STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 6: Mixture Models and EM algorithm

Jingshu Wang

The University of Chicago

Outline

From hard clustering to soft assignments

Key idea: move from geometric clustering to probabilistic modeling.

K-means (hard assignment)

- each point belongs to exactly one cluster;
- based on squared Euclidean distance;
- no generative model for data.

Mixture model (soft assignment)

- each point has a distribution over clusters;
- based on likelihood (probability);
- explicit generative model.

Connection: K-means can be viewed as a special case of a mixture model.

A simple example: mixture in one dimension

Suppose $x_i \in \mathbb{R}$ comes from two latent groups:

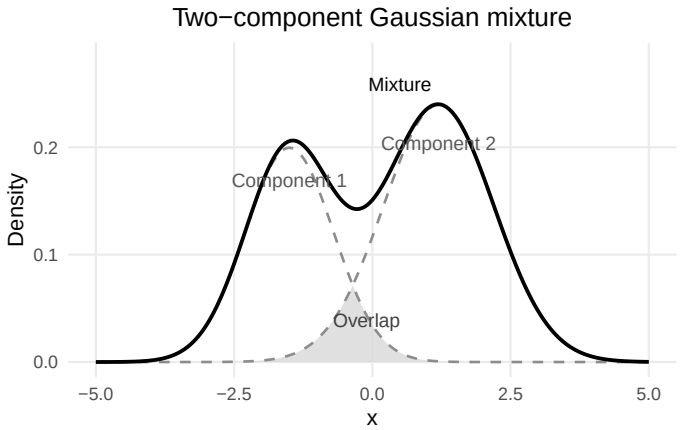
$$x_i \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{with probability } \pi, \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{with probability } 1-\pi, \end{cases} \qquad 0 < \pi < 1.$$

Each observation comes from one group, but we do not observe which one.

- Group 1: bell curve centered at μ_1.
- Group 2: bell curve centered at μ_2.
- Observed data: a weighted combination of the two.

Remark: even a single-peaked histogram may hide multiple latent subpopulations.

Mixture of two Gaussians (visualization)



Overlap ⇒ uncertainty ⇒ soft assignment

1-D soft assignment (posterior membership)

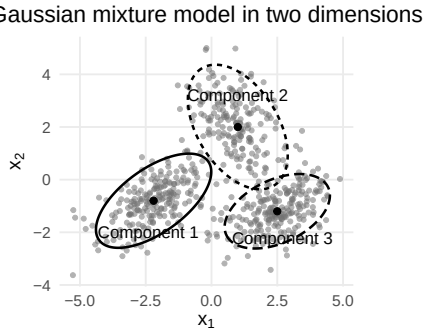
For two components, define the probability that x_i belongs to component 1 (responsibility for component 1):

$$\gamma_i = \Pr(Z_i = 1 \mid x_i) = \frac{\pi\, \phi(x_i \mid \mu_1, \sigma_1^2)}{\pi\, \phi(x_i \mid \mu_1, \sigma_1^2) + (1-\pi)\, \phi(x_i \mid \mu_2, \sigma_2^2)}.$$

- If component 1 assigns much higher density to x_i than component 2, then γ_i is close to 1.
- If component 2 assigns much higher density to x_i than component 1, then γ_i is close to 0.
- If both components assign similar density to x_i, then $0 < \gamma_i < 1$ reflects uncertainty.

From 1-D to multivariate GMM

- Same latent-variable idea, but now $\mathbf{x}_i \in \mathbb{R}^p$.
- Components become multivariate Gaussians (elliptical shapes).
- Responsibilities are computed with the same Bayes-rule structure.



Finite mixture model setup

Each observation $\mathbf{x}_i \in \mathbb{R}^p$ is generated from a mixture:

$$p(\mathbf{x}_i \mid \Theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i \mid \theta_k),$$

where

- $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$ are mixing proportions;
- $f_k(\cdot \mid \theta_k)$ is component density k;
- $\Theta = \{\pi_k, \theta_k\}_{k=1}^{K}$ are parameters.

Introduce latent class variable $Z_i \in \{1, \ldots, K\}$:

$$\Pr(Z_i = k) = \pi_k, \qquad \mathbf{x}_i \mid (Z_i = k) \sim f_k(\cdot \mid \theta_k).$$

- $Z_i = k$ means observation i came from component k.
- first draw $Z_i \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$, then draw $\mathbf{x}_i \mid Z_i$;
- Z_i is unobserved (missing data), which motivates EM.

Gaussian mixture model (GMM)

A common and important choice:

$$f_k(\mathbf{x}) = \phi_p(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k),$$

so

$$p(\mathbf{x}_i \mid \Theta) = \sum_{k=1}^{K} \pi_k \, \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k).$$

Parameter set:

$$\Theta = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K}.$$

- $\phi_p(\cdot)$ denotes the p-dimensional Gaussian density;
- each component has elliptical contours determined by Σ_k;
- flexible shape if each Σ_k is unrestricted;
- can also use constrained covariance forms for stability/interpretability.

Role of covariance matrices in GMM

- Σ_k determines shape and orientation of each component;
- spherical $\Sigma_k \Rightarrow$ round clusters;
- general $\Sigma_k \Rightarrow$ elliptical clusters;
- different Σ_k allow clusters with different geometry.
- This flexibility goes beyond K-means (more explanations in later slides).

Observed-data log-likelihood

For i.i.d. $\mathbf{x}_1, \ldots, \mathbf{x}_n$,

$$\ell(\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \, \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k) \right).$$

Why direct optimization is hard:

- log of a sum couples all components;
- no closed-form solution from setting derivatives to zero;
- objective is non-convex with multiple local maxima.

Strategy: optimize an easier surrogate repeatedly via EM.

EM general principle

E-step: compute posterior membership probabilities under current parameters,

$$\gamma_{ik} = \Pr\left(Z_i = k \mid \mathbf{x}_i, \Theta^{(t)}\right).$$

M-step: maximize expected complete-data log-likelihood,

$$Q(\Theta \mid \Theta^{(t)}) = \mathbb{E}_{Z|X,\Theta^{(t)}}[\log p(X, Z \mid \Theta)].$$

Repeat until convergence in likelihood or parameters.

Key point: EM turns difficult incomplete-data MLE into alternating tractable subproblems.

Step 1: complete-data likelihood

If latent indicators z_{ik} were observed,

$$p(X, Z \mid \Theta) = \prod_{i=1}^{n} \pi_{Z_i} \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_{Z_i}, \Sigma_{Z_i}).$$

Taking log:

$$\log p(X, Z \mid \Theta) = \sum_{i=1}^{n} \left[\log \pi_{Z_i} + \log \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_{Z_i}, \Sigma_{Z_i})\right].$$

Compared to the observed likelihood:

- avoids the log of sums structure;
- separates across components k.

Step 2: derive E-step responsibilities

Use Bayes' rule under current estimate $\Theta^{(t)}$:

$$\gamma_{ik}^{(t)} = \Pr\left(Z_i = k \mid \mathbf{x}_i, \Theta^{(t)}\right) = \frac{\pi_k^{(t)} \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}.$$

- $\gamma_{ik}^{(t)} \in [0, 1]$ and $\sum_k \gamma_{ik}^{(t)} = 1$;
- interpreted as *soft labels*;
- effective cluster size: $N_k^{(t)} = \sum_{i=1}^{n} \gamma_{ik}^{(t)}$.

Step 3: build the Q-function

Replace missing indicators with responsibilities:

$$Q(\Theta \mid \Theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{(t)} \left[\log \pi_k + \log \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k) \right].$$

Each Gaussian component is,

$$\log \phi_p(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + \text{const.}$$

So the M-step will optimize for:

- $(\boldsymbol{\mu}_k, \Sigma_k)$: weighted Gaussian MLE with weights γ_{ik};
- π_k: the effective proportions of samples for the component k.

M-step derivation: update for π_k

We maximize

$$\sum_{k=1}^{K}\left(\sum_{i=1}^{n}\gamma_{ik}^{(t)}\right)\log\pi_k \quad \text{s.t.} \quad \sum_{k=1}^{K}\pi_k = 1.$$

Lagrangian:

$$\mathcal{L} = \sum_{k=1}^{K} N_k^{(t)}\log\pi_k + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

FOC gives

$$\frac{N_k^{(t)}}{\pi_k} + \lambda = 0 \Longrightarrow \pi_k^{(t+1)} = \frac{N_k^{(t)}}{n}.$$

M-step derivation: update for $\boldsymbol{\mu}_k$

For fixed Σ_k, maximize

$$-\frac{1}{2}\sum_{i=1}^{n}\gamma_{ik}^{(t)}(\mathbf{x}_i-\boldsymbol{\mu}_k)^\top\Sigma_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k).$$

Equivalent to minimizing weighted quadratic form. Set gradient to zero:

$$\sum_{i=1}^{n}\gamma_{ik}^{(t)}\Sigma_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)=0.$$

Hence

$$\boldsymbol{\mu}_k^{(t+1)}=\frac{1}{N_k^{(t)}}\sum_{i=1}^{n}\gamma_{ik}^{(t)}\mathbf{x}_i.$$

M-step derivation: update for Σ_k

Using matrix calculus, maximize w.r.t. Σ_k:

$$-\frac{1}{2}\sum_{i=1}^{n}\gamma_{ik}^{(t)}\left[\log|\Sigma_k| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right].$$

Resulting update:

$$\Sigma_k^{(t+1)} = \frac{1}{N_k^{(t)}}\sum_{i=1}^{n}\gamma_{ik}^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^\top.$$

Interpretation: weighted sample covariance within component k.

EM algorithm summary for GMM

Initialize $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$.

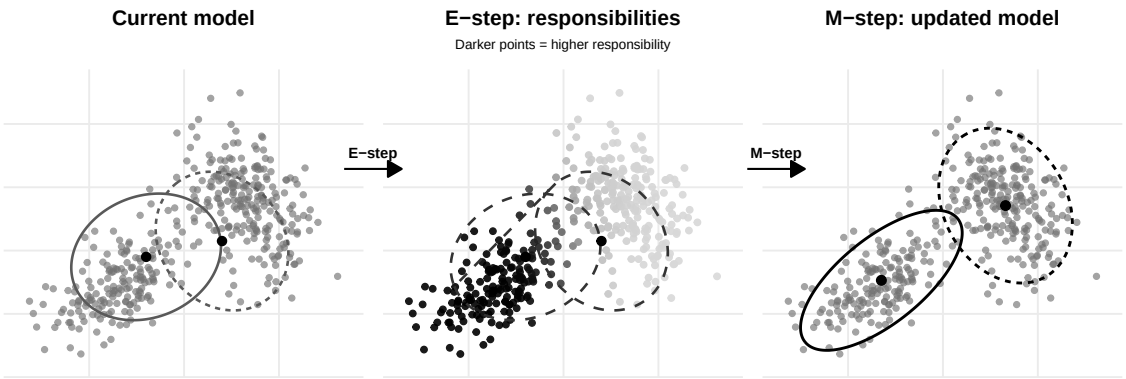
1. E-step: compute all γ_{ik}.
2. M-step:

$$N_k = \sum_i \gamma_{ik}, \qquad \pi_k \leftarrow \frac{N_k}{n}, \qquad \boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_i \gamma_{ik} \mathbf{x}_i,$$

$$\Sigma_k \leftarrow \frac{1}{N_k} \sum_i \gamma_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top.$$

3. Evaluate log-likelihood and stop if increase is small.

Visualizing one EM iteration



Why EM increases likelihood

Recall: in EM we maximize

$$Q(\Theta \mid \Theta^{(t)}) = \mathbb{E}_{Z|X,\Theta^{(t)}}[\log p(X, Z \mid \Theta)].$$

- E-step: compute $Q(\Theta \mid \Theta^{(t)})$ using current parameters;
- M-step: choose $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta \mid \Theta^{(t)})$;

Key fact:

$$\ell(\Theta) \;\geq\; Q(\Theta \mid \Theta^{(t)}) + \text{const}(\Theta^{(t)}),$$

with equality at $\Theta = \Theta^{(t)}$.

- M-step increases Q;
- therefore $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$.

Important: EM increases likelihood monotonically but may converge to a local optimum.

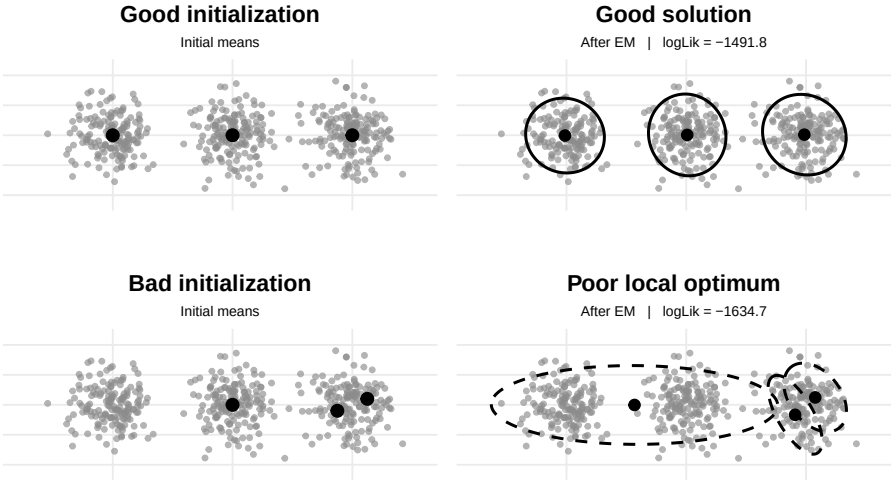
Initialization and local optima

- EM is sensitive to initialization because the likelihood is non-convex;
- Common starts: random assignments or random parameters, K-means centers;
- Use multiple random starts and keep solution with largest final log-likelihood.

Diagnostics:

- compare final log-likelihood values across runs;
- check whether clustering results are stable across starts.

Good v.s. bad initialization



Singularity in Gaussian mixtures

Key issue:

- The MLE problem is actually ill-posed: the likelihood is unbounded above.
- A component can collapse onto a single data point: if $\mu_k = x_i$ and $\Sigma_k \to 0$, then

$$\phi_p(x_i \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \;\to\; \infty.$$

- This drives the likelihood to infinity.

What happens in practice:

- although the MLE is ill-posed, EM typically converges to a reasonable local optimum;
- this type of degeneracy is possible but uncommon in well-behaved data.

If needed, we can stabilize the model:

- add small regularization (e.g., $\Sigma_k \leftarrow \Sigma_k + \epsilon I$);
- or constrain covariance structure.

Covariance parameterization and cluster shapes

Model	Covariance form	Shape intuition
Spherical	$\Sigma_k = \sigma_k^2 I$	same spread in all directions
Diagonal	$\Sigma_k = \mathrm{diag}(\cdot)$	stretched along axes
Shared full	$\Sigma_k = \Sigma$	same shape, different centers
Full by cluster	unrestricted Σ_k	arbitrary ellipses

More models are implemented in the R/Python software for GMM. Tradeoff: flexibility vs. number of parameters and numerical stability.

K-means as a special case of GMM

Assume a K-component spherical GMM with:

$$\Sigma_k = \sigma^2 I_p, \quad \pi_k = \frac{1}{K} \text{ for all } k.$$

Then posterior responsibility is

$$\gamma_{ik} = \frac{\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2\right)}.$$

As σ^2 becomes small, the exponential strongly favors the closest mean:

$$\gamma_{ik} \approx \begin{cases} 1, & k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2, \\ 0, & \text{otherwise.} \end{cases}$$

So K-means is the hard-assignment/small-variance limit of this spherical GMM.

Choosing number of components K

Typical criteria:

- **BIC**: trade-off between fit and complexity

$$\mathrm{BIC} = -2\ell(\hat{\Theta}) + d \log n \quad \text{(lower is better)}$$

d = *number of free parameters*
Example (spherical GMM):

$$d = Kp \text{ (means)} \; + \; K \text{ (variances)} \; + \; (K-1) \text{ (weights)}$$

More flexible covariance $\Rightarrow$ *larger* d *(stronger penalty)*

- **Predictive performance:** held-out log-likelihood
- interpretability and stability across initializations.

No single "best" K—check robustness.

Live R notebook

Notebook file: `Lecture6_demo.nb.html`

Lecture 6 summary

- Mixture models provide a probabilistic framework for heterogeneous populations.
- GMM introduces latent class indicators and soft memberships.
- EM alternates between posterior responsibility computation and weighted MLE updates.
- Detailed derivations give closed-form updates for π_k, $\boldsymbol{\mu}_k$, and Σ_k.
- In practice, initialization, covariance constraints, and model selection are central.

Suggested reading

- Bishop, *Pattern Recognition and Machine Learning*, Ch. 9 (mixture models and EM). Book link: `https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf`
- Johnson & Wichern (6th Edition), Chapter 12.5.