Tutorial 2: PCA

SYD 522

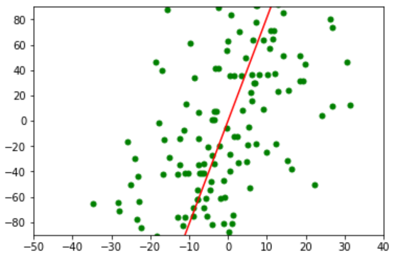
Shivam Kalra

University of Waterloo

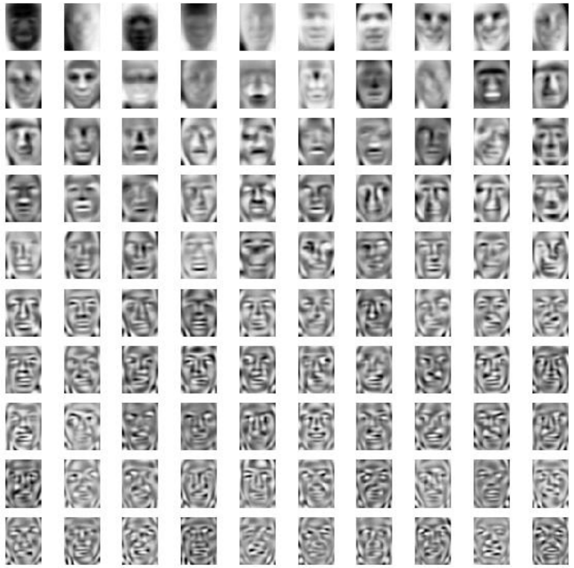
Administrative

- **Four** Tutorial Assignments
  - 2% Assignment 1 (posted this week)
    - PCA and Basic Probability Question
    - Due within a week of the posting
  - 3% Assignment 2
  - 3% Assignment 3
  - **12% Kaggle Challenge**

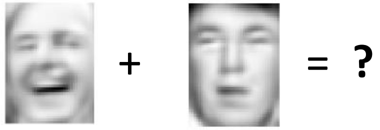
Today Agenda



1. Understand PCA



2. Apply PCA on Faces



Paul Martin Donald Trump

3. Generate Faces

4. Constructing an external face

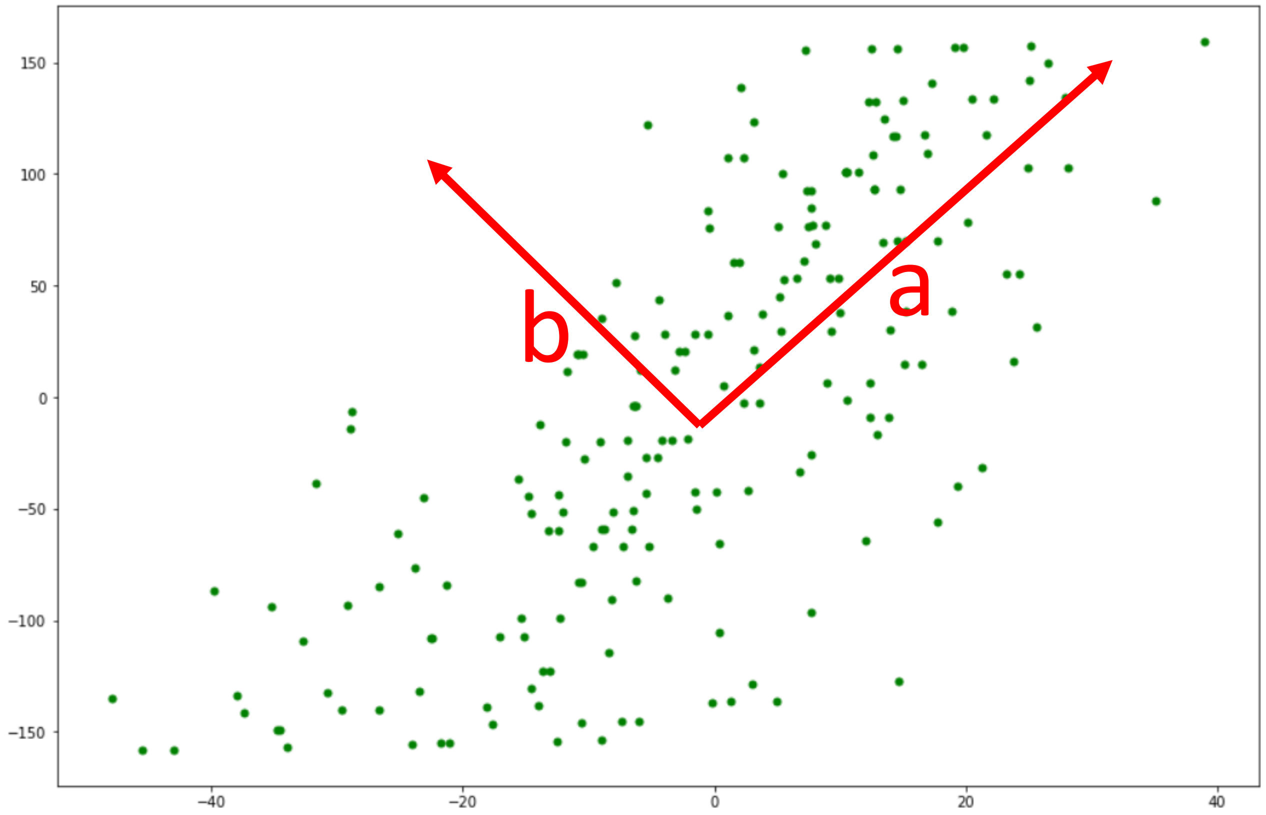
5. Searching similar faces

What is PCA?

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

Ref: [Jolliffe, Pricipal Component Analysis, 2 nd edition]

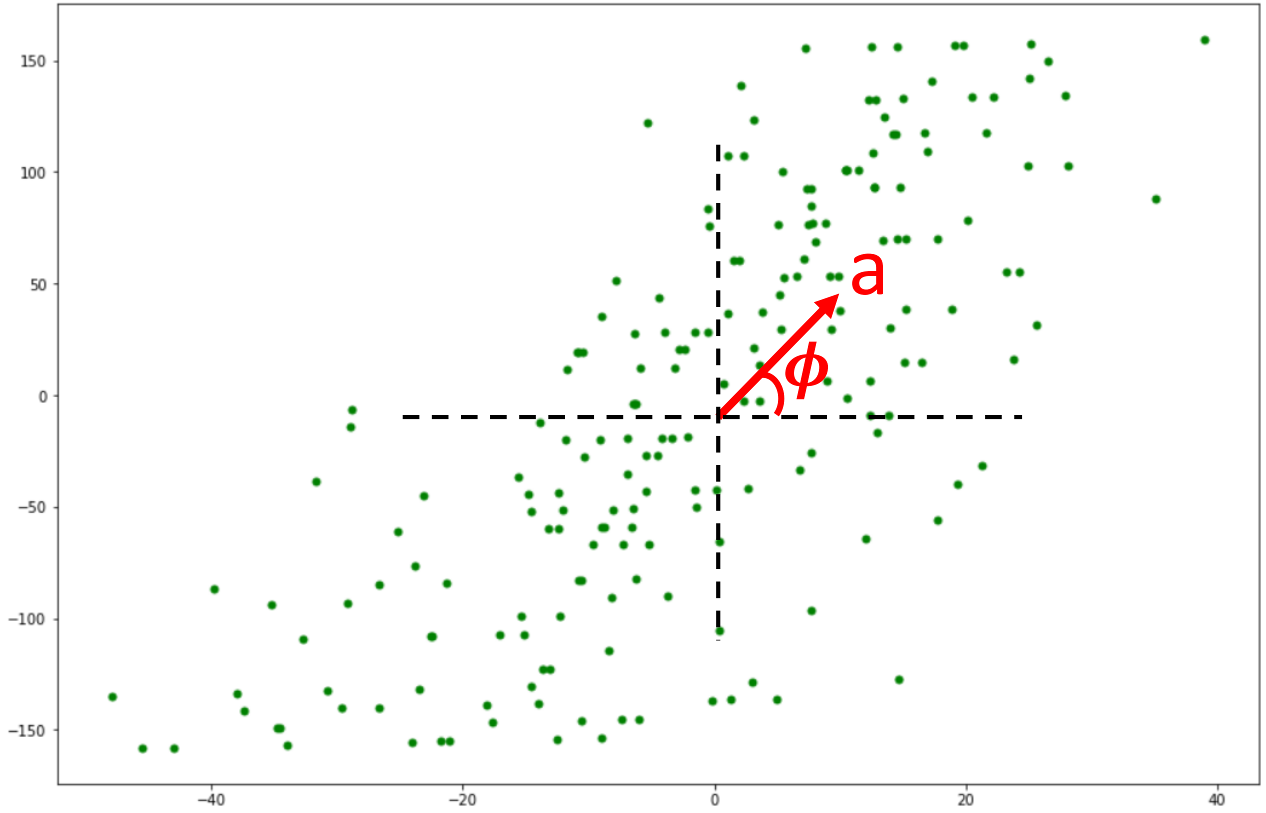
PCA (Kahoot!)



Which component captures higher variance in the data?

1) a
2) b

Iterative algorithm for PCA

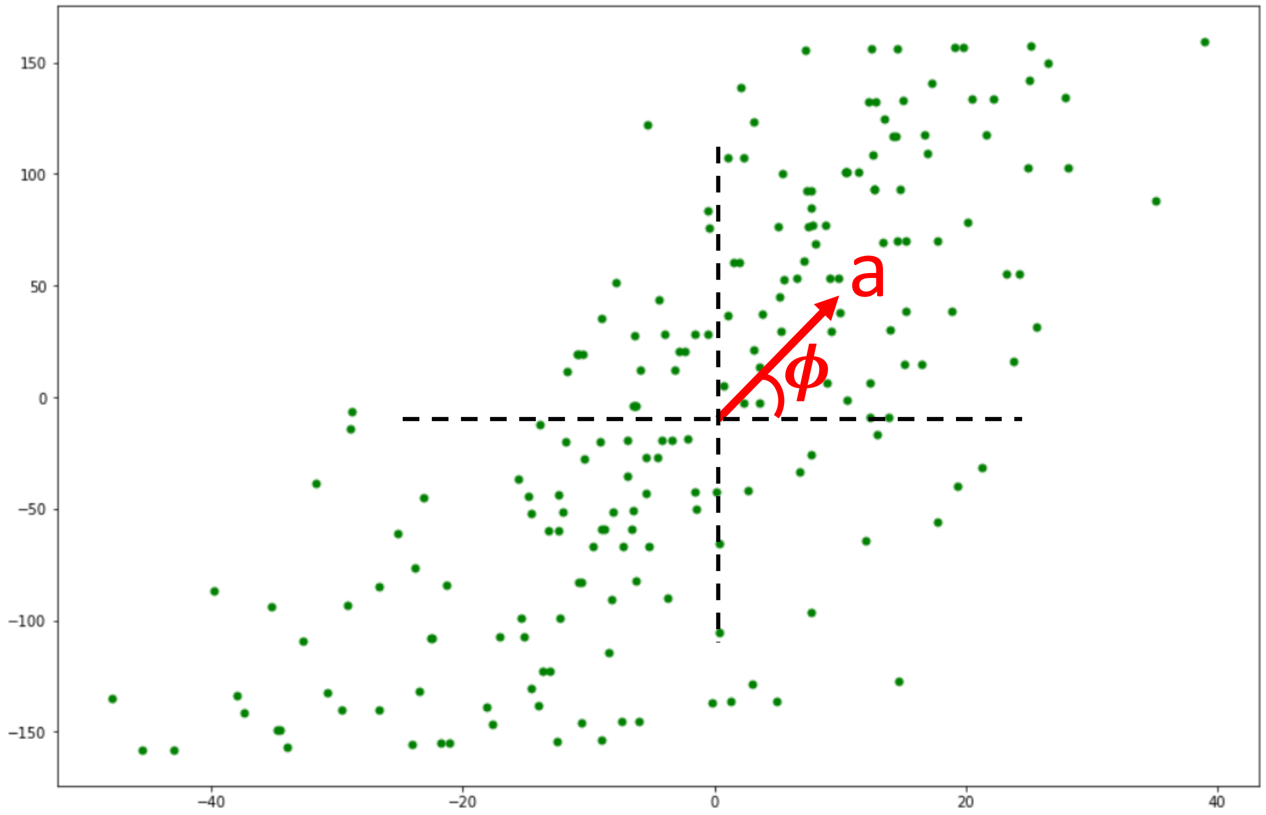


- Center the data first
- From 0^0 to 180^0
- Find ϕ that results in maximum variance when data is projected to $\bar{a}$

OR

$$arg\max_{\phi} var(data.\bar{a})$$

Iterative algorithm for PCA (kahoot!)



- What is **a** (it is a unit vector)?

Iterative algorithm for PCA

On the notebook!

Better algorithm for PCA

- Given centered data X and covariance matrix $\sum$.

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T$$

- PCA vectors are eigenvectors of $\sum$.
- Larger eigenvalue means more important PCA.

Better algorithm for PCA

- **Homework:** Find yourself why eigenvalue decomposition of covariance matrix yield principal components sorted according to their eigenvalues?

- Notebook time!

PCA: Reconstruction error

- Lets say $\mathrm{PCA} = [\overrightarrow{c_1}; \overrightarrow{c_2}; \dots; \overrightarrow{c_n}]$ where $\overrightarrow{c_n} = [c_1, c_2, \dots, c_m]$
- A data point $\vec{d} = [d_1, d_2, \dots, d_m]$can be transformed into m length using

$$d_{new} = \vec{d}.PCA$$

- Similarly, it can be reconstructed using d_{new}

$$d_{rec} = \vec{d}_{new}(1).\overrightarrow{c_1} + \vec{d}_{new}(2).\overrightarrow{c_1} + \cdots + \vec{d}_{new}(m).\overrightarrow{c_m}$$

- You can use numpy broadcasting for above, no for loops :)
- Notebook!

PCA: Reconstruction error (Kahoot!)

I’ve 500 2D points normally distributed with mean=0 and variance=1.

What is reconstruction error if I use all 2 principal components?

PCA: Similar Face Search

- We will build a very simple face search algorithm using PCA.
- It will find similar faces for a given query face's image.
- Notebook time!

Thank you