STA 414/2104: Machine Learning

Russ Salakhutdinov

Department of Computer Science
Department of Statistics
rsalakhu@cs.toronto.edu
http://www.cs.toronto.edu/~rsalakhu/

Lecture 7

Mixture Models

• We will look at the mixture models, including Gaussian mixture models.

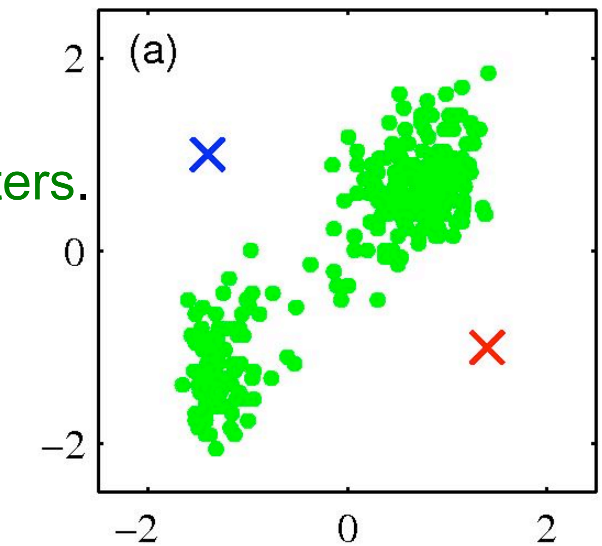
• The key idea is to introduce latent variables, which allows complicated distributions to be formed from simpler distributions.

• We will see that mixture models can be interpreted in terms of having discrete latent variables (in a directed graphical model).

• Later in class, we will also look at the continuous latent variables.

K-Means Clustering

- Let us first look at the following problem: Identify clusters, or groups, of data points in a multidimensional space.
- We observe the dataset $\{\mathbf{x_1}, ..., \mathbf{x}_N\}$ consisting of N D-dimensional observations
- We would like to partition the data into K clusters, where K is given.
- We next introduce D-dimensional vectors, prototypes, $\boldsymbol{\mu}_k, k = 1, ..., K.$
- We can think of μ_k as representing cluster centers.
- Our goal:
  - Find an assignment of data points to clusters.
  - Sum of squared distances of each data point to its closest prototype is at the minimum.

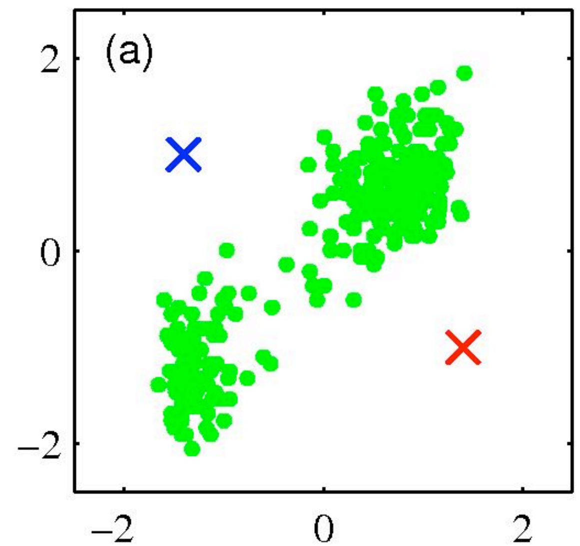


K-Means Clustering

- For each data point $\mathbf{x}_n$ we introduce a binary vector $\mathbf{r}_n$ of length K (1-of-K encoding), which indicates which of the K clusters the data point $\mathbf{x}_n$ is assigned to.
- Define objective (distortion measure):

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

- It represents the sum of squares of the distances of each data point to its assigned prototype μ_k.
- Our goal it find the values of r_{nk} and the cluster centers μ_k so as to minimize the objective J.



Iterative Algorithm

- Define iterative procedure to minimize:

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

- Given μ_k, minimize J with respect to r_{nk} (**E-step**):

Hard assignments of points to clusters.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

which simply says assign n^{th} data point $\mathbf{x}_n$ to its closest cluster center.

- Given r_{nk}, minimize J with respect to μ_k (**M-step**):

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}.$$

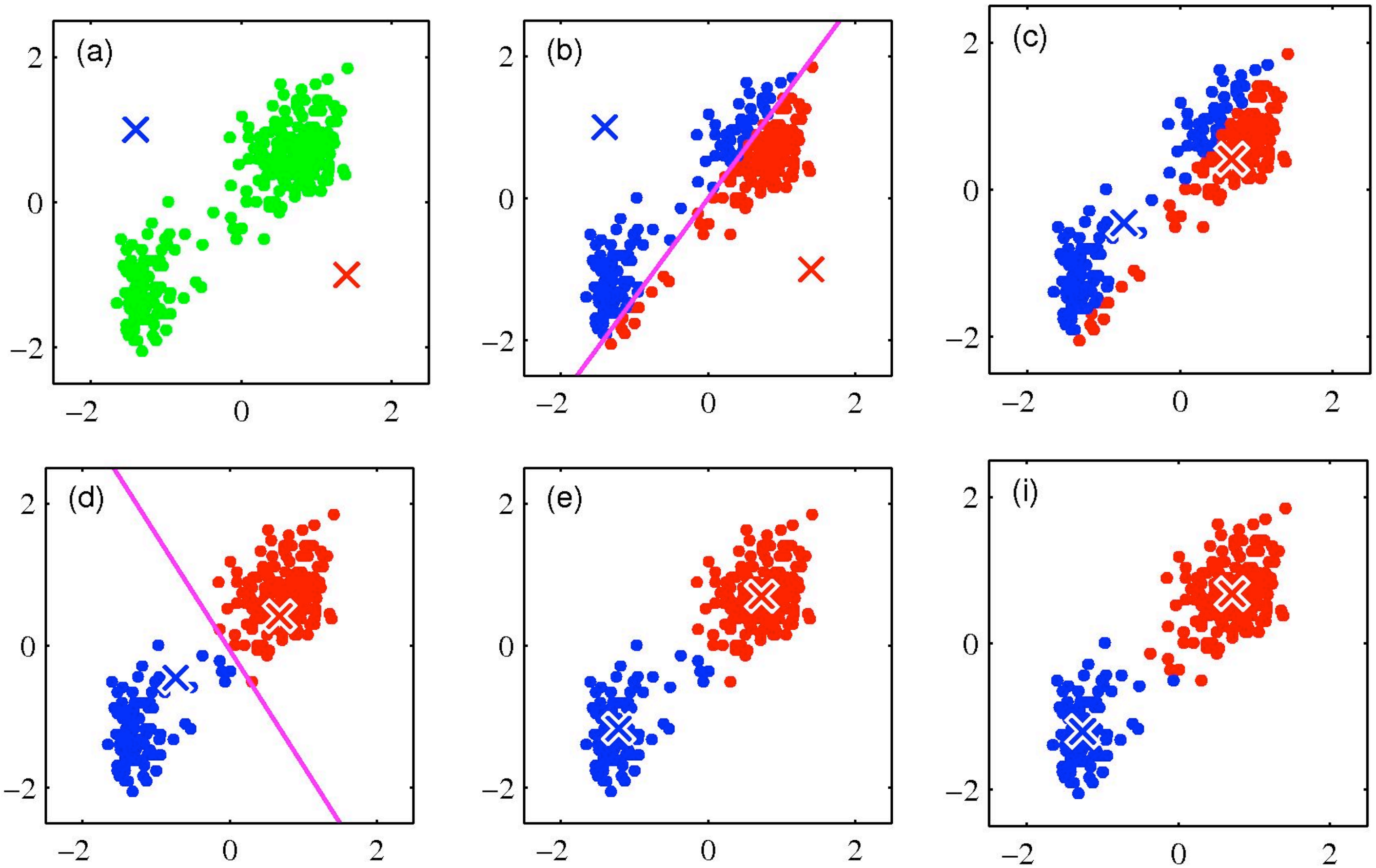
Number of points assigned to cluster k.

Set μ_k equal to the mean of all the data points assigned to cluster k.

- Guaranteed convergence to local minimum (not global minimum).

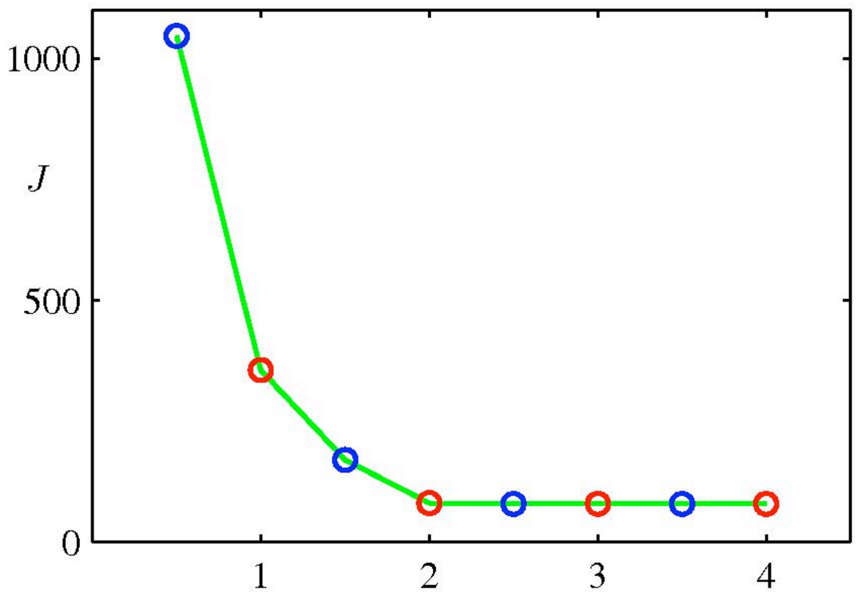
Example

- Example of using K-means (K=2) on Old Faithful dataset.



Convergence

• Plot of the cost function after each E-step (blue points) and M-step (red points)



The algorithm has converged after 3 iterations.

• K-means can be generalized by introducing a more general dissimilarity measure:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} K(\mathbf{x}_n, \boldsymbol{\mu}_k).$$

Image Segmentation

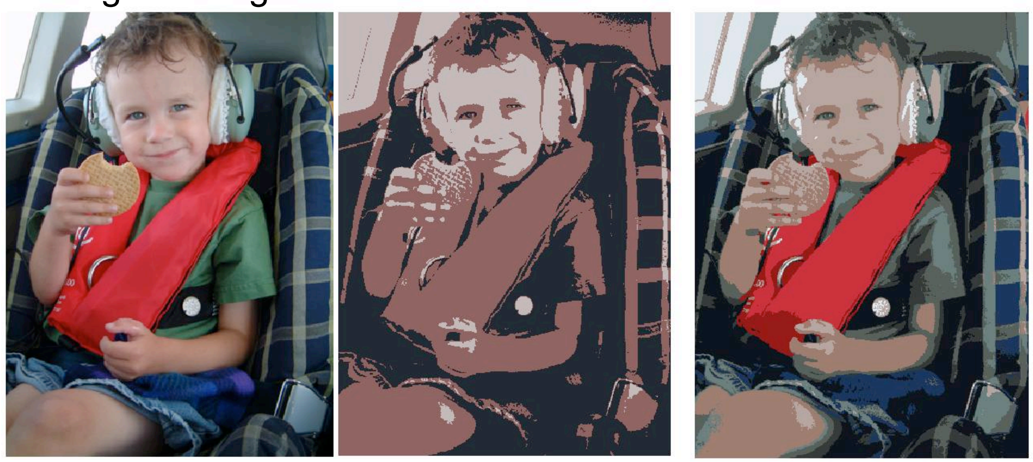
- Another application of K-means algorithm.
- Partition an image into regions corresponding, for example, to object parts.
- Each pixel in an image is a point in 3-D space, corresponding to R,G,B channels.

- For a given value of K, the algorithm represent an image using K colors.

- Another application is image compression.

Image Compression

- For each data point, we store only the identity k of the assigned cluster.
- We also store the values of the cluster centers μ_k.
- Provided K $\ll$ N, we require significantly less data.



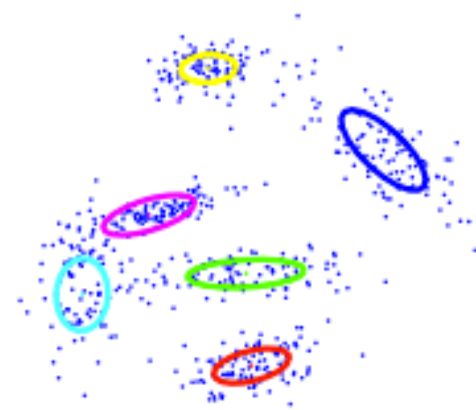
- The original image has 240 $\times$ 180 = 43,200 pixels.
- Each pixel contains {R,G,B} values, each of which requires 8 bits.

- Requires 43,200 $\times$ 24 = 1,036,800 bits to transmit directly.
- With K-means, we need to transmit K code-book vectors μ_k -- 24K bits.
- For each pixel we need to transmit $\log_2$K bits (as there are K vectors).
- Compressed image requires 43,248 (K=2), 86,472 (K=3), and 173,040 (K=10) bits, which amounts to compression rations of 4.2%, 8.3%, and 16.7%.

Mixture of Gaussians

- We will look at mixture of Gaussians in terms of discrete latent variables.
- The Gaussian mixture can be written as a linear superposition of Gaussians:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_K).$$



- Introduce K-dimensional binary random variable **z** having a 1-of-K representation:

$$z_k \in \{0, 1\}, \quad \sum_k z_k = 1.$$

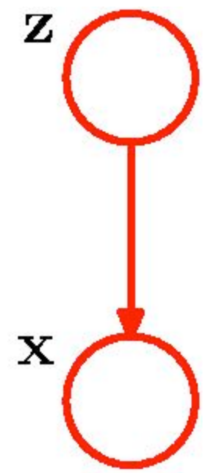
- We will specify the distribution over **z** in terms of mixing coefficients:

$$p(z_k = 1) = \pi_k, \quad 0 \leq \pi_k \leq 1, \quad \sum_k \pi_k = 1.$$

Mixture of Gaussians

- Because **z** uses 1-of-K encoding, we have:

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$



- We can now specify the conditional distribution:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \text{ or } p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

- We have therefore specified the joint distribution:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}).$$

- The marginal distribution over **x** is given by:

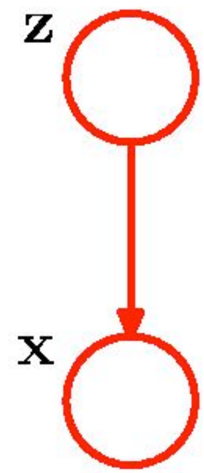
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- The marginal distribution over **x** is given by a Gaussian mixture.

Mixture of Gaussians

- The marginal distribution:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$



- If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, it follows that for every observed data point $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$.

- Let us look at the conditional p(**z**|**x**), responsibilities, which we will need for doing inference:

$$\gamma(z_k) = p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} =$$

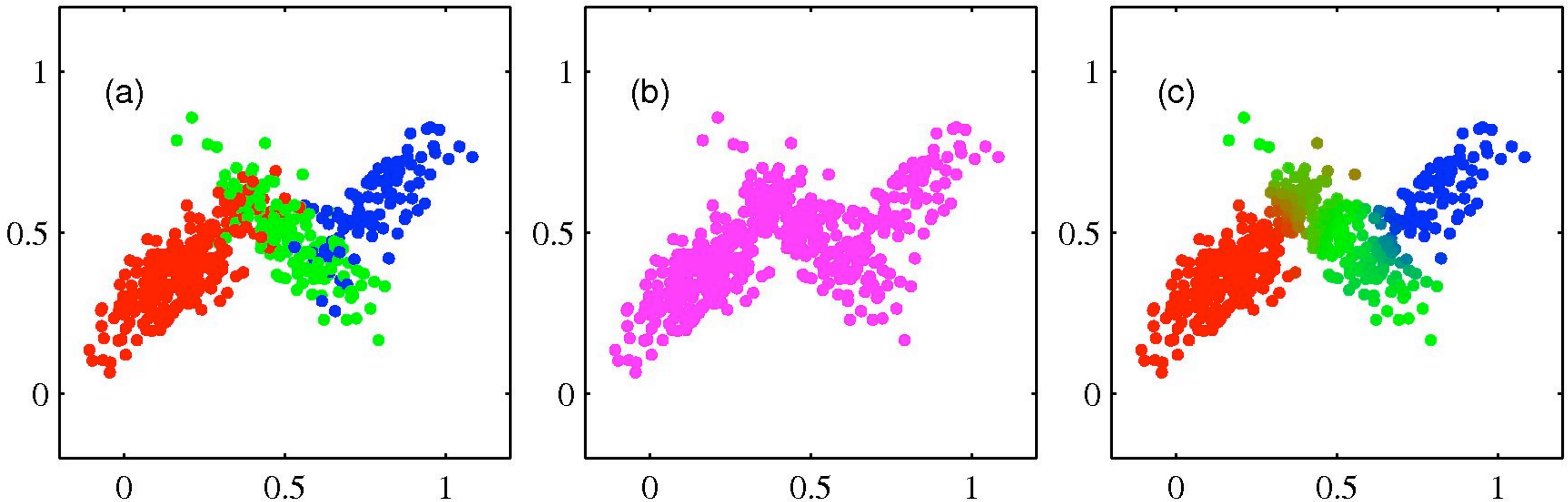
$$= \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

responsibility that component k takes for explaining the data x

- We will view π_k as prior probability that z_k=1, and $\gamma(z_k)$ is the corresponding posterior once we have observed the data.

Example

- 500 points drawn from a mixture of 3 Gaussians.



Samples from the joint distribution p(**x**,**z**).

Samples from the marginal distribution p(**x**).

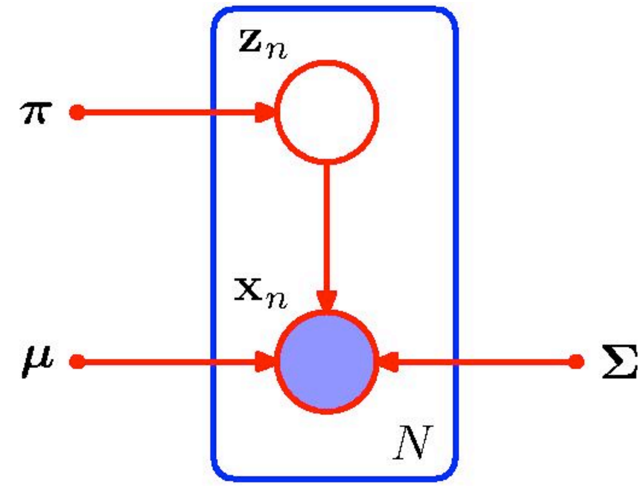
Same samples where colors represent the value of responsibilities.

Maximum Likelihood

- Suppose we observe a dataset $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, and we model the data using mixture of Gaussians.
- We represent the dataset as an N by D matrix **X**.
- The corresponding latent variables will be represented and an N by K matrix **Z**.
- The log-likelihood takes form:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Model parameters



Graphical model for a Gaussian mixture model for a set of i.i.d. data point $\{\mathbf{x}_n\}$, and corresponding latent variables $\{\mathbf{z}_n\}$.

Maximum Likelihood

- The log-likelihood:

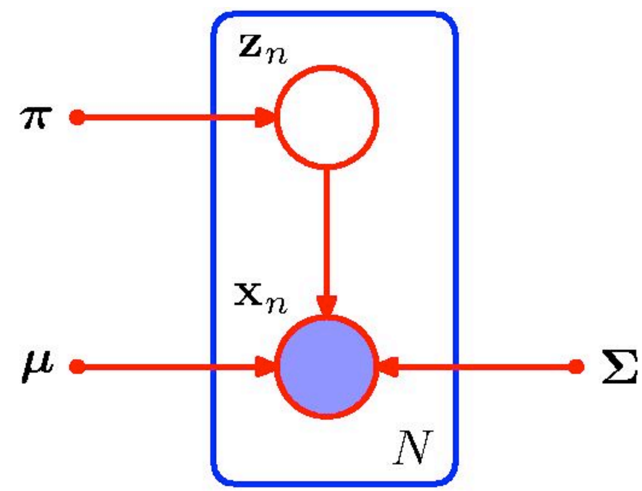
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Differentiating with respect to μ_k and setting to zero:

$$0 = \sum_n \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_K^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k).$$

Soft assignment

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) \mathbf{x}_n, \quad N_k = \sum_n \gamma(z_{nk}).$$



- We can interpret N_k as effective number of points assigned to cluster k.

- The mean μ_k is given by the mean of all the data points weighted by the posterior $\gamma(z_{nk})$ that component k was responsible for generating $\mathbf{x}_n$.

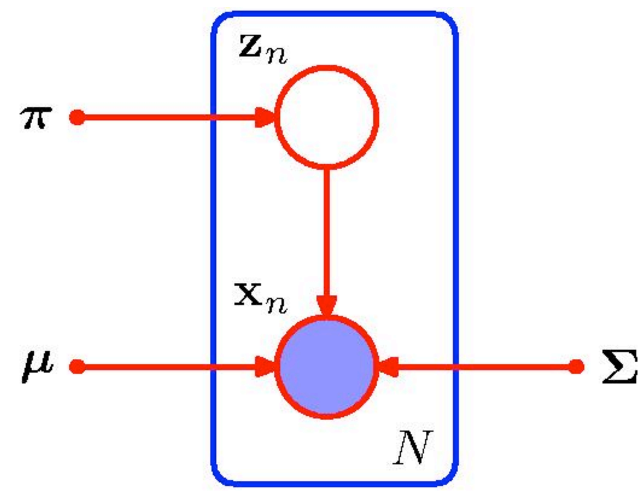
Maximum Likelihood

- The log-likelihood:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Differentiating with respect to Σ_k and setting to zero:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T.$$



- Note that the data points are weighted by the posterior probabilities.

- Maximizing log-likelihood with respect to mixing proportions:

$$\pi_k = \frac{N_k}{N}.$$

- Mixing proportion for the kth component is given by the average responsibility which that component takes for explaining the data.

Maximum Likelihood

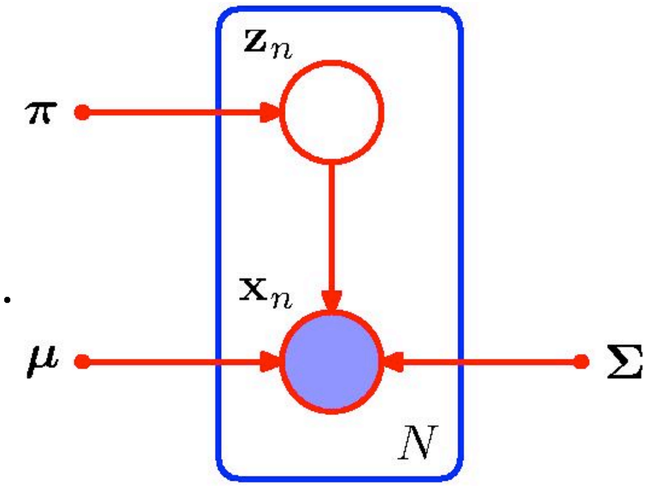
- The log-likelihood:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Note that the maximum likelihood does not have a closed form solution.

- Parameter updates depend on responsibilities $\gamma(z_{nk})$, which themselves depend on those parameters:

$$\gamma(z_{nk}) = p(z_{nk} = 1|\mathbf{x}) = \frac{\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$



- Iterative Solution:

  E-step: Update responsibilities $\gamma(z_{nk})$.
  M-step: Update model parameters π_k, μ_k, Σ_k, for k=1,…,K.

EM algorithm

- Initialize the means μ_k, covariances Σ_k, and mixing proportions π_k.
- E-step: Evaluate responsibilities using current parameter values:

$$\gamma(z_{nk}) = p(z_{nk} = 1|\mathbf{x}) = \frac{\pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- M-step: Re-estimate model parameters using the current responsibilities:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_n \gamma(z_{nk})\mathbf{x}_n, \quad N_k = \sum_n \gamma(z_{nk}),$$

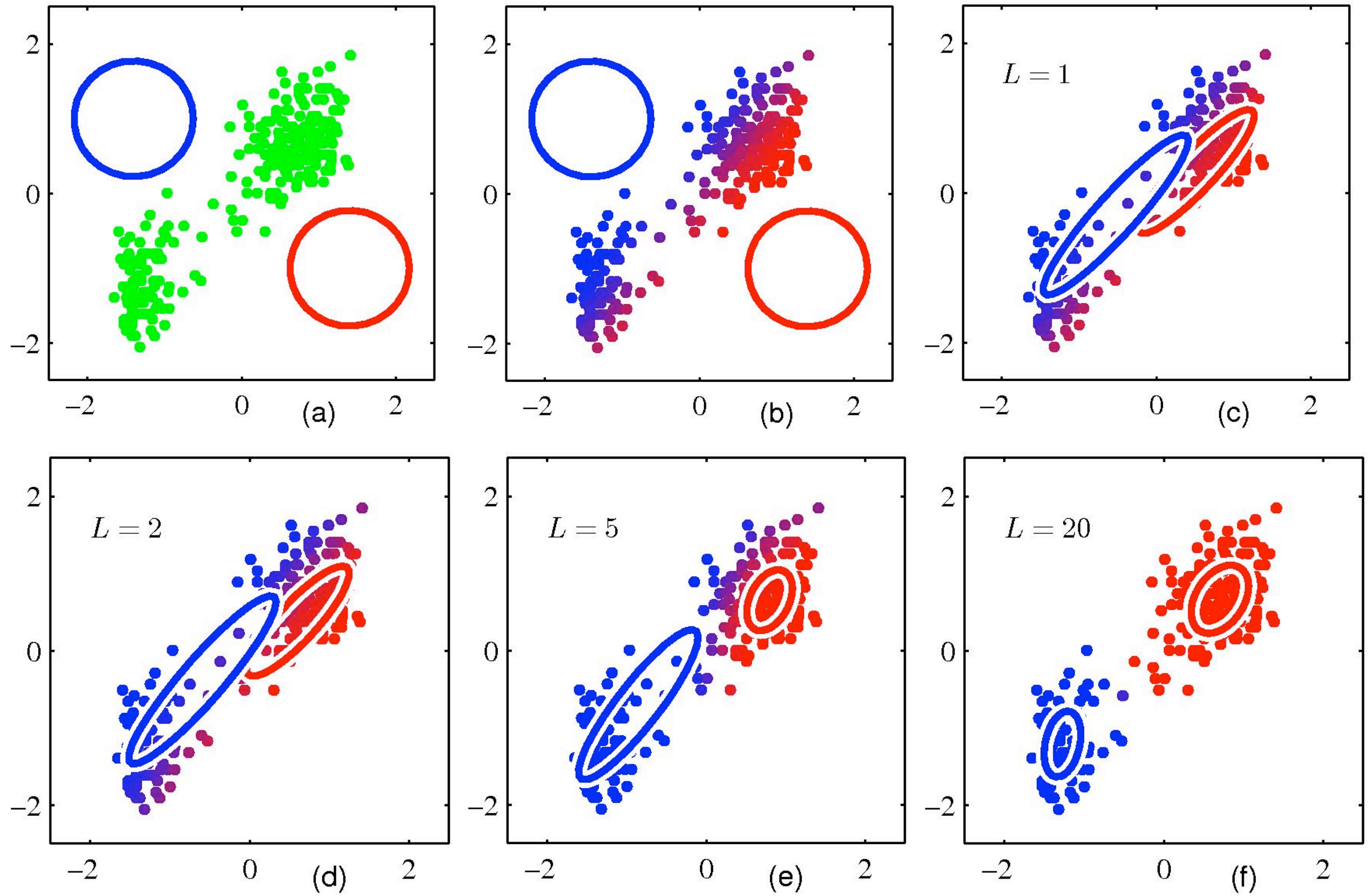
$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T,$$

$$\pi_k^{new} = \frac{N_k}{N}.$$

- Evaluate the log-likelihood and check for convergence.

Mixture of Gaussians: Example

• Illustration of the EM algorithm (much slower convergence compared to K-means)



An Alternative View of EM

• The goal of EM is to find maximum likelihood solutions for models with latent variables.

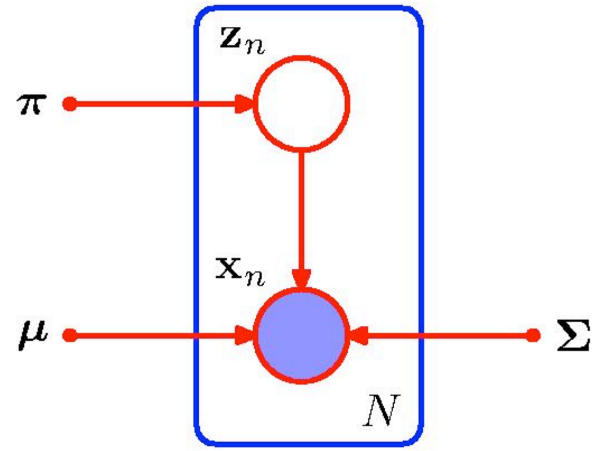
• We represent the observed dataset as an N by D matrix **X**.

• Latent variables will be represented and an N by K matrix **Z**.

• The set of all model parameters is denoted by θ.

• The log-likelihood takes form:

$$\ln p(\mathbf{X}|\theta) = \ln \left[\sum_Z p(\mathbf{X}, \mathbf{Z}|\theta) \right].$$



• Note: even if the joint distribution belongs to exponential family, the marginal typically does not!

• We will call:

$\{\mathbf{X}, \mathbf{Z}\}$ as complete dataset.

$\{\mathbf{X}\}$ as incomplete dataset.

An Alternative View of EM

- In practice, we are not given a complete dataset {**X**,**Z**}, but only incomplete dataset {**X**}.
- Our knowledge about the latent variables is given only by the posterior distribution p(**Z**|**X**,θ).
- Because we cannot use the complete data log-likelihood, we can consider expected complete-data log-likelihood:

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

- In the E-step, we use the current parameters θ^{old} to compute the posterior over the latent variables p(**Z**|**X**,θ^{old}).
- We use this posterior to compute expected complete log-likelihood.
- In the M-step, we find the revised parameter estimate θ^{new} by maximizing the expected complete log-likelihood:

$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}).$$

Tractable

The General EM algorithm

• Given a joint distribution p(**Z**,**X**|θ) over observed and latent variables governed by parameters θ, the goal is to maximize the likelihood function p(**X**|θ) with respect to θ.

• Initialize parameters θ^{old}.

• E-step: Compute posterior over latent variables: p(**Z**|**X**,θ^{old}).

• M-step: Find the new estimate of parameters θ^{new}:

$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}).$$

where

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

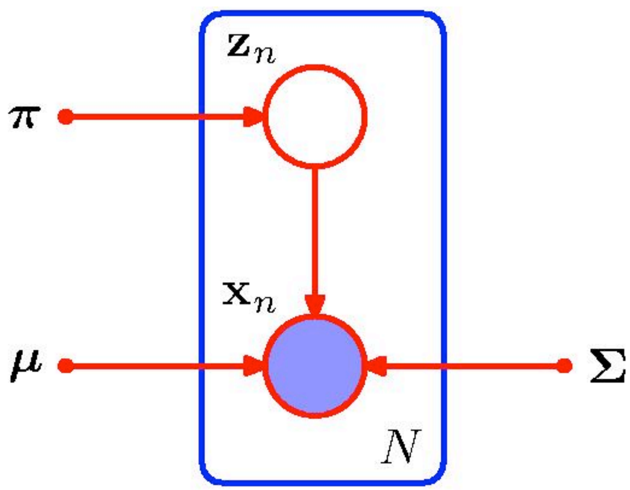
• Check for convergence of either log-likelihood or the parameter values. Otherwise:

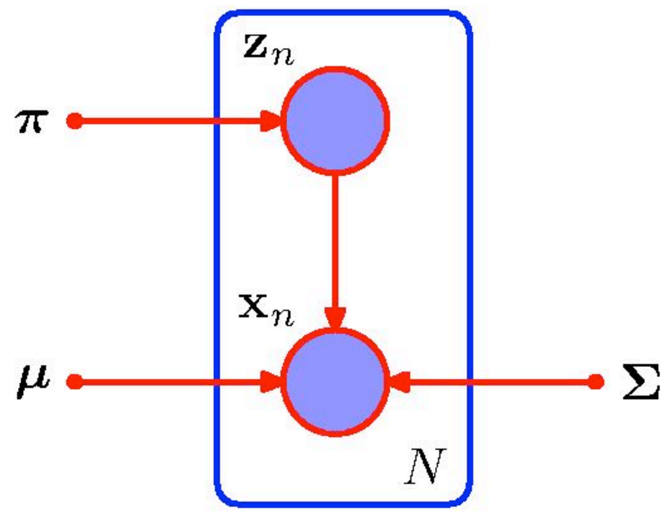
$$\theta^{new} \leftarrow \theta^{old}, \qquad \text{and iterate.}$$

Gaussian Mixtures Revisited

- We now consider the application of the latent variable view of EM the case of Gaussian mixture model.
- Recall:

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k).$$





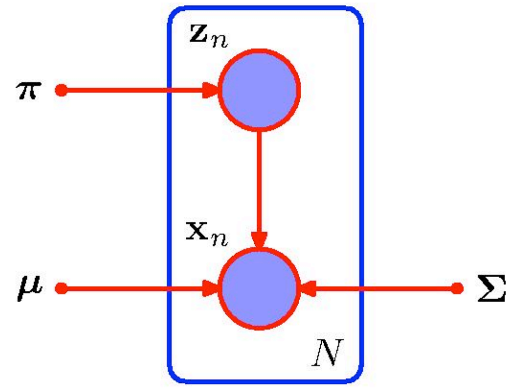
$\{\mathbf{X}\}$ -- incomplete dataset. $\{\mathbf{X},\mathbf{Z}\}$ -- complete dataset.

Maximizing Complete Data

- Consider the problem of maximizing the likelihood for the complete data:

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_{nk}}.$$

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \left[\underbrace{\sum_{n=1}^{N} z_{nk} \ln \pi_k + z_{nk} \ln \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Sum of K independent contributions, one for each mixture component.}} \right].$$



$\{\mathbf{X}, \mathbf{Z}\}$

-- complete dataset.

- Maximizing with respect to mixing proportions yields:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}.$$

- And similarly for the means and covariances.

Posterior Over Latent Variables

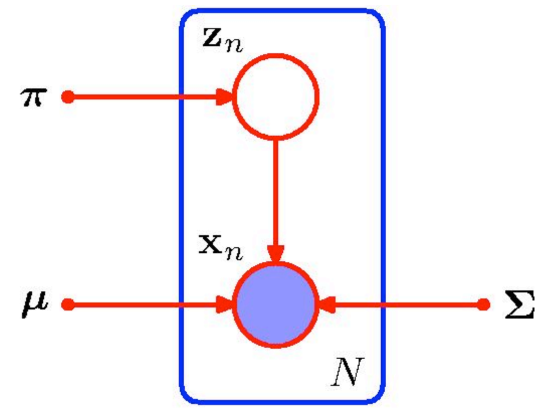
- Remember:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}, \quad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

- The posterior over latent variables takes form:

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_k}.$$

- Note that the posterior factorizes over n points, so that under the posterior distribution {$\mathbf{z}_n$} are independent.



Expected Complete Log-Likelihood

- The expected value of indicator variable z_{nk} under the posterior distribution is:

$$\mathbb{E}[z_{nk}] = \frac{\sum_{\mathbf{z}_n} z_{nk} \prod_j \left[\pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right]^{z_{nj}}}{\sum_{\mathbf{z}_n} \prod_j \left[\pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right]^{z_{nj}}}$$

$$= \frac{\pi_k N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk}).$$

- This represent the responsibility of component k for data point x_n.

- The complete-data log-likelihood:

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big].$$

- The expected complete data log-likelihood is:

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Big[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big].$$

Expected Complete Log-Likelihood

- The expected complete data log-likelihood is:

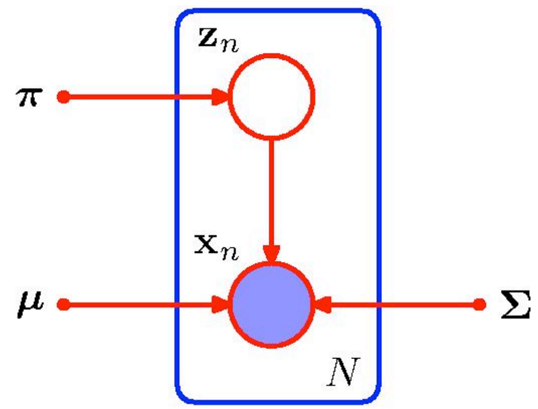
$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Big[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big].$$

- Maximizing the respect to model parameters we obtain:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_n \gamma(z_{nk}) \mathbf{x}_n, \quad N_k = \sum_n \gamma(z_{nk}),$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(y_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T,$$

$$\pi_k^{new} = \frac{N_k}{N}.$$



Relationship to K-Means

- Consider a Gaussian mixture model in which covariances are shared and are given by $\epsilon \boldsymbol{I}$.

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left[-\frac{1}{2\epsilon}||\mathbf{x} - \boldsymbol{\mu}_k||^2\right].$$

- Consider EM algorithm for a mixture of K Gaussians, in which we treat ϵ as a fixed constant. The posterior responsibilities take form:

$$\gamma(z_{nk}) = \frac{\pi_k \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_k||^2/2\epsilon)}{\sum_{j=1}^{K} \pi_j \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_j||^2/2\epsilon)}.$$

- Consider the limit $\epsilon \rightarrow 0$.
- In the denominator, the term for which $||\mathbf{x}_n - \boldsymbol{\mu}_j||^2$ is smallest will go to zero most slowly. Hence $\gamma(z_{nk}) \rightarrow r_{nk}$, where

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

Relationship to K-Means

- Consider EM algorithm for a mixture of K Gaussians, in which we treat ϵ as a fixed constant. The posterior responsibilities take form:

$$\gamma(z_{nk}) = \frac{\pi_k \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_k||^2/2\epsilon)}{\sum_{j=1}^{K} \pi_j \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_j||^2/2\epsilon)}.$$

- Finally, in the limit $\epsilon \to 0$, the expected complete log-likelihood becomes:

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 + \text{const.}$$

- Hence in the limit, maximizing the expected complete log-likelihood is equivalent to minimizing the distortion measure J for the K-means algorithm.