Overview of Principal Component Analysis

And Low-Rank Models

Probability and Statistics for Data Science

Carlos Fernandez-Granda

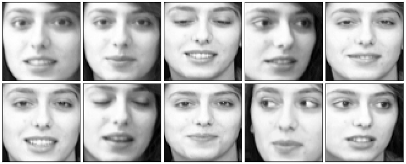
NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Motivation

Model data with multiple features



Motivation

Model data associated with two entities

$$\begin{array}{cc} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \begin{pmatrix} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{pmatrix} & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Plan

- Covariance matrix
- Principal component analysis
- Dimensionality reduction
- Low-rank models
- Matrix completion

Goal

Describe data with multiple features

Model: d-dimensional random vector

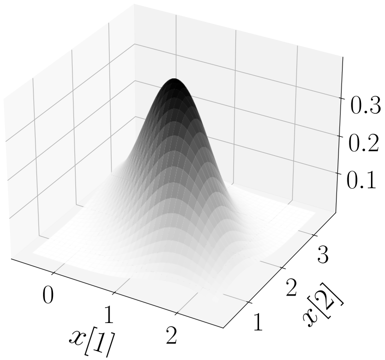
$$\tilde{x} := \begin{bmatrix} \tilde{x}[1] \\ \tilde{x}[2] \\ \cdots \\ \tilde{x}[d] \end{bmatrix}$$

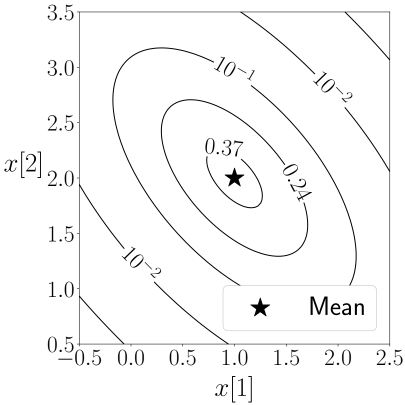
Mean of a random vector

The d-dimensional mean of a random vector $\tilde{x}$ is

$$\mathrm{E}\left[\tilde{x}\right] := \begin{bmatrix} \mathrm{E}\left[\tilde{x}[1]\right] \\ \mathrm{E}\left[\tilde{x}[2]\right] \\ \cdots \\ \mathrm{E}\left[\tilde{x}[d]\right] \end{bmatrix}$$

Random vector



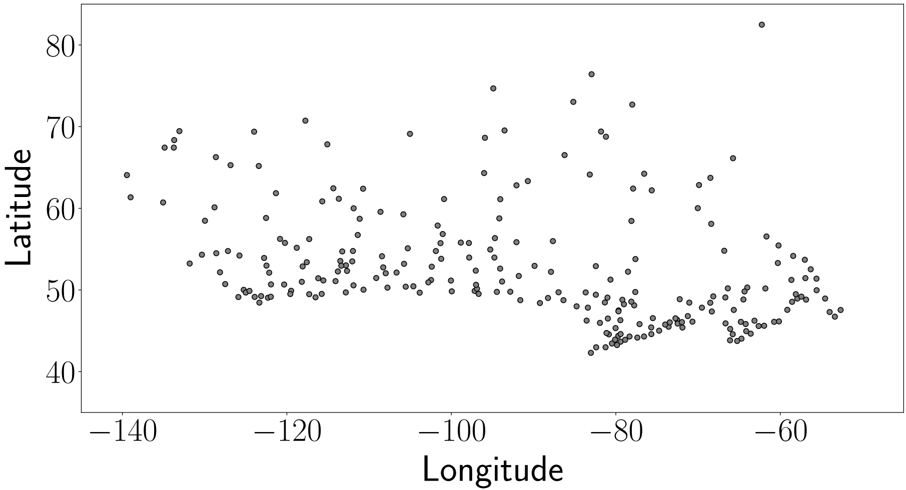


Sample mean

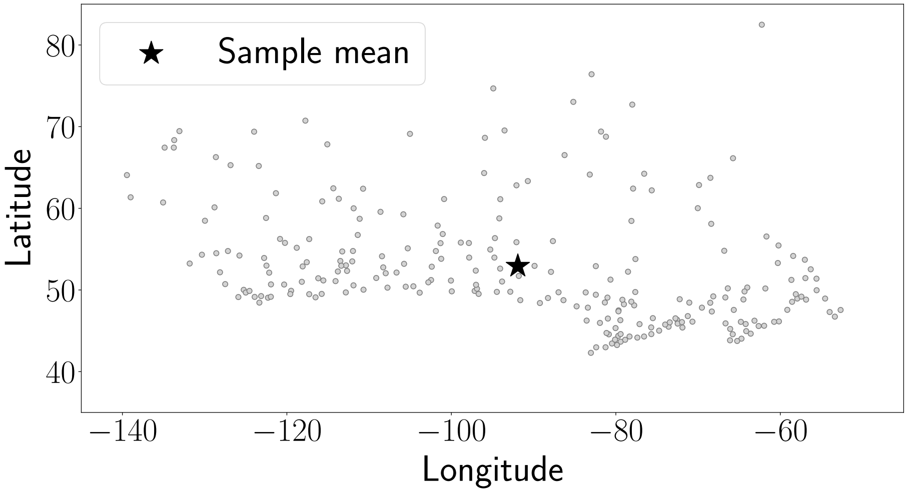
Dataset with d features: $X := \{x_1, x_2, \ldots, x_n\}$

$$m(X) := \frac{1}{n} \sum_{i=1}^{n} x_i$$

Canadian cities



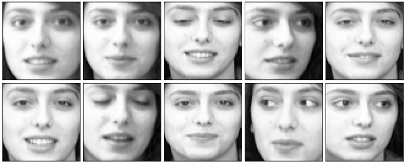
Canadian cities



Faces

64×64 images from 40 subjects

Vectorized images interpreted as vectors in $\mathbb{R}^{4096}$



Sample mean

Variance

The variance characterizes average variation of a random variable

How can we characterize fluctuations of a random vector?

Variance of linear combinations of the entries

Covariance matrix

The covariance matrix of a random vector $\tilde{x}$ is

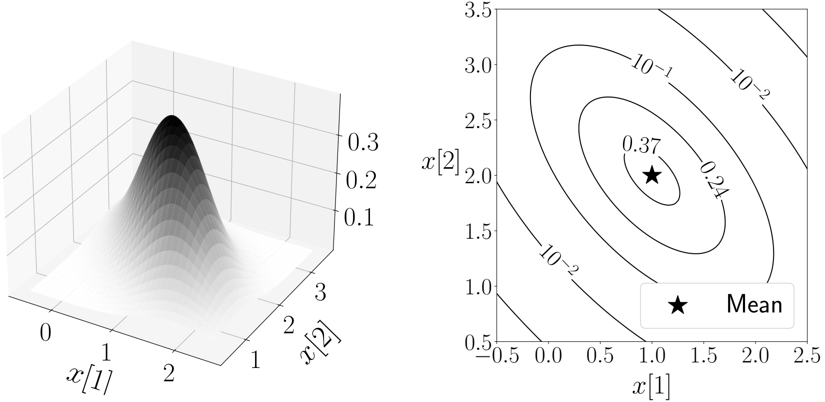
$$\Sigma_{\tilde{x}} := \begin{bmatrix} \mathrm{Var}\left[\tilde{x}[1]\right] & \mathrm{Cov}\left[\tilde{x}[1], \tilde{x}[2]\right] & \cdots & \mathrm{Cov}\left[\tilde{x}[1], \tilde{x}[d]\right] \\ \mathrm{Cov}\left[\tilde{x}[1], \tilde{x}[2]\right] & \mathrm{Var}\left[\tilde{x}[2]\right] & \cdots & \mathrm{Cov}\left[\tilde{x}[2], \tilde{x}[d]\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left[\tilde{x}[1], \tilde{x}[d]\right] & \mathrm{Cov}\left[\tilde{x}[2], \tilde{x}[d]\right] & \cdots & \mathrm{Var}\left[\tilde{x}[d]\right] \end{bmatrix}$$

Variance of linear combination $a^T \tilde{x}$

For any deterministic vector a

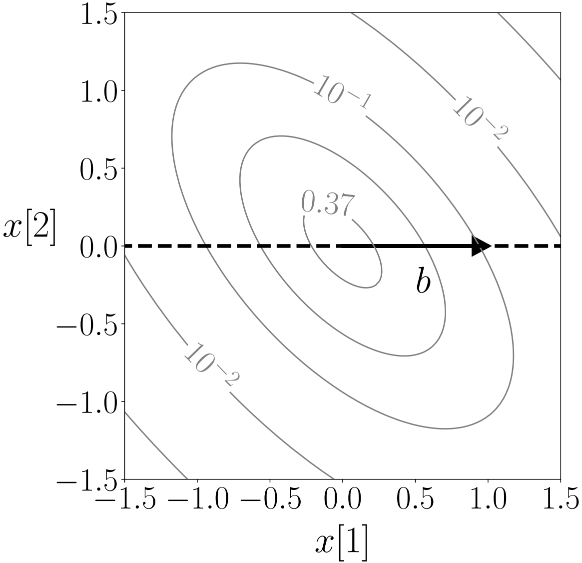
$$\mathrm{Var}\left[a^T \tilde{x}\right] = a^T \Sigma_{\tilde{x}} a$$

Gaussian random vector

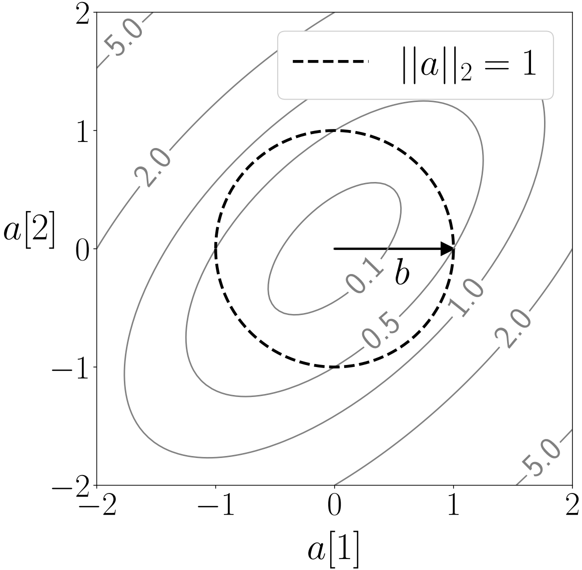


$$\Sigma_{\tilde{x}} := \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}$$

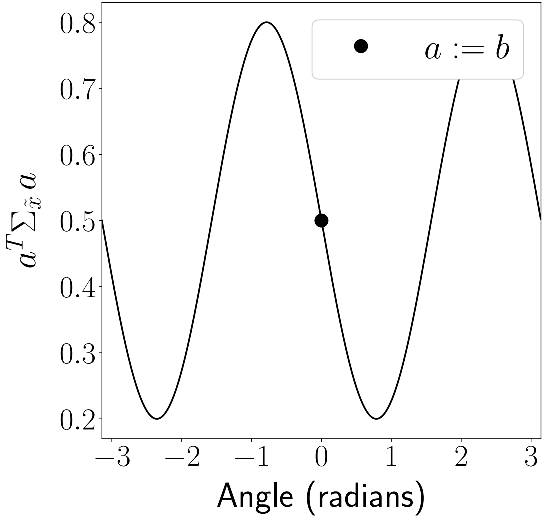
Variance in a certain direction?



Directional variance $\mathrm{Var}[a^T \tilde{x}] = a^T \Sigma_{\tilde{x}} a$



$a^T \Sigma_{\tilde{x}} a$ on the unit circle



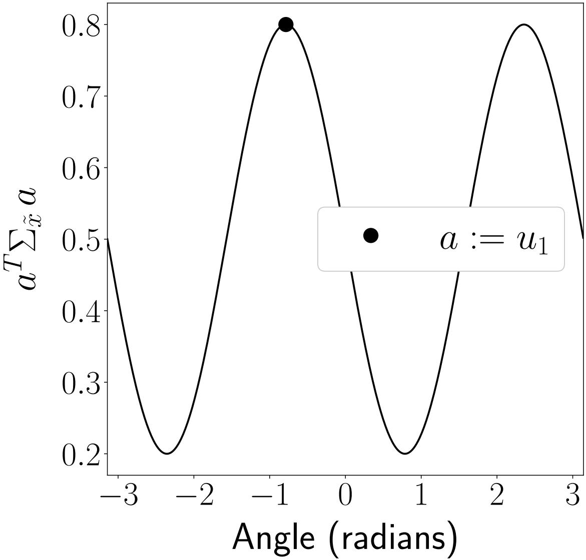
Maximum is direction of maximum variance

Principal directions

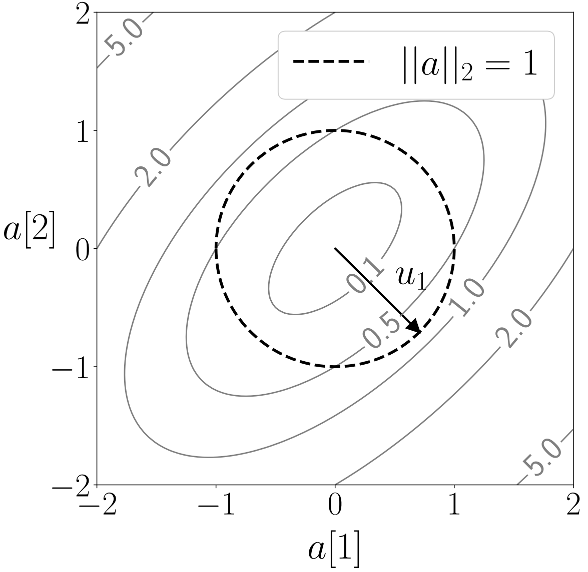
The eigenvectors of the covariance matrix are directions of maximum variance

The components in those directions are the principal components

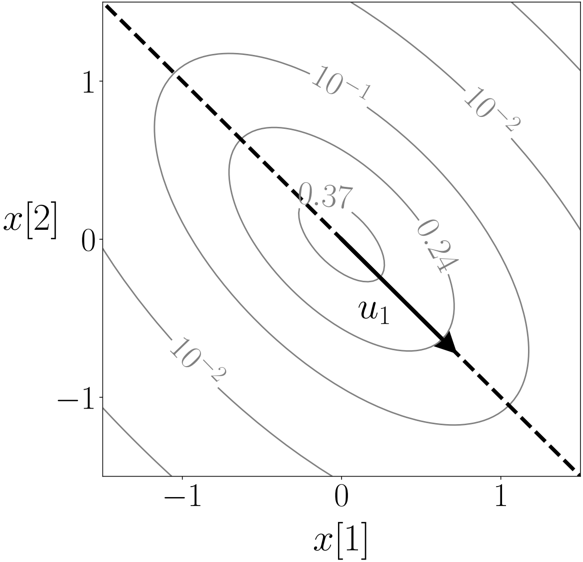
First principal direction



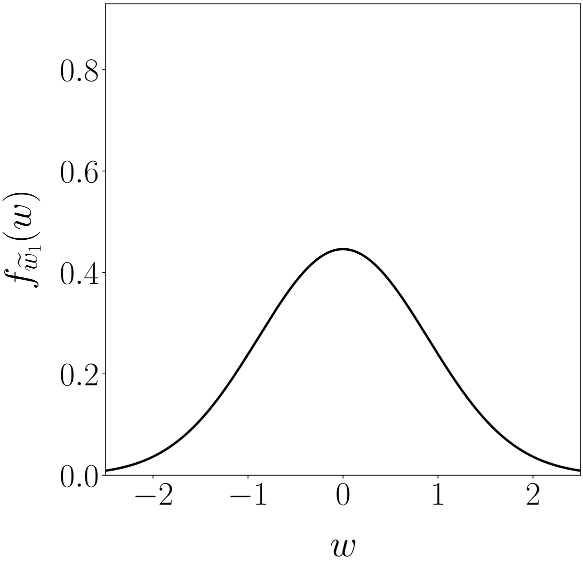
First principal direction



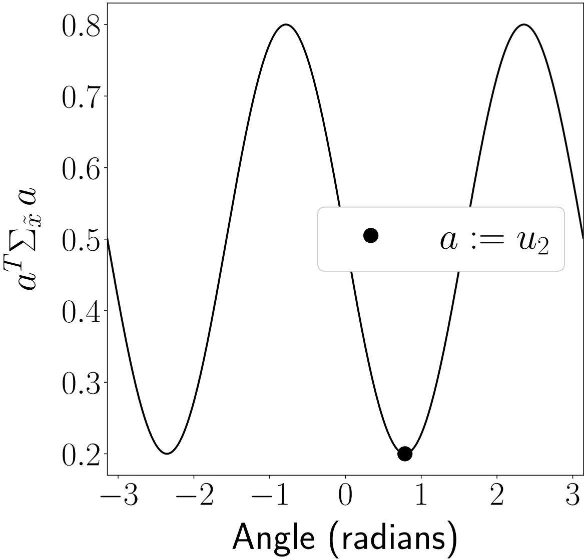
Joint pdf of $\tilde{x}$



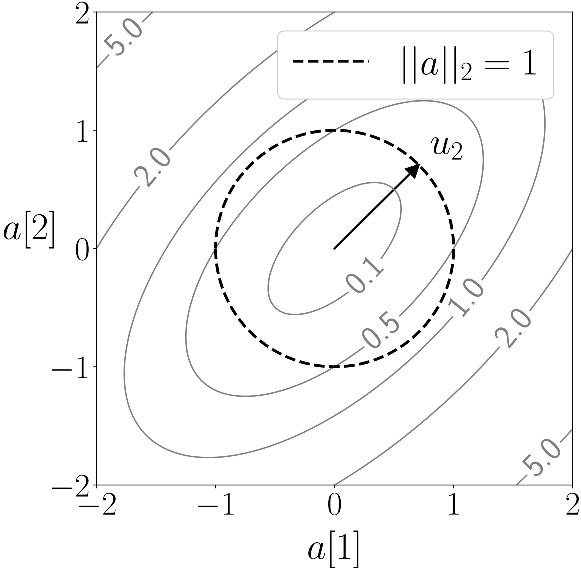
First principal component



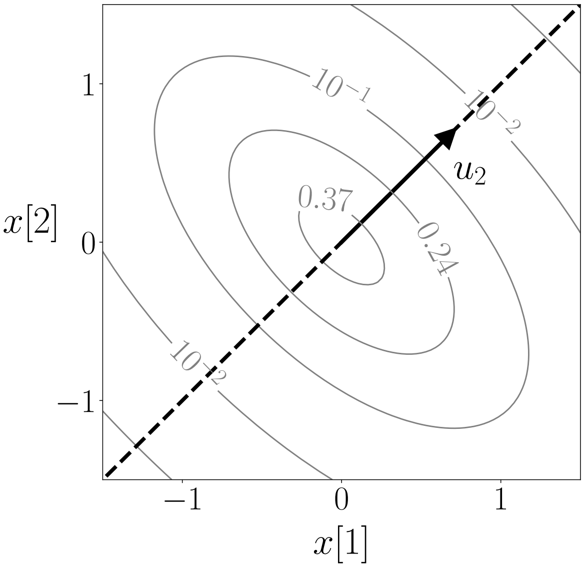
Second principal direction



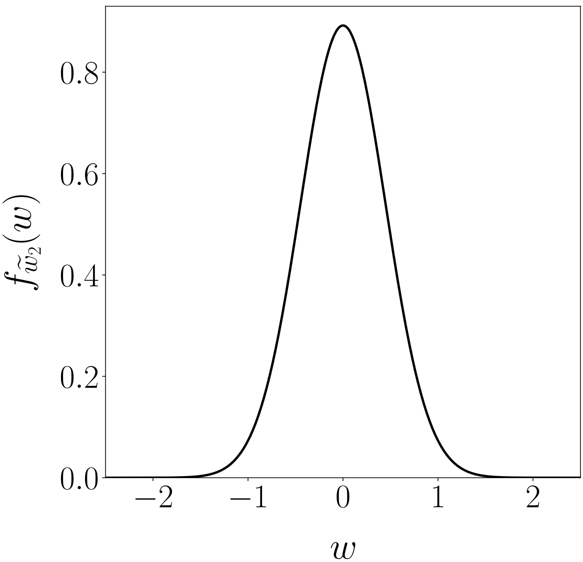
Second principal direction



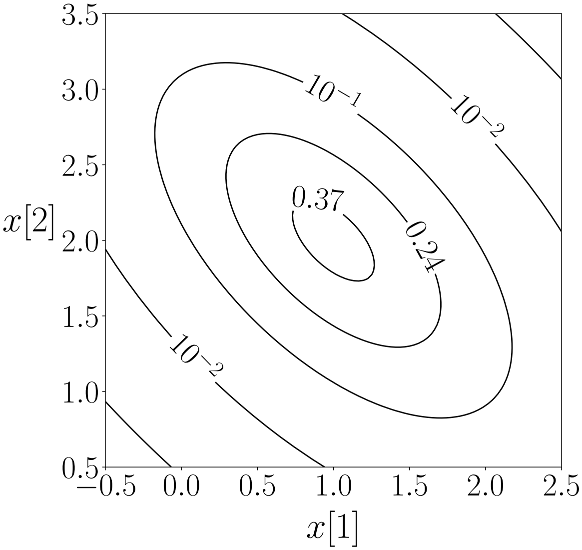
Joint pdf of $\tilde{x}$



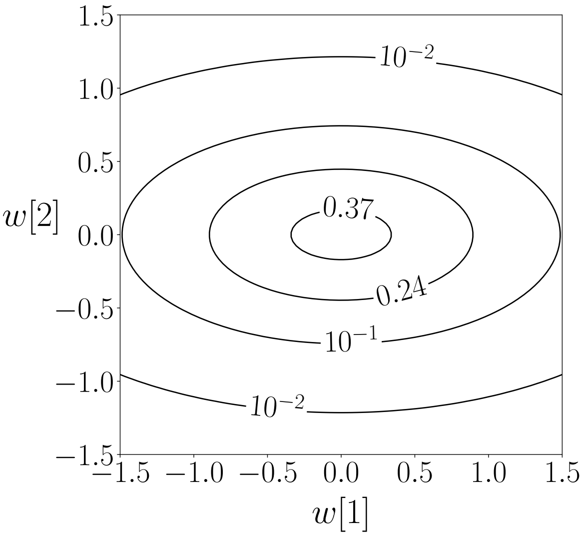
Second principal component



Original joint pdf



Joint pdf of principal components



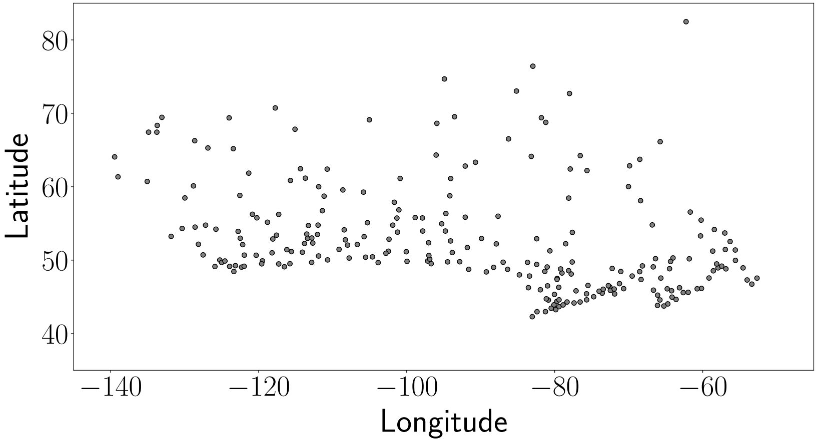
Covariance matrix of a dataset

Data with d features: $X := \{x_1, x_2, \ldots, x_n\}$

Sample covariance matrix of X:

$$\Sigma_X := \begin{bmatrix} v(X[1]) & c(X[1], X[2]) & \cdots & c(X[1], X[d]) \\ c(X[1], X[2]) & v(X[2]) & \cdots & c(X[2], X[d]) \\ \vdots & \vdots & \ddots & \vdots \\ c(X[1], X[d]) & c(X[2], X[d]) & \cdots & v(X[d]) \end{bmatrix}$$

Cities in Canada



Sample covariance matrix:

$$\Sigma_X = \begin{bmatrix} 524.9 & -59.8 \\ -59.8 & 53.7 \end{bmatrix}$$

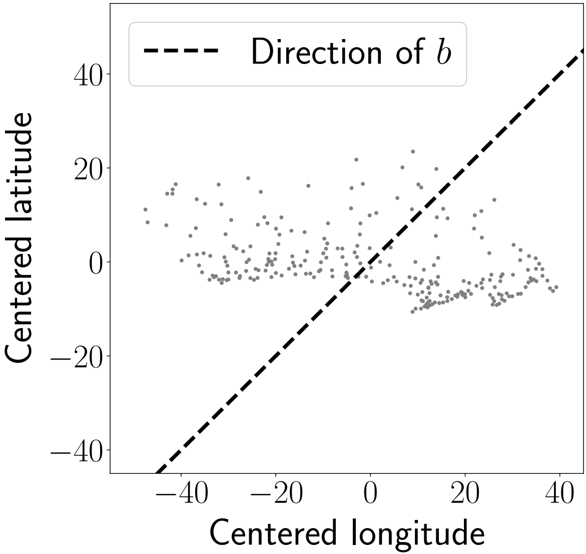
Sample variance of linear combination

Dataset: $X = \{x_1, \ldots, x_n\}$

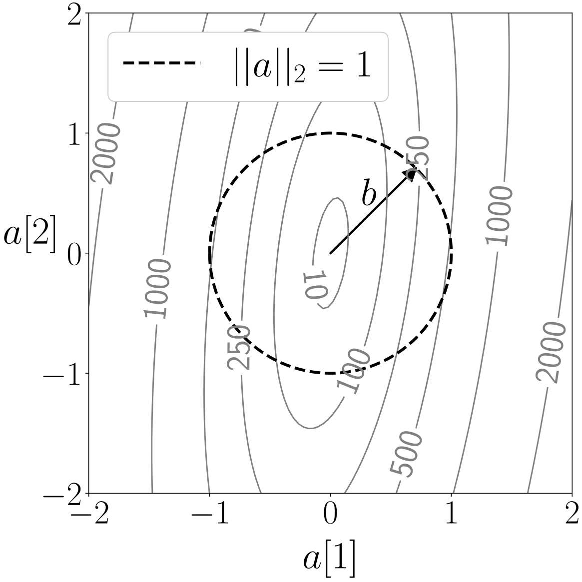
$$X_a := \left\{a^T x_1, \ldots, a^T x_n\right\}$$

$$v(X_a) = a^T \Sigma_X a$$

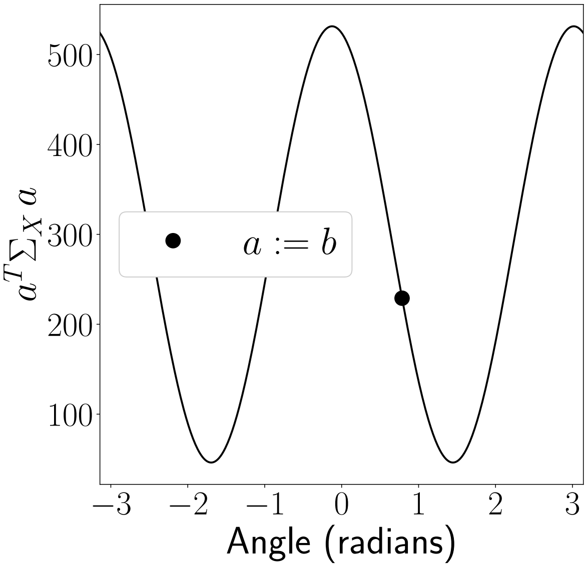
Sample variance in a certain direction?



Sample directional variance $a^T \Sigma_X a = v(X_a)$



$a^T \Sigma_X a$ on the unit circle

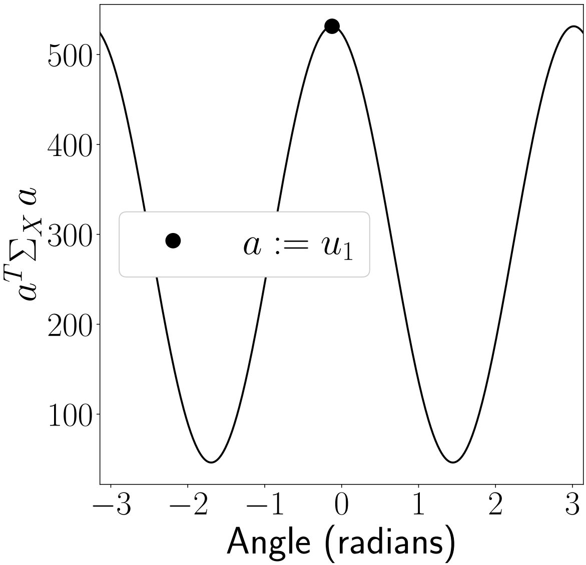


Principal directions

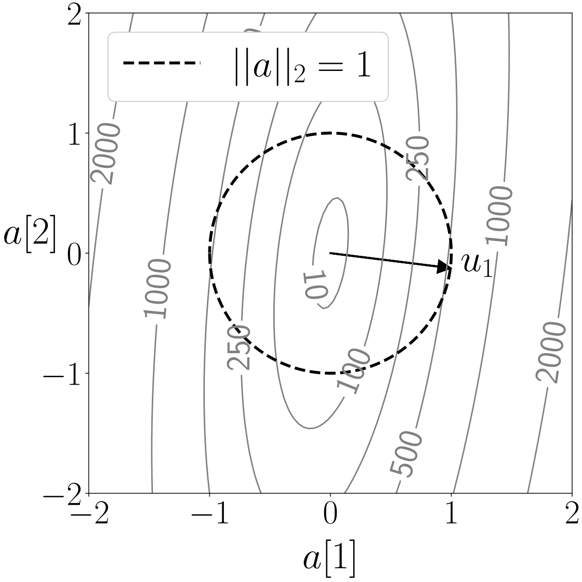
The eigenvectors of the sample covariance matrix are directions of maximum variance

The components in those directions are the principal components

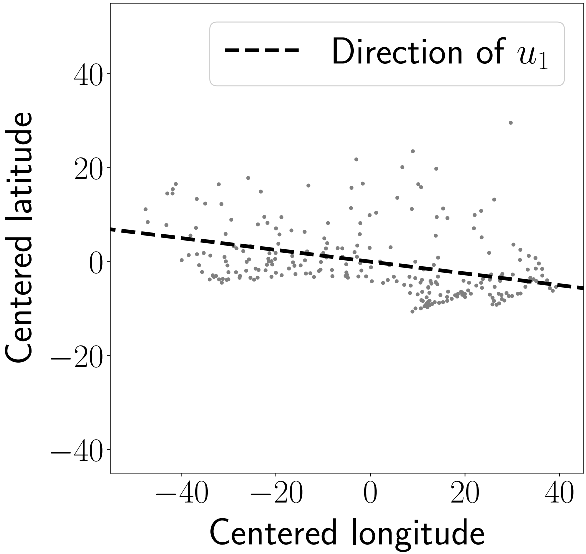
First principal direction



First principal direction

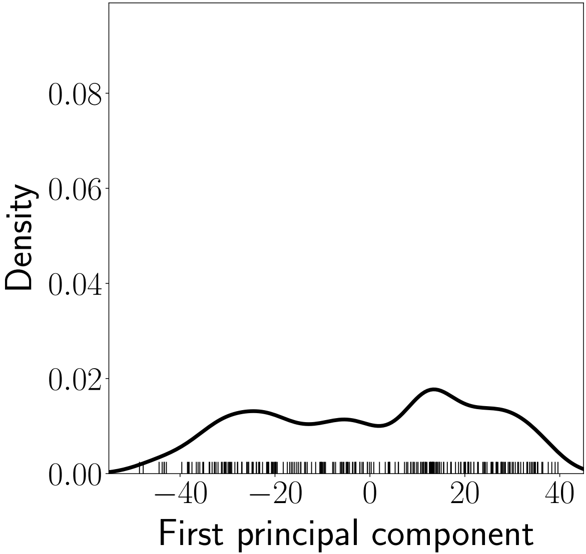


Data

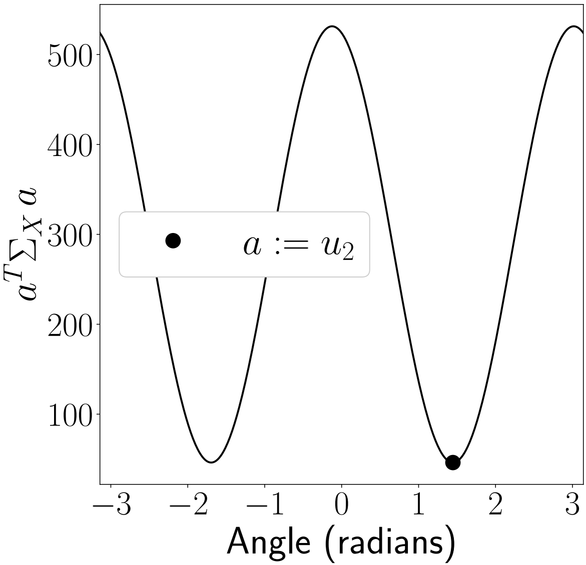


Direction of u_1
40
20
0
−20
−40
Centered latitude
−40
−20
0
20
40
Centered longitude

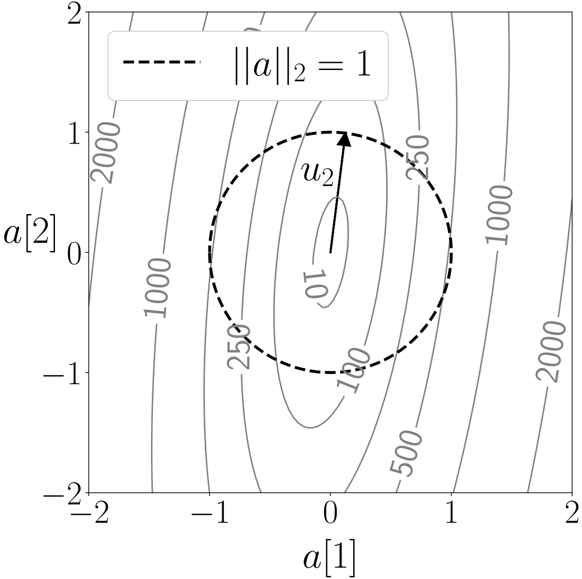
First principal component



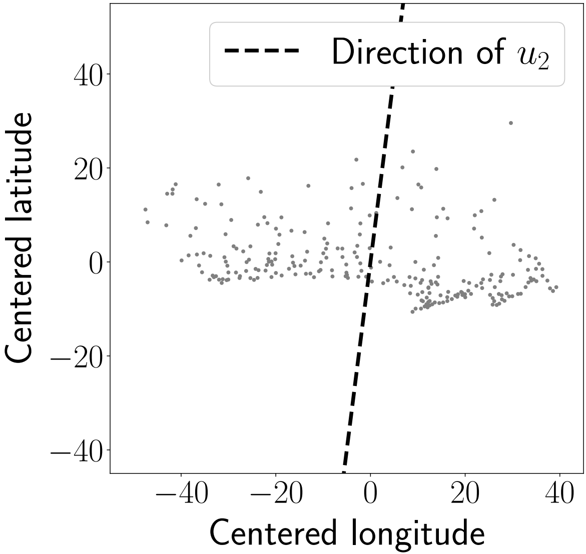
Second principal direction



Second principal direction

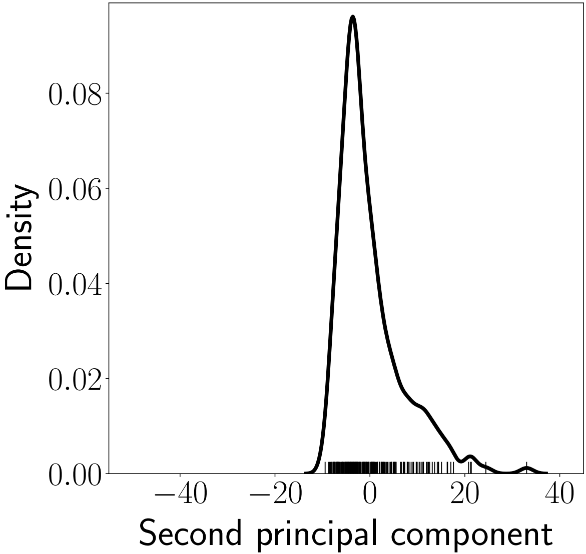


Data

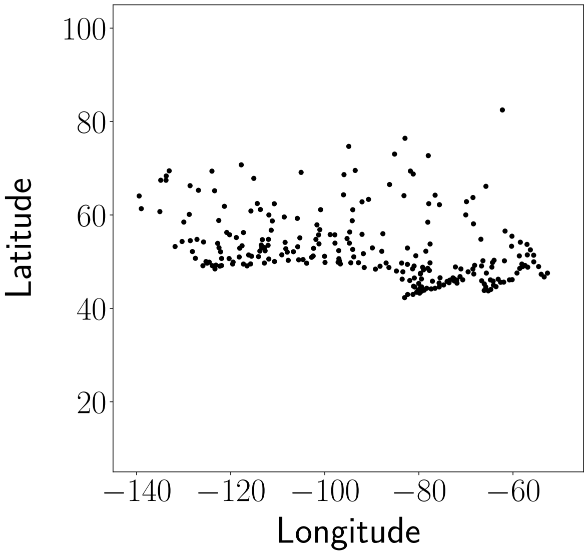


Direction of u_2
40
20
0
−20
−40
Centered latitude
−40
−20
0
20
40
Centered longitude

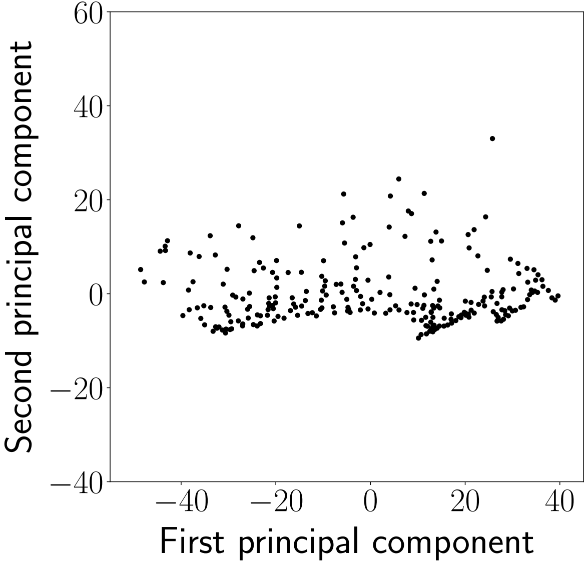
Second principal component



Data



Principal components



Dimensionality reduction

Data with a large number of features can be difficult to analyze/process

Solution: Reduce dimensionality while preserving as much information as possible

Important preprocessing step in many applications

The first k principal directions span the subspace that captures the most variance in the data

Wheat seeds

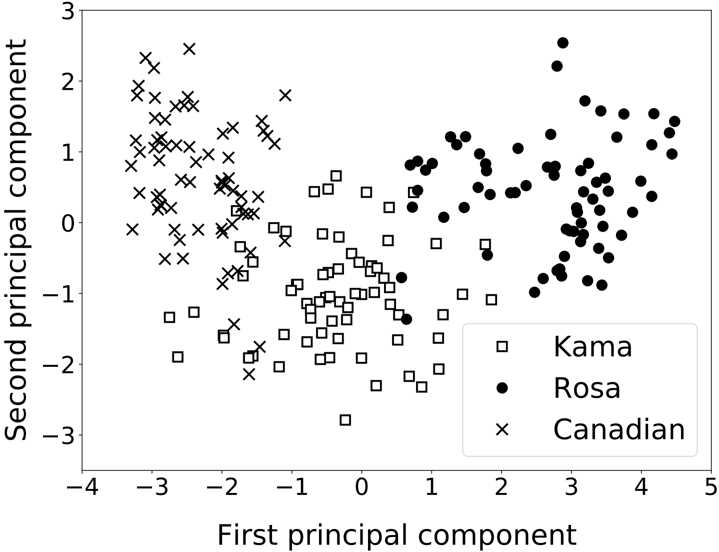
3 varieties: Kama, Rosa and Canadian

Features:

- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

Challenge: How to visualize the data in two dimensions?

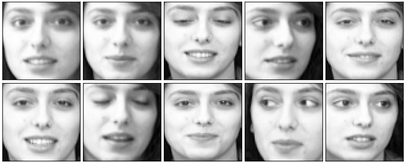
Two first principal components



Faces

64×64 images from 40 subjects

Vectorized images interpreted as vectors in $\mathbb{R}^{4096}$



Sample mean

Principal directions

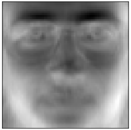
u_1

18.8

u_2

11.1

u_3



6.30

u_4

3.95

u_5

2.86

Principal directions

u_{10}	u_{20}	u_{30}	u_{40}	u_{50}

1.32	0.591	0.349	0.217	0.162

Principal directions

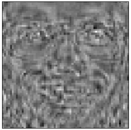
u_{100}

0.061

u_{200}

0.019

u_{250}

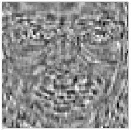


0.011

u_{300}

0.008

u_{350}



0.004

$k = 5$

$\underset{u_1,\dots,u_5}{\mathrm{approx}}(x_i)$

$=$

$m(X)$

w_1

- 1.89

u_1

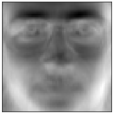
w_2

+ 0.92

u_2

- 1.08

w_3



u_3

- 1.51

w_4

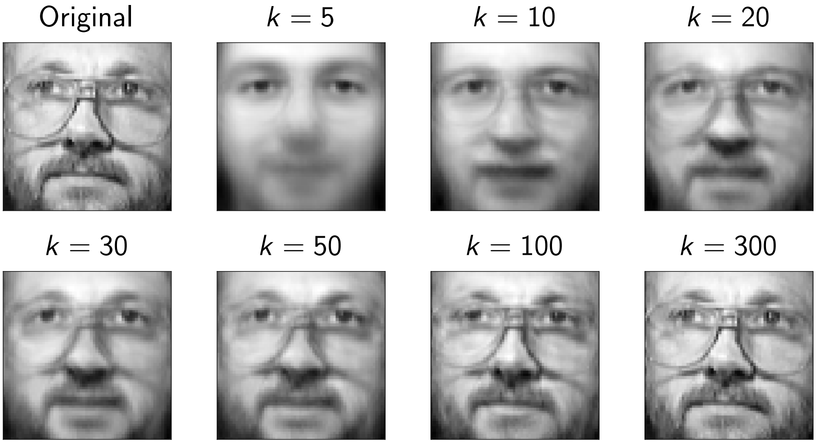
u_4

- 0.73

w_5

u_5

Approximation



Matrix-valued data

$$D := \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array}$$

$$D := \left(\begin{array}{cccc} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{array} \right) \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array}$$

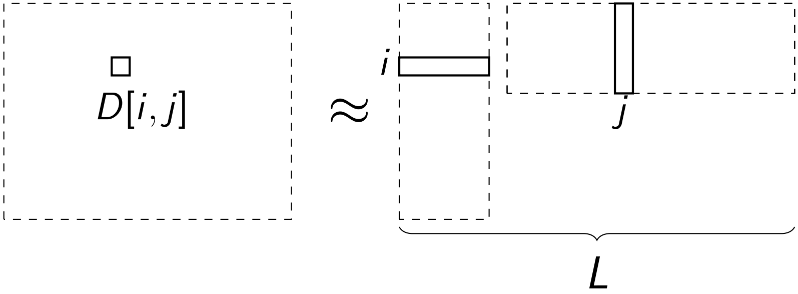
Rank-1 model a[movie]b[user]

Ratings $\approx$ Mean rating +

$$\begin{array}{r}\text{Dark Knight}\\ \text{Spiderman 3}\\ \text{Love Actually}\\ \text{BJ's Diary}\\ \text{Pretty Woman}\\ \text{Superman 2}\end{array}\begin{pmatrix}-0.45\\ -0.39\\ 0.39\\ 0.38\\ 0.38\\ -0.45\end{pmatrix}\quad\begin{array}{cccc}\text{Bob} & \text{Molly} & \text{Mary} & \text{Larry}\\ (3.74 & 4.05 & -3.74 & -4.05)\end{array}$$

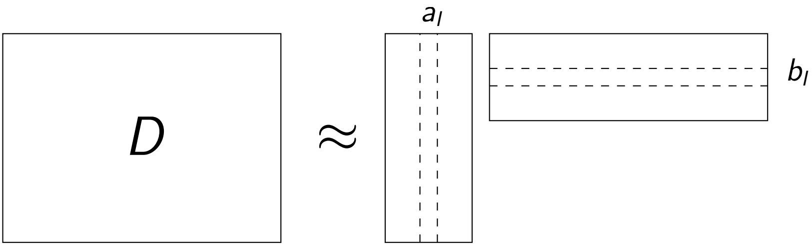
Low-rank model

$$D[i,j] \approx L[i,j] := \sum_{l=1}^{r} a_l[i] b_l[j]$$



Dimensionality reduction of rows and columns

Rank of L is r



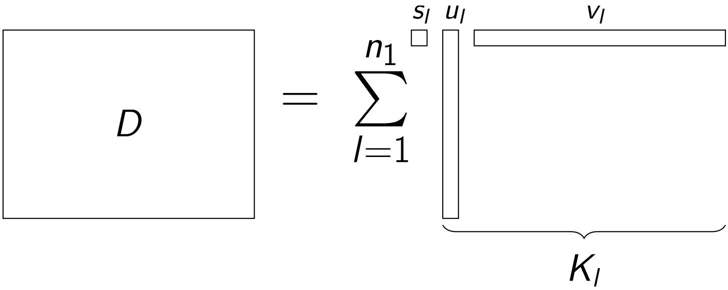
Singular value decomposition

All matrices have an SVD ($n_1 \leq n_2$)

$$D = \underbrace{\begin{bmatrix} u_1 & u_2 & \cdots & u_{n_1} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & s_{n_1} \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} v_1 & v_2 & \cdots & v_{n_1} \end{bmatrix}^T}_{V^T}$$

- Singular values $s_1 \geq s_2 \geq \cdots \geq s_r \geq 0$
- Left singular vectors $u_1, u_2, \ldots u_{n_1} \in \mathbb{R}^{n_1}$ are orthonormal
- Right singular vectors $v_1, v_2, \ldots v_{n_1} \in \mathbb{R}^{n_2}$ are orthonormal

SVD as a superposition of rank-1 components



$K_1, \ldots, K_{n_1}$ are rank 1, orthogonal, unit norm

Norm of $D = \sqrt{\sum_{l=1}^{l} s_l^2}$

Truncated SVD

$$\boxed{L_{\mathrm{SVD}}} := \sum_{l=1}^{r} s_l \square \boxed{K_l}$$

Equivalent to principal component analysis of columns / rows

$$L_{\mathrm{SVD}} = \arg \min_{\mathrm{rank}(L)=r} ||D - L||_{\mathrm{F}}$$

Movie ratings

$$D := \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \\ \end{array}$$

$$D := \begin{array}{cc} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\ \left(\begin{array}{cccc} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{array}\right) & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \end{array}$$

Rank-1 model a[movie]b[user]

Ratings $\approx$ Mean rating +

$$
\begin{array}{r}
\text{Dark Knight} \\
\text{Spiderman 3} \\
\text{Love Actually} \\
\text{BJ's Diary} \\
\text{Pretty Woman} \\
\text{Superman 2}
\end{array}
\begin{pmatrix}
-0.45 \\
-0.39 \\
0.39 \\
0.38 \\
0.38 \\
-0.45
\end{pmatrix}
\begin{array}{c}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\
\begin{pmatrix} 3.74 & 4.05 & -3.74 & -4.05 \end{pmatrix}
\end{array}
$$

Rank-1 model

$$\begin{array}{cl}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\
\begin{pmatrix}
1.34\,(1) & 1.19\,(1) & 4.66\,(5) & 4.81\,(4) \\
1.55\,(2) & 1.42\,(1) & 4.45\,(4) & 4.58\,(5) \\
4.45\,(4) & 4.58\,(5) & 1.55\,(2) & 1.42\,(1) \\
4.43\,(5) & 4.56\,(4) & 1.57\,(2) & 1.44\,(1) \\
4.43\,(4) & 4.56\,(5) & 1.57\,(1) & 1.44\,(2) \\
1.34\,(1) & 1.19\,(2) & 4.66\,(5) & 4.81\,(5)
\end{pmatrix} &
\begin{array}{l}
\text{The Dark Knight} \\
\text{Spiderman 3} \\
\text{Love Actually} \\
\text{B. Jones's Diary} \\
\text{Pretty Woman} \\
\text{Superman 2}
\end{array}
\end{array}$$

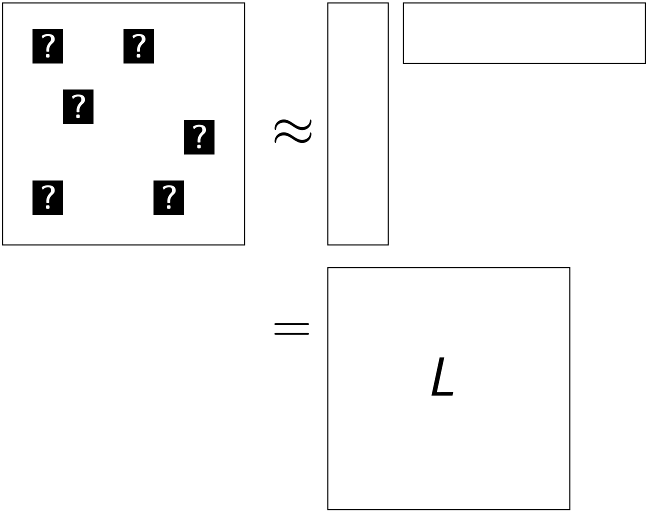
What if some entries are missing?

$$D := \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\ \begin{pmatrix} ? & ? & 5 & 4 \\ ? & 1 & 4 & ? \\ 4 & 5 & 2 & ? \\ ? & 4 & 2 & 1 \\ 4 & ? & 1 & 2 \\ 1 & 2 & ? & 5 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array}$$

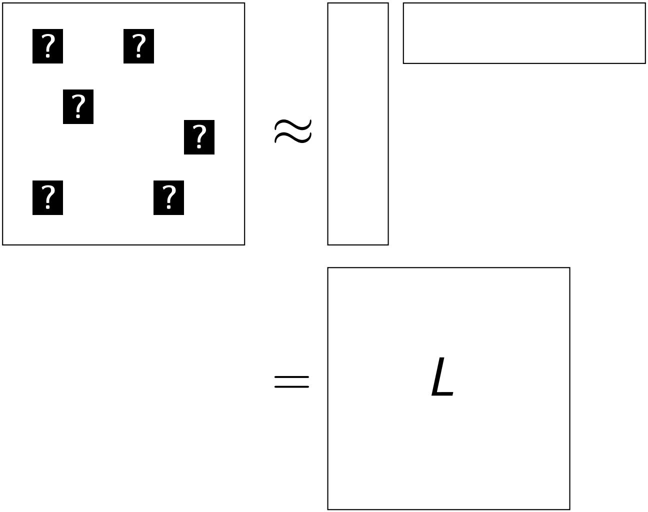
Matrix completion problem

We can insert any values!

Assumption: Matrix is low rank



Low-rank matrix completion



In practice

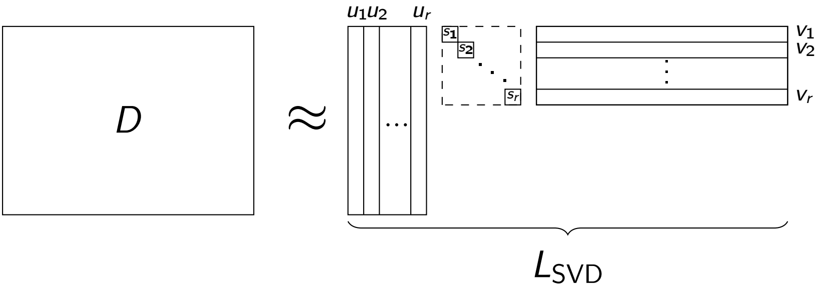
Data cannot be expected to be exactly low rank

Goal: Find low-rank matrix that is closest to the data

$$\sum_{(i,j) \in \text{observed}} \left(D[i,j] - \sum_{l=1}^{r} a_l[i] b_l[j] \right)^2$$

Problem: Nonconvex cost function that is difficult to optimize

Truncated SVD



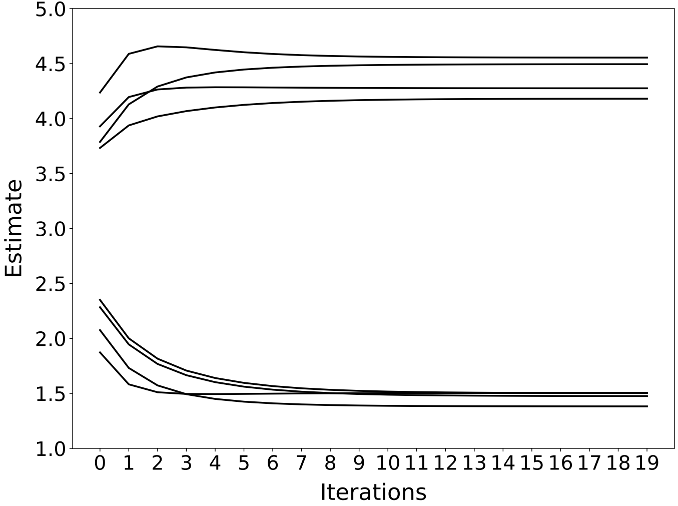
Optimal if no entries are missing

Movie ratings

$$
\begin{array}{cc}
 & \begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} \\
D := & \left(\begin{array}{cccc} ? & ? & 5 & 4 \\ ? & 1 & 4 & ? \\ 4 & 5 & 2 & ? \\ ? & 4 & 2 & 1 \\ 4 & ? & 1 & 2 \\ 1 & 2 & ? & 5 \end{array} \right) \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array}
\end{array}
$$

Idea: Alternate between imputing missing entries and fitting low-rank model

Missing entries (mean observed rating = 2.94)



Final estimate

$$\begin{array}{cl}
\begin{array}{cccc} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{array} & \\
\begin{pmatrix}
\color{blue}{1.48}\ (1) & \color{blue}{1.38}\ (1) & 4.45\ (5) & 4.52\ (4) \\
\color{blue}{1.50}\ (2) & 1.41\ (1) & 4.42\ (4) & \color{blue}{4.50}\ (5) \\
4.26\ (4) & 4.34\ (5) & 1.57\ (2) & \color{blue}{1.51}\ (1) \\
\color{blue}{4.18}\ (5) & \color{blue}{4.26}\ (4) & 1.65\ (2) & 1.59\ (1) \\
4.2\ (4) & \color{blue}{4.28}\ (5) & 1.64\ (1) & 1.57\ (2) \\
1.37\ (1) & 1.27\ (2) & \color{blue}{4.55}\ (5) & 4.63\ (5)
\end{pmatrix} & \begin{array}{l} \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array}
\end{array}$$

What have we learned?

- Covariance matrix
- Principal component analysis
- Dimensionality reduction
- Low-rank models
- Matrix completion