PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 9: MIXTURE MODELS AND EM

Learning Objectives

1、What are the differences between supervised and unsupervised learning schemes?

2、What is K-means clustering?

3、What are Gaussian Mixture Models?

4、What are Bernoulli Mixture Models?

5、What is the EM learning scheme?

6、How to understand EM from the perspective of likelihood?

7、How to generalize the EM scheme via decomposition?

Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

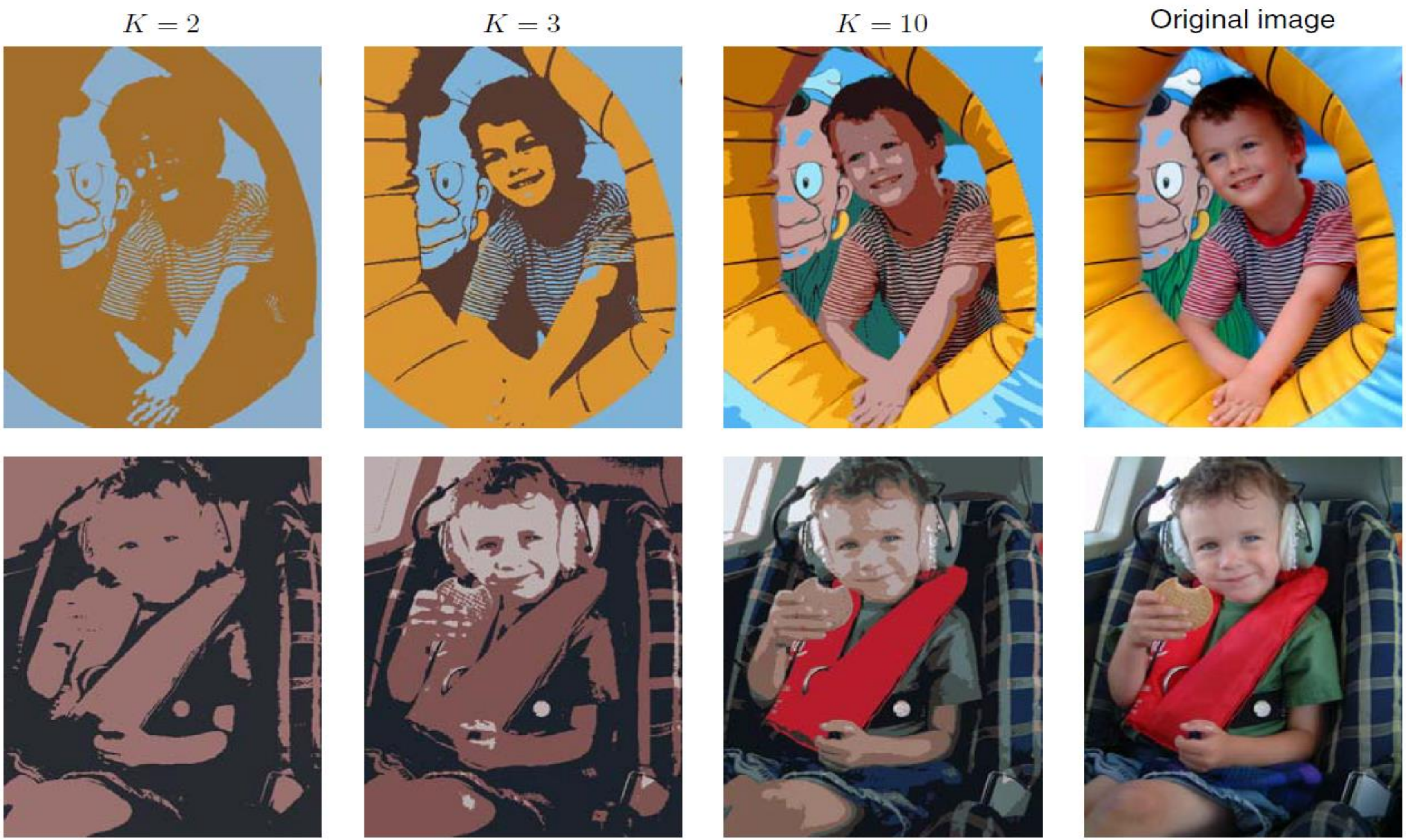
Supervised vs Unsupervised Learning

- Supervised learning
  - ✓ Training data have labels (complete data)
  - ✓ To learn the mapping between data and labels
  - ✓ Regression, classification
  - ✓ Detection, semantic/instance segmentation
  - ✓ KNNs, SVMs, decision trees, neural networks
  - ✓ Deep neural networks are good at supervised learning

Supervised vs Unsupervised Learning

- Unsupervised learning
  - ✓ Training data have no labels (incomplete data)
  - ✓ To learn the intrinsic structures of data
  - ✓ Clustering, data dimension reduction
  - ✓ Segmentation, compression
  - ✓ K-means, GMMs, PCA, ICA, NMF
  - ✓ GAN is a kind of unsupervised learning

Unsupervised Learning



Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

K-means Clustering (I)

- Problem of identifying groups, or clusters, of data points in a multidimensional space
  - ✓ Partitioning the data set into some number K of clusters
  - ✓ Cluster: a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster
  - ✓ Goal: an assignment of data points to clusters such that the sum of the squares of the distances to each data point to its closest vector (the center of the cluster) is a minimum

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

K-means Clustering (II)

- Two-stage optimization
  - ✓ In the 1st stage: minimizing $\mathcal{J}$ with respect to the r_{nk}, keeping the μ_k fixed

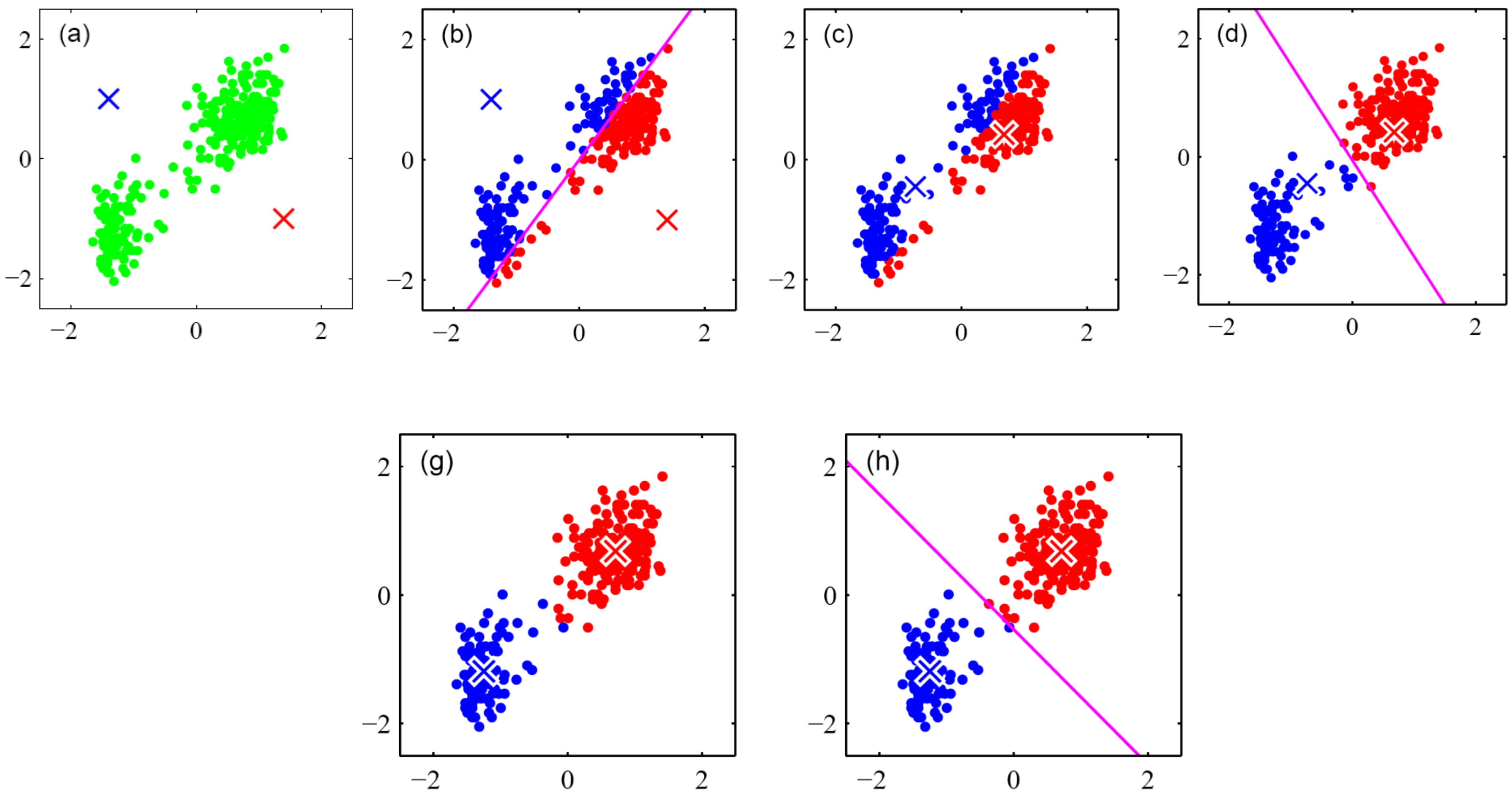
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

  - ✓ In the 2nd stage: minimizing $\mathcal{J}$ with respect to the μ_k, keeping r_{nk} fixed

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} \Leftarrow \boxed{2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0}$$

The mean of all of the data points assigned to cluster k

K-means Clustering (III)



Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

Gaussian Mixture Model (I)

- Gaussian mixture distribution can be written as a linear superposition of Gaussian

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

- random variable **z** having a 1-of-K distribution

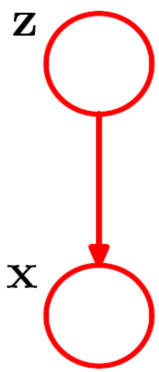
$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad \sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1 \qquad p(z_k = 1) = \pi_k$$

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

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

Gaussian Mixture Model (II)

- An equivalent formulation of the Gaussian mixture involving an explicit latent variable
  - ✓ Graphical representation of a mixture model
  - ✓ The marginal distribution of **x** is a Gaussian mixture (for every observed data point $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$, that is, the cluster label)



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

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

Gaussian Mixture Model (III)

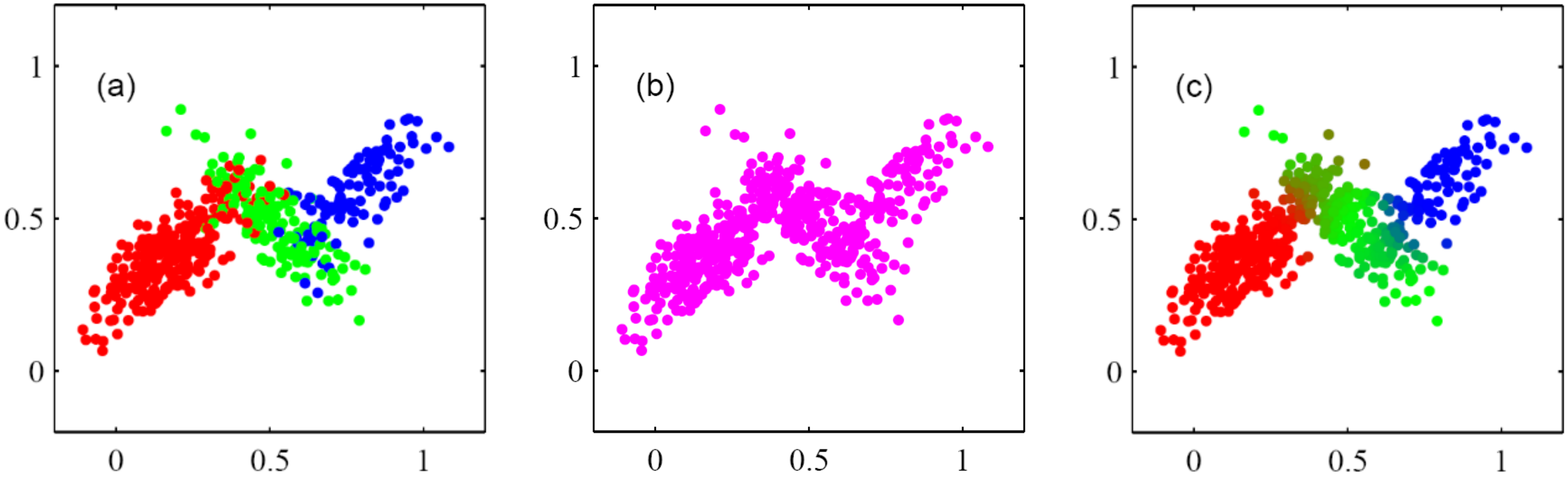
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

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

- $\gamma(z_k)$ can also be viewed as the responsibility that component *k* takes for explaining the observation **x**

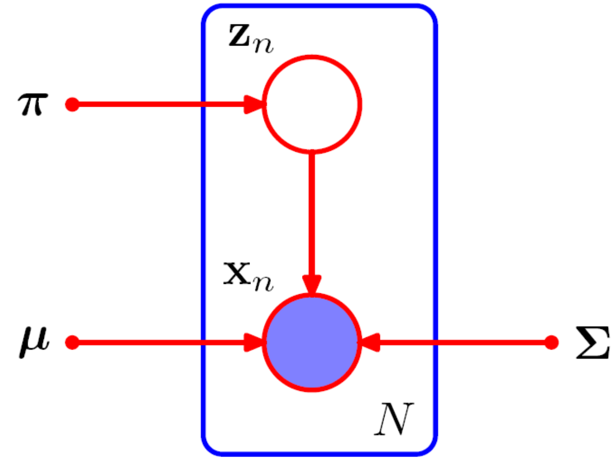
Gaussian Mixture Model (IV)

- Generating random samples distributed according to the Gaussian mixture model
  - ✓ Generating a value for **z**, which denoted as $\widehat{\mathbf{z}}$ from the marginal distribution p(**z**) and then generate a value for **x** from the conditional distribution $p(\mathbf{x}|\widehat{\mathbf{z}})$



Maximum Likelihood (I)

- Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{x_n\}$, with corresponding latent points $\{z_n\}$



- The log of the likelihood function

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

Maximum Likelihood (II)

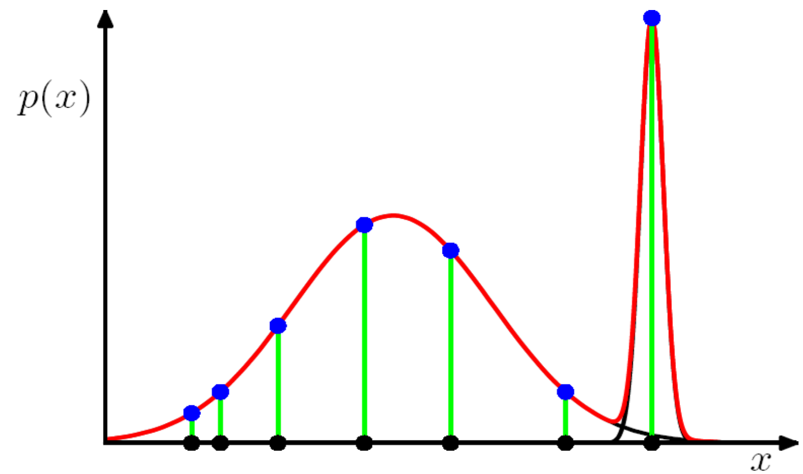
- For simplicity, consider a Gaussian mixture whose components have covariance matrices given by

$$\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$$

  - ✓ Suppose that one of the components of the mixture model has its mean μ_j exactly equal to one of the data points so that $\mu_j = \mathbf{x}_n$
  - ✓ This data point will contribute a term in the likelihood function of the form

$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

  - ✓ over-fitting problem



Maximum Likelihood (III)

- **Over-fitting** problem
  - ✓ Example of the over-fitting in a maximum likelihood approach
  - ✓ This problem does not occur in the case of Bayesian approach
  - ✓ In applying maximum likelihood to a Gaussian mixture models, there should be heuristics to seek local minima of the likelihood function that are well behaved
- **Identifiability** problem
  - ✓ A K-component mixture will have a total of K! equivalent solutions corresponding to the K! ways of assigning K sets of parameters to K components
- **Difficulty** of maximizing the log likelihood function → the presence of the summation over k that appears inside the logarithm gives **no closed form solution** as in the single case

Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

EM for Gaussian Mixtures (I)

① **Initialization**:

Initialize values for means, covariances, and mixing coefficients

② **Expectation or E step**

Using the current values for the parameters to evaluate the posterior probabilities or *responsibilities*

③ **Maximization or M step**

Using the results of ② to re-estimate the means, covariances, and mixing coefficients

- It is common to run the K-means algorithm in order to find a suitable initial values
  - ✓ The covariance matrices → the sample covariances of the clusters found by the K-means algorithm
  - ✓ Mixing coefficients → the fractions of data points assigned to the respective clusters

EM for Gaussian Mixtures (II)

- Goal: to maximize the likelihood function with respect to the parameters
  1. Initialize the means μ_k, covariance Σ_k and mixing coefficients π_k
  2. E step
  $$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$
  3. M step
  $$\begin{aligned}
  \boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \\
  \boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
  \pi_k^{\text{new}} &= \frac{N_k}{N}
  \end{aligned}$$
  4. Evaluate the log likelihood
  $$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

EM for Gaussian Mixtures (III)

- Setting the derivatives of likelihood with respect to the means of the Gaussian components to zero →

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

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

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

- Setting the derivatives of likelihood with respect to the covariance of the Gaussian components to zero →

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

  - ✓ Each data point weighted by the corresponding posterior probability
  - ✓ The denominator given by the effective # of points associated with the corresponding component

EM for Gaussian Mixtures (IV)

- Setting the derivatives of likelihood with respect to mixing coefficients to zero, subject to their sum equal to 1 →

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \quad \Longrightarrow \quad \boxed{\lambda = -N}$$

multiply π_k and sum over k

$$\Longrightarrow \quad \boxed{\pi_k = \frac{N_k}{N}}$$

EM for Gaussian Mixtures (V)