

# General EM

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 9

# Gaussian mixture model (GMM)

- Let  $\mathbf{z}$  be a random variable that denotes the clustering.
  - $\mathbf{z}$  is one-hot and  $z_k = 1$  implies choosing the  $k$ -th cluster.

$$\Downarrow \mathbf{z} = \mathbf{e}_k$$

- The marginal distribution over  $\mathbf{z}$  is given by

$$p(z_k = 1) = \pi_k$$

where the parameters satisfy

$$0 \leq \pi_k \leq 1$$

$$\sum_{k=1}^K \pi_k = 1$$

# Gaussian mixture model (GMM)

- Similar to multi-class classification, we can write

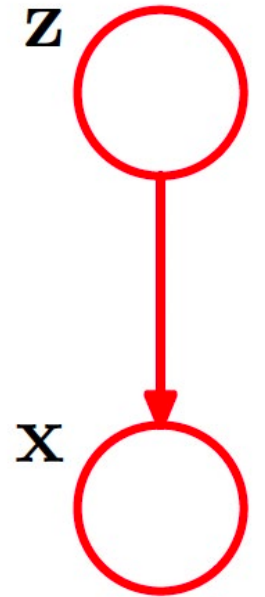
$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

- In a **Gaussian mixture model (GMM)**, the conditional distribution  $p(\mathbf{x}|\mathbf{z})$  satisfies

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- That is, each cluster is a Gaussian. We can write

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



# Gaussian mixture model (GMM)

- Therefore,

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

- By Bayes' Theorem,  $p(z_k = 1|\mathbf{x})$

“responsibility”  
that  $z_k$  takes in  
explaining  $\mathbf{x}$

$$\begin{aligned} \gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \end{aligned}$$

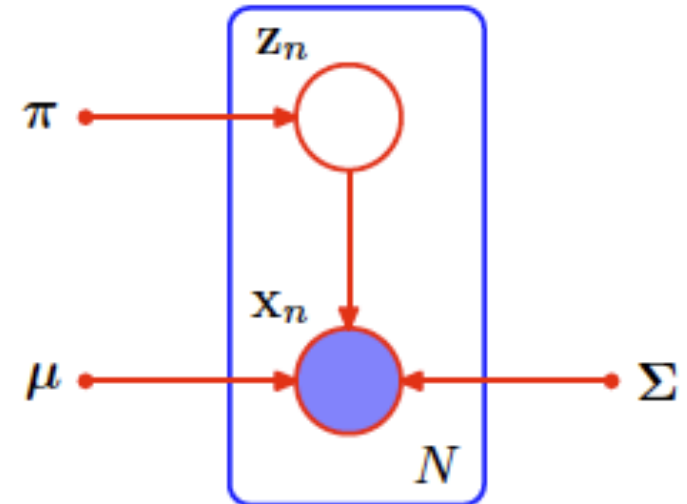
# MLE for GMM

- Question: if we are given a sample  $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ , how do we derive the **MLE** of the underlying GMM?

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

||

$\ln \prod_{n=1}^N p(\mathbf{x}_n | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$



$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^N \underbrace{\ln \left( \sum_{k=1}^K \pi_k N(x_n | \mu_k, \Sigma_k) \right)}_{f(\underbrace{\{\pi_k\}_{k=1}^K}_{\pi}, \underbrace{\{\mu_k\}_{k=1}^K}_{\mu}, \underbrace{\{\Sigma_k\}_{k=1}^K}_{\Sigma})}$$

$$\frac{\partial f}{\partial \mu_k} = \sum_{n=1}^N \frac{\pi_k \frac{\partial}{\partial \mu_k} N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)}$$

$$\begin{aligned} & \frac{\partial}{\partial \mu_k} N(x_n | \mu_k, \Sigma_k) \\ &= \frac{\partial}{\partial \mu_k} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right) \\ &= N(x_n | \mu_k, \Sigma_k) \cdot \frac{\partial}{\partial \mu_k} \left( -\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right) \\ &= N(x_n | \mu_k, \Sigma_k) \Sigma_k^{-1} (x_n - \mu_k) \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial f}{\partial \mu_k} &= \sum_{n=1}^N \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)} \Sigma_k^{-1} (x_n - \mu_k) \\ &= \sum_{n=1}^N \gamma(z_{nk}) \Sigma_k^{-1} (x_n - \mu_k) \end{aligned}$$

Setting  $\frac{\partial f}{\partial \mu_k} = 0$  yields

$$\sum_{n=1}^N \gamma(z_{nk}) \cancel{\Sigma_k^{-1}} (x_n - \mu_k) = 0$$

That is,  $\left[ \sum_{n=1}^N \gamma(z_{nk}) \right] \mu_k = \sum_{n=1}^N \gamma(z_{nk}) x_n$

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

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

Next,

$$\frac{\partial f}{\partial \Sigma_k} = \frac{\sum_{n=1}^N \pi_k \frac{\partial}{\partial \Sigma_k} N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)}$$

$$\frac{\partial}{\partial \Sigma_k} N(x_n | \mu_k, \Sigma_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \left[ \frac{1}{(2\pi)^{\frac{p}{2}}} \cdot \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right) \right]$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}} \left[ \left( \frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \right) \cdot \exp(\dots) + \right.$$

$$\left. \left(-\frac{1}{2}\right) \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \cdot \exp(\dots) \left( \frac{\partial}{\partial \Sigma_k} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right) \right]$$

$$\frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}}$$

$$= - \frac{1}{2|\Sigma_k|^{\frac{3}{2}}} \frac{\partial}{\partial \Sigma_k} |\Sigma_k|$$

$$= - \frac{1}{2|\Sigma_k|^{\frac{3}{2}}} |\Sigma_k| (\Sigma_k^{-1})^T \quad \left( \begin{array}{l} \text{by HW1, Problem \#4;} \\ \text{or [B:] Appendix C} \\ \text{Eq (C.22)} \end{array} \right)$$

$$= - \frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \Sigma_k^{-1}$$

$$\frac{\partial}{\partial \Sigma_k} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \text{tr} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \text{tr} (\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T)$$

$$= - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}$$

$$\left( \begin{array}{l} \text{by e.g.3 in the} \\ \text{previous lecture} \\ \frac{\partial}{\partial A} \text{tr}(A^{-1}B) \\ = - (A^{-1}B A^{-1})^T \end{array} \right)$$



Therefore,

$$\begin{aligned}
 & \frac{\partial}{\partial \Sigma_k} N(x_n | \mu_k, \Sigma_k) \\
 &= \frac{1}{(2\pi)^{\frac{D}{2}}} \left[ \left( -\frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \Sigma_k^{-1} \right) \cdot \exp(\dots) + \right. \\
 & \quad \left. \left( -\frac{1}{2} \right) \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp(\dots) \left( -\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) \right] \\
 &= -\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \exp(\dots) \Sigma_k^{-1} \cdot \\
 & \quad \left( I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) \\
 &= -\frac{1}{2} N(x_n | \mu_k, \Sigma_k) \Sigma_k^{-1} \left( I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)
 \end{aligned}$$

and

$$\begin{aligned}
 \frac{\partial f}{\partial \Sigma_k} &= \sum_{n=1}^N \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)} \left( -\frac{1}{2} \right) \cdot \\
 & \quad \Sigma_k^{-1} \left( I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) \\
 &= \sum_{n=1}^N \gamma(z_{nk}) \left( -\frac{1}{2} \right) \Sigma_k^{-1} \left( I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)
 \end{aligned}$$

Setting this to zero yields

$$\sum_{n=1}^N \gamma(z_{nk}) \left( -\frac{1}{\Sigma_k} \right) \Sigma_k^{-1} (I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}) = 0$$

Right-multiplying with  $\Sigma_k$ , we have

$$\left( \sum_{n=1}^N \gamma(z_{nk}) \right) \Sigma_k = \sum_{n=1}^N \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$$

Next, for  $\pi$ , we need to

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

Setting

$$\frac{\partial g}{\partial \pi_k} = \sum_{n=1}^N \frac{N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)} + \lambda = 0$$

yields

$$\sum_{n=1}^N \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | \mu_j, \Sigma_j)} = -\pi_k \lambda$$

Summing over  $k$ , we have

$$\sum_{n=1}^N 1 = -\lambda$$

Therefore,

$$\lambda = -N$$

Hence,

$$\pi_k = \frac{\sum_{n=1}^N \delta(z_{nk})}{-\lambda} = \frac{N_k}{N}$$

# MLE for GMM

- Note that it is **not a closed-form solution** that

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

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

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

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

$$\underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)}}_{\gamma(z_{nk})}$$

# MLE for GMM

- Note that it is **not a closed-form solution** that

The diagram illustrates the relationship between the parameters of a Gaussian Mixture Model (GMM) and the Expectation-Maximization (EM) algorithm's E-step. The parameters are defined as:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$
$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$
$$\pi_k = \frac{N_k}{N}$$
$$N_k = \sum_{n=1}^N \gamma(z_{nk})$$

On the right, the E-step formula is shown, with parameters highlighted in yellow:

$$\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)}$$

A red bracket under the denominator indicates that the entire denominator represents the responsibility  $\gamma(z_{nk})$ . Blue arrows show the flow of information from the E-step output back to the parameter equations: from the denominator to  $N_k$ , from the numerator to  $\mu_k$  and  $\Sigma_k$ , and from the entire fraction to  $\pi_k$ . Red arrows show the forward flow of parameters into the E-step: from  $\pi_k$  to the numerator, from  $\mu_k$  and  $\Sigma_k$  to the Gaussian term in the numerator, and from  $N_k$  to the denominator.

# EM for GMM

(E-step): expectation

for fixed parameters, find the responsibilities  $\gamma(z_{nk})$

$$\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

~~for fixed  $\mu_1, \dots, \mu_K$ , find  $\{r_{nk}\}$  that minimize  $J$~~

(M-step): maximization

for fixed responsibilities, find the corresponding

~~for fixed  $\{r_{nk}\}$ , find  $\mu_1, \dots, \mu_K$  that minimize  $J$~~

$\left\{ \begin{array}{l} \mu_k \\ \Sigma_k \\ \pi_k \end{array} \right.$

**general EM**

# complete dataset with latent variables

- Suppose  $\mathbf{X}$  is the data matrix, and  $\mathbf{Z}$  the corresponding latent variables (assumed to be discrete). Then

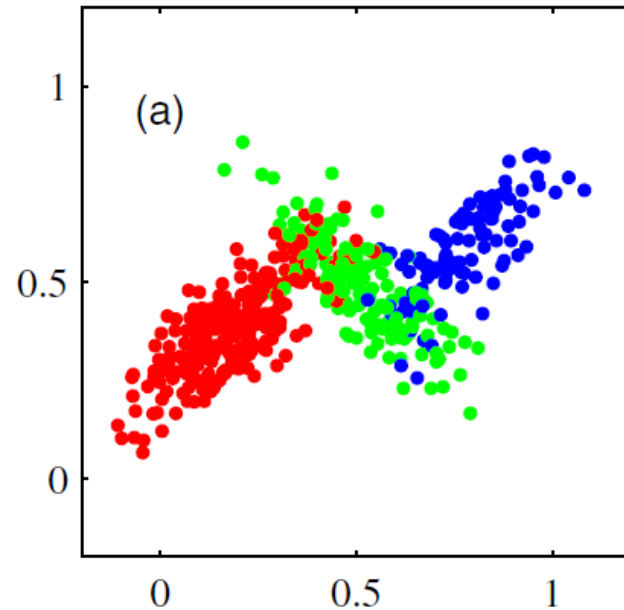
$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

- $\{\mathbf{X}, \mathbf{Z}\}$  is called the **complete** data set;  $\mathbf{X}$  is **incomplete**
- In practice, we are not given the complete data set; the only way we estimate  $\mathbf{Z}$  is by the **posterior**

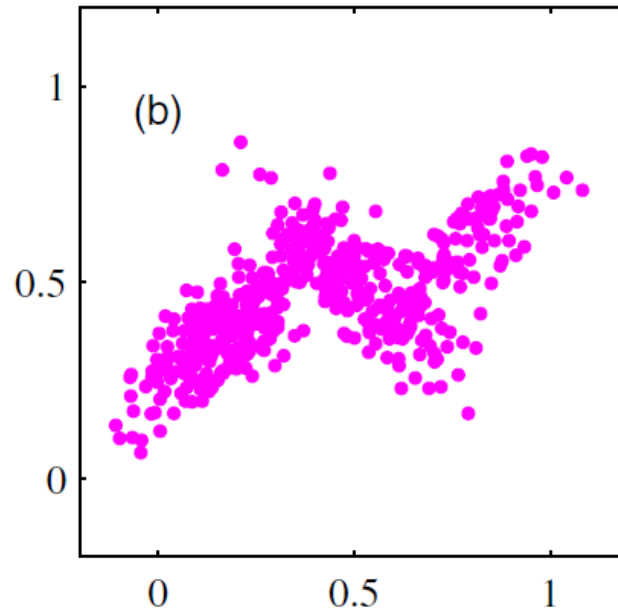
$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$



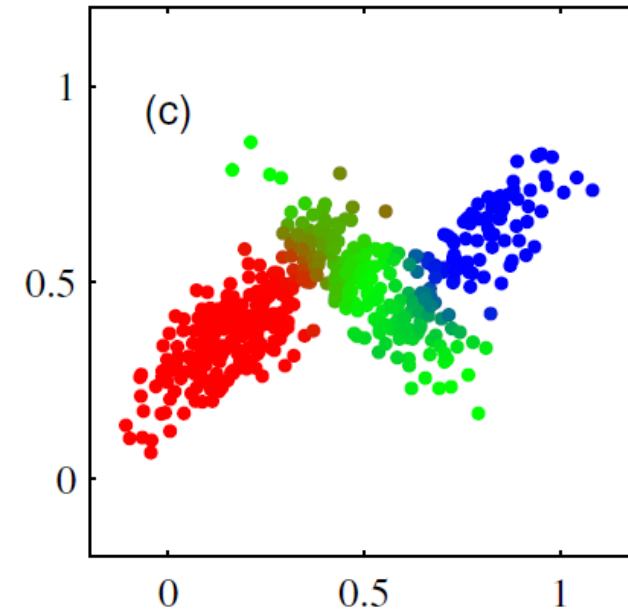
# complete dataset with latent variables



complete data set  
with both  $\{X, Z\}$



incomplete data  
set with only  $X$



$X$  with the  
posterior  $p(Z|X, \theta)$

# general EM

(E-step): expectation

1. for fixed parameters, find  $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
2. calculate the expectation  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

(M-step): maximization

solve for  $\boldsymbol{\theta}^{\text{new}} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$



# Questions?

---

## *Reference*

- *K-means*:
  - *[Al]* Ch.7.3
  - *[HaTF]* Ch.13.2.1
  - *[Bi]* Ch.9.1
- *EM*:
  - *[Al]* Ch.7.2, 7.4
  - *[HaTF]* Ch.13.2.3
  - *[Bi]* Ch.9.2-9.4

