KMEANS & PCA

AARHUS
UNIVERSITY
DEPARTMENT OF CHEMISTRY

21 OCTOBER 2017

LISBETH RAVNKILDE KJØLBYE
RESEARCH ASSISTENT

UNSUPERVISED LEARNING

- K-means Clustering
- Principal Component Analysis (PCA) – with focus on MD simulations

K-MEANS CLUSTERING (PARTITIONAL)

To partition a data set N, with each data point being a collection of features, into K clusters by minimizing a mean-square-error (MSE) function.

$$J_{MSE} = \sum_{i=1}^{K} \sum_{x_t \in C_i} \|x_t - m_i\|^2$$

$$C_i, i = 1, \ldots, K$$

Where m_i is the geometrical centroid of cluster C_i
and x_t is a vector of the t-th data point in cluster C_i

K-MEANS CLUSTERING

Step 1: Randomly assign the centroid $m_1, \ldots m_K$

Step 2: Each data point is assigned to a cluster C_i based on the squared minimum distance $argmin(\|x_t - m_i\|^2)$ between the data point and the centroid m_i.

Step 3: For all K clusters the centroid m_i is set to the center of mass of all the points in cluster C_i

Step 2 and 3 is repeated for all data points and K clusters until a convergence criteria is reached.

Convergence criteria can be a maximum set of iterations, no data points change clusters, or the sum of distances is minimized.

K-MEANS CLUSTERING

Drawbacks:

- The number of K clusters must be know and fixed
  - Several ways to estimate K

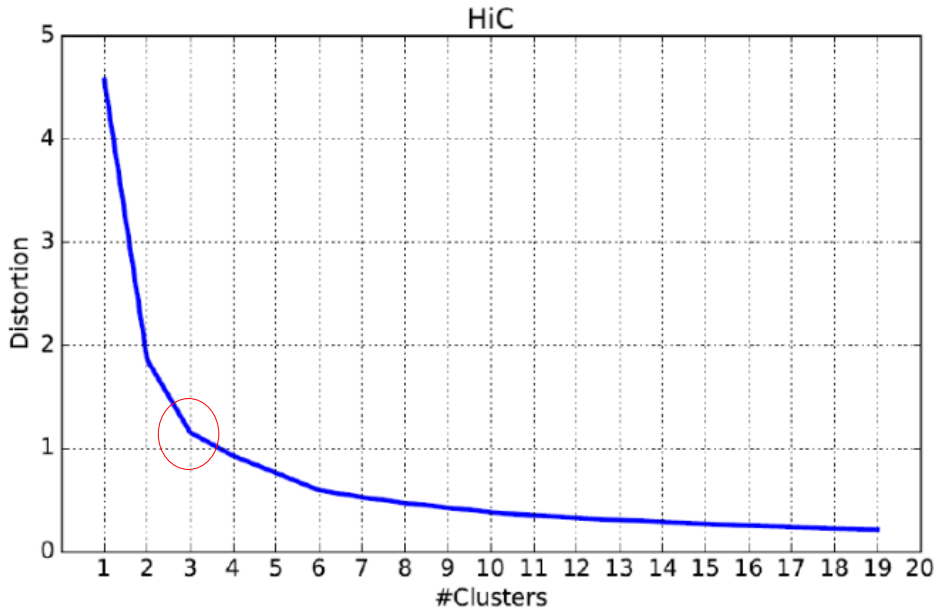
- The results of the algorithm depend on the initial selection of centroids m_i
  - Different results for each run – perform several runs

- The algorithm can be caught in a local minima

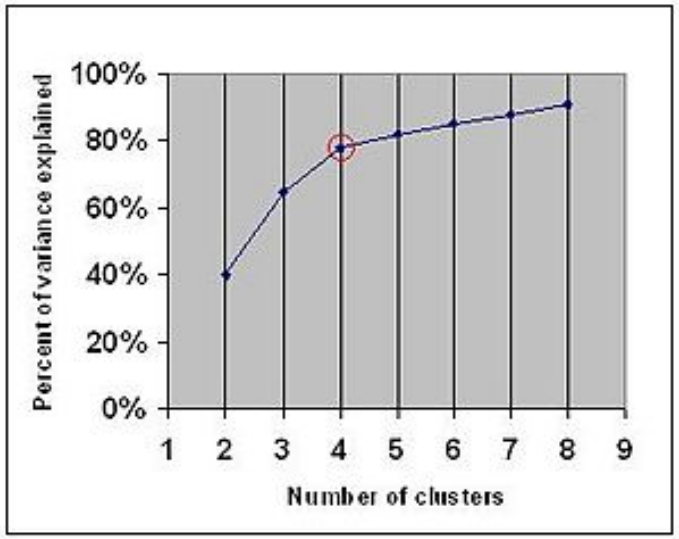
- It contains the dead-unit problem
  - If a centroid is selected far away from the rest of the input data

K-MEANS CLUSTERING

One method for estimation K: The “Elbow method”, “Elbow point” or “Elbow criteria”



The distortion is the sum of the squared difference between the observations and the corresponding centroid.



Percentage of variance explained as the ratio of the between-group variance to the total variance

K-MEANS CLUSTERING

Other methods for estimating K:

- Cross-validation
- G-means algorithm
- The silhouette method
- Genetic algorithm

K-MEANS CLUSTERING

Treatment of input data:

- It is common practice to “whiten” / normalizing the date before applying K-means algorithm.
- Principal Component Analysis (PCA) for feature reduction or extraction and visualization

PCA

A statistical method developed by Karl Pearson in 1901

A set of possibly correlated variables → Linear combination of uncorrelated Principal Components

PCA

From X a matrix of n samples and m measurements containing the data, the covariance matrix C can be calculated:

$C = X^T X$

Eigenvalue decomposition of C can be performed:.

$C = T \Lambda T^T$

where T is a matrix with the eigenvectors as columns &

Λ is a diagonal matrix with the corresponding eigenvalues.

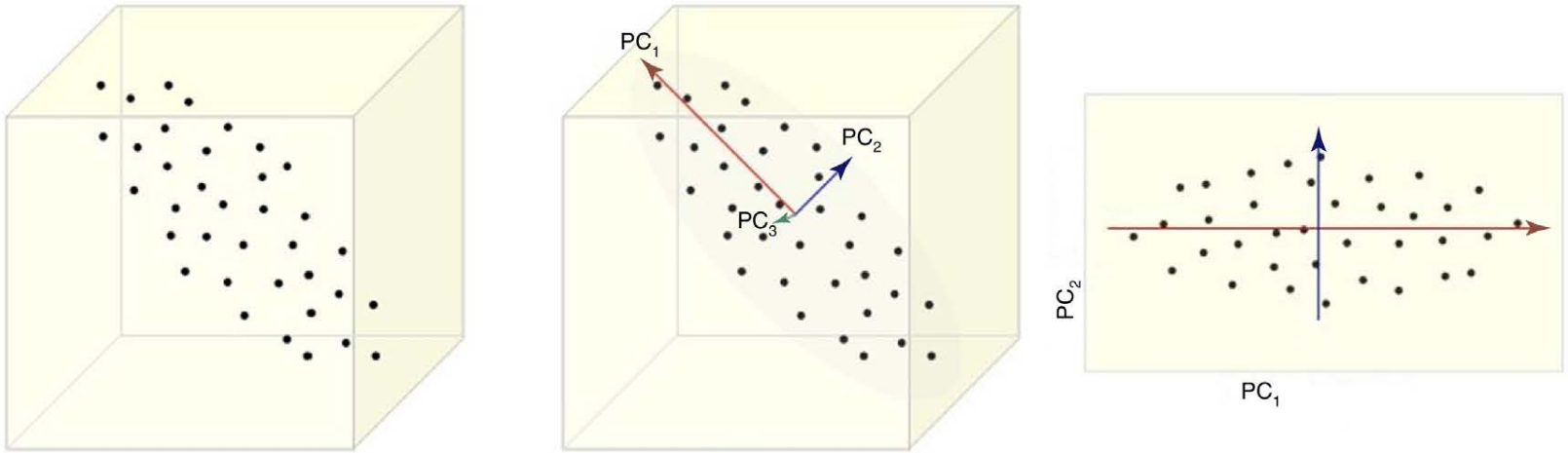
The eigenvectors are ordered according to the corresponding eigenvalue.

The eigenvalues are indicators of the variance along each eigenvector.

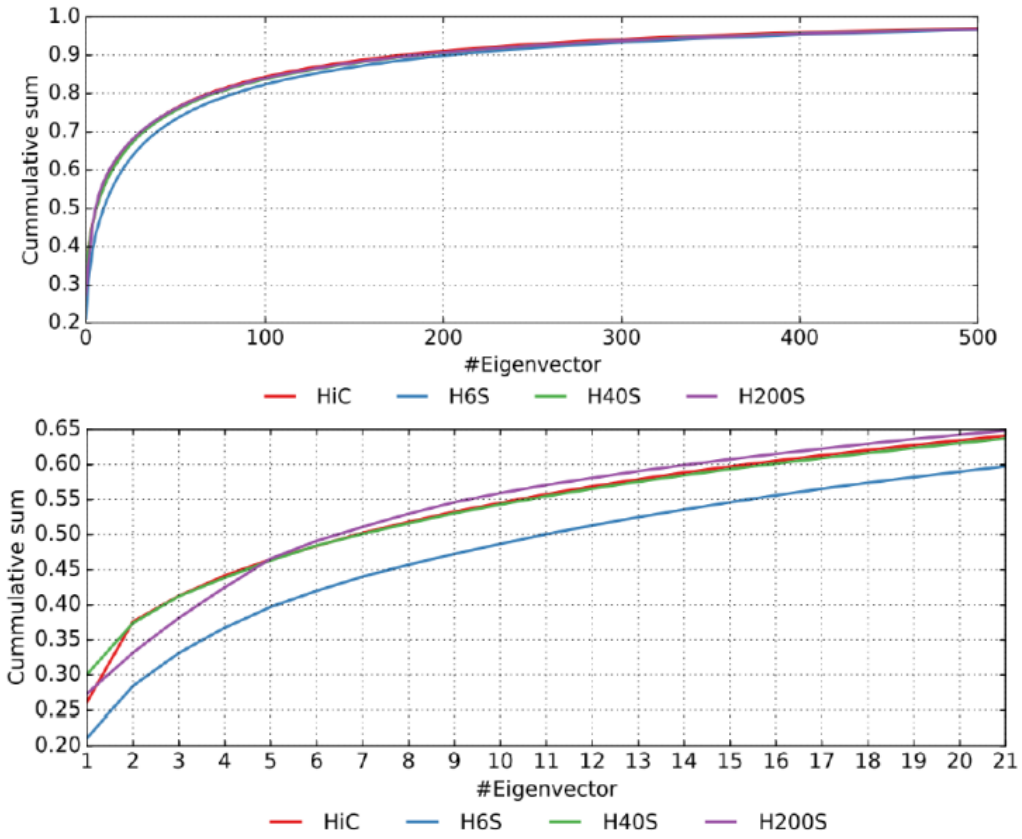
$P = XT$

X is thereby transformed into a new orthogonal basis set composed of eigenvectors

PCA



PCA



Scree plot represents the fraction of the total variance of the data as a function of eigenvectors.

WHAT I DID

Investigate the essential motion of a protein along with clustering of the sampled structures.

- Calculated the PCs and projected the trajectory along the first 4 eigenvectors

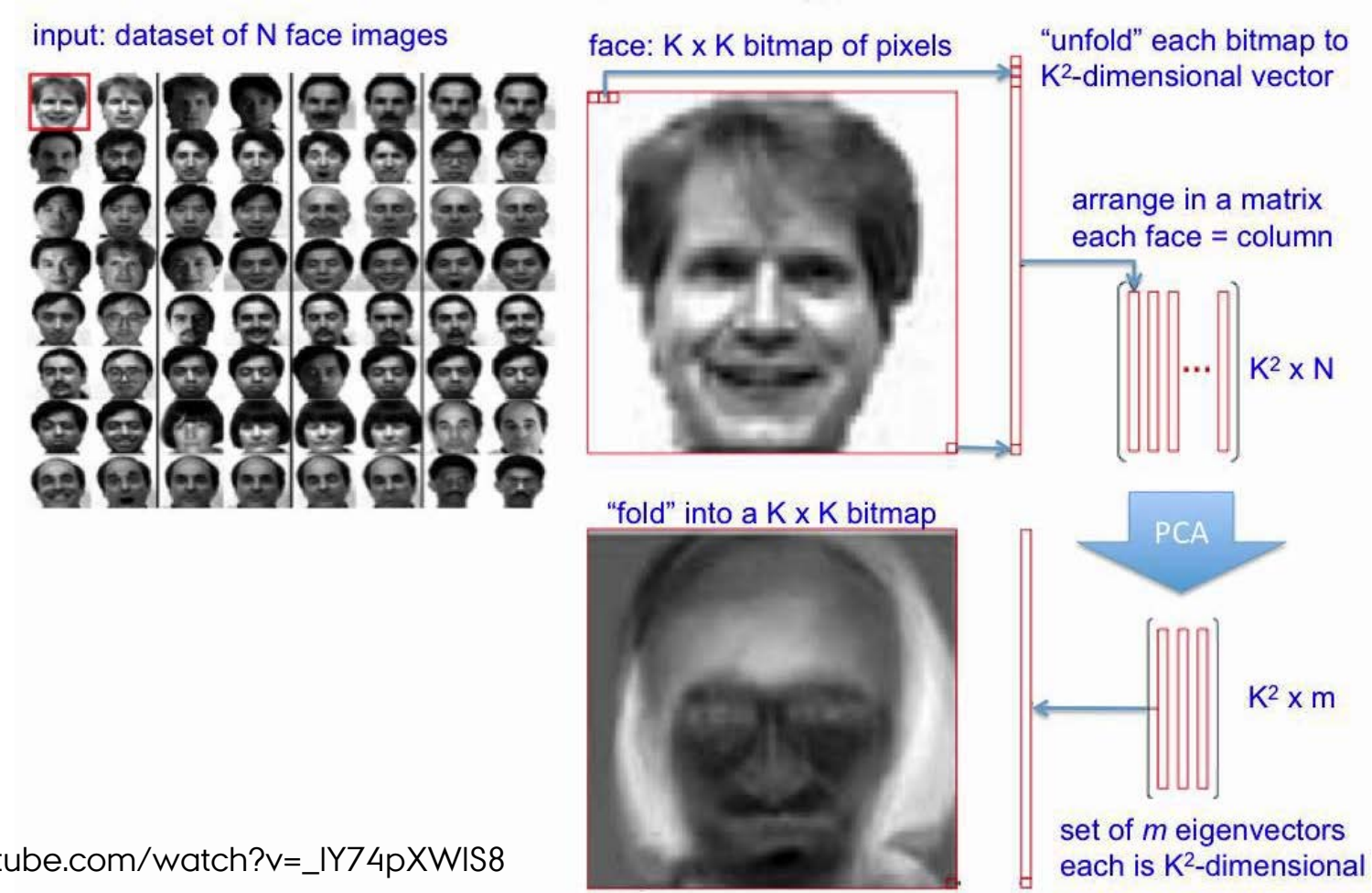
- Performed K-means clustering on the projection

- ”Back traced” the clustering to the space of my measurements (X) thereby clustering the sampled structures of the protein

PCA - DRAWBACKS

- In case of MD simulation:
  - Depend on the sampling – cosine content
  - Concatenated trajectories might result in artefacts
- Considered the data you use – The first PC might not be of relevance
  - For MD remove any translation and rotation of the protein, in order for the PCA to capture the internal motions

PCA FOR FACIAL RECOGNITION



https://www.youtube.com/watch?v=_lY74pXWIS8

AARHUS
UNIVERSITY