Topic modeling

Singular value decomposition (SVD)

& Non-negative matrix factorization (NMF)

Singular value decomposition

- **Factorization of a real or complex matrix.**
- **Generalizes the eigendecomposition of a square normal matrix with an orthonormal eigenbasis to any $\boldsymbol{\Lambda}$ matrix.**
- **SVD of an $\mathbf{m \times n}$ matrix A is a factorization of the form**

$$\mathbf{A = U\Lambda V^T}$$

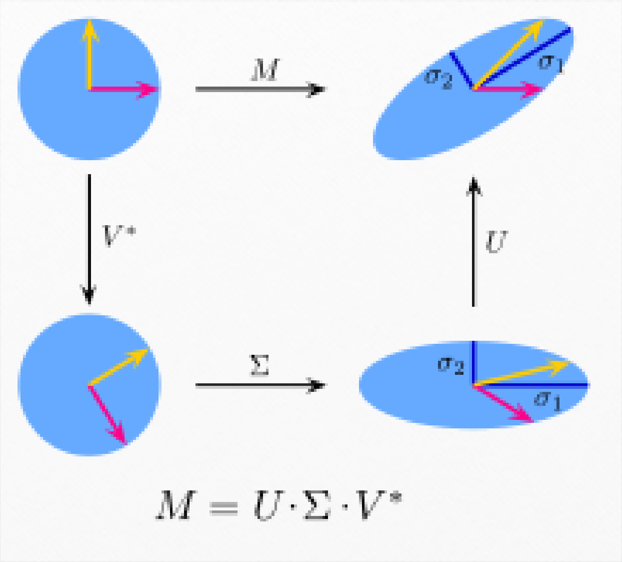
  - **U** is an $m \times n$ unitary matrix, $\mathbf{UU^T} = \boldsymbol{I}$
  - $\boldsymbol{\Lambda}$ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal,
  - **V** is an $n \times n$ unitary matrix,
  - $\mathbf{V^T}$ is the conjugate transpose of **V**. $\mathbf{VV^T} = \boldsymbol{I}$

Results

- $\boldsymbol{U}$ is the eigenvector of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. **Thus,** $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}})\boldsymbol{U}^{\boldsymbol{T}}$ **(left singular vector of** A**)**
- $\boldsymbol{V}$ is the eigenvector of $\mathbf{A}^{\mathrm{T}}\boldsymbol{A}$.
- **Thus,** $\mathbf{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{V}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma})\boldsymbol{V}^{\boldsymbol{T}}$ **(right singular vector of** A**)**
- $\boldsymbol{\sigma_i}$ is the square root of the eigenvalues of $\mathbf{A}^{\mathrm{T}}\boldsymbol{A}$ or $\mathbf{A}\mathbf{A}^{\mathrm{T}}$.

$$\boldsymbol{A} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{s}} \sigma_i \boldsymbol{u_i} \boldsymbol{v_i^T}$$

Geometric interpretation of SVD

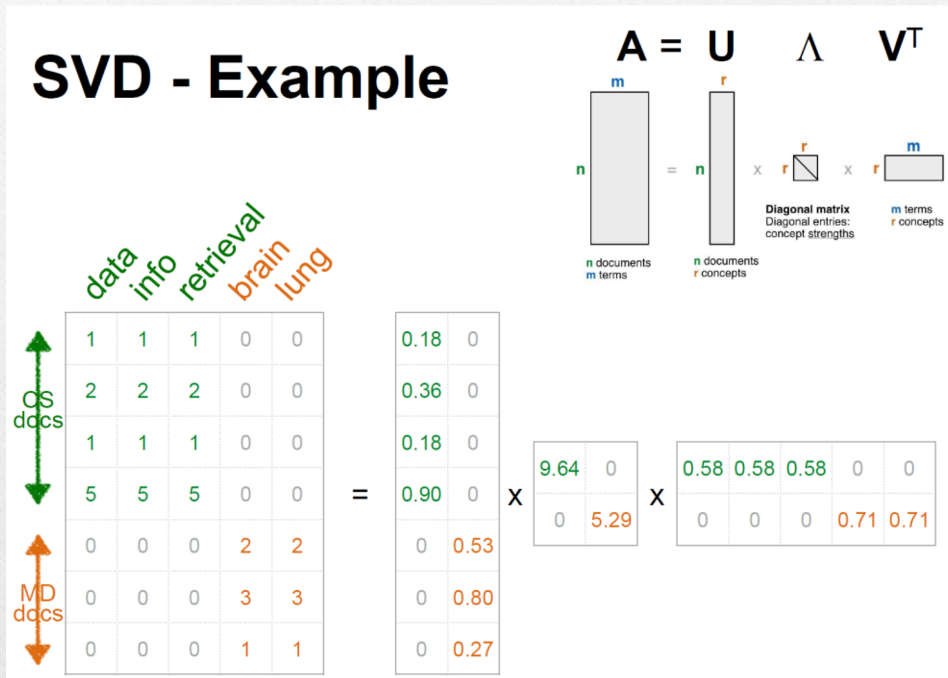


Source: https://en.wikipedia.org/wiki/Singular_value_decomposition

R code

```
library(OpenImageR)
x<-readImage("C:/Users/user/Documents/pansy.jpeg")
r<-rgb_2gray(x)
imageShow(r)
 r.svd<-svd(r)
plot(r.svd$d)
u<-r.svd$u
v<-r.svd$v
d<-diag(r.svd$d)
 depth<-50
us<-as.matrix(u[,1:depth])
vs<-as.matrix(v[,1:depth])
ds<-as.matrix(d[1:depth,1:depth])
ls<-us%*%ds%*%t(vs)
imageShow(ls)
```

Topic modeling using SVD



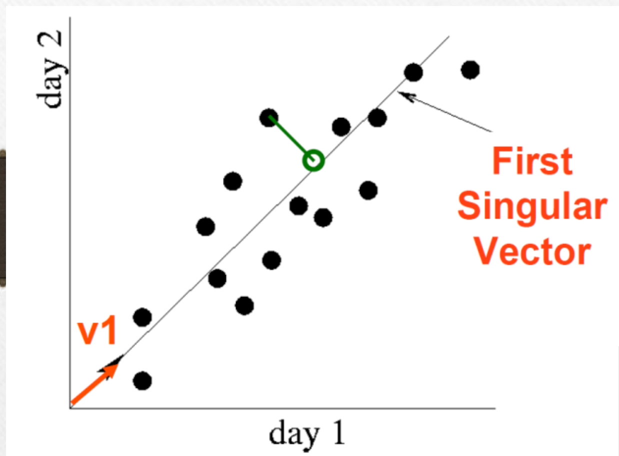
SVD -Interpretation #1

- ‘documents’, ‘terms’ and ‘concepts’:

- **U**: document-concept similarity matrix
- **V**: term-concept similarity matrix
- Λ: diagonal elements: concept “strengths”

- $\mathbf{A}^{\mathbf{T}}\boldsymbol{A} = ?$
- $\mathbf{A}\mathbf{A}^{\mathbf{T}} = ?$

SVD -Interpretation #2



SVD is closely related to PCA

variance ('spread') on the v1 axis

v1

$$\underbrace{\begin{bmatrix}1&1&1&0&0\\2&2&2&0&0\\1&1&1&0&0\\5&5&5&0&0\\0&0&0&2&2\\0&0&0&3&3\\0&0&0&1&1\end{bmatrix}}_{A} = \underbrace{\begin{bmatrix}0.18&0\\0.36&0\\0.18&0\\0.90&0\\0&0.53\\0&0.80\\0&0.27\end{bmatrix}}_{U} \times \underbrace{\begin{bmatrix}9.64&0\\0&5.29\end{bmatrix}}_{\Lambda} \times \underbrace{\begin{bmatrix}\boxed{0.58\;\;0.58\;\;0.58\;\;0\;\;0}\\0\;\;0\;\;0\;\;0.71\;\;0.71\end{bmatrix}}_{V^T}$$

How do we determine # of topics?

More details

Q: how exactly is dim. reduction done?

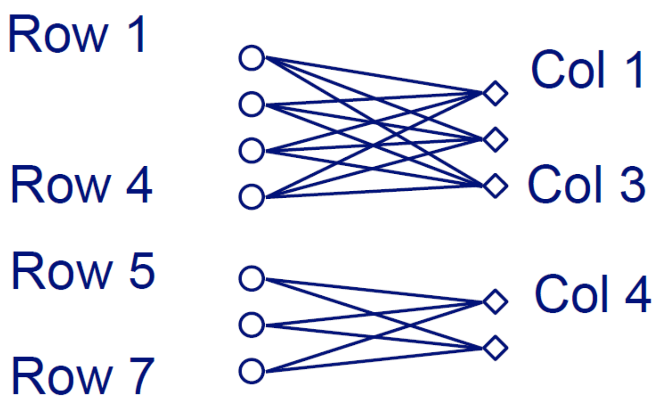
A: set the smallest singular values to zero:

$$\begin{bmatrix}1&1&1&0&0\\2&2&2&0&0\\1&1&1&0&0\\5&5&5&0&0\\0&0&0&2&2\\0&0&0&3&3\\0&0&0&1&1\end{bmatrix} = \begin{bmatrix}0.18&0\\0.36&0\\0.18&0\\0.90&0\\0&0.53\\0&0.80\\0&0.27\end{bmatrix} \times \begin{bmatrix}9.64&0\\0&\cancel{5.29}\end{bmatrix} \times \begin{bmatrix}0.58&0.58&0.58&0&0\\0&0&0&0.71&0.71\end{bmatrix}$$

SVD -Interpretation #3

- finds non-zero ‘blobs’ in a data matrix =
- ‘communities’ (bi-partite cores, here)

$$
\left[\begin{array}{ccc|cc}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
\hline
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{array}\right]
$$



NMF

- Given $X \in \mathbb{R}^{m \times n}$, compute an approximation $\mathrm{X} \approx WH$ for some matrix $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ where W and H are nonnegative matrices

$$\min_{W \in \mathbb{R}^{m \times k},\ H \in \mathbb{R}^{k \times n}} \|X - WH\|_F^2$$

subject to

$$W \geq 0$$

$$H \geq 0$$

Often positive factors will be **more easily interpretable**

Example

users

items	1		3			5			5		4	
			5		?		4			2	1	3
	2	4		1	2		3		4	3	5	
		2	4		5			4			2	
			4	3	4	2					2	5
	1		3		3			2			4	

~

items

.1	-.4	.2
-.5	.6	.5
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

●

users

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	.1

Applications to Topic modeling