Dimensionality reduction with SVD

Dimensionality reduction

- Dataset **X** consisting of **n** points in a **d**-dimensional space
- Data point $\mathbf{x_i} \in \mathbf{R^d}$ (**d**-dimensional real vector):

  $\mathbf{x_i = [x_{i1}, x_{i2}, \ldots, x_{id}]}$

Datasets in the form of matrices

We are given **n** objects and **d** features describing the objects. (Each object has **d** numeric values describing it.)

<u>**Dataset**</u>
An **n-by-d** matrix **A**, A_{ij} shows the "**importance**" of feature **j** for object **i**.
Every row of **A** represents an object.

<u>Goal</u>
1. **Understand** the structure of the data, e.g., the underlying process generating the data.
2. **Reduce the number of features** representing the data

Market basket matrices

d products
(e.g., milk, bread, wine, etc.)

n customers

$$\left(\quad A \quad \right)$$

A_{ij} = quantity of **j**-th product purchased by the **i**-th customer

Find a subset of the products that characterize customer behavior

Social-network matrices

d groups
(e.g., BU group, opera, etc.)

n users

$$A$$

A_{ij} = partiticipation of the **i**-th user in the **j**-th group

Find a subset of the groups that accurately clusters social-network users

Document matrices

$$
\begin{array}{cc}
 & \textbf{d}\text{ terms} \\
 & \text{(e.g., theorem, proof, etc.)} \\
\begin{array}{l}\textbf{n} \\ \text{documents}\end{array} & \begin{pmatrix} \\ \\ A \\ \\ \\ \end{pmatrix}
\end{array}
$$

A_{ij} = frequency of the **j**-th term in the **i**-th document

Find a subset of the terms that accurately clusters the documents

Recommendation systems

$$\begin{array}{cc} & \textbf{d} \text{ products} \\ \textbf{n} \text{ customers} & \begin{pmatrix} \\ & A & \\ \\ \end{pmatrix} \end{array}$$

A_{ij} = frequency of the j-th product is bought by the i-th customer

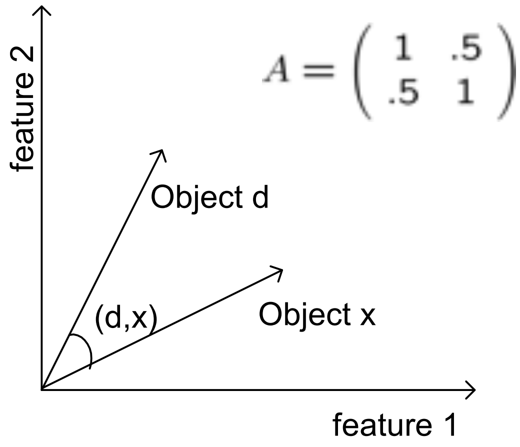
Find a subset of the products that accurately describe the behavior or the customers

The Singular Value Decomposition (SVD)

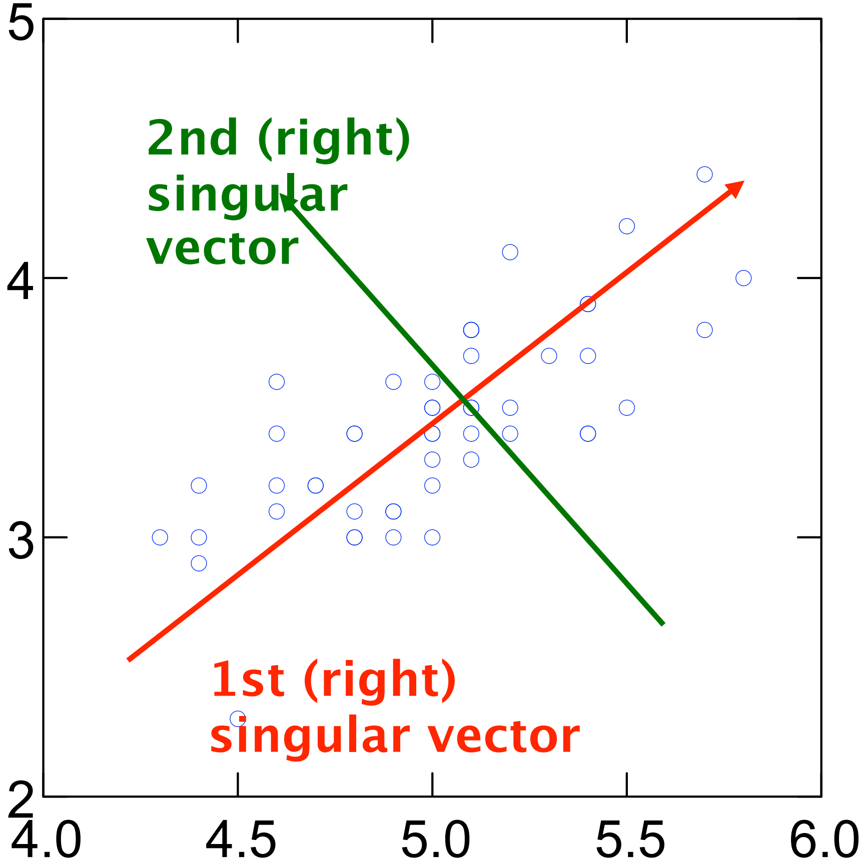
Data matrices have **n** rows (one for each object) and **d** columns (one for each feature).

Rows: vectors in a Euclidean space,

Two objects are "**close**" if the angle between their corresponding vectors is small.



SVD: Example



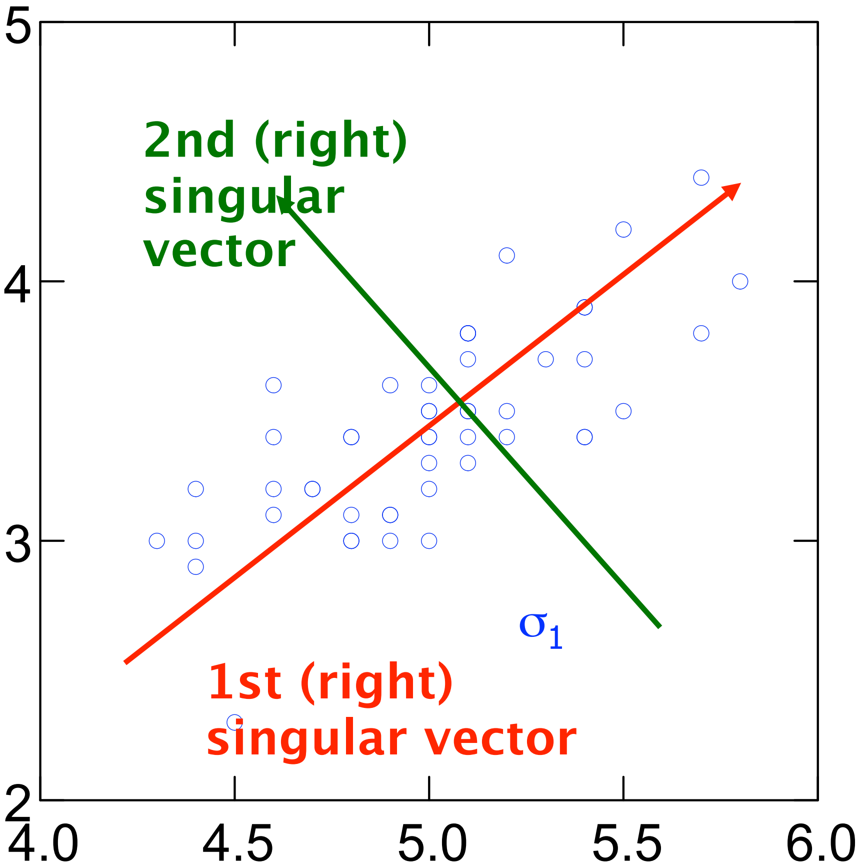
Input: **2-d** dimensional points

Output:

1st (right) singular vector: direction of maximal variance,

2nd (right) singular vector: direction of maximal variance, after removing the projection of the data along the first singular vector.

Singular values



σ_1: measures how much of the data variance is explained by the first singular vector.

σ_2: measures how much of the data variance is explained by the second singular vector.

SVD decomposition

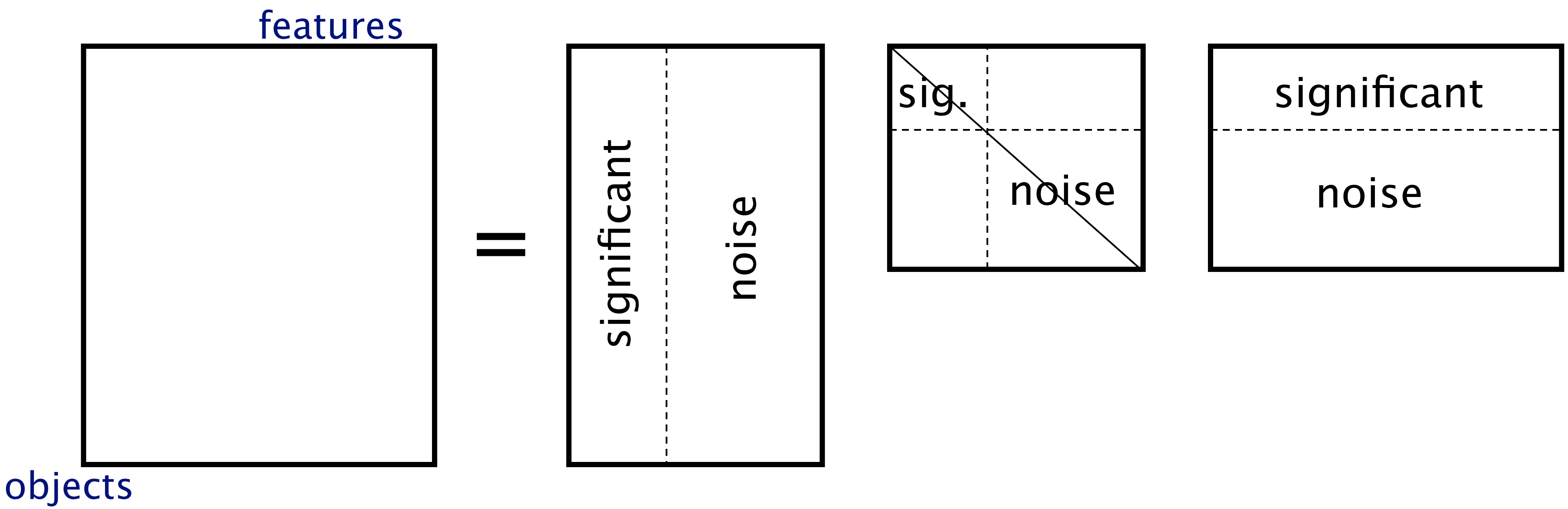
$$\underset{\mathbf{n \times d}}{\begin{pmatrix} \quad A \quad \end{pmatrix}} = \underset{\mathbf{n \times \ell}}{\begin{pmatrix} \quad U \quad \end{pmatrix}} \cdot \underset{\mathbf{\ell \times \ell}}{\begin{pmatrix} \Sigma & \\ 0 & \end{pmatrix}} \cdot \underset{\mathbf{\ell \times d}}{\begin{pmatrix} \quad V \quad \end{pmatrix}^T}$$

U (V): orthogonal matrix containing the left (right) singular vectors of **A**.
Σ: diagonal matrix containing the **singular values** of **A**:
($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_\ell$)

Exact computation of the SVD takes **O(min{mn^2 , m^{2}n})** time.
The top k left/right singular vectors/values can be **computed faster** using Lanczos/Arnoldi methods.

SVD and Rank-k approximations

$$A = U \Sigma V^T$$



Rank-k approximations (A_k)

$$A_k = U_k \cdot \Sigma_k \cdot V_k^T$$

n x d   n x k   k x d

A_k is the **best** approximation of A

U_k (V_k): orth
(right) singu
Σ_k: diagonal m
values of A

A_k is an approximation of A

PCA and SVD

- PCA is SVD done on **centered** data
- PCA looks for such a direction that the data projected to it has the maximal variance
- PCA/SVD continues by seeking the next direction that is orthogonal to all previously found directions
- All directions are orthogonal

How to compute the PCA

- Data matrix A, **rows = data points**, **columns = variables** (attributes, features, parameters)

1. Center the data by subtracting the mean of each column
2. Compute the SVD of the centered matrix A' (i.e., find the first k singular values/vectors)
   $A' = U\Sigma V^T$
3. The principal components are the columns of V, the coordinates of the data in the basis defined by the principal components are $U\Sigma$

Singular values tell us something about the variance

- The variance in the direction of the **k**-th principal component is given by the corresponding singular value σ_k^2

- Singular values can be used to estimate how many components to keep

- **Rule of thumb:** keep enough to explain **85%** of the variation:

$$\frac{\sum_{j=1}^{k} \sigma_j^2}{\sum_{j=1}^{n} \sigma_j^2} \approx 0.85$$