

# Gaussian Mixture Models and the EM Algorithm

## Machine Learning (CE 40717) — Spring 2025

Ali Sharifi-Zarchi

CE Department  
Sharif University of Technology

November 18, 2025



- ① Introduction & Motivation
- ② Mixture Models
- ③ Gaussian Mixture Models (GMM)
- ④ Maximum Likelihood and the EM Algorithm
- ⑤ Convexity, Concavity, and EM Lower Bound
- ⑥ Examples and Visualizations
- ⑦ Convergence, Practical Issues, and Failure Modes
- ⑧ Summary and References

- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References

# Motivation

- Many real-world densities are **multi-modal**. These arise from heterogeneous sources, sub-populations, or hidden structure.
- We want a **flexible probabilistic model** to:
  - model complex densities,
  - perform soft clustering,
  - generate new samples,
  - evaluate likelihood.
- Gaussian Mixture Models (GMMs) achieve this with a simple generative structure.

## Outline (full version)

- Mixture models and Gaussian mixture definition
- Maximum Likelihood for incomplete-data likelihood
- EM Algorithm:
  - E-step: responsibility computation
  - M-step: closed-form updates
  - full derivations (means, covariances, mixing weights)
- Variational lower bound, KL divergence
- Convex vs. concave functions (full explanation)
- Full proof of EM monotonicity
- Examples, figures, EM iterations
- Practical notes and failure modes

- 1 Introduction & Motivation
- 2 Mixture Models**
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References

# Multivariate Gaussian Distribution

A **multivariate Gaussian** (normal) distribution over  $\mathbf{x} \in \mathbb{R}^d$  is defined by a mean vector  $\boldsymbol{\mu} \in \mathbb{R}^d$  and a positive definite covariance matrix  $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ :

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

**Density function:**

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

**Key properties:**

- Elliptical level sets defined by the quadratic form  $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ .
- $\boldsymbol{\Sigma}$  encodes variances and pairwise correlations.
- Fully characterized by first and second moments.

This distribution will serve as the building block for Gaussian mixture models.

## Variance–Covariance Matrix

For a multivariate Gaussian  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \cdots \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

encodes how each dimension varies and how pairs of dimensions interact.

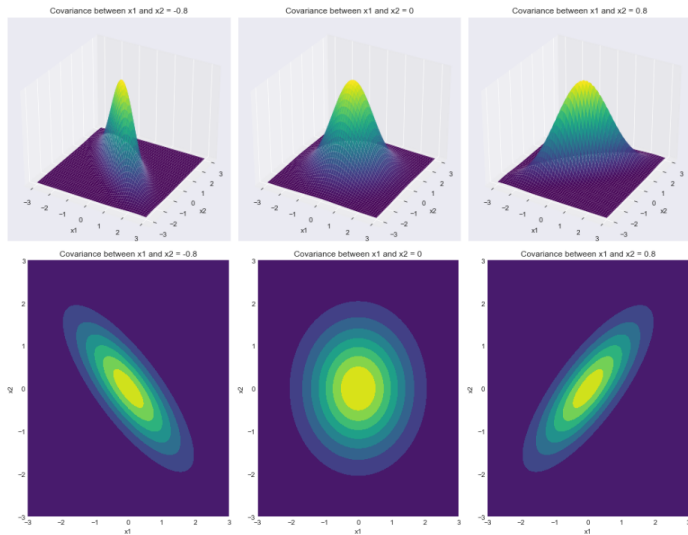
### Interpretation of entries:

- Diagonal entries:  $\Sigma_{ii} = \text{Var}(x_i)$  — how spread out the distribution is along dimension  $i$ .
- Off-diagonal entries:  $\Sigma_{ij} = \text{Cov}(x_i, x_j)$  — how dimensions  $i$  and  $j$  move together.
- Positive covariance:  $x_i$  increases when  $x_j$  increases  $\Rightarrow$  elongated ellipse along the diagonal direction.
- Negative covariance:  $x_i$  increases when  $x_j$  decreases  $\Rightarrow$  ellipse tilted in the opposite direction.





# Multivariate Gaussian Distribution



# What is a Mixture Model?

A mixture model represents a probability density as a convex combination of simpler component densities:

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

where:

- $\mathbf{x} \in \mathbb{R}^d$  is a **feature vector**,
- $z$  is a latent discrete variable,
- $\pi_j \geq 0, \sum_j \pi_j = 1$ .

The latent variable  $z$  indicates which component generated  $\mathbf{x}$ .

# Gaussian Mixture Model (GMM)

$$\mathbb{P}(\mathbf{x}) = \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma_j)$$

The multivariate Gaussian density:

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

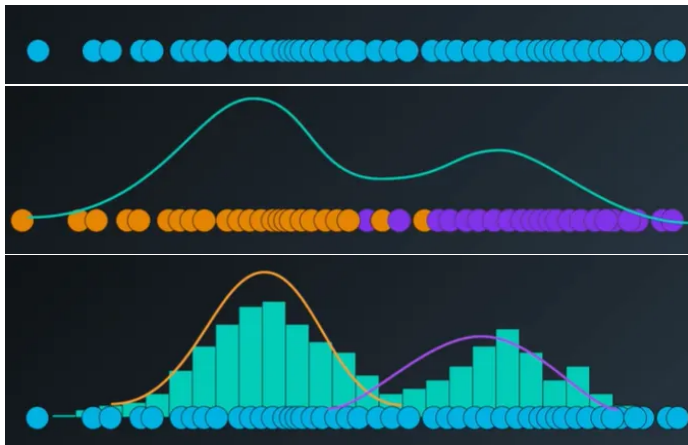
Parameters:

$$\boldsymbol{\theta} = \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K.$$

## Notation (vector-corrected)

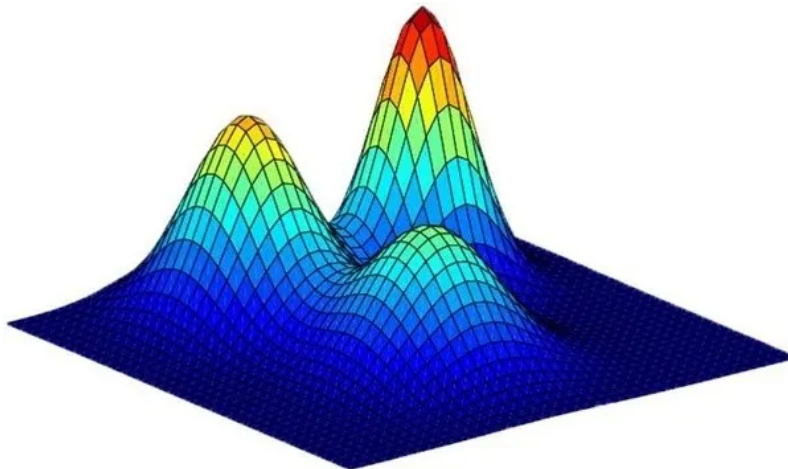
- $\mathbf{x}^{(i)}$ : feature vector for sample  $i$ .
- $x_t^{(i)}$ : scalar component  $t$  of  $\mathbf{x}^{(i)}$ .
- $z_j^{(i)}$ : latent component label (categorical). One-hot form:  $z_j^{(i)} \in \{0, 1\}$ .
- $\pi_j$ : mixing weight (scalar).
- $\mu_j$ : mean vector of component  $j$ .
- $\Sigma_j$ : covariance matrix of component  $j$ .
- $N$ : number of samples;  $d$ : dimension.

# GMM



Source: Tilak Mudgal, "Gaussian Mixture Modeling (GMM)," Medium.

# Multivariate GMM



Source: Tilak Mudgal, "Gaussian Mixture Modeling (GMM)," Medium.

- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)**
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References



# Why GMMs?

- Can approximate arbitrary continuous densities (universal approximators).
- Soft clustering via posterior probabilities.
- Generative: can sample new data.
- More expressive than k-means due to covariance matrices.

- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm**
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References

# Incomplete-data Log-Likelihood

Given data  $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ , the likelihood is:

$$\mathbb{P}(\mathcal{X} | \boldsymbol{\theta}) = \prod_{i=1}^N \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(i)} | \mu_j, \Sigma_j).$$

Hence the incomplete-data log-likelihood is:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \ln \left( \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(i)} | \mu_j, \Sigma_j) \right).$$

- The main difficulty:  $\ln(\sum_j \cdot)$  makes direct maximization impossible.
- EM solves this by introducing latent variables.

## Latent Variables and Complete-data Likelihood

Let  $z_j^{(i)} \in \{0, 1\}$  indicate which component generated  $\mathbf{x}^{(i)}$ :

$$\sum_{j=1}^K z_j^{(i)} = 1.$$

The complete-data likelihood factorizes:

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

Complete-data log-likelihood:

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

## E-step: Responsibilities

In E-step we compute the posterior:

$$\gamma_j^{(i)} = \mathbb{P}(z^{(i)} = j | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}) = \frac{\pi_j^{(t)} \mathcal{N}(\mathbf{x}^{(i)} | \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{k=1}^K \pi_k^{(t)} \mathcal{N}(\mathbf{x}^{(i)} | \mu_k^{(t)}, \Sigma_k^{(t)})}.$$

Define:

$$N_j = \sum_{i=1}^N \gamma_j^{(i)}.$$

This represents the soft count for component  $j$ .

## Q-function (Expectation of Complete-data LL)

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{Z|\mathcal{X}, \boldsymbol{\theta}^{(t)}} [\ln \mathbb{P}(\mathcal{X}, Z | \boldsymbol{\theta})] \\ &= \sum_{i=1}^N \sum_{j=1}^K \gamma_j^{(i)} \left( \ln \pi_j + \ln \mathcal{N}(\mathbf{x}^{(i)} | \mu_j, \Sigma_j) \right). \end{aligned}$$

Maximizing  $Q$  w.r.t.  $\boldsymbol{\theta}$  gives M-step updates.

## M-step: Update for $\mu_j$ (Full Derivation)

We maximize the part of  $Q$  involving  $\mu_j$ :

$$Q_{\mu_j} = \sum_{i=1}^N \gamma_j^{(i)} \left( -\frac{1}{2} (\mathbf{x}^{(i)} - \mu_j)^\top \Sigma_j^{-1} (\mathbf{x}^{(i)} - \mu_j) \right) + \text{const.}$$

Take derivative:

$$\frac{\partial Q_{\mu_j}}{\partial \mu_j} = \sum_{i=1}^N \gamma_j^{(i)} \Sigma_j^{-1} (\mathbf{x}^{(i)} - \mu_j).$$

Set to zero:

$$\sum_{i=1}^N \gamma_j^{(i)} \mathbf{x}^{(i)} = \left( \sum_{i=1}^N \gamma_j^{(i)} \right) \mu_j.$$

Thus:

$$\mu_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^N \gamma_j^{(i)} \mathbf{x}^{(i)}$$

## M-step: Update for $\Sigma_j$ (Full Derivation)

We maximize:

$$Q_{\Sigma_j} = \sum_{i=1}^N \gamma_j^{(i)} \left[ -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x}^{(i)} - \mu_j)^\top \Sigma_j^{-1} (\mathbf{x}^{(i)} - \mu_j) \right].$$

Derivative identity:

$$\frac{\partial}{\partial \Sigma} \ln |\Sigma| = (\Sigma^{-1})^\top, \quad \frac{\partial}{\partial \Sigma} \mathbf{a}^\top \Sigma^{-1} \mathbf{a} = -\Sigma^{-1} \mathbf{a} \mathbf{a}^\top \Sigma^{-1}.$$

Setting derivative to zero yields:

$$\Sigma_j^{\text{new}} = \frac{1}{N_j} \sum_{i=1}^N \gamma_j^{(i)} (\mathbf{x}^{(i)} - \mu_j^{\text{new}})(\mathbf{x}^{(i)} - \mu_j^{\text{new}})^\top.$$

For numerical stability, in practice:

$$\Sigma_j \leftarrow \Sigma_j + \epsilon I.$$



## M-step: Update for $\pi_j$ - Part 1

We maximize:

$$Q_{\pi} = \sum_{j=1}^K N_j \ln \pi_j.$$

Subject to:

$$\sum_{j=1}^K \pi_j = 1.$$

Using a Lagrange multiplier:

$$\mathcal{L} = \sum_j N_j \ln \pi_j + \lambda \left( \sum_j \pi_j - 1 \right).$$

## M-step: Update for $\pi_j$ - Part 2

Derivative:

$$\frac{\partial \mathcal{L}}{\partial \pi_j} = \frac{N_j}{\pi_j} + \lambda = 0.$$

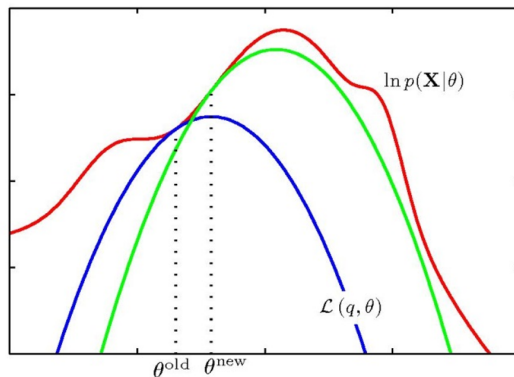
Solve:

$$\pi_j = -\frac{N_j}{\lambda}.$$

Apply constraint  $\rightarrow$  final update:

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

# EM Algorithm Illustration



- EM alternates between computing posteriors (E-step)
- and maximizing expected complete-data LL (M-step)
- until convergence.

- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound**
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References

# Convex vs. Concave Functions (Full Explanation)

A function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is:

- **Convex** if

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}), \quad 0 \leq \alpha \leq 1.$$

Its epigraph forms a convex set.

- **Concave** if

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \geq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}).$$

- **Strictly convex** if inequality is strict.

**Key fact:**  $-\ln x$  is convex;  $\ln x$  is concave.

This property is used to derive Jensen's inequality:

$$\ln \mathbb{E}[X] \geq \mathbb{E}[\ln X].$$

# Variational Lower Bound via Jensen's Inequality

We start with:

$$\ln \mathbb{P}(\mathcal{X} \mid \boldsymbol{\theta}) = \ln \sum_Z \mathbb{P}(\mathcal{X}, Z \mid \boldsymbol{\theta}).$$

Introduce any distribution  $Q(Z)$ :

$$= \ln \sum_Z Q(Z) \frac{\mathbb{P}(\mathcal{X}, Z \mid \boldsymbol{\theta})}{Q(Z)}.$$

Apply Jensen:

$$\ln \mathbb{P}(\mathcal{X} \mid \boldsymbol{\theta}) \geq \sum_Z Q(Z) \ln \frac{\mathbb{P}(\mathcal{X}, Z \mid \boldsymbol{\theta})}{Q(Z)}$$

Define:

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

And:

$$\ln \mathbb{P}(\mathcal{X} \mid \boldsymbol{\theta}) = F(\boldsymbol{\theta}, Q) + \text{KL}(Q(Z) \parallel \mathbb{P}(Z \mid \mathcal{X}, \boldsymbol{\theta})).$$

# Why EM Increases Likelihood (Full Proof) - Part 1

At iteration  $t$ :

$$Q^{(t)}(Z) = \mathbb{P}(Z | \mathcal{X}, \boldsymbol{\theta}^{(t)}).$$

Then:

$$\ln \mathbb{P}(\mathcal{X} | \boldsymbol{\theta}^{(t)}) = F(\boldsymbol{\theta}^{(t)}, Q^{(t)}).$$

M-step chooses:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} F(\boldsymbol{\theta}, Q^{(t)}).$$

Thus:

$$F(\boldsymbol{\theta}^{(t+1)}, Q^{(t)}) \geq F(\boldsymbol{\theta}^{(t)}, Q^{(t)}).$$

## Why EM Increases Likelihood (Full Proof) - Part 2

Since KL divergence is always non-negative:

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

Combine them:

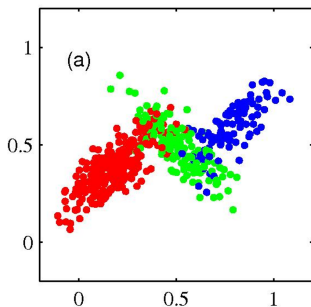
$$\boxed{\ln \mathbb{P}(\mathcal{X} | \boldsymbol{\theta}^{(t+1)}) \geq \ln \mathbb{P}(\mathcal{X} | \boldsymbol{\theta}^{(t)})}$$

→ EM **always increases likelihood.**

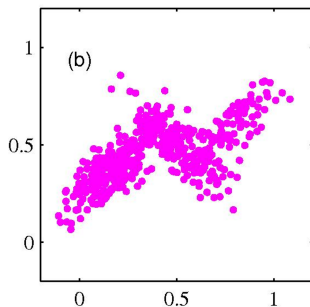


- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations**
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References

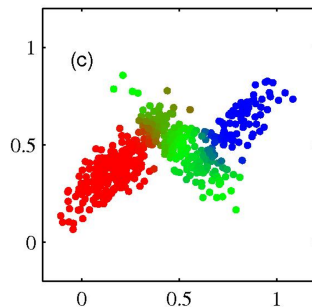
## EM & GMM Example



(a)



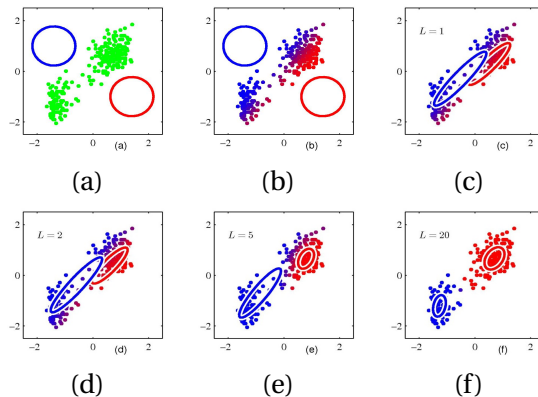
(b)



(c)

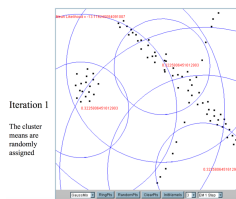
- Dataset contains two overlapping Gaussian components.
- Colors illustrate **soft assignments** (responsibilities).
- Means gradually move toward cluster centers as EM iterates.

# EM & GMM Example

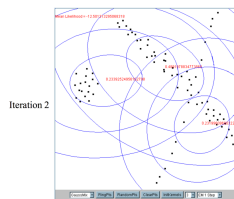


- Six-step visualization of EM evolution in 2D.
- Components reshape (covariance) according to local density.
- Shows **different initializations** → **different convergence paths**.

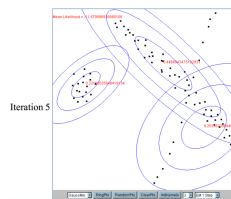
# EM Iterations — Evolution of Parameters



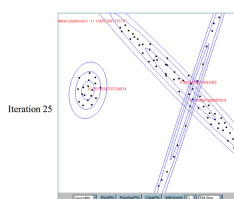
Iter 1



Iter 2



Iter 5



Iter 25

- Early steps: drastic changes in means/covariances.
- Later steps: parameters stabilize and likelihood increases slowly.
- EM often converges smoothly after a few iterations.

## GMM 1-D Example — Detailed

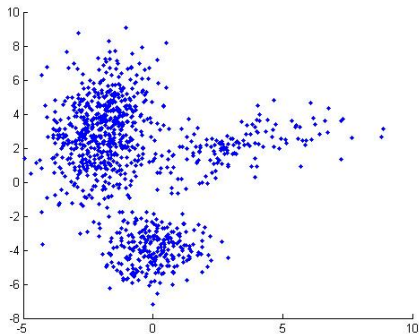
- 37 / 48

## GMM 2-D Example — Detailed

- Synthetic dataset with  $K = 3$  components.
- Components have **anisotropic covariances**:

$\Sigma_j$  not diagonal

- EM visually fits ellipses corresponding to covariance contours.
- Soft assignments shown by color gradient.



- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes**
- 8 Summary and References

# Convergence: Monotonicity and Local Optima

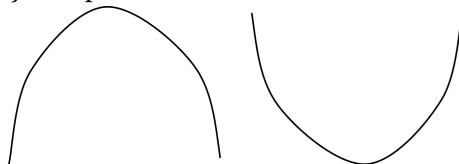
- EM **always increases** the log-likelihood:

$$\ell(\boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{\theta}^{(t)}).$$

- EM converges to a stationary point:

$$\nabla \ell(\boldsymbol{\theta}) = 0.$$

- This stationary point may be a:
  - local maximum,
  - saddle point.
- Initialization has major impact on final solution.



Concave LL

Convex LL



## Practical Tips for Using EM in GMMs

- **Initialization:**

- k-means
- k-means++ for more robustness
- multiple random restarts

- **Covariance choices:**

full, diagonal, tied, spherical

- **Stopping criteria:**

- $\Delta \ell < \epsilon$
- small parameter change
- max iterations

- **Regularization:**

$$\Sigma_j \leftarrow \Sigma_j + \epsilon I \quad (\epsilon \approx 10^{-5})$$

## Failure Modes and Solutions

- **Component collapse:**

$$\Sigma_j \rightarrow 0$$

Solution: add regularization, reinitialize collapsed component.

- **Singular covariance:** happens when one component takes a single point.
- **Slow convergence:** Solutions:
  - better initialization
  - annealing/tempering EM
  - variational EM
  - stochastic EM
- **Overfitting:** use MAP estimation or Bayesian GMM.

## Comparison: EM for GMM vs. k-means

- k-means:
  - hard assignments
  - spherical clusters
  - minimizes within-cluster square distances
- EM + GMM:
  - soft assignments ( $\gamma_j^{(i)}$ )
  - full covariance modeling
  - probabilistic interpretation (likelihood)
- EM is more flexible—but requires more parameters and computation.

- 1 Introduction & Motivation
- 2 Mixture Models
- 3 Gaussian Mixture Models (GMM)
- 4 Maximum Likelihood and the EM Algorithm
- 5 Convexity, Concavity, and EM Lower Bound
- 6 Examples and Visualizations
- 7 Convergence, Practical Issues, and Failure Modes
- 8 Summary and References**

# Summary

- GMMs provide flexible density estimation with multiple Gaussian components.
- EM algorithm:
  - E-step computes responsibilities.
  - M-step updates  $\mu_j, \Sigma_j, \pi_j$ .
- EM maximizes a lower bound and increases likelihood monotonically.
- Initialization and regularization are critical to good performance.
- GMMs outperform k-means when clusters are anisotropic or overlapping.

# References

- C. M. Bishop, *Pattern Recognition and Machine Learning*, Chapter 9.
- Original slides from:

Hamid R. Rabiee & Zahra Dehghanian (Spring 2025)

- Additional illustrations from course-provided materials.

# Acknowledgement

Special Thanks

**Soheil Sayah Varg**

for contributions to preparing, organizing, and refining this slide deck.

# Thank You!

## Questions?