Gaussian Mixture Models and the EM Algorithm

Machine Learning (CE 40717) — Spring 2025

Ali Sharifi-Zarchi

CE Department
Sharif University of Technology

November 13, 2025

Mixture Models

$$\mathbb{P}(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j \mathbb{P}(\mathbf{x} \mid z = j; \boldsymbol{\theta}_j), \qquad 0 \le \pi_j \le 1, \ \sum_j \pi_j = 1$$

Gaussian Mixture Model (GMM):

$$\mathbb{P}(\mathbf{x}) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

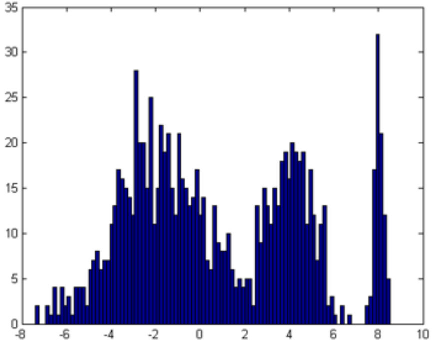
where

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

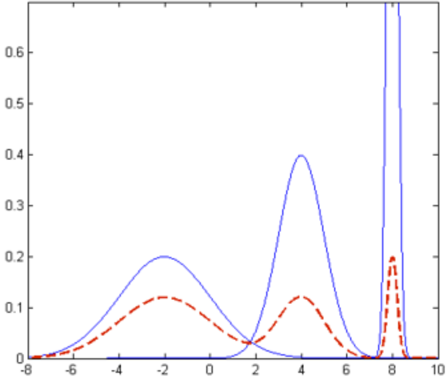
Why GMMs?

- Can model multi-modal densities beyond k-means.
- Provide a probabilistic generative framework.
- Parameters: $\boldsymbol{\theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{K}$.
- Maximum likelihood has no closed-form $\rightarrow$ use EM.

1-D GMM Examples

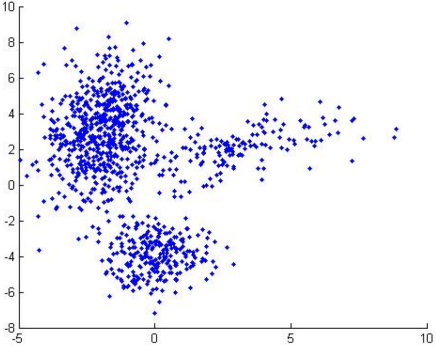


(1)

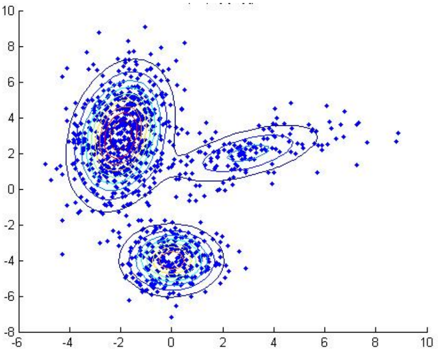


(2)

2-D GMM Examples

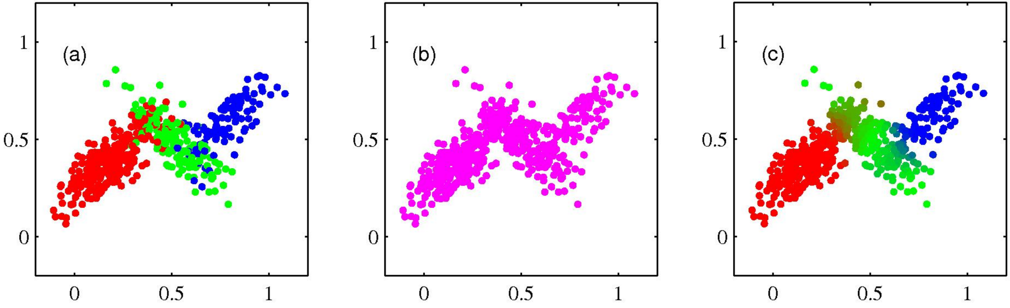


(1)



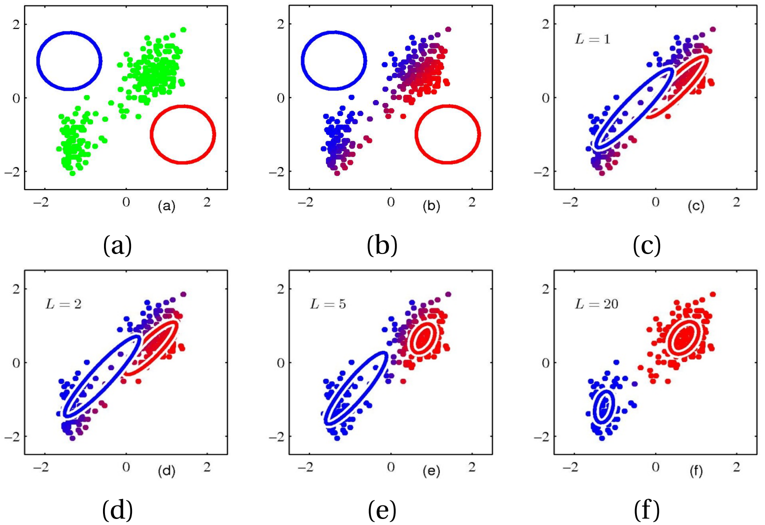
(2)

EM & GMM Examples (Set 1)

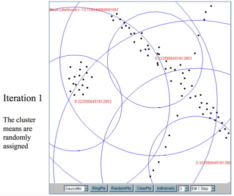


(a) (b) (c)

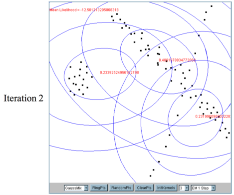
EM & GMM Examples (Set 2)



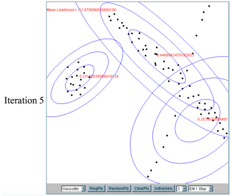
EM & GMM Iterations



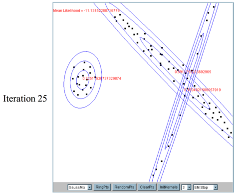
Iter 1



Iter 2



Iter 5



Iter 25

Maximum Likelihood for GMM

Given data $\mathscr{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$:

$$\ell(\boldsymbol{\theta}) = \ln \mathbb{P}(\mathscr{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln\left(\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right).$$

No closed form due to the inner summation inside the log.

Complete-data Likelihood

Introduce latent variables $Z = \{z^{(i)}\}$:

$$\mathbb{P}(\mathscr{X}, Z \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left[\pi_j \mathcal{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right]^{z_j^{(i)}}.$$

Thus:

$$\ln \mathbb{P}(\mathscr{X}, Z \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_j^{(i)} \left(\ln \pi_j + \ln \mathcal{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right).$$

E-step: Responsibilities

$$\gamma_j^{(i)} = \mathbb{P}(z^{(i)} = j \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}) = \frac{\pi_j^{(t)} \mathcal{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}.$$

Interpretation: soft assignment of each $\mathbf{x}^{(i)}$ to component j.

M-step: Maximization of Expected Log-likelihood

Define expected complete log-likelihood:

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{Z|\mathscr{X},\boldsymbol{\theta}^{(t)}}[\ln \mathbb{P}(\mathscr{X}, Z \mid \boldsymbol{\theta})].$$

Substitute expectations:

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{K}\gamma_j^{(i)}\Big(\ln \pi_j + \ln \mathscr{N}(\mathbf{x}^{(i)} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\Big).$$

Update for $\boldsymbol{\mu}_j$

Maximize Q w.r.t. $\boldsymbol{\mu}_j$:

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_j} = \sum_{i=1}^{N} \gamma_j^{(i)} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_j) = 0$$

giving:

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_j^{(i)} \mathbf{x}^{(i)}, \qquad N_j = \sum_{i=1}^{N} \gamma_j^{(i)}.$$

Update for Σ_j

$$\boldsymbol{\Sigma}_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^{N} \gamma_j^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_j^{\text{new}})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_j^{\text{new}})^\top .$$

(In practice add small ϵI for numerical stability.)

Update for π_j

Maximize under $\sum_j \pi_j = 1$:

$$\pi_j^{\text{new}} = \frac{N_j}{N}.$$

Using Lagrange multiplier ensures normalization.

Algorithm Summary

Algorithm 1 EM for Gaussian Mixture Models

1: Initialize $\{\pi_j^{(0)}, \boldsymbol{\mu}_j^{(0)}, \boldsymbol{\Sigma}_j^{(0)}\}$
2: **repeat**
3:     **E-step:** Compute $\gamma_j^{(i)}$
4:     **M-step:** Update parameters as above
5: **until** log-likelihood convergence

Variational View of EM

$$\ln \mathbb{P}(\mathscr{X} \mid \boldsymbol{\theta}) = F(\boldsymbol{\theta}, Q) + \mathrm{KL}(Q(Z) \,\|\, \mathbb{P}(Z \mid \mathscr{X}, \boldsymbol{\theta})),$$

where

$$F(\boldsymbol{\theta}, Q) = \sum_{Z} Q(Z) \ln \frac{\mathbb{P}(\mathscr{X}, Z \mid \boldsymbol{\theta})}{Q(Z)}.$$

EM maximizes F alternately over Q and $\boldsymbol{\theta}$.

EM Monotonicity

$$\ln \mathbb{P}(\mathscr{X} \mid \boldsymbol{\theta}^{(t+1)}) \geq F(\boldsymbol{\theta}^{(t+1)}, Q) \geq F(\boldsymbol{\theta}^{(t)}, Q) = \ln \mathbb{P}(\mathscr{X} \mid \boldsymbol{\theta}^{(t)}).$$

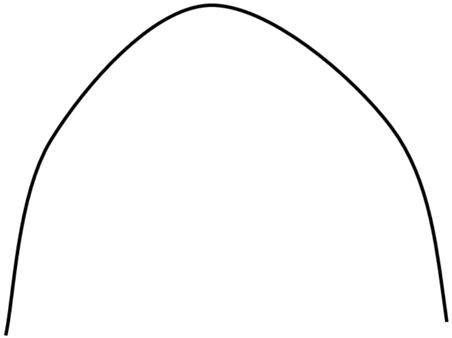
Thus the likelihood never decreases.

Practical Notes

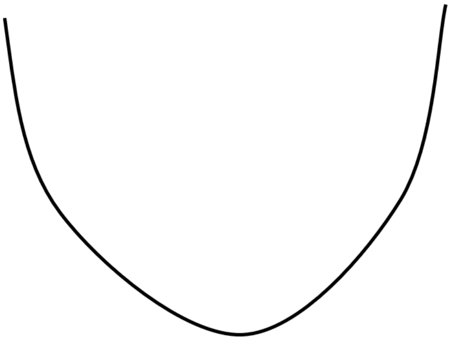
- Initialization via k-means.
- Regularize $\boldsymbol{\Sigma}_j \leftarrow \boldsymbol{\Sigma}_j + \epsilon I$.
- Convergence: monitor $\Delta\ell < 10^{-6}$.

Convergence and Local Optima

- EM converges to a stationary (local) optimum.
- Run multiple initializations.



(Left) Concave function



(Right) Convex function

k-means vs EM+GMM

- k-means: hard assignments, equal variance clusters.
- EM: soft probabilistic assignments.
- EM generalizes k-means when covariances are isotropic.

Summary

- GMM = flexible probabilistic mixture model.
- EM = iterative ML method with guaranteed non-decrease.
- Use careful initialization and covariance regularization.

References

- C. M. Bishop, *Pattern Recognition and Machine Learning*, Ch. 9.
- Lecture slides: Hamid R. Rabiee & Zahra Dehghanian, Spring 2025.

Contributions

- **This slide deck was prepared thanks to:**
  - Soheil Sayah Varg