CSE 6740 Lecture 5 Notes - Sep 3
Gaussian Mixture Models(GMMs)

Hui Qiao
GT Username:hqiao38

Students' in class questions & Clarifications

1. **What does the mixing weight π_k represent? Does $\sum_k \pi_k = 1$?**
   π_k is the prior probability that a sample is drawn from component k. Yes, $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ so that $p(x)$ is a valid density.

2. **How is K chosen?**
   K is user-specified for a (parametric) GMM. In practice, pick via model selection (e.g., BIC/AIC, held-out likelihood) or domain knowledge.

3. **Joint vs. marginal: why do we sometimes have a sum over k inside a log?**
   The marginal likelihood $p(x_i) = \sum_k \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$ includes a sum (unknown component). The joint $p(x_i, z_i = k) = \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$ has no sum.

4. **What are the indicator variables τ_{ik}?**
   $\tau_{ik} = \mathbf{1}[z_i = k]$ is 1 if sample i comes from component k, else 0. In EM we replace them by soft responsibilities $\gamma_{ik} \in [0, 1]$ with $\sum_k \gamma_{ik} = 1$.

5. **Can a component receive zero points (all $\tau_{ik} = 0$ or $N_k = 0$)?**
   Yes, it can happen (especially with large K). Then $\pi_k = N_k/n = 0$ and the component can effectively die. Good initialization or small regularization can mitigate this.

6. **Where do the posterior responsibilities γ_{ik} come from?**
   Bayes' rule:
   $$\gamma_{ik} = p(z_i = k \mid x_i, \theta) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i \mid \mu_{k'}, \Sigma_{k'}}.$$

7. **Updates in the M-step look like "weighted" estimates—why?**
   They are the MLEs for a fully observed mixture if we treat γ_{ik} as fractional counts:
   $$N_k = \sum_i \gamma_{ik}, \quad \mu_k = \frac{1}{N_k} \sum_i \gamma_{ik} x_i, \quad \Sigma_k = \frac{1}{N_k} \sum_i \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^\top, \quad \pi_k = \frac{N_k}{n}.$$

8. **What is the relation between GMM-EM and k-means?**
   If all $\Sigma_k = \sigma^2 I$ and $\sigma^2 \to 0$, the E-step becomes hard assignment to the nearest mean and the M-step recomputes centroids—recovering k-means as a limiting case.

9. **Do we assume anything about $p(x)$ directly?**
   No. We posit the mixture form $p(x) = \sum_k \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$ and learn its parameters; we do not assume a closed form for $p(x)$ beyond the mixture.

10. **What is $H(q)$ in the bound? Does it depend on θ?**
    $H(q)$ is the entropy of $q(z)$. It depends on q (from the E-step), not on θ (in the M-step), so it is constant during the M-step optimization.

11. **Notation check: x vs. x_i and k vs. k'?**
    x denotes a generic variable; x_i is the i-th sample. Index k identifies a particular component; k' is used as a dummy index inside sums.

12. **Why do we regularize Σ_k?**
    If a mean collapses onto a data point, Σ_k can become singular and the likelihood can blow up. Adding ϵI or constraining covariance structure (e.g., diagonal) improves stability.

13. **How do I know EM has converged?**
    Stop when the increase in log-likelihood between iterations is below a tolerance, or when parameter changes are small, or after a max number of iterations.

14. **Any advice on initialization?**
    Use k-means++ or several random restarts and choose the run with highest final log-likelihood. Poor initialization can lead to bad local optima.

Notes

1. Motivation for seeking a more flexible model (GMMs)
   The previous parametric models are too restricted (they must fit a single Gaussian), while nonparametric models (histogram, KDE) often need a lot of data. We want a model that can fit multi-modal densities with relatively few parameters.

2. Definition of GMMs
   A Gaussian Mixture Model with K components has density

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\big(x \mid \mu_k, \Sigma_k\big), \quad \pi_k \geq 0, \; \sum_{k=1}^{K} \pi_k = 1.$$

   It is parametric if K is fixed in advance (parameters are $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$). If K is allowed to grow with data, the family becomes effectively nonparametric.

3. Suppose we have a GMM—how do we sample from it? (Generative view)
   Introduce a latent variable $z \in \{1, \ldots, K\}$ for the component index.

   (a) Sample $z \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$.

   (b) Given $z = k$, sample $x \sim \mathcal{N}(x \mid \mu_k, \Sigma_k)$.

This implies the joint and marginal:

$$p(x, z) = \pi_z \mathcal{N}\big(x \mid \mu_z, \Sigma_z\big), \qquad p(x) = \sum_{z=1}^{K} p(x, z) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(x \mid \mu_k, \Sigma_k\big).$$

4. Learning the parameters by Maximum Likelihood (MLE)
Given data $D = \{x_i\}_{i=1}^n$, maximize

$$\ell(\theta; D) = \sum_{i=1}^{n} \log p(x_i \mid \theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\big(x_i \mid \mu_k, \Sigma_k\big),$$

where $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$. The $\log \sum$ makes the objective non-convex and hard to optimize directly.

5. If the latent assignments were known (complete-data log-likelihood)
Introduce binary indicators $\tau_{ik} = \mathbf{1}[z_i = k]$. Then

$$\ell(\theta; D, \tau) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \Big(\log \pi_k - \tfrac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) - \tfrac{1}{2} \log |\Sigma_k| - c \Big).$$

Maximizing w.r.t. θ (with $\sum_k \pi_k = 1$) yields the closed forms:

$$\pi_k = \frac{\sum_i \tau_{ik}}{n}, \quad \mu_k = \frac{\sum_i \tau_{ik} x_i}{\sum_i \tau_{ik}}, \quad \Sigma_k = \frac{\sum_i \tau_{ik} (x_i - \mu_k)(x_i - \mu_k)^\top}{\sum_i \tau_{ik}}.$$

6. When z is unknown: posterior responsibilities (soft assignments)
Define the responsibility $\gamma_{ik} \equiv p(z_i = k \mid x_i, \theta)$ via Bayes' rule:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_i \mid \mu_{k'}, \Sigma_{k'}}, \qquad \sum_{k=1}^{K} \gamma_{ik} = 1.$$

We will use $\tau_{ik} \leftarrow \gamma_{ik}$ as a soft version of the unknown labels.

7. Expectation-Maximization (EM) for GMMs
Initialize $\{\pi_k, \mu_k, \Sigma_k\}$. Iterate until convergence:

E-step: Compute responsibilities

$$\gamma_{ik} \leftarrow \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i \mid \mu_{k'}, \Sigma_{k'}}.$$

M-step: Update parameters using weighted MLE

$$N_k = \sum_{i=1}^{n} \gamma_{ik}, \quad \pi_k \leftarrow \frac{N_k}{n}, \quad \mu_k \leftarrow \frac{1}{N_k} \sum_i \gamma_{ik} x_i, \quad \Sigma_k \leftarrow \frac{1}{N_k} \sum_i \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^\top.$$

Each EM iteration does not decrease the data log-likelihood; in practice it monotonically increases it until reaching a stationary point.

8. EM as lower-bound maximization (intuition)
   EM can be seen as maximizing a variational lower bound

$$\ell(\theta; D) \;\geq\; \mathcal{L}(\theta, q) = \mathbb{E}_{q(z)}\big[\log p(x, z \mid \theta)\big] + H(q),$$

   where $q(z)$ is any distribution over latent variables.
   E-step chooses $q(z) = p(z \mid x, \theta^{(t)})$ (tightest bound via Jensen).
   M-step maximizes $\mathcal{L}(\theta, q)$ w.r.t. θ with q fixed.

9. Relation to k-means
   If we constrain $\Sigma_k = \sigma^2 I$ (equal, spherical) and send $\sigma^2 \to 0$, the E-step approaches hard assignments to the nearest mean and the M-step reduces to recomputing cluster centroids—recovering k-means as a limit case (EM is a soft clustering generalization).

10. Practical tips
    Initialization matters: use k-means or multiple random restarts.
    Regularize Σ_k (e.g., add ϵI) to avoid singular covariances.
    Stop when the increase in $\ell(\theta; D)$ is below a tolerance.
    Choosing K: compare BIC/AIC across fits; visualize with PCA for intuition.

11. (Demo mentioned)
    Toy examples (e.g., wine data) show EM iteratively reweights points (via γ_{ik}) and adjusts (μ_k, Σ_k) until modes are captured.