GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 13: DIMENSIONALITY REDUCTION

YUCHEN ZHAO / DAT-14

LAST TIME:

- SVM'S
- HARD/SOFT MARGIN CLASSIFIERS
- KERNEL METHODS FOR NONLINEAR CLASSIFICATION

I. DIMENSIONALITY REDUCTION
II. PRINCIPAL COMPONENTS ANALYSIS
III. SINGULAR VALUE DECOMPOSITION

EXERCISE:
IV. DIMENSIONALITY REDUCTION IN SCIKIT-LEARN

I. DIMENSIONALITY REDUCTION

Q: What is dimensionality reduction?

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

In general, the idea is to regard the dataset is a matrix and to decompose the matrix into simpler, meaningful pieces.

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

In general, the idea is to regard the dataset is a matrix and to decompose the matrix into simpler, meaningful pieces.

Dimensionality reduction is frequently performed as a pre-processing step before another learning algorithm is applied.

Q: What are the motivations for dimensionality reduction?

Q: What are the motivations for dimensionality reduction?

The number of features in our dataset can be difficult to manage, or even misleading (eg, if the relationships are actually simpler than they appear).

For example, suppose we have a dataset with some features that are related to each other.

For example, suppose we have a dataset with some features that are related to each other.

Ideally, we would like to eliminate this redundancy and consolidate the number of variables we're looking at.

For example, suppose we have a dataset with some features that are related to each other.

Ideally, we would like to eliminate this redundancy and consolidate the number of variables we're looking at.

If these relationships are linear, then we can use well-established techniques like PCA/SVD.

EXAMPLE: 1D HARMONIC OSCILLATOR

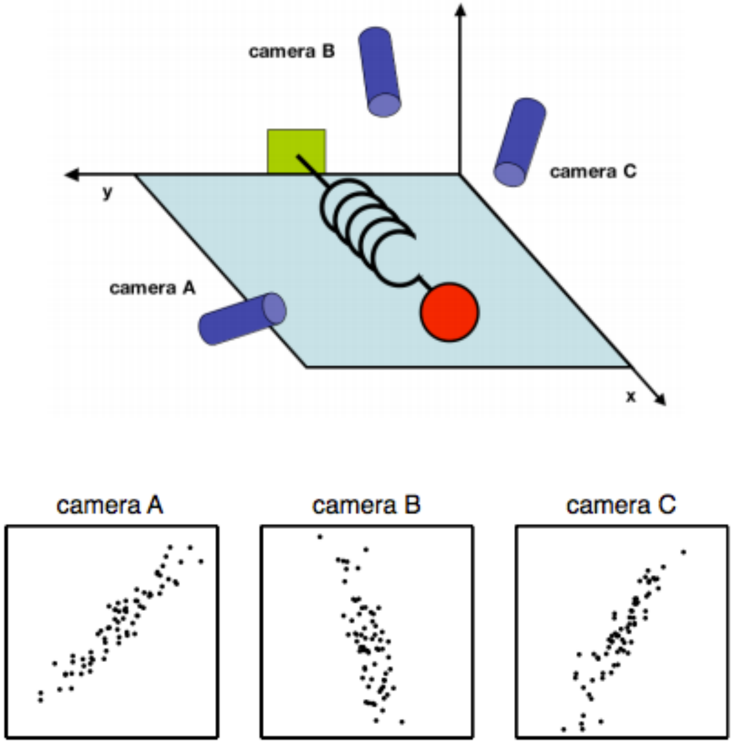


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

source: http://www.snl.salk.edu/~shlens/pca.pdf

ASIDE: CURSE OF DIMENSIONALITY

The complexity that comes with a large number of features is due in part to the **curse of dimensionality**.

ASIDE: CURSE OF DIMENSIONALITY

The complexity that comes with a large number of features is due in part to the **curse of dimensionality**.

Namely, the sample size needed to accurately estimate a random variable taking values in a d*-dimensional feature space grows exponentially with* d *(almost).*

ASIDE: CURSE OF DIMENSIONALITY

Another way of characterizing this is to say that high-dimensional spaces are inherently **sparse***.*

ASIDE: CURSE OF DIMENSIONALITY

In high-dimensional spaces, most of the points are "far" from each other.

ASIDE: CURSE OF DIMENSIONALITY

In high-dimensional spaces, most of the points are "far" from each other.

This illustrates the fact that local methods will break down in these circumstances (eg, in order to collect enough neighbors for a given point, you need to expand the radius of the neighborhood so far that locality is not preserved).

In high-dimensional spaces, most of the points are "far" from each other.

This illustrates the fact that local methods will break down in these circumstances (eg, in order to collect enough neighbors for a given point, you need to expand the radius of the neighborhood so far that locality is not preserved).

The bottom line is that high-dimensional spaces can be problematic.

Q: What is the goal of dimensionality reduction?

Q: What is the goal of dimensionality reduction?

We'd like to analyze the data using the most meaningful basis (or **coordinates***) possible.*

Q: What is the goal of dimensionality reduction?

We'd like to analyze the data using the most meaningful basis (or **coordinates***) possible.*

More precisely: given an n *x* d *matrix* A *(encoding* n *observations of a* d*-dimensional random variable), we want to find a* k*-dimensional representation of* A *(*$k < d$*) that captures the information in the original data, according to some criterion.*

Q: What is the goal of dimensionality reduction?

- *reduce computational expense*
- *reduce susceptibility to overfitting*
- *reduce noise in the dataset*
- *enhance our intuition*

Q: How is dimensionality reduction performed?

Q: How is dimensionality reduction performed?

A: There are two approaches: feature selection and feature extraction.

Q: How is dimensionality reduction performed?
A: There are two approaches: feature selection and feature extraction.

feature selection – *selecting a subset of features using an external criterion (filter)*

feature extraction – *mapping the features to a lower dimensional space*

Q: How is dimensionality reduction performed?

A: There are two approaches: feature selection and featur

NOTE

We've already seen one example of feature selection for regression: backward elimination.

feature selection – *selecting a subset of features using an external criterion (filter)*

feature extraction – *mapping the features to a lower dimensional space*

Feature selection is important, but typically when people say dimensionality reduction, they are referring to feature extraction.

Feature selection is important, but typically when people say dimensionality reduction, they are referring to feature extraction.

The goal of feature extraction is to create a new set of coordinates that simplify the representation of the data.

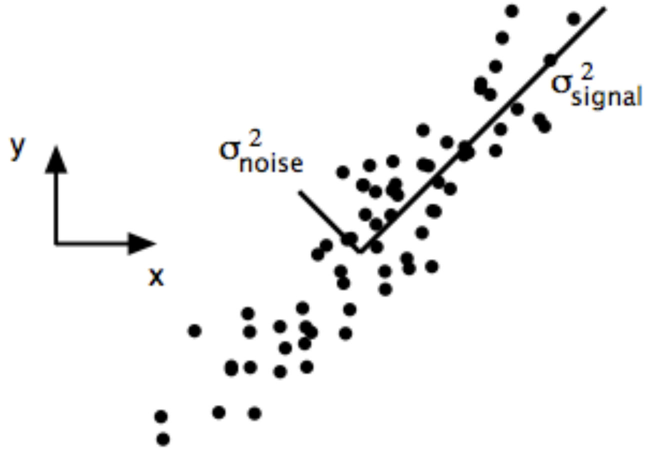


FIG. 2 Simulated data of (x, y) for camera A. The signal and noise variances σ^2_{signal} and σ^2_{noise} are graphically represented by the two lines subtending the cloud of data. Note that the largest direction of variance does not lie along the basis of the recording (x_A, y_A) but rather along the best-fit line.

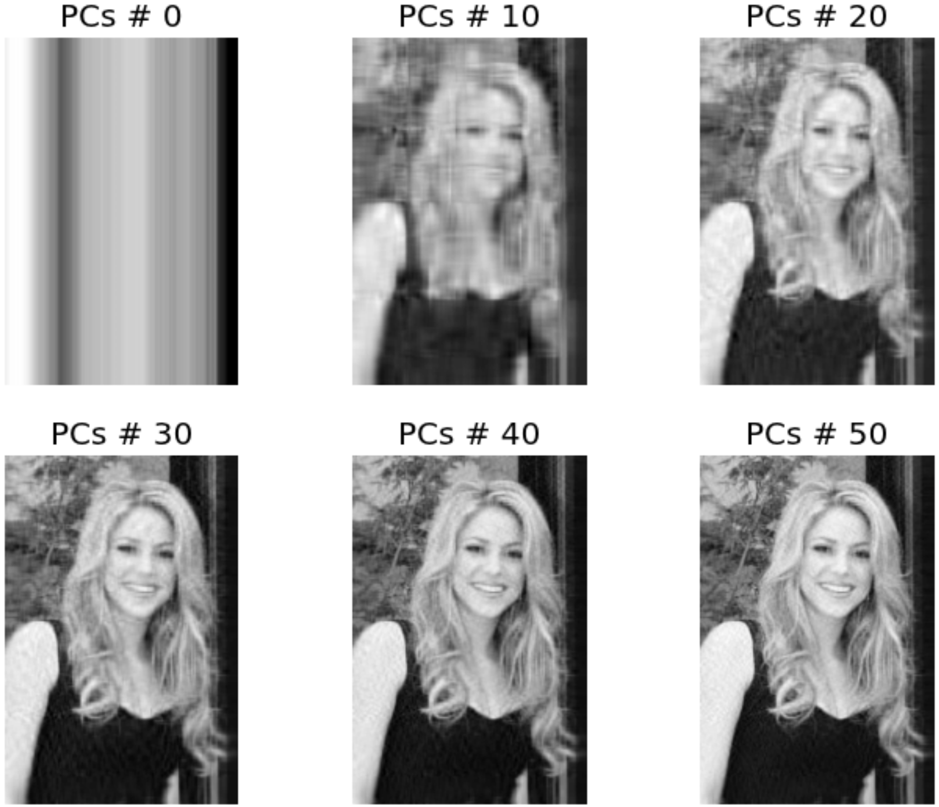
source: http://www.snl.salk.edu/~shlens/pca.pdf

Q: What are some applications of dimensionality reduction?

Q: What are some applications of dimensionality reduction?

- *topic models (document clustering)*
- *image recognition/computer vision*
- *bioinformatics (microarray analysis)*
- *speech recognition*
- *astronomy (spectral data analysis)*
- *recommender systems*

DIMENSIONALITY REDUCTION



source: http://glowingpython.blogspot.it/2011/07/pca-and-image-compression-with-numpy.html

II. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

This procedure produces a new basis, each of whose components retain as much variance from the original data as possible.

Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

This procedure produces a new basis, each of whose components retain as much variance from the original data as possible.

The PCA of a matrix $\mathbf{A}$ *boils down to the* **eigenvalue decomposition** *of the* **covariance matrix** *of* $\mathbf{A}$.

what is variance?

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Variance is the average distance from the mean of a data set to a point in that data set.

In other words, it is a measure of the *spread* of the data. Recall that standard deviation is the square root of variance.

what is covariance?

covariance is a measure of how much two random variables change together

Variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \qquad var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

Covariance:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

covariance is a measure of how much two random variables change together

Variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \qquad var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

Covariance:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

The covariance matrix C *of a matrix* A *is always square:*

$$\mathrm{C} = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

- *off-diagonal elements* $\mathrm{C_{ij}}$ *give the covariance between* $\mathrm{X_i}, \mathrm{X_j}$ $(\mathrm{i} \neq \mathrm{j})$
- *diagonal elements* $\mathrm{C_{ii}}$ *give the variance of* $\mathrm{X_i}$

ASIDE: EIGENVALUE DECOMPOSITION

The eigenvalue decomposition of a square matrix A *is given by:*

$$\mathrm{A} = \mathrm{Q}\Lambda\mathrm{Q}^{-1}$$

The eigenvalue decomposition of a square matrix A *is given by:*

$$\mathrm{A} = \mathrm{Q}\Lambda\mathrm{Q}^{-1}$$

The columns of Q *are the* **eigenvectors** *of* A*, and the values in* Λ *are the associated* **eigenvalues** *of* A*.*

$$\begin{aligned} A &= \begin{bmatrix} -1/2 & 3/2 \\ 3/2 & -1/2 \end{bmatrix} \\ &= \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right)\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right)^T \end{aligned}$$

The eigenvalue decomposition of a square matrix A *is given by:*

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$$

The columns of Q *are the* **eigenvectors** *of* A*, and the values in* Λ *are the associated* **eigenvalues** *of* A*.*

For an eigenvector v *of* A *and its eigenvalue* λ*, we have the important relation:*

$$\mathrm{Av} = \lambda \mathrm{v}$$

The eigenvalue decomposition of a square matrix A *is given by:*

$$\mathrm{A} = \mathrm{Q}\Lambda\mathrm{Q}^{-1}$$

NOTE

This relationship defines what it means to be an eigenvector of A.

The columns of Q *are the* **eigenvectors** *of* A*, and the valu*

the associated **eigenvalues** *of* A.

For an eigenvector v *of* A *and its eigenvalue* λ*, we have the important relation:*

$$\mathrm{Av} = \lambda\mathrm{v}$$

*The eigenvectors form a **basis** of the vector space on which $\mathbb{A}$ acts (eg, they are orthogonal).*

The eigenvectors form a basis of the vector space on which $\mathbb{A}$ *acts (eg, they are orthogonal).*

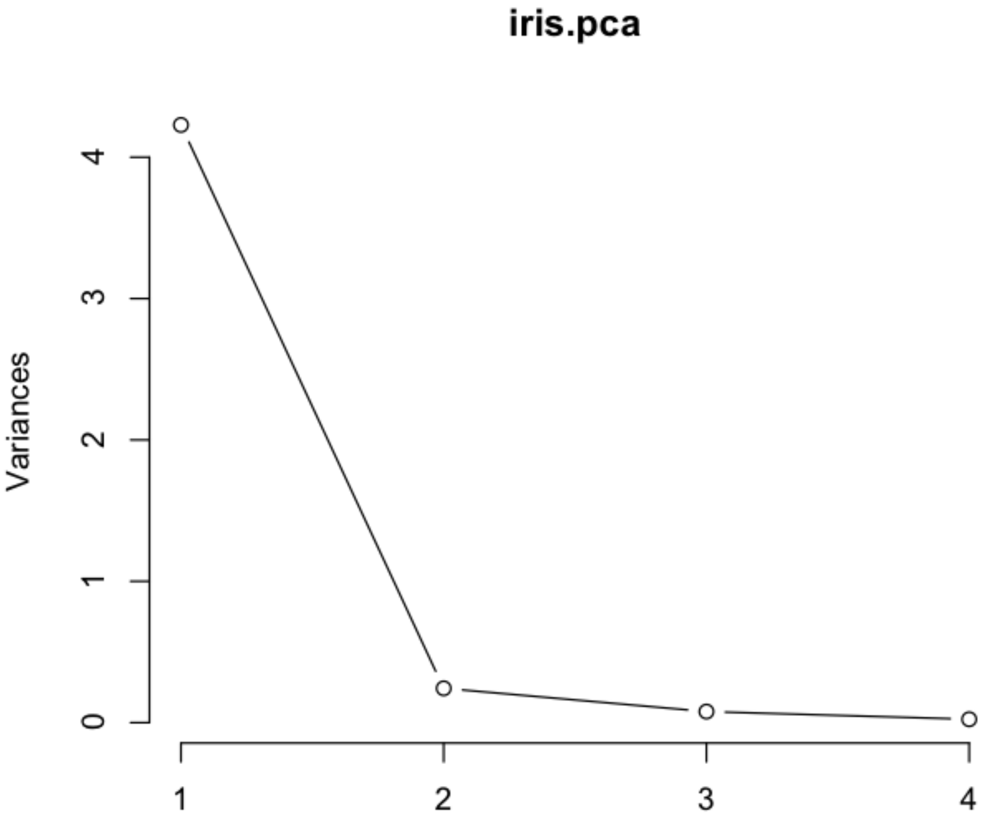
Furthermore the basis elements are ordered by their eigenvalues (from largest to smallest), and these eigenvalues represent the amount of variance explained by each basis element.

The eigenvectors form a basis of the vector space on which $\mathbf{A}$ *acts (eg, they are orthogonal).*

Furthermore the basis elements are ordered by their eigenvalues (from largest to smallest), and these eigenvalues represent the amount of variance explained by each basis element.

This can be visualized in a **scree plot***, which shows the amount of variance explained by each basis vector.*

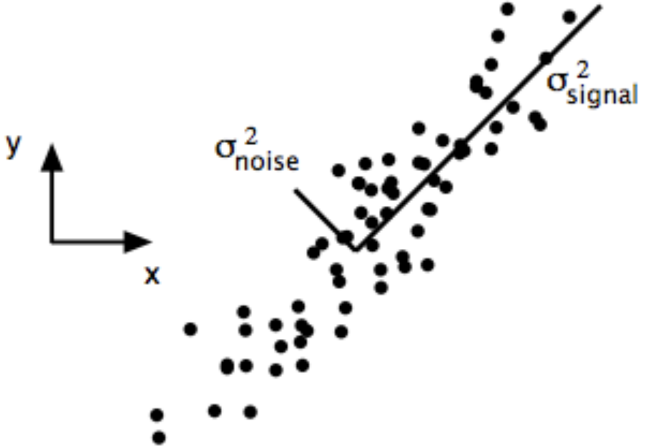
PRINCIPAL COMPONENT ANALYSIS



1. Linearity *– The change in basis is a linear projection*

2. Large variances have important structure

we assume that principal components with larger associated variances are signal, while those with lower variances represent noise.



*3. The principal components are **orthogonal***

III. SINGULAR VALUE DECOMPOSITION

Consider a matrix M *with* m *rows and* n *features.*

Consider a matrix M *with* m *rows and* n *features.*

The **singular value decomposition** *of* A *is given by:*

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

Consider a matrix M *with* m *rows and* n *features.*

The **singular value decomposition** *of* A *is given by:*

$$\underset{(m \times n)}{\mathbf{M}} = \underset{(m \times r)}{\mathbf{U}} \; \underset{(r \times r)}{\Sigma} \; \underset{(r \times n)}{\mathbf{V}^{\mathbf{T}}}$$

Consider a matrix M *with* m *rows and* n *features.*

The **singular value decomposition** *of* A *is given by:*

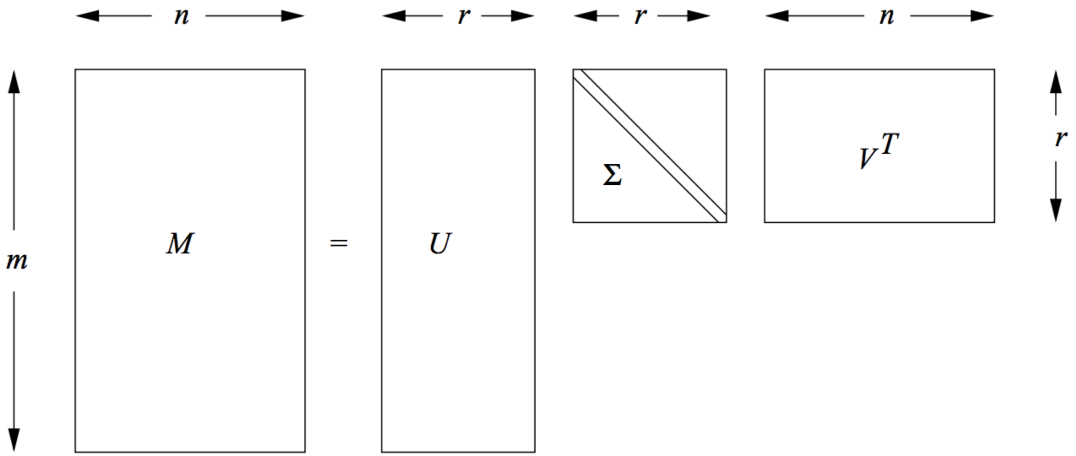
$$\underset{(m \times n)}{\mathbf{M}} = \underset{(m \times r)}{\mathbf{U}} \; \underset{(r \times r)}{\Sigma} \; \underset{(r \times n)}{\mathbf{V}^{\mathbf{T}}}$$

st. U, V *are* **orthogonal** *matrices and* Σ *is a* **diagonal** *matrix.*

The **singular value decomposition** *of* M *is given by:*

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

(m x n) (m x r) (r x r) (r x n)



source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION – EXAMPLE

Ratings of movies by users:

	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

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION – EXAMPLE

Ratings of movies by users:

	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

there are two "concepts" underlying the movies:

science-fiction and romance

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION - EXAMPLE

Ratings of movies by users:

	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

All the boys rate only science-fiction

All the girls rate only romance

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION - EXAMPLE

	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

Ratings of movies by users:

$$\underset{M}{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}} = \underset{U}{\begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}} \underset{V^{\mathrm{T}}}{\begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}}$$

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION - EXAMPLE

	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

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}}_{M} = \underbrace{\begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}}_{V^{\mathrm{T}}}$$

M: people -> movies

U: people -> concepts

V: concepts -> movies

Σ: the strength of each of the concepts

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION – A MORE REALISTIC EXAMPLE

$$
\begin{array}{r|ccccc|}
 & \text{Matrix} & \text{Alien} & \text{Star Wars} & \text{Casablanca} & \text{Titanic} \\
\hline
\text{Joe} & 1 & 1 & 1 & 0 & 0 \\
\text{Jim} & 3 & 3 & 3 & 0 & 0 \\
\text{John} & 4 & 4 & 4 & 0 & 0 \\
\text{Jack} & 5 & 5 & 5 & 0 & 0 \\
\text{Jill} & 0 & 0 & 0 & 4 & 4 \\
\text{Jenny} & 0 & 0 & 0 & 5 & 5 \\
\text{Jane} & 0 & 0 & 0 & 2 & 2 \\
\hline
\end{array}
$$

$$
\underbrace{\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}}_{M'} =
$$

$$
\underbrace{\begin{bmatrix}
.13 & .02 & -.01 \\
.41 & .07 & -.03 \\
.55 & .09 & -.04 \\
.68 & .11 & -.05 \\
.15 & -.59 & .65 \\
.07 & -.73 & -.67 \\
.07 & -.29 & .32
\end{bmatrix}}_{U}
\underbrace{\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}}_{\Sigma}
\underbrace{\begin{bmatrix}
.56 & .59 & .56 & .09 & .09 \\
.12 & -.02 & .12 & -.69 & -.69 \\
.40 & -.80 & .40 & .09 & .09
\end{bmatrix}}_{V^{\mathrm{T}}}
$$

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION - EXAMPLE

How to reduce dimensions?

<u>Drop Low Singular Values</u> -> eliminate corresponding rows of U and V

$$\underset{M'}{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}} =$$

$$\underset{U}{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}} \underset{V^{\mathrm{T}}}{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}$$

$$\underset{\Sigma}{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}}$$

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

SINGULAR VALUE DECOMPOSITION – EXAMPLE

How to reduce dimensions?

Drop Low Singular Values

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

$$M'$$

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

$$U \qquad \Sigma \qquad V^{\mathrm{T}}$$

$$\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .07 & -.73 \\ .07 & -.29 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \end{bmatrix}$$

$$= \begin{bmatrix} 0.93 & 0.95 & 0.93 & .014 & .014 \\ 2.93 & 2.99 & 2.93 & .000 & .000 \\ 3.92 & 4.01 & 3.92 & .026 & .026 \\ 4.84 & 4.96 & 4.84 & .040 & .040 \\ 0.37 & 1.21 & 0.37 & 4.04 & 4.04 \\ 0.35 & 0.65 & 0.35 & 4.87 & 4.87 \\ 0.16 & 0.57 & 0.16 & 1.98 & 1.98 \end{bmatrix}$$

source: http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

IV. LAB