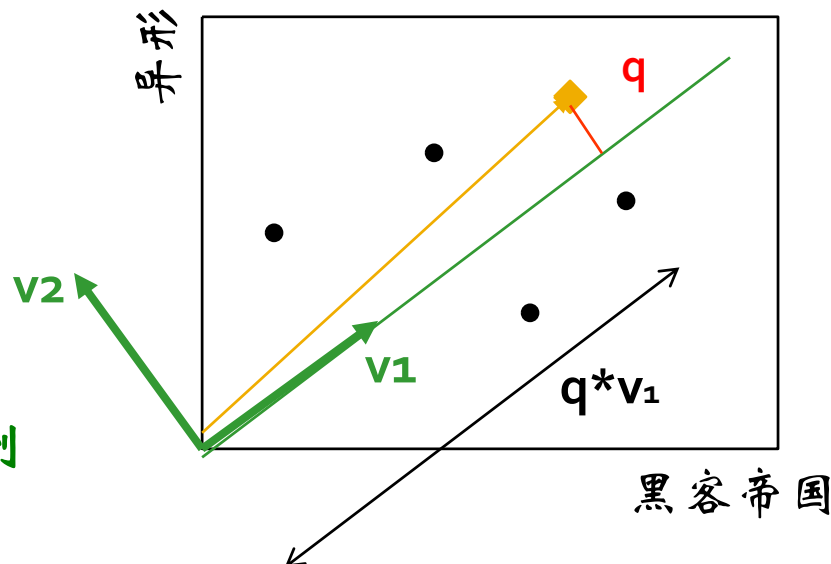
□ A: 将用户的搜索统计映射到“主题空间”

■ —— 怎么做？

黑客帝国 异形 冲出宁静号 卡萨布兰卡 天使爱美丽

$$q = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

与主题向量的基底做内积，投影到“主题空间”中。



案例：给用户推荐电影

□ 也就是说，我们有

□ $q_{\text{主题}} = qV$

$$q = \begin{matrix} & \begin{matrix} \text{黑} & \text{异形} & \text{冲出} & \text{卡萨} & \text{天使} \\ \text{帝} & & \text{宁静} & \text{布兰} & \text{爱} \\ \text{国} & & \text{号} & \text{兰} & \text{美} \\ & & & & \text{丽} \end{matrix} \\ \begin{matrix} 5 & 0 & 0 & 0 & 0 \end{matrix} \end{matrix} \times \begin{bmatrix} 0.56 & 0.12 \\ 0.59 & -0.02 \\ 0.56 & 0.12 \\ 0.09 & -0.69 \\ 0.09 & -0.69 \end{bmatrix} = \begin{bmatrix} 2.8 & 0.6 \end{bmatrix}$$

电影-主题近似
转换矩阵(V)

科幻-主题
↓



案例：给用户推荐电影

□ 对于喜欢“异形”、“冲出宁静号”的用户d

□ $d_{\text{主题}} = dV$

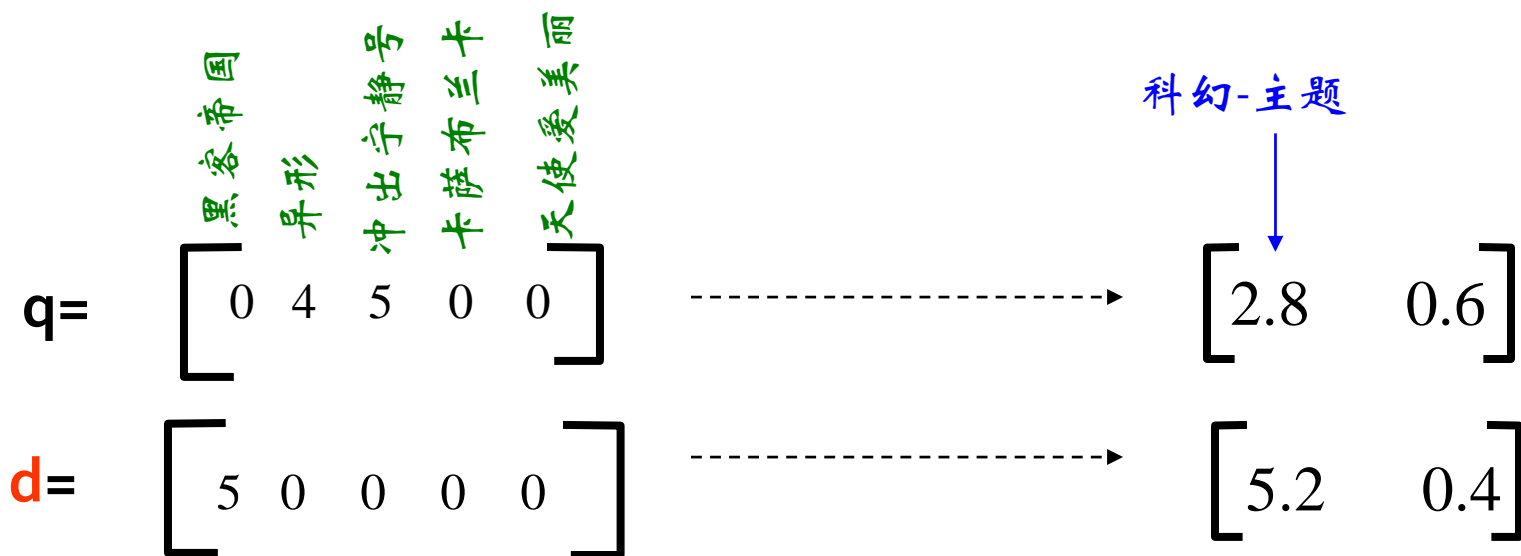
$$d = \begin{matrix} & \begin{matrix} \text{黑} & \text{异} & \text{冲} & \text{卡} & \text{天} \\ \text{客} & \text{形} & \text{出} & \text{萨} & \text{使} \\ \text{帝} & & \text{宁} & \text{布} & \text{爱} \\ \text{国} & & \text{静} & \text{兰} & \text{美} \\ & & \text{号} & & \text{丽} \end{matrix} \\ \begin{bmatrix} 0 & 4 & 5 & 0 & 0 \end{bmatrix} & \mathbf{X} & \begin{bmatrix} 0.56 & 0.12 \\ 0.59 & -0.02 \\ 0.56 & 0.12 \\ 0.09 & -0.69 \\ 0.09 & -0.69 \end{bmatrix} & = & \begin{bmatrix} 5.2 & 0.4 \end{bmatrix} \end{matrix}$$

科幻-主题
↓
电影-主题近似
转换矩阵(V)



案例：给用户推荐电影

□ 观察：用户d与用户q是相似的，虽然他们在原始坐标下是正交的！

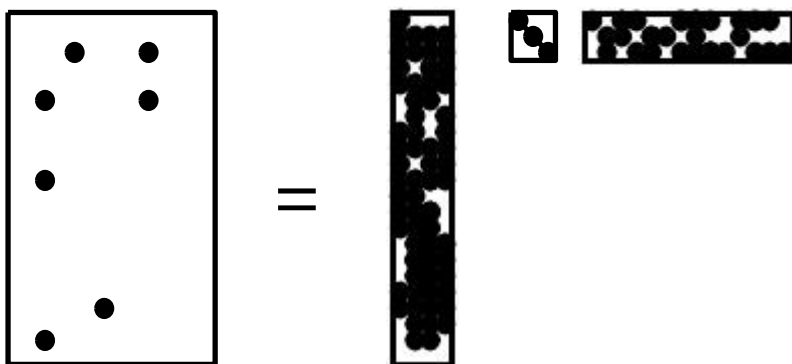


SVD降维的特点

□ 奇异向量

- 每一个奇异向量是所有输入矩阵的行向量或列向量的线性组合

□ 奇异向量是稠密的



CUR分解简介

□ 将矩阵A表示成C、U、R三个矩阵相乘

■ 使得 $\|A - C \cdot U \cdot R\|_F$ 变得很小。

$$\begin{pmatrix} \text{red bar} & \text{blue bar} & \text{dark red bar} \end{pmatrix} \begin{matrix} A \end{matrix} \approx \begin{pmatrix} \text{red bar} & \text{red bar} & \text{red bar} & \text{blue bar} & \text{dark red bar} & \text{dark red bar} \end{pmatrix} \cdot \begin{pmatrix} U \end{pmatrix} \cdot \begin{pmatrix} R \end{pmatrix}$$

$A \qquad C \qquad U \qquad R$



CUR分解简介

□ 将矩阵A表示成C、U、R三个矩阵相乘

■ 使得 $\|A - C \cdot U \cdot R\|_F$ 变得很小。

The diagram illustrates the CUR decomposition of matrix A. Matrix A is shown on the left as a vertical stack of three horizontal bars: red (top), dark brown (middle), and blue (bottom). It is followed by an approximation symbol \approx . Matrix C is a vertical stack of three dark brown horizontal bars. Matrix U is a light gray square with a red arrow pointing to its bottom row. Matrix R is a horizontal stack of six horizontal bars: three red (top), two dark brown (middle), and one blue (bottom). The equation is labeled A, C, U, and R below the respective matrices.

$$\begin{pmatrix} \text{red bar} \\ \text{dark brown bar} \\ \text{blue bar} \end{pmatrix} \approx \begin{pmatrix} \text{dark brown bar} \\ \text{dark brown bar} \\ \text{dark brown bar} \end{pmatrix} \cdot \begin{pmatrix} \text{gray square} \end{pmatrix} \cdot \begin{pmatrix} \text{red bar} \\ \text{red bar} \\ \text{red bar} \\ \text{dark brown bar} \\ \text{dark brown bar} \\ \text{blue bar} \end{pmatrix}$$

A C U R

C和R的交集的伪逆



SVD vs. CUR

$$\text{SVD: } A = U \Sigma V^T$$

Diagram illustrating the SVD decomposition $A = U \Sigma V^T$ with annotations:

- A : 大却稀疏 (Large but sparse)
- U : 大且稠密 (Large and dense)
- Σ : 稀疏且小 (Sparse and small)
- V^T : 大且稠密 (Large and dense)

$$\text{CUR: } A = C U R$$

Diagram illustrating the CUR decomposition $A = C U R$ with annotations:

- A : 大却稀疏 (Large but sparse)
- C : 大且稠密 (Large and dense)
- U : 稠密却小 (Dense but small)
- R : 大且稠密 (Large and dense)



感谢大家！

恳请大家批评指正！

