UNIVERSITY OF GOTHENBURG

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Statistical Methods in Natural Language Processing (NLP)

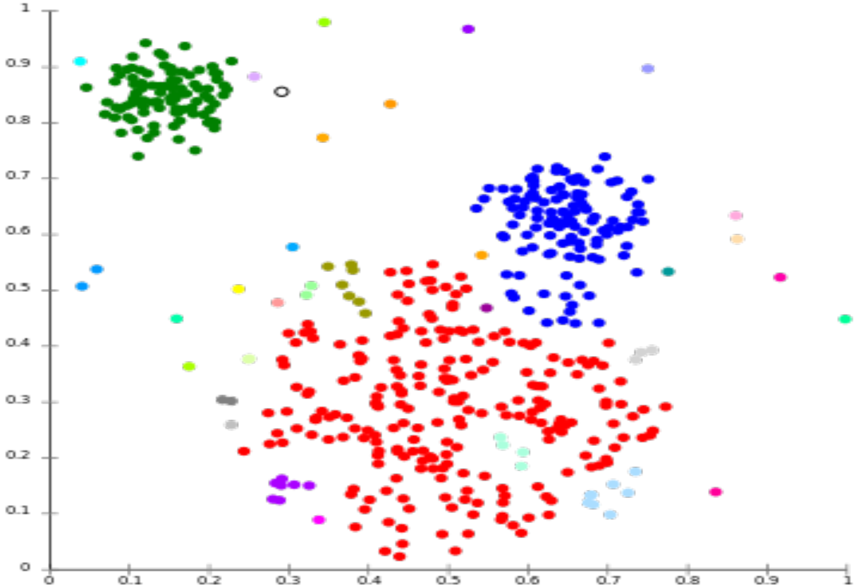
Chatrine Qwaider

Clustering Algorithms

K-means
Gaussian Mixture Model

Clustering

- Kind of Unsupervised Learning problem
- Given : N unlabelled examples : $\{x_1, \ldots, x_N\}$, each with D dimensions (features) : $\{f_1, \ldots, f_D\}$, Number of clusters K
- **Target**: how to group all the examples into the K portions.
- Example:
  - K-means
  - Gaussian Mixture Model



K-mean Algorithm

- Nonprobabilistic technique
- Hard assignment : each example must belongs to one cluster
- **Cluster** a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.

K-mean Algorithm

- Given
  - X data set $\{x_1, \ldots, x_N\}$ consisting of N observations of a random D-dimensional Euclidean variable x.
  - Goal is to partition the data set into some number K of clusters
- Define
  - set of D-dimensional vectors μ_k, where $k = 1, \ldots, K$, in which μ_k is a prototype associated with the k^{th} cluster.
- **Goal** :
  - find an assignment of data points to clusters
  - such that the sum of the squares of the distances of each data point to its closest vector μ_k, is a minimum.

K-mean (*Objective Function*)

- For each data point x_n, we introduce a corresponding set of binary indicator variables $r_n^k \in \{0, 1\}$, where $k = 1, \ldots, K$ describing which of the K clusters the data point x_n is assigned to

- objective function, sometimes called a ***distortion measure***,

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

- Goal: find values for the $\{ r_n^k \}$ and the $\{ \mu_k, \}$ to minimize J

K mean algorithm

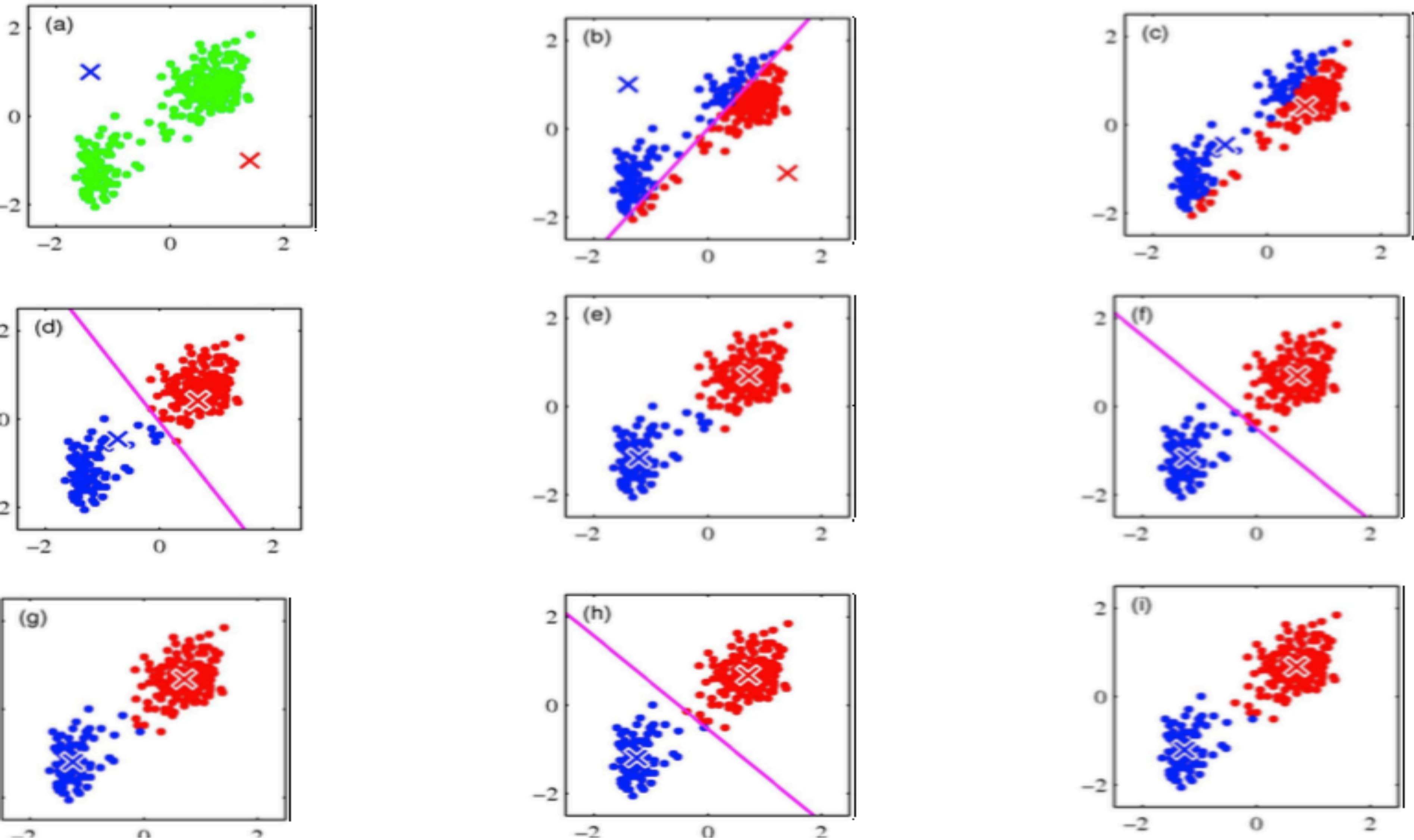
- Randomly initialize *k* centers
  - $\mu^{(0)} = \mu_1^{(0)}, \ldots, \mu_k^{(0)}$
- **Assignment step**: Assign each point $i \in \{1, \ldots, m\}$ to nearest center:

$$C^{(t)}(j) \leftarrow \arg\min_i ||\mu_i - x_j||^2$$

- **Update step**: μ_i becomes centroid of its point:

$$\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j:C(j)=i} ||\mu - x_j||^2$$

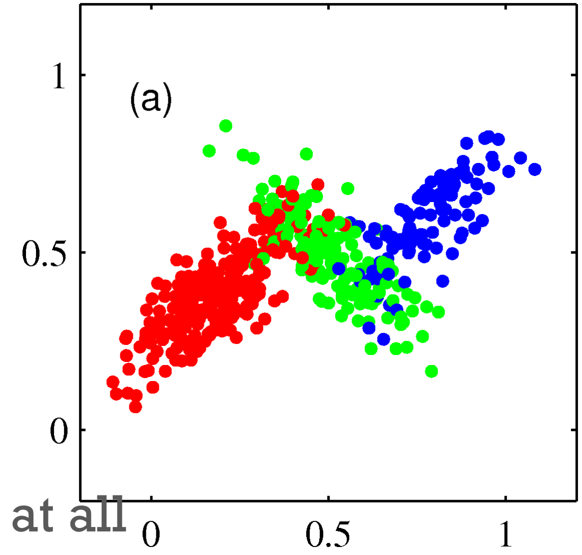
K-mean Example



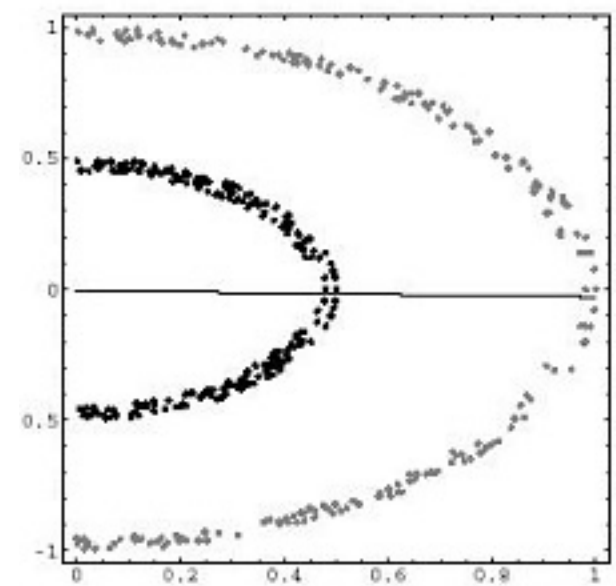
K mean Advantages

- Fast, robust and easier to understand.
- Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.
- Gives best result when data set are distinct or well separated from each other.

Limitation



- Makes hard assignments of points to clusters
  A point either totally belongs to a cluster or not at all
- No notion of a soft/fractional assignment (i.e., probability of being assigned to each cluster: say K = 3 and for some point x_n, p1 = 0.7,p2 = 0.2,p3 = 0.1)
- K-means often doesn't work when clusters are not round shaped, and/or may overlap, and/or are unequal
- Unable to handle nonlinear or noisy data and outliers.



K mean DEMO

http://syskall.com/kmeans.js/

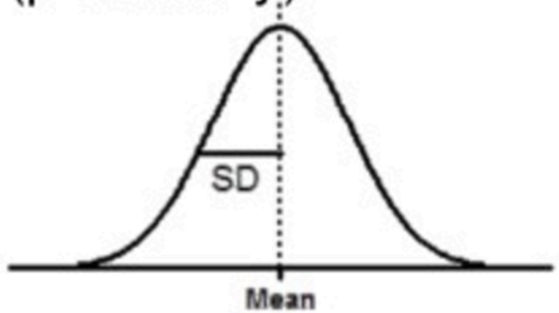
Soft Assignment

- **K-means algorithm** :
  - every data point is assigned uniquely to one, and only one, of the clusters.
  - Some data points that lie roughly midway between cluster centers.
- **Solution**: adopting a probabilistic approach, we obtain '**soft**' assignments of data points to clusters in a way that reflects the level of uncertainty over the most appropriate assignment.

Gaussian Mixture Model GMM

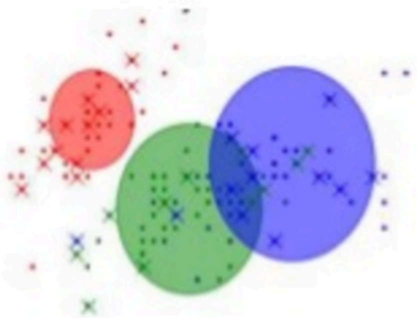
- Gaussian

"Gaussian is a characteristic symmetric "bell curve" shape that quickly falls off towards 0 (practically)"



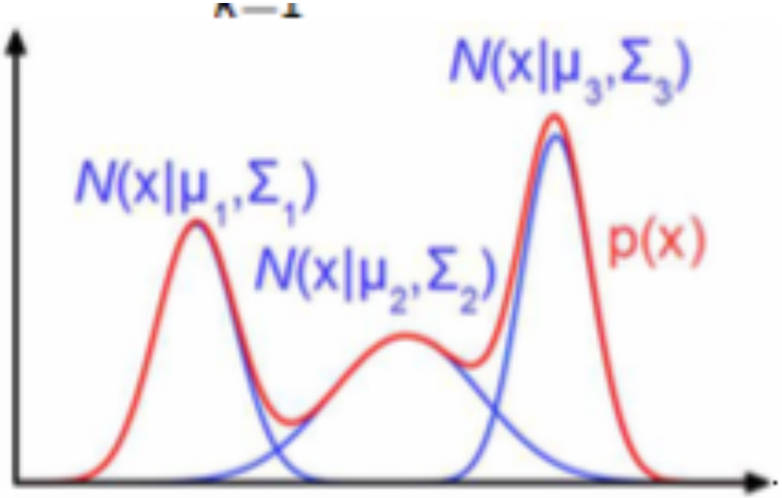
- Mixture Model

"mixture model is a probabilistic model which assumes the underlying data to belong to a mixture distribution"



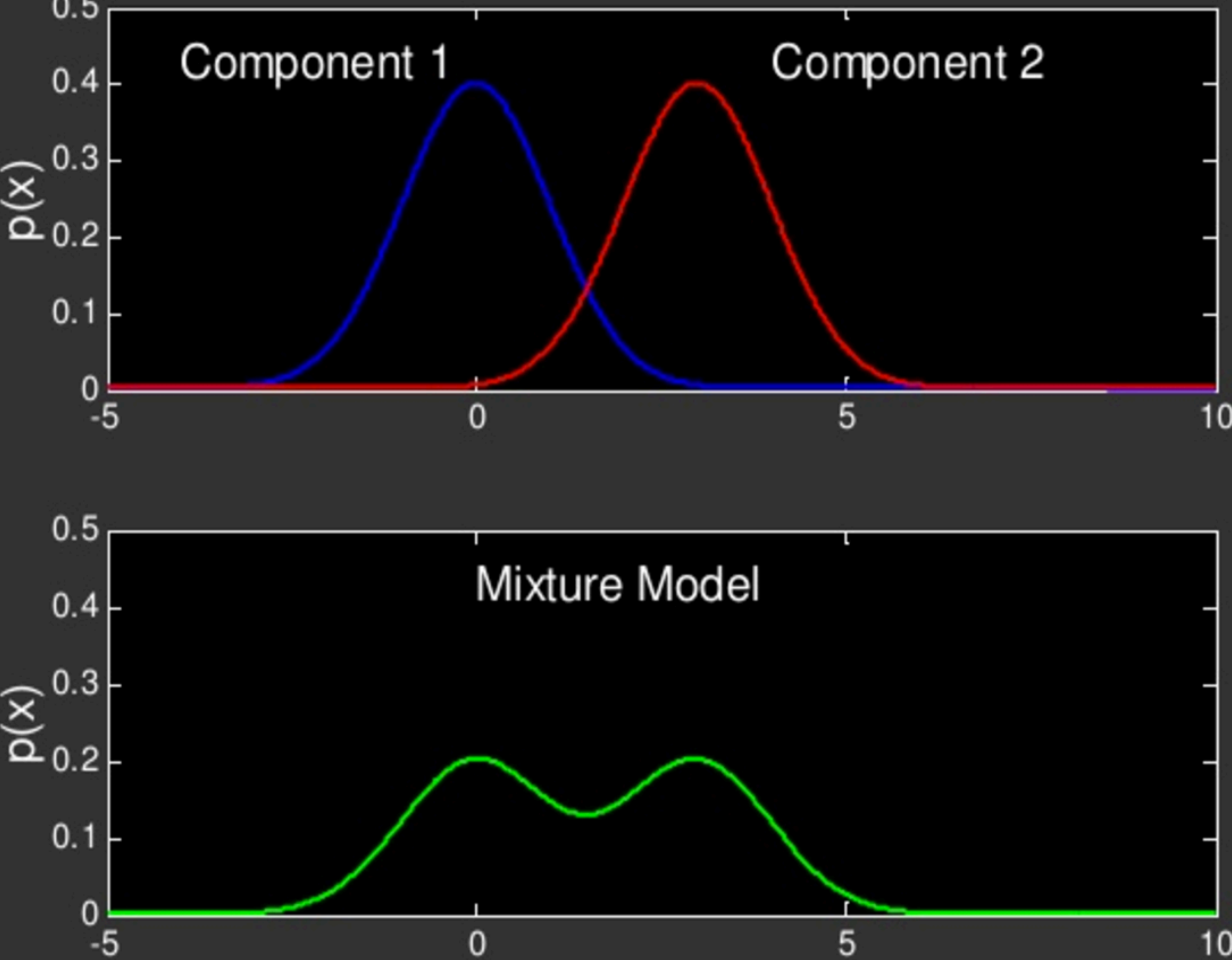
Gaussian Mixture Model GMM

- **Gaussian (normal) distribution**: model for the distribution of continuous variables
- **Mixture Distribution** : probability distribution of a random variable that is derived from a collection of other random variable.
- A **probability density p**(x) represents a mixture distribution or mixture model

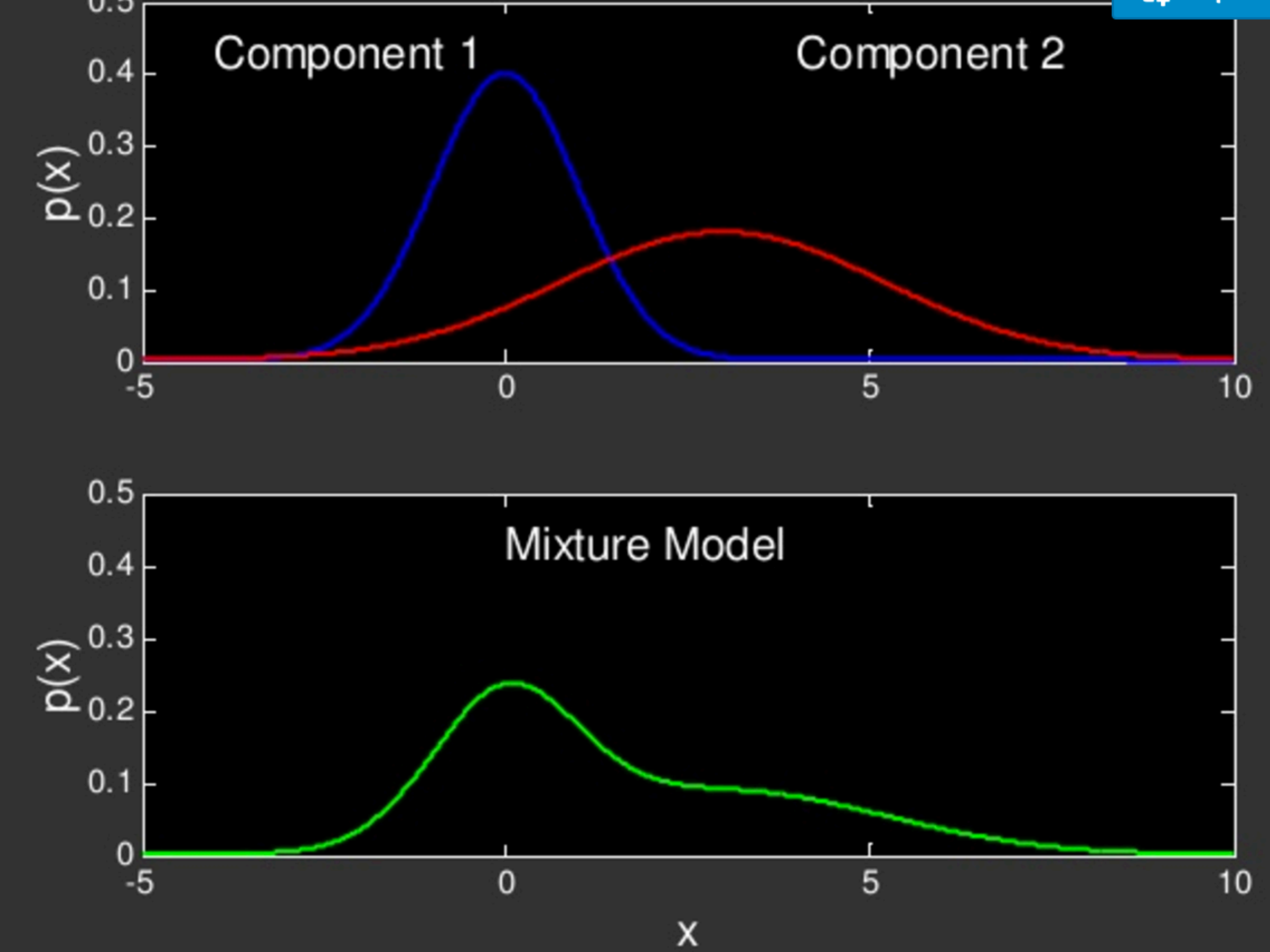


Gaussian Mixture Model GMM

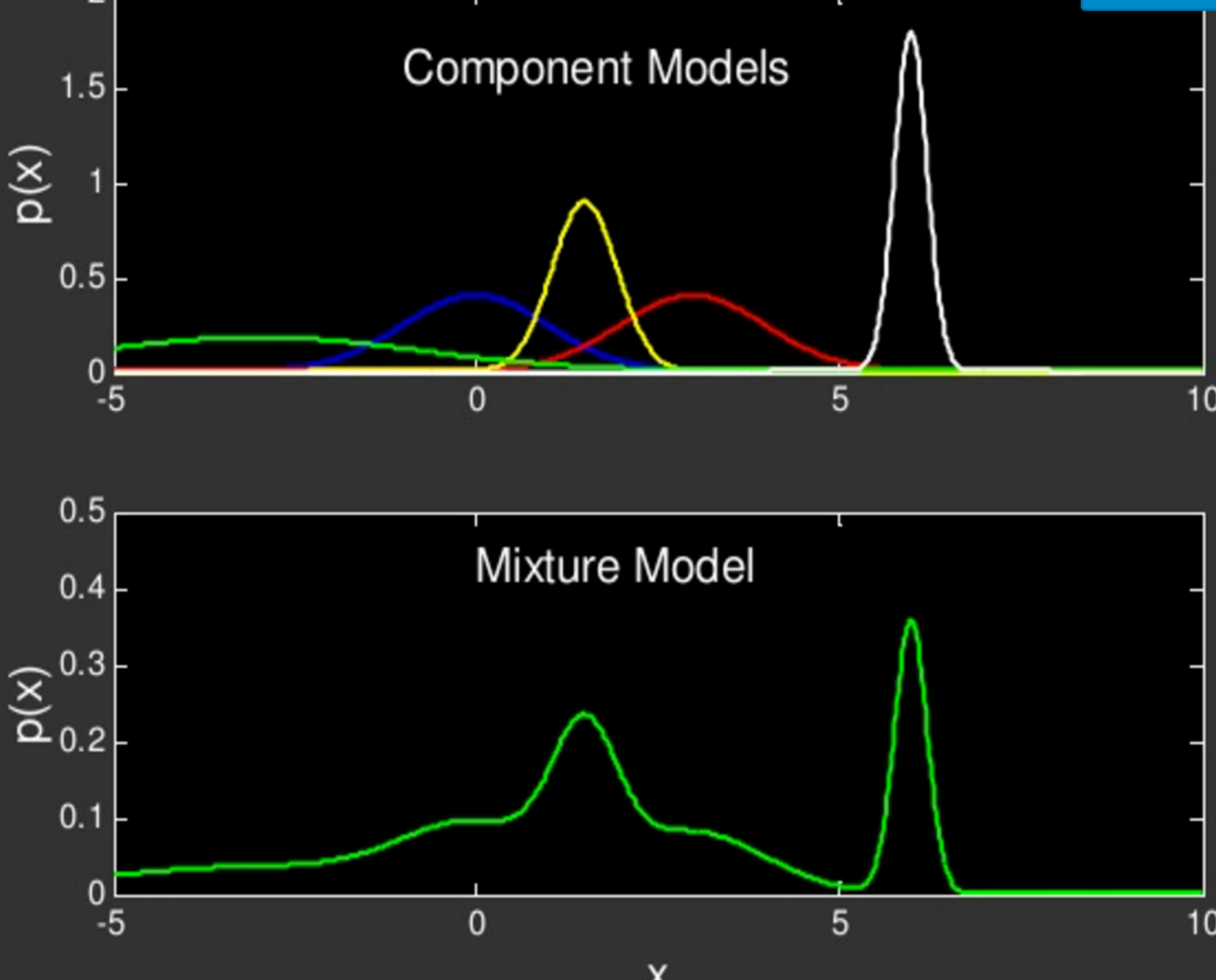
- A Gaussian mixture model is a **probabilistic model** that assumes all the data points are generated from a **mixture** of a finite number of Gaussian distributions with unknown parameters.



Component 1
Component 2
p(x)
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
Mixture Model
p(x)
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10



Component 1
Component 2
p(x)
0
0.1
0.2
0.3
0.4
-5
0
5
10
Mixture Model
p(x)
0
0.1
0.2
0.3
0.4
0.5
-5
0
5
10
x


Component Models
p(x)
1.5
1
0.5
0
-5
0
5
10
Mixture Model
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
x

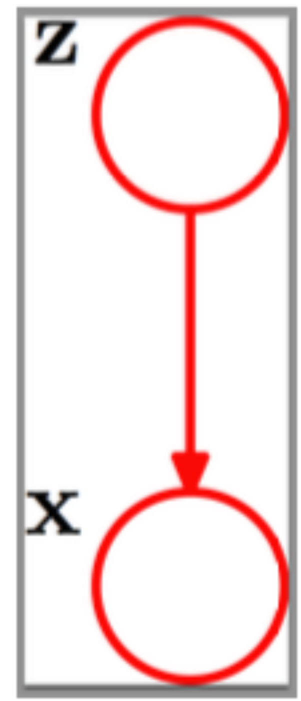
Gaussian Mixture Model: Joint Distribution

- Factorize joint according to Bayes net:

$$p(x,z) = p(z)p(x \mid z) = \pi_z N(x \mid \mu_z, \Sigma_z)$$

- π_z is probability of choosing cluster z.

- $X \mid Z = z$ has distribution $N(\mu_z, \Sigma_z)$.



Learning the Gaussian Mixture Model - likelihood function

- we only observe X, we need the **marginal distribution**:

$$
\boxed{
\begin{aligned}
p(x) &= \sum_{z=1}^{k} p(x, z) \\
&= \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z)
\end{aligned}
}
$$

Learning the Gaussian Mixture Model - likelihood function

- The model **likelihood** for $D = \{x_1, \ldots, x_n\}$ is

$$
\begin{aligned}
L(\pi, \mu, \Sigma) &= \prod_{i=1}^{n} p(x_i) \\
&= \prod_{i=1}^{n} \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z).
\end{aligned}
$$

- As usual, we'll take **objective function** to be the log of this:

$$
J(\pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right\}
$$

Gaussian Mixture Model: Conditional Distribution

- We observe X = x, the **conditional distribution** of the cluster Z given X = x is

$$\mathbf{p(z \mid X = x) = p(x,z)/p(x)}$$

- The conditional distribution is a **soft assignm**ent to clusters.
- $z \in \{1,\ldots,k\}$, Cluster assignment Z is called a **hidden variable**

Learning the Gaussian Mixture Model

- Build GMM

  1. Cluster probabilities $\pi = (\pi_1,\ldots,\pi_k)$
  2. Cluster means $\mu = (\mu_1,\ldots,\mu_k)$
  3. Cluster covariance matrices: $\Sigma = (\Sigma_1,\ldots\Sigma_k)$

- We have a probability model: let's find the MLE.

- Suppose we have data $D = \{x_1,\ldots,x_n\}$.
  We need the **model likelihood** for D.

Learning the Gaussian Mixture Model

- Denote the probability that observed value x_i comes from cluster j by $\gamma_{ji} = P(Z = j \mid X = x_i)$.
- The responsibility that cluster j takes for observation xi . Computationally,
  - $\gamma_{ji} = P(Z = j \mid X = x_i)$
  - $\gamma_{ji} = p(Z=j, X = x_i)/p(x) = \pi_j N(xi \mid \mu j, \Sigma j) /p(x)$
- The vector $\gamma_1,\ldots,\gamma_k$ is exactly the soft assignment for x_i.

Exception Maximization Learning

1. Initialize parameters μ, π, Σ.

2. "E step". Evaluate the responsibilities using current parameters

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)},$$

   - for i = 1,…,n and j = 1,…,k.

3. "M step". Re-estimate the parameters using responsibilities

$$\mu_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c x_i$$

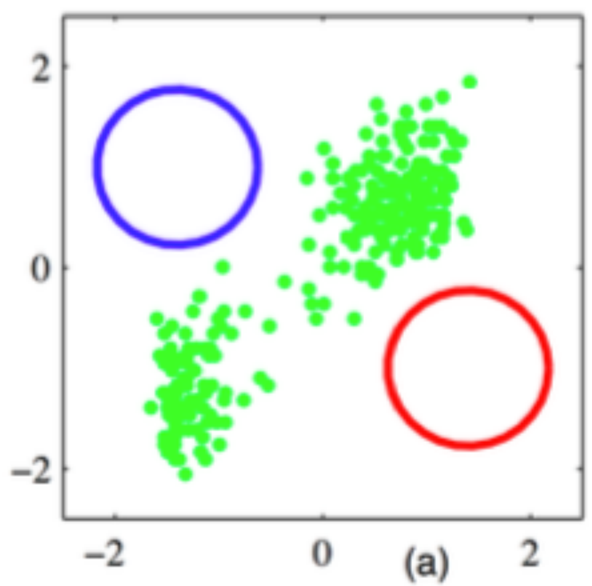
$$\Sigma_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (x_i - \mu_{\text{MLE}})(x_i - \mu_{\text{MLE}})^T$$

$$\pi_c^{\text{new}} = \frac{n_c}{n},$$

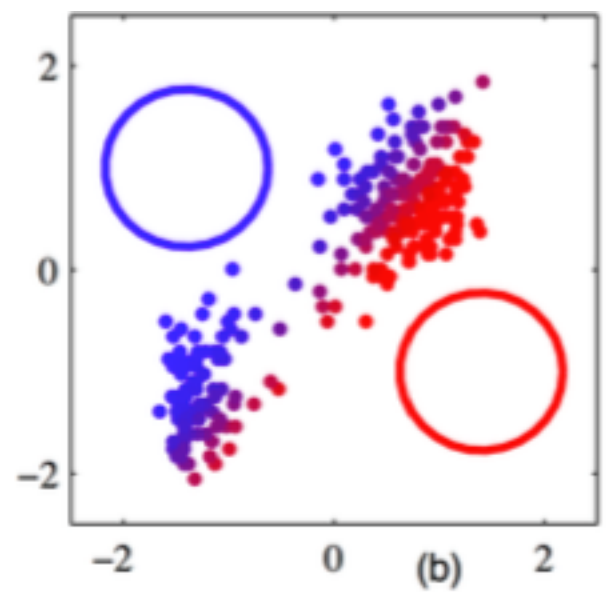
4. Repeat from Step 2, until log-likelihood converges.

EM Example

1. Initialization

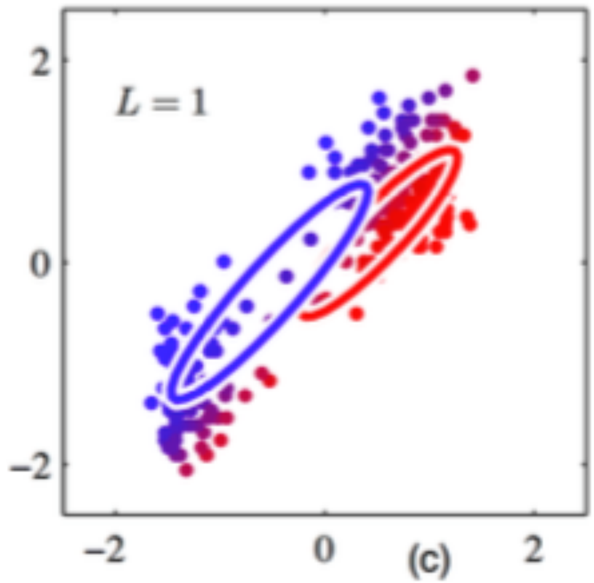


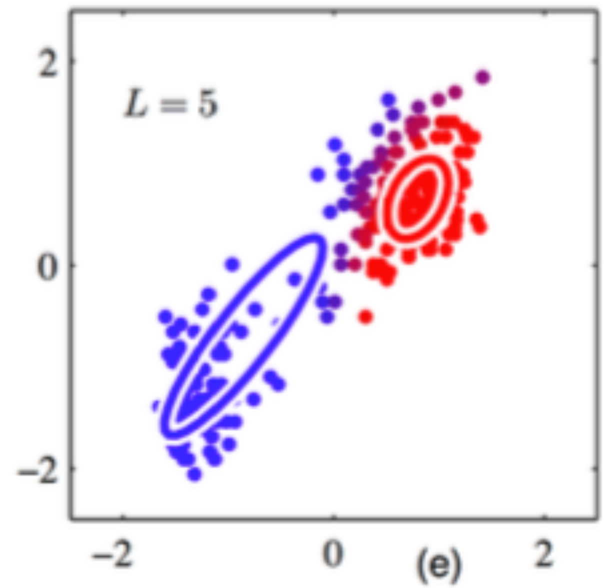
2. First soft assignment



EM Example

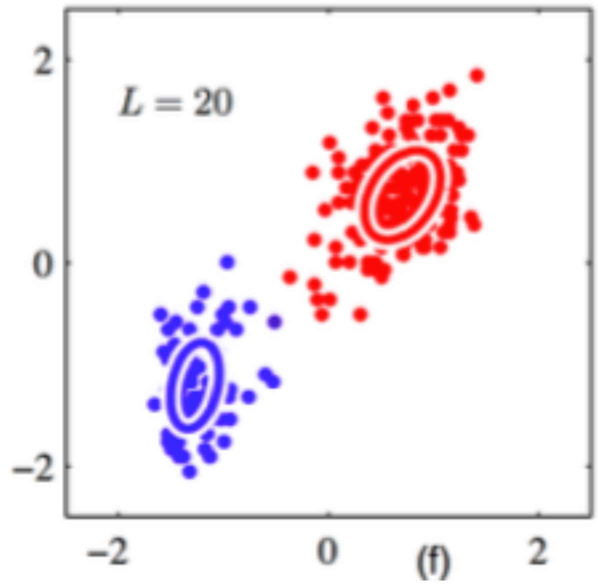
3. Updating Parameters:
4. Repeat Assignment





EM Example

5. After 20 rounds:



Advantages/Dis of GMM

Strength

- ✓ GMM model accommodates mixed membership
- ✓ Flexible in terms of cluster covariance

Weakness:

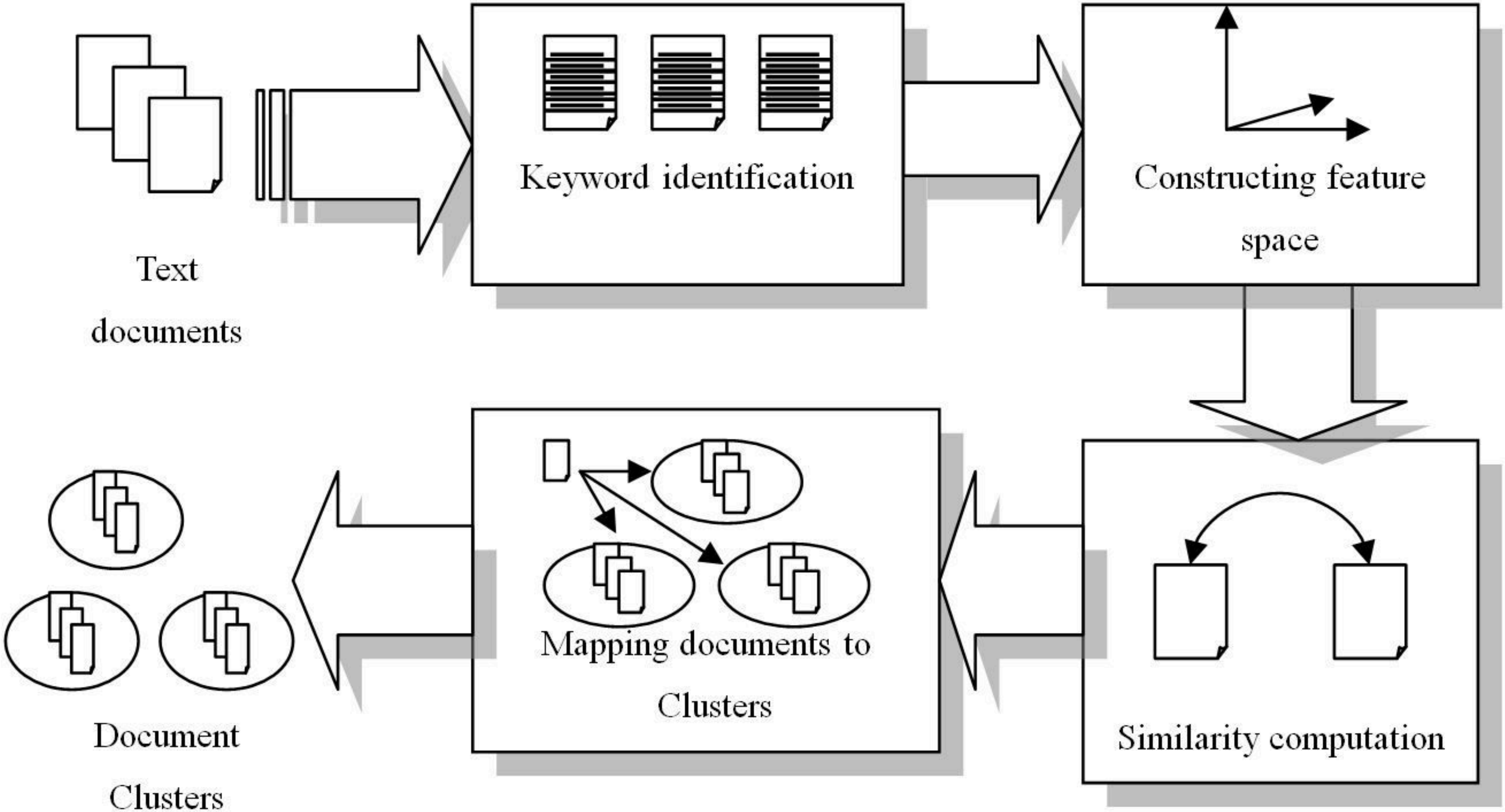
- x Computationally expensive if the number of distribution is large, or the data set contains very few observed data points
- x Hard to estimate number of clusters

GMM Demo

https://lovasoa.github.io/expectation-maximization/dist/

+ Clustering– NLP example

- Document clustering
  - The similarity measure is the key point to the clustering problem.
  - represent each document as an array of numbers Using VSM
    - Count up the number of times each word appears in the document.
    - Choose a set of "feature" words that will be included in your vector. This should exclude extremely common words ("stopwords") like "the", "a", etc.
    - Make a vector for each document based on the counts (. TF X IDF) of the feature words.
  - Apply k-means , EM to cluster documents by maximizing the likelihood of the unlabeled documents


Text documents
Keyword identification
Constructing feature space
Similarity computation
Mapping documents to Clusters
Document Clusters

+

Clusetring related documents

[PDF] A survey on various approaches in **document clustering**
…, V Preamsudha, MP **Scholar** - International …, 2011 - pdfs.semanticscholar.org
Abstract Document clustering is the process of segmenting a particular collection of texts into subgroups including content based similar ones. The purpose of document clustering is to meet human interests in information searching and understanding. Nowadays all paper
Cited by 15 Related articles All 3 versions Cite Save More

A new unsupervised method for **document clustering** by using WordNet lexical and conceptual relations
DR Recupero - Information Retrieval, 2007 - Springer
… improve the preprocessing since it is the most critical step for the generation of an appropriate **document** representation. The other one is how the use of WordNet benefits the cluster labeling task: having **documents** represented by concepts … Frequent term-based text **clustering**. …
Cited by 66 Related articles All 7 versions Web of Science: 16 Cite Save

Ant-based and swarm-based **clustering**
J Handl, B Meyer - Swarm Intelligence, 2007 - Springer
… Homogeneous ants for web **document** similarity modeling and categorization. … A stochastic heuristic for visualising graph **clusters** in a bi-dimensional space prior to partitioning. Journal of Heuristics, 5(3), 327–351. … Antclust: ant **clustering** and web usage mining. …
Cited by 134 Related articles All 5 versions Cite Save

+ NLP example

- D1: there is a dog who chased a cat
- D2: someone ate pizza for lunch
- D3: the dog and a cat walk down the street toward another dog
- **feature words** are [dog, cat, street, pizza, lunch]
- **[1, 1, 0, 0, 0]** // dog 1 time, cat 1 time
- [0, 0, 0, 1, 1] // pizza 1 time, lunch 1 time
- **[2, 1, 1, 0, 0]** // dog 2 times, cat 1 time, street 1 time
- **Apply Clustering Algorithm Cosine similarity between vectors**

Document clustering using K-means

- Assuming we have data with no labels for **news** and **technical** data
- We want to be able to categorize a new document into **one** of the 2 classes (K=2)
- We can extract represent document as feature vectors
  - Features can be word id or other NLP features such as POS tags, word context etc (D=total dimension of Feature vectors)
  - N documents are available
- Randomly initialize 2 class means
- Compute square distance of each point (x_n)(D) to class means (μ_k)
- Assign the point to K for which μ_k is lowest
- Re-compute μ_k and re-iterate until converge

References

1. Bishop , C. M. (2006). *Pattern Recognition and Machine Learning.* Cambridge : Springer Science+Business Media .
2. Do, C. B., & Batzoglou, S. (2008). *What is the expectation maximization algorithm?* ature Publishing Group .
3. *Gaussian mixture models*. (u.d.). Hämtat från scikit-learn: http://scikit-learn.org/stable/modules/mixture.html den 6 3 2017
4. *K Means Clustering with Tf-idf Weights*. (2 2013). Hämtat från Jonathanzong: http://jonathanzong.com/blog/2013/02/02/k-means-clustering-with-tfidf-weights den 5 3 2017
5. *K-means clustering* . (u.d.). Hämtat från OnMyPhD: http://www.onmyphd.com/?p=k-means.clustering#h3_goodexample den 8 3 2017
6. Reynolds, D. *Gaussian Mixture Models .* Lexington : MIT Lincoln Laboratory.
7. Jamnejad, M. I., Heidarzadegan, A., & Meshki, M. (2014). Text Recognition with k-means Clustering. *Research in Computing Science*, *84*, 29-40.
8. Kaur, R., & Kaur, A. (2016). Text Document Clustering and Classification using K-Means Algorithm and Neural Networks. *Indian Journal of Science and Technology*, *9*(40).
9. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.
10. (Gaussian mixture models) http://scikit-learn.org/stable/modules/mixture.html