Knowledge Discovery and Data Mining 1 (VO) (707.003)

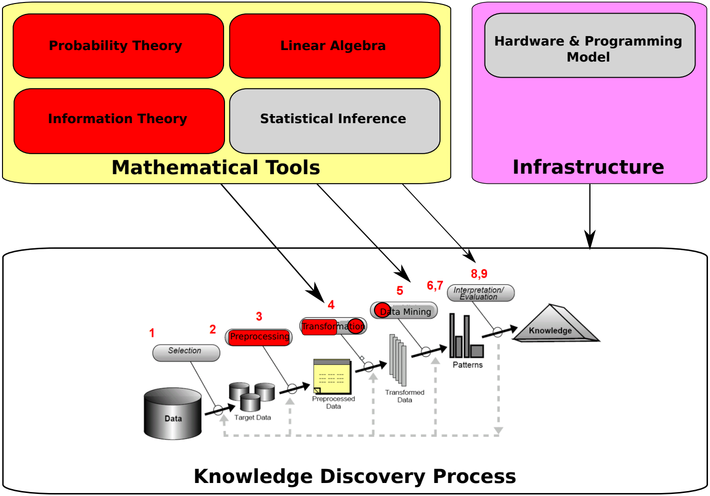
Singular Value Decomposition and Latent Semantic Analysis

Denis Helic

KTI, TU Graz

Nov 20, 2014

Big picture: KDDM



Outline

Slides

Slides are partially based on "Mining Massive Datasets" Chapter 11, "Introduction to Information Retrieval" by Manning, Raghavan and Schütze, and Melanie Martin AI Seminar

Recap

Recap

Review of data matrices

Recap – PCA: Algorithm

- Organize data as an $n \times m$ matrix, with n data points and m features
- **Subtract the average for each feature** to obtain centered data matrix $\mathbf{X}$
- Calculate the covariance matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- Calculate the eigenvalues and the eigenvectors of the covariance matrix
- Select the top r eigenvectors
- Project the data to the new space spanned by those r eigenvectors: $\mathbf{XE} \in \mathbb{R}^{n \times r}$, where $\mathbf{E} \in \mathbb{R}^{m \times r}$

Recap – PCA: Interpretation

- You can interpret the first couple of principal components to learn something about the dataset
- Data mining
- However, be very careful: you can not generalize from a single dataset
- PCA transforms the set of correlated observations into a set of linearly uncorrelated observations
- I.e. the goal of the analysis is to decorrelate the data

SVD

- We investigate now a second form of matrix analysis called **Singular Value Decomposition**
- It allows an exact representation of any matrix
- It decomposes a matrix into a product of three matrices
- It also provides an elegant way of dimensionality reduction
- It is easy to eliminate the less important parts of the representation

SVD

- SVD is based on the idea that there exist a small number of "concepts" that connect the rows and columns of the matrix
- SVD can rank the "concepts" from the most to the least important
- This ranking may be used to remove the less important "concepts" from the matrix
- This process closely approximates the original matrix

SVD: Definition

Singular Value Decomposition

Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be a matrix and let r be the rank of $\mathbf{M}$ (the rank of a matrix is the largest number of linearly independent rows or columns). Then we can find matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$ with the following properties:

- $\mathbf{U} \in \mathbb{R}^{m \times r}$ is a column-orthonormal matrix
- $\mathbf{V} \in \mathbb{R}^{n \times r}$ is a column-orthonormal matrix
- $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix.

The matrix $\mathbf{M}$ can be then written as:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

SVD form

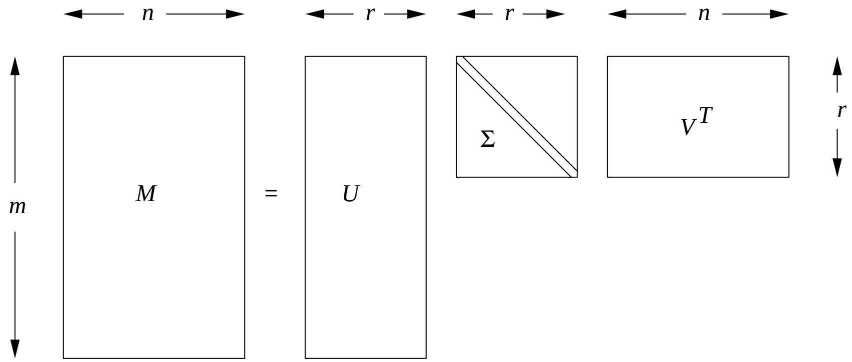


Figure : Figure from "Mining Massive Datasets"

SVD: Simple example

- Let us decompose a utility matrix of a movie recommender system
- Thus, we have users who rate movies
- Let there be two “concepts” which underlie the movies and steer the rating process
- E.g. let these concepts represent two movie genres: science fiction and romance
- Let all the boys rate only science fiction and all the girls only romance

SVD: Simple example

Movie / User	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

SVD: Simple example

- This strict adherence to those two concepts gives the matrix a rank of 2
- E.g. we may pick one of the first four rows and one of the last three rows and we can not find a nonzero linear combination that gives $\mathbf{0}$
- But we can not pick three independent rows
- E.g. if we pick rows 1, 2 and 7 then three times the first minus the second plus zero times the seventh gives $\mathbf{0}$

SVD: Simple example

- Similarly for columns
- We may pick one of the first three and one of the last two and they will be independent
- But we can not pick three independent columns
- E.g. if we pick columns 1, 2, and 5 then the first minus the second plus zero times the fifth gives $\mathbf{0}$
- Thus, the rank is indeed $r = 2$ and $\mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$

SVD: Simple example

- We will see later how to calculate the decomposition

$$\mathbf{U} = \begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}$$

SVD: Interpretation

- The key to understanding SVD is in viewing the r columns of $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ as representing concepts that are hidden or *latent* in the original matrix $\mathbf{M}$
- In our example these concepts are clear
- One is science fiction
- The other one is romance

SVD: Interpretation

- The rows of **M** are people
- The columns of **M** are movies
- Then the rows of **U** are people
- The columns of **U** are concepts
- **U** connects people to concepts

SVD: Interpretation

- For example, the person Joe (the first row in $\mathbf{M}$) likes only science fiction

$$
\mathbf{U} = \begin{pmatrix}
0.14 & 0 \\
0.42 & 0 \\
0.56 & 0 \\
0.70 & 0 \\
0 & 0.60 \\
0 & 0.75 \\
0 & 0.30
\end{pmatrix}
$$

SVD: Interpretation

- The value of 0.14 in the first row and first column of **U** indicates this fact
- However, this value is smaller than some of other values in the first column of **U**

$$\mathbf{U} = \begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}$$

SVD: Interpretation

- Because while Joe watches only science fiction he does not rate these movies highly
- Thus, Joe contributes to the concept of science fiction but not as much as e.g. Jack who rated these movies highly

$$\mathbf{U} = \begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}$$

SVD: Interpretation

- On the other hand, the second column of the first row of $\mathbf{U}$ is zero

$$\mathbf{U} = \begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}$$

SVD: Interpretation

- Joe does not rate romance movies at all and does not contribute anything to that concept

$$
\mathbf{U} = \begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}
$$

SVD: Simple example

$$\mathbf{V}^T = \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

SVD: Interpretation

- The rows of $\mathbf{M}$ are people
- The columns of $\mathbf{M}$ are movies
- Then the rows of $\mathbf{V}^T$ are concepts
- The columns of $\mathbf{V}^T$ are movies
- $\mathbf{V}$ connects movies to concepts

SVD: Interpretation

- For example, the 0.58 in the first three columns of the first row of $\mathbf{V}^T$ indicates that the first three movies are of science fiction genre
- Matrix, Alien and Star Wars

$$\mathbf{V}^T = \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

SVD: Interpretation

- On the other hand, the last two movies have nothing to do with science fiction
- Casablanca and Titanic

$$\mathbf{V}^T = \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

SVD: Interpretation

- Also, Matrix, Alien and Star Wars do not partake of the concept of romance at all
- As indicated by 0's in the first three columns of the second row

$$\mathbf{V}^T = \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

SVD: Interpretation

- Whereas, Casablanca and Titanic are romance movies
- The 0.71 in the last two columns of the second row

$$\mathbf{V}^T = \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

SVD: Simple example

$$\boldsymbol{\Sigma} = \begin{pmatrix} 12.4 & 0 \\ 0 & 9.5 \end{pmatrix}$$

SVD: Interpretation

- Finally, the matrix $\mathbf{\Sigma}$ gives the strength of each concept
- In our example the strength of science fiction is 12.4
- The strength of romance is 9.4
- Intuitively, science fiction is a stronger concept because the data provides more movies of that genre and more people who rate these movies

SVD: Simple example

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{pmatrix}}_{\mathbf{M}} = \underbrace{\begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix}}_{\mathbf{U}} \times \underbrace{\begin{pmatrix} 12.4 & 0 \\ 0 & 9.5 \end{pmatrix}}_{\mathbf{\Sigma}} \times \underbrace{\begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}}_{\mathbf{V}^T}$$

SVD: Interpretation

- In general, the concepts will not be so clearly composed
- There will be fewer 0's in $\mathbf{U}$ and $\mathbf{V}$
- $\boldsymbol{\Sigma}$ is always diagonal
- Typically the entities represented by the rows and columns of $\mathbf{M}$ will contribute to several different concepts to varying degrees

SVD: Interpretation

- The decomposition of the simple example was especially simple because the rank of the matrix $\mathbf{M}$ was equal to the number of concepts
- In practice that is rarely the case
- The rank r will be in many cases greater than the number of the concepts and some of the columns in $\mathbf{U}$ are harder to interpret

SVD: Another example

User \ Movie	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

SVD: Another example

- In this (more realistic) example Jill and Jane rated "Alien"
- Neither liked it much, but nevertheless they rated it
- This gives the matrix a rank of 3
- E.g. we may pick the first, sixth, and seventh rows and check that they are independent
- However no four rows are independent

SVD: Another example

- Thus, in our decomposition we have $r = 3$
- We will have three columns in $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$
- The first column corresponds to science fiction
- The second column corresponds to romance
- The interpretation of the third column is not easy (it is a linear combination of the users)
- Nice property: the third columns is the least important one

SVD: Another example

$$\mathbf{U} = \begin{pmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{pmatrix}$$

SVD: Another example

$$\mathbf{V}^T = \begin{pmatrix} 0.56 & 0.59 & 0.56 & 0.9 & 0.9 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{pmatrix}$$

SVD: Another example

$$\mathbf{\Sigma} = \begin{pmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{pmatrix}$$

SVD: Another example

$$
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{pmatrix} =
$$

$$
\begin{pmatrix}
0.13 & 0.02 & -0.01 \\
0.41 & 0.07 & -0.03 \\
0.55 & 0.09 & -0.04 \\
0.68 & 0.11 & -0.05 \\
0.15 & -0.59 & 0.65 \\
0.07 & -0.73 & -0.67 \\
0.07 & -0.29 & 0.32
\end{pmatrix} \times
\begin{pmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{pmatrix} \times
\begin{pmatrix}
0.56 & 0.59 & 0.56 & 0.9 & 0.9 \\
0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
0.40 & -0.80 & 0.40 & 0.09 & 0.09
\end{pmatrix}
$$

Matrix diagonalization theorem

Theorem

Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ *be a square matrix with n linearly independent eigenvectors. Then there exists an eigen decomposition:*

$$\mathbf{S} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^{-1}$$

where the columns of $\mathbf{U}$ *are the eigenvectors of* $\mathbf{S}$ *and* $\boldsymbol{\Lambda}$ *is a diagonal matrix whose diagonal entries are the eigenvalues of* $\mathbf{S}$ *in decreasing order*

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \dots & \\ & & & \lambda_n \end{pmatrix}, \lambda_i \geq \lambda_{i+1}.$$

If the eigenvalues are distinct, then this decomposition is unique.

Matrix diagonalization theorem

- How does this theorem work?
- $\mathbf{U}$ has eigenvectors of $\mathbf{S}$ as its columns

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots \mathbf{u}_n \end{pmatrix}$$

Matrix diagonalization theorem

- Then we have

$$
\begin{aligned}
\mathbf{SU} &= \mathbf{S} \times \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots \mathbf{u}_n \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 \mathbf{u}_1 & \lambda_2 \mathbf{u}_2 & \ldots \lambda_n \mathbf{u}_n \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots \mathbf{u}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ldots & \\ & & & \lambda_n \end{pmatrix} \\
&= \mathbf{U\Lambda}
\end{aligned}
$$

Matrix diagonalization theorem

- Thus we have

$$\mathbf{SU} = \mathbf{U\Lambda}$$

$$\mathbf{S} = \mathbf{U\Lambda U}^{-1}$$

Symmetric diagonalization theorem

Theorem

Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be a square symmetric matrix with n linearly independent eigenvectors. Then there exists a symmetric diagonal decomposition:

$$\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

where the columns of $\mathbf{Q}$ are the orthogonal and normalized eigenvectors of $\mathbf{S}$ (i.e. $\mathbf{Q}$ is an orthonormal matrix) and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{S}$. Further, all entries of $\mathbf{Q}$ are real and we have $\mathbf{Q}^{-1} = \mathbf{Q}^T$.

SVD

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

- Let us calculate $\mathbf{M}^T$

$$\begin{aligned}
\mathbf{M}^T &= (\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T)^T \\
&= (\mathbf{V}^T)^T\boldsymbol{\Sigma}^T\mathbf{U}^T \\
&= \mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T \\
&= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T
\end{aligned}$$

- The last equality since $\boldsymbol{\Sigma}$ is diagonal and thus $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$

SVD

$$\mathbf{M} = \mathbf{U\Sigma V}^T$$

$$\mathbf{M}^T = \mathbf{V\Sigma U}^T$$

- Let us calculate $\mathbf{MM}^T$

$$\begin{aligned}\mathbf{MM}^T &= \mathbf{U\Sigma V}^T\mathbf{V\Sigma U}^T \\ &= \mathbf{U\Sigma}^2\mathbf{U}^T\end{aligned}$$

SVD

$$\mathbf{M}\mathbf{M}^T = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T$$

- Thus, we have $\mathbf{M}\mathbf{M}^T = \mathbf{S}$ and $\boldsymbol{\Sigma}^2 = \boldsymbol{\Lambda}$
- That is $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{M}\mathbf{M}^T$
- $\boldsymbol{\Sigma}$ is the matrix of square roots of the eigenvalues of $\mathbf{M}\mathbf{M}^T$

SVD

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

$$\mathbf{M}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$$

- Let us calculate $\mathbf{M}^T\mathbf{M}$

$$\begin{aligned} \mathbf{M}^T\mathbf{M} &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \end{aligned}$$

SVD

$$\mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$

- Thus, we have $\mathbf{M}^T\mathbf{M}\mathbf{V}$ and $\mathbf{\Sigma}^2 = \mathbf{\Lambda}$
- That is $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{M}^T\mathbf{M}\mathbf{V}$
- $\mathbf{\Sigma}$ is the matrix of square roots of the eigenvalues of $\mathbf{M}^T\mathbf{M}\mathbf{V}$

SVD

- What is the relationship between eigenvalues of $\mathbf{MM}^T$ and $\mathbf{M}^T\mathbf{M}$
- Suppose $\mathbf{e}$ is an eigenvector of $\mathbf{MM}^T$

$$\mathbf{MM}^T\mathbf{e} = \lambda\mathbf{e}$$

SVD

- We multiply both sides of the equation by $\mathbf{M}^T$ on the left

$$
\begin{aligned}
\mathbf{M}^T\mathbf{M}\mathbf{M}^T\mathbf{e} &= \mathbf{M}^T\lambda\mathbf{e} \\
\mathbf{M}^T\mathbf{M}(\mathbf{M}^T\mathbf{e}) &= \lambda(\mathbf{M}^T\mathbf{e})
\end{aligned}
$$

- As long as $(\mathbf{M}^T\mathbf{e})$ is not the zero vector $\mathbf{0}$ it will be an eigenvector of $\mathbf{M}^T\mathbf{M}$

SVD

- The converse holds as well
- Suppose $\mathbf{e}$ is an eigenvector of $\mathbf{M}^T\mathbf{M}$

$$\mathbf{M}^T\mathbf{M}\mathbf{e} = \lambda\mathbf{e}$$

SVD

- We multiply both sides of the equation by $\mathbf{M}$ on the left

$$
\begin{aligned}
\mathbf{M}\mathbf{M}^T\mathbf{M}\mathbf{e} &= \mathbf{M}\lambda\mathbf{e} \\
\mathbf{M}\mathbf{M}^T(\mathbf{M}\mathbf{e}) &= \lambda(\mathbf{M}\mathbf{e})
\end{aligned}
$$

- As long as $(\mathbf{M}\mathbf{e})$ is not the zero vector $\mathbf{0}$ it will be an eigenvector of $\mathbf{M}\mathbf{M}^T$

SVD

- What happens when e.g. $\mathbf{Me} = \mathbf{0}$

$$\begin{aligned}\mathbf{M}^T\mathbf{Me} &= \mathbf{0}\\ \mathbf{M}^T\mathbf{Me} &= \lambda\mathbf{e}\\ \lambda\mathbf{e} &= \mathbf{0}\end{aligned}$$

- Since $\mathbf{e}$ is not $\mathbf{0}$ it must be $\lambda = 0$

SVD

- Conclusion: eigenvalues of $\mathbf{M}^T\mathbf{M}$ are eigenvalues of $\mathbf{M}\mathbf{M}^T$ plus additional zeros
- If the dimension of $\mathbf{M}^T\mathbf{M}$ were less than the dimension of $\mathbf{M}\mathbf{M}^T$
- If the dimension of $\mathbf{M}^T\mathbf{M}$ were greater than the dimension of $\mathbf{M}\mathbf{M}^T$ than opposite is true
- Eigenvalues of $\mathbf{M}\mathbf{M}^T$ are eigenvalues of $\mathbf{M}^T\mathbf{M}$ plus additional zeros

SVD

- $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{MM}^T$
- $\mathbf{\Sigma}$ is the matrix of square roots of the non-zero eigenvalues of $\mathbf{MM}^T$
- $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{M}^T\mathbf{M}$
- $\mathbf{\Sigma}$ is the matrix of square roots of the non-zero eigenvalues of $\mathbf{M}^T\mathbf{M}$
- These are equal values
- This gives also the algorithm for calculating the decomposition: eigenvalues and eigenvectors of $\mathbf{M}^T\mathbf{M}$ and $\mathbf{MM}^T$

SVD Interpretation

- What does $\mathbf{M}\mathbf{M}^T$ represent?
- It is a square matrix with rows and columns corresponding to e.g. people
- Each element measures the overlap between the people based on their co-ratings of the movies
- It is the sum of the products of their movie ratings

SVD Interpretation

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{pmatrix}$$

SVD Interpretation

$$\mathbf{MM}^T = \begin{pmatrix} 3 & 9 & 12 & 15 & 2 & 0 & 1 \\ 9 & 27 & 36 & 45 & 6 & 0 & 3 \\ 12 & 36 & 48 & 60 & 8 & 0 & 4 \\ 15 & 45 & 60 & 75 & 10 & 0 & 5 \\ 2 & 6 & 8 & 10 & 36 & 40 & 18 \\ 0 & 0 & 0 & 0 & 40 & 50 & 20 \\ 1 & 3 & 4 & 5 & 18 & 20 & 9 \end{pmatrix}$$

SVD Interpretation

- What does $\mathbf{M}^T\mathbf{M}$ represent?
- It is a square matrix with rows and columns corresponding to e.g. movies
- Each element measures the overlap between the movies based on their co-ratings by people
- It is the sum of the products of the ratings that they got from different people

SVD Interpretation

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{pmatrix}$$

SVD Interpretation

$$\mathbf{M}^T\mathbf{M} = \begin{pmatrix} 51 & 51 & 51 & 0 & 0 \\ 51 & 56 & 51 & 10 & 10 \\ 51 & 51 & 51 & 0 & 0 \\ 0 & 10 & 0 & 45 & 45 \\ 0 & 10 & 0 & 45 & 45 \end{pmatrix}$$

SVD dimensionality reduction

- Can we use SVD for dimensionality reduction?
- Suppose we want to represent a very large matrix $\mathbf{M}$ by its SVD components $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$
- We interpreted the entries in $\mathbf{\Sigma}$ as the measure of importance of concepts
- Thus, we might set the s smallest entries in $\mathbf{\Sigma}$ to zero
- With this we eliminate the s rows of $\mathbf{U}$ and $\mathbf{V}$

SVD dimensionality reduction

$$
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{pmatrix} =
$$

$$
\begin{pmatrix}
0.13 & 0.02 & -0.01 \\
0.41 & 0.07 & -0.03 \\
0.55 & 0.09 & -0.04 \\
0.68 & 0.11 & -0.05 \\
0.15 & -0.59 & 0.65 \\
0.07 & -0.73 & -0.67 \\
0.07 & -0.29 & 0.32
\end{pmatrix} \times
\begin{pmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{pmatrix} \times
\begin{pmatrix}
0.56 & 0.59 & 0.56 & 0.9 & 0.9 \\
0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
0.40 & -0.80 & 0.40 & 0.09 & 0.09
\end{pmatrix}
$$

SVD dimensionality reduction

- Now we set the smallest value in $\mathbf{\Sigma}$ to 0 and eliminate the corresponding rows and columns in **U** and **V**

$$\begin{pmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{pmatrix} \times \begin{pmatrix} 12.4 & 0 \\ 0 & 9.5 \end{pmatrix} \times \begin{pmatrix} 0.56 & 0.59 & 0.56 & 0.9 & 0.9 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \end{pmatrix}$$

SVD dimensionality reduction

$$\begin{pmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{pmatrix} \times \begin{pmatrix} 12.4 & 0 \\ 0 & 9.5 \end{pmatrix} \times \begin{pmatrix} 0.56 & 0.59 & 0.56 & 0.9 & 0.9 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \end{pmatrix} =$$

$$\begin{pmatrix} 0.93 & 0.95 & 0.93 & 0.014 & 0.014 \\ 2.93 & 2.99 & 2.93 & 0.000 & 0.000 \\ 3.92 & 4.01 & 3.92 & 0.026 & 0.026 \\ 4.84 & 4.96 & 4.84 & 0.040 & 0.040 \\ 0.37 & 1.21 & 0.37 & 4.04 & 4.04 \\ 0.35 & 0.65 & 0.35 & 4.87 & 4.87 \\ 0.16 & 0.57 & 0.16 & 1.98 & 1.98 \end{pmatrix}$$

Advanced: SVD dimensionality reduction

- The resulting matrix is quite close to the original matrix
- Thus, this approach to dimensionality reduction seems to work quite well
- However, since we approximate $\mathbf{M} \rightarrow$ we need to measure the approximation error
- We can pick among several measures for this error
- For SVD decomposition we might pick Frobenius norm, which is proportional to RMSE

Advanced: SVD dimensionality reduction

- Frobenius norm $||\mathbf{M}||$ of a matrix $\mathbf{M}$ is the square root of the sum of the squares of the elements of $\mathbf{M}$

$$||\mathbf{M}|| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} m_{ij}^2}$$

Advanced: SVD dimensionality reduction

- It can be shown that:

$$
\begin{aligned}
||\mathbf{M}|| &= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} m_{ij}^2} \\
&= \sqrt{tr(\mathbf{M}^T\mathbf{M})} \\
&= \sqrt{\sum_{i=1}^{min(m,n)} \lambda_i} \\
&= \sqrt{\sum_{i=1}^{min(m,n)} \sigma_i^2}
\end{aligned}
$$

Advanced: SVD dimensionality reduction

- Now suppose we want to approximate $\mathbf{M}$ with a matrix $\mathbf{M}'$ of the rank $k < r$ such that $||\mathbf{M} - \mathbf{M}'||$ is minimal
- Thus, we minimize the Frobenius norm of the difference between the original matrix and its approximation
- Eckart-Young theorem states that the solution to this problem is given by

$$\mathbf{M}' = \mathbf{U}\mathbf{\Sigma}'\mathbf{V}^T$$

- $\mathbf{\Sigma}'$ is the same matrix as $\mathbf{\Sigma}$ except that it contains k largest singular values and $r - k$ of the smallest values are replaces by zero

Advanced: SVD dimensionality reduction

- General proof is complicated
- However, if we assume that the optimal solution is of the form $\mathbf{M}' = \mathbf{U}\mathbf{\Sigma}'\mathbf{V}^T$
- Then we can easily show that setting the smallest singular values to zero reduces the Frobenius norm of the difference at most

Advanced: SVD dimensionality reduction

$$\mathbf{M} - \mathbf{M}' = \mathbf{U}(\mathbf{\Sigma} - \mathbf{\Sigma}')\mathbf{V}^T$$

$$||\mathbf{M} - \mathbf{M}'|| = \sqrt{\sum_{i=1}^{min(m,n)} (\sigma_i - \sigma_i')^2}$$

- The singular values of this matrix are kept in $\mathbf{\Sigma} - \mathbf{\Sigma}$'
- These are zeros for $r - k$ singular values that we choose to keep
- They are non-zeros for all singular values that we set to zero
- Thus, to minimize the Frobenius norm we should set the smallest values to zero

SVD dimensionality reduction

- How many singular values should we keep?
- A useful rule of the thumb is to keep enough singular values to make up 90% of *energy* in $\mathbf{\Sigma}$
- We define energy as the sum of squares of singular values

$$\sum_{i=1}^{min(m,n)} \sigma_i^2$$

SVD dimensionality reduction

- In the example the total energy:

$$12.4^2 + 9.5^2 + 1.3^2 = 245.7$$

- By removing the smallest singular value: $(12.4^2 + 9.5^2 = 244.01)$ we keep 99% of the energy
- By also removing the second smallest singular value: $(12.4^2 = 153.76)$ we would keep only 63% of the energy

SVD of a term-document matrix

- Vector Space Model: documents are represented as term vectors
- The complete document collection is represented as a large term-document matrix
- This has many advantages especially in the field of information retrieval
- Both documents and queries are treated uniformly
- Cosine similarity to compute scores
- The ability to weight different terms differently, e.g. tf-idf

SVD of a term-document matrix

- Vector Space Model can not cope with two classic problems arising in natural languages
- *Synonymy*: two words having the same meaning
- E.g. “car” and “automobile”
- Those synonym words get separate dimensions in the VSM
- The model would underestimate the similarity of a document containing both “car” and “automobile” to a query containing only “car”

SVD of a term-document matrix

- The second problem
- *Polysemy*: one word having multiple meanings
- E.g. "bank" may mean a financial institution or a river bank
- The VSM would overestimate the similarity of a query containing "bank" to a document that contains the word "bank" in both senses
- Can we use co-occurrences of the terms to distinguish between those two cases?
- "Bank" co-occurs in a document with "money" vs. it co-occurs in a document with "dam"

SVD of a term-document matrix

- Another problem is the dimension of the term-document matrix
- In latent semantic analysis (LSA) or latent semantic indexing (LSI) we use SVD to create a low-rank approximation of the term-document matrix
- We select k largest singular values and create $\mathbf{M}_k$ approximation to the original matrix
- We thus map each term/document to a k-dimensional space of "concepts"

SVD of a term-document matrix

- These concepts are hidden (latent) in the collection
- They represent the semantic of the terms and documents
- E.g. the topics of terms and documents
- In practice, however the interpretation is rather difficult

SVD of a term-document matrix

- By computing low-rank approximation of the original term-document matrix the SVD brings together the terms with similar co-occurrences
- Retrieval quality may actually be improved by the approximation!
- Confirmed by experiments
- Retrieval by folding the query into the low-rank space

$$\mathbf{q}_k = \mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{q}$$

SVD of a term-document matrix

- Computational cost is significant
- As we reduce k recall improves
- A value of k in low hundreds tend to increase precision as well (this suggests that a suitable k addresses some of the challenges of synonymy)
- LSI works best in applications where there is little overlap between documents and the query
- LSI can be viewed as a soft clustering method
- Each concept is a cluster and the value that a document has at that concept is its fractional membership in that concept

LSA: Example

- Technical memo titles
- Two different collections
- The first about HCI
- The second about graph theory

Example

Example from Melanie Martin AI Seminar

LSA: Example

c1: Human machine interface for ABC computer applications

c2: A survey of user opinion of computer system response time

c3: The EPS user interface management system

c4: System and human system engineering testing of EPS

c5: Relation of user perceived response time to error measurement

LSA: Example

m1: The generation of random, binary, ordered trees

m2: The intersection graph of paths in trees

m3: Graph minors IV: Widths of trees and well-quasi-ordering

m4: Graph minors: A survey

LSA: example

Term \ Title	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

LSA: example

- We would expect that human is similar to user but not to minors in this context
- Correlation coefficient (covariance normalized to interval $[-1, 1]$)
- $r(human, user) = -0.37796$
- $r(human, minors) = -0.28571$

LSA: example

	1	2	3	4	5	6	7	8	9
human	-0.22	-0.11	0.29	-0.41	-0.11	-0.34	-0.52	0.06	0.41
interface	-0.20	-0.07	0.14	-0.55	0.28	0.50	0.07	0.01	0.11
computer	-0.24	0.04	-0.16	-0.60	-0.11	-0.26	0.30	-0.06	-0.49
user	-0.40	0.06	-0.34	0.10	0.33	0.38	-0.00	0.00	-0.01
system	-0.64	-0.17	0.36	0.33	-0.16	-0.21	0.17	-0.03	-0.27
response	-0.27	0.11	-0.43	0.07	0.08	-0.17	-0.28	0.02	0.05
time	-0.27	0.11	-0.43	0.07	0.08	-0.17	-0.28	0.02	0.05
EPS	-0.30	-0.14	0.33	0.19	0.11	0.27	-0.03	0.02	0.17
survey	-0.21	0.27	-0.18	-0.03	-0.54	0.08	0.47	0.04	0.58
trees	-0.01	0.49	0.23	0.02	0.59	-0.39	0.29	-0.25	0.23
graph	-0.04	0.62	0.22	0.00	-0.07	0.11	-0.16	0.68	-0.23
minors	-0.03	0.45	0.14	-0.01	-0.30	0.28	-0.34	-0.68	-0.18

Table : **U**

LSA: example

	1	2	3	4	5	6	7	8	9
1	3.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	2.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	2.35	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	1.64	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.31	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36

Table : $\mathbf{\Sigma}$

LSA: example

	1	2	3	4	5	6	7	8	9
c1	-0.20	-0.61	-0.46	-0.54	-0.28	-0.00	-0.01	-0.02	-0.08
c2	-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
c3	0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
c4	-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.02
c5	0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
m1	-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
m2	-0.18	0.43	0.24	-0.26	-0.67	0.34	0.15	-0.25	-0.04
m3	0.01	-0.05	-0.01	0.02	0.06	-0.45	0.76	-0.45	0.07
m4	0.06	-0.24	-0.02	0.08	0.26	0.62	-0.02	-0.52	0.45

Table : **V**

LSA: example

Term \ Title	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graphs	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

LSA: example

- $r(human, user) = 0.9385$
- $r(human, minors) = -0.8309$
- LSA brought together human and user through co-occurrences
- Also, the dissimilarity between human and minors is now stronger