# Low-Rank Decompositions of Matrices

# Objectives

- Define low-rank decomposition (matrix factorization)

- Explore applications

# Matrices represent many types of information 2

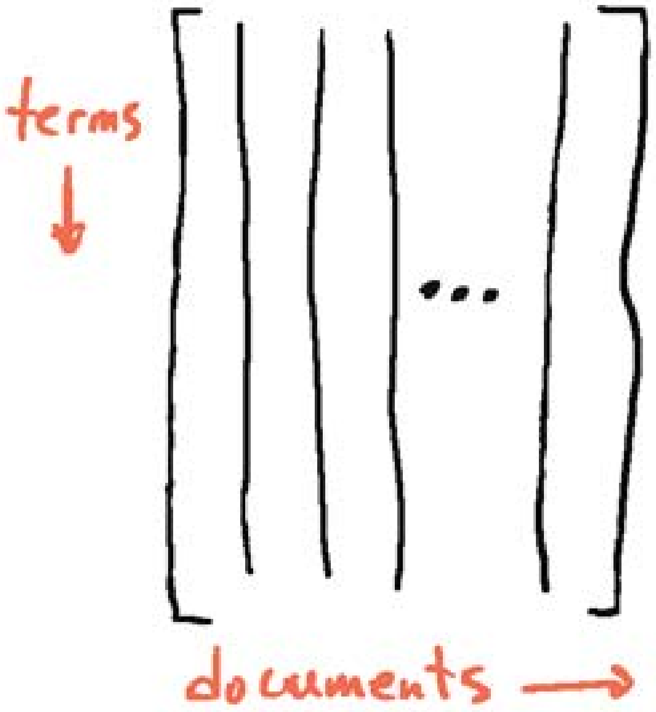
1) Features in classification or modeling
2) User ratings
3) Collections of documents

Bag of words model


"Apple trees blossom in May."

$$\begin{array}{l|c}
\text{word} & \text{frequency} \\
\hline
\text{agency} & 0 \\
\text{apple} & 1 \\
\text{blossom} & 1 \\
\text{car} & 0 \\
\text{currency} & 0 \\
\text{happy} & 0 \\
\text{May} & 1 \\
\text{politics} & 0 \\
\text{road} & 0 \\
\text{tree} & 1
\end{array}$$

Term-document matrix



# Low-rank decompositions emphasize patterns



$$\underset{N\times M}{\boxed{\underline{A}}} \approx \underset{N\times P}{\boxed{\underline{T}}}\ \underset{P\times M}{\boxed{\underline{W}^T}}$$

$P<N$, $P<M$ $\quad$ $\text{rank}\,\underline{T} = \text{rank}\,\underline{W} = P$

$\Rightarrow \text{rank}(\underline{T}\,\underline{W}^T) = P$

Let $\underline{T} = [\underline{t}_1\ \underline{t}_2 \cdots \underline{t}_P]$, $\underline{W} = [\underline{w}_1 \cdots \underline{w}_P]$

$$\underline{T}\,\underline{W}^T = \underbrace{\begin{bmatrix} | & | & & | \\ \underline{t}_1 & \underline{t}_2 & \cdots & \underline{t}_P \\ | & | & & | \end{bmatrix}}_{\text{column patterns}} \underbrace{\begin{bmatrix} - \underline{w}_1^T - \\ - \underline{w}_2^T - \\ \vdots \\ - \underline{w}_P^T - \end{bmatrix}}_{\text{row patterns}} = \sum_{i=1}^{P} \underbrace{\underline{t}_i \underline{w}_i^T}_{\substack{\text{rank-one}\\ \text{patterns}}} = \sum_{i=1}^{P} \Big|_{N}\ \overline{\phantom{xxxx}}^{\,M} = \sum_{i=1}^{P} \boxed{N\times M}$$

## Finding patterns –

1) $\min\limits_{\underline{T},\underline{W}} \|\underline{A} - \underline{T}\,\underline{W}^T\|$

singular value decomposition

2) $\underline{A} \approx \underline{T}\,\underline{W}^T$, $\underline{T}, \underline{W} \geq 0$

nonnegative matrix factorization

3) $\underline{A} \approx \underline{T}\,\underline{W}^T$

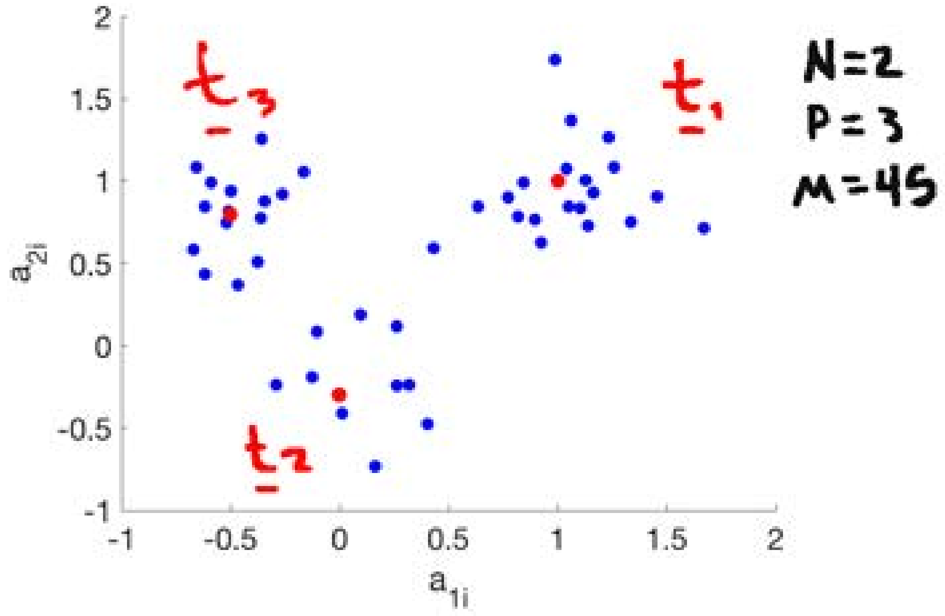
each col $\underline{W}^T$ all 0 w. single 1

clustering

# Clustering groups similar columns


$$\begin{bmatrix} \underline{a}_1 & \underline{a}_2 & \underline{a}_3 & \underline{a}_4 \cdots \underline{a}_M \end{bmatrix} \approx \begin{bmatrix} \underline{t}_1 & \underline{t}_2 & \underline{t}_3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 & & 0 \\ 1 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 1 & & 0 \end{bmatrix}$$

$$\Rightarrow \underline{a}_1 \approx \underline{t}_2, \; \underline{a}_M \approx \underline{t}_2, \; \underline{a}_2 \approx \underline{t}_1, \; \underline{a}_3 \approx \underline{t}_1, \; \underline{a}_4 \approx \underline{t}_3 \cdots$$




Group similar documents, customers, products, etc

Many algorithms –

k-means

# Low rank models "complete" missing data

$$
\begin{array}{l} \\ \text{Star Trek} \\ \text{Pride + Prejudice} \\ \text{The Martian} \\ \text{Sense + Sensibility} \\ \text{Empire Strikes Back} \end{array}
\begin{array}{c} \text{Jill} \\ \left[\begin{array}{c} 8 \\ 3 \\ 7 \\ 4 \\ ? \end{array}\right] \end{array}
$$

Suppose $\underline{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} w_1 + \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} w_2$

Use known ratings to solve $w_1, w_2$

$$
\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 3 \\ 7 \\ 4 \end{bmatrix} \Rightarrow \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 5.5 \\ 2 \end{bmatrix}
$$

Predict ratings using $w_1, w_2$

$$
\hat{\underline{a}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} 5.5 + \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} 2 = \begin{bmatrix} 7.5 \\ 3.5 \\ 7.5 \\ 3.5 \\ 7.5 \end{bmatrix}
$$

# Use of "patterns" can mitigate noise

Noisy data $\underline{A}_m = \underset{\text{strong patterns}}{\underline{A}_t} + \underset{\text{no dominant patterns}}{\underline{E}}$

Low-rank model $\hat{\underline{A}}_m = \underline{T}\,\underline{W}^T$ can be closer to $\underline{A}_t$ than $\underline{A}_m$

## Low rank classifier / model fit

$$\hat{\underline{A}}_m \underline{w} = \underline{d}$$

$$\underline{T}\,\underbrace{\underline{W}^T \underline{w}}_{\underline{w}'\ p\times 1} = \underline{d}$$

$$\underline{T}\,\underline{w}' = \underline{d}$$

$\underline{T}$: transformed features $\underline{x}_i'^T$

$$x_i'^T = \underline{x}_i^T \underline{W}(\underline{W}^T\underline{W})^{-1}$$

New feature $\underline{x}^T$:

1) $\underline{x}'^T = \underline{x}^T \underline{W}(\underline{W}^T\underline{W})^{-1}$ 2) $\hat{d} = \operatorname{sign}(\underline{x}'^T \underline{w}')$