Introduction and learning objectives

In Lecture 9, k-means assigned each point to its nearest centre and updated each centre to its group's mean. This can fail when groups have different spreads (amounts of variability around their centres). A Gaussian mixture model (GMM) uses a Gaussian distribution for each group and gives probabilities for the component labels.

By the end of the tutorial, you should be able to:

- Construct a GMM using an unobserved component label and derive its marginal density.
- Derive its likelihood and posterior component probabilities, and relate them to k-means assignments.
- Explain the advantages and limitations of GMMs compared with k-means.

Optional exercise: Derive gradient descent and SGD updates in Problem 3.

1 Gaussian Mixture Models

Problem 1(a). Each weight is a probability, so $0 \leq \pi_k \leq 1$. The component labels are mutually exclusive and exhaustive, so $\sum_k \pi_k = 1$.

Problem 1(b). Each $\boldsymbol{\Sigma}_k$ is a symmetric positive-definite $D \times D$ matrix. Symmetry follows from the definition of covariance, and $\mathbf{v}^\top \boldsymbol{\Sigma}_k \mathbf{v} = \mathrm{Var}(\mathbf{v}^\top \mathbf{x} \mid z = k) \geq 0$. General covariances can be singular, but the Gaussian density used here requires an inverse. We therefore require positive definiteness.

Problem 1(c). The product rule gives

$$p(\mathbf{x}_n, z_n = k \mid \theta) = p(z_n = k \mid \theta) p(\mathbf{x}_n \mid z_n = k, \theta) = \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Apply the sum rule, then substitute the joint density:

$$\begin{aligned} p(\mathbf{x}_n \mid \theta) &= \sum_{k=1}^{K} p(\mathbf{x}_n, z_n = k \mid \theta) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned}$$

This density is nonnegative. Each Gaussian integrates to one, so $\int p(\mathbf{x} \mid \theta)\, d\mathbf{x} = \sum_k \pi_k = 1$.

Problem 1(d). The covariance describes spread (variability around the mean) and correlations. It can distinguish the wide group in Lecture 9 from the compact groups. The weight π_k is the probability of selecting component k before observing a point, and describes its relative prevalence.

Problem 2(a). Independence gives a product over observations. Taking a logarithm turns this *product* into a sum:

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \left[\sum_k \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right],$$
$$\log p(\mathcal{D} \mid \theta) = \sum_{n=1}^{N} \log \left[\sum_k \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].$$

The logarithm is strictly increasing, so it preserves the maximiser. Multiplying by $-1/N$ changes maximisation to minimisation:

$$\boxed{\mathcal{L}(\theta) = -\frac{1}{N} \sum_n \log \left[\sum_k \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right].}$$

In general, $\log(a+b) \neq \log a + \log b$; the right side is $\log(ab)$. The sum over component labels remains inside each logarithm. We can evaluate the likelihood and differentiate it, even though direct analytic maximisation is difficult.

Problem 2(b). Bayes' rule gives

$$\boxed{\gamma_{nk} = \frac{p(\mathbf{x}_n, z_n = k \mid \theta)}{p(\mathbf{x}_n \mid \theta)} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.}$$

Summing the numerator over k gives the denominator, hence $\sum_k \gamma_{nk} = 1$. Each γ_{nk} is a posterior probability in $[0, 1]$. K-means instead chooses a one-hot assignment $r_{nk} \in \{0, 1\}$.

The latent label z_n is a single discrete label. Responsibilities describe our uncertainty about it. They are determined by the observation and the current parameters.

Problem 2(c). With $\pi_k = 1/K$ and $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$, each numerator is

$$\frac{1}{K} (2\pi\sigma^2)^{-D/2} \exp\left(-\frac{\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2}{2\sigma^2} \right).$$

The factors $K^{-1}(2\pi\sigma^2)^{-D/2}$ are shared by all components and cancel:

$$\boxed{\gamma_{nk} = \frac{\exp(-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/(2\sigma^2))}{\sum_j \exp(-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/(2\sigma^2))}.}$$

The exponential is increasing in its argument, so the smallest squared distance gives the largest responsibility. With a unique nearest centre, taking the most probable component gives the same label as k-means. The GMM retains the other posterior probabilities as well. Unequal weights or covariances can change which component is most probable.

Problem 2(d). No. The denominator in γ_{nk} is shared by all components, so compare $\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. A component with zero weight has zero responsibility. For positive weights, compare the log scores

$$\log \pi_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k),$$

where the common term $-\frac{D}{2} \log(2\pi)$ has been omitted. The score depends on the component's weight, variability and orientation as well as the point's displacement from its mean.

In Lecture 9, some points from the wide group were closer to another group's centre. A GMM can give such a point a higher probability under the wide component if its weighted density is larger. Greater variability does not always give a higher responsibility: the determinant term also changes the density. The nearest-centre rule follows under the equal-weight, equal-spherical-covariance assumptions in part (c).

Optional extension. Let $d_{nk} = \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ and let k_* be the unique nearest centre. Divide the numerator and denominator of each responsibility by $e^{-d_{nk_*}/(2\sigma^2)}$:

$$\gamma_{nk} = \frac{\exp(-(d_{nk} - d_{nk_*})/(2\sigma^2))}{\sum_j \exp(-(d_{nj} - d_{nk_*})/(2\sigma^2))}.$$

The term for k_* is 1, and all other terms tend to zero as $\sigma^2 \to 0$. Thus $\gamma_{nk_*} \to 1$ and $\gamma_{nk} \to 0$ for $k \neq k_*$. If q centres tie for the smallest distance, each has limiting responsibility $1/q$ under the equal-weight assumption. K-means still needs a rule that chooses one of them.

A connection to the loss (optional). For equal weights and fixed spherical covariance,

$$\ell_n = \log K + \frac{D}{2} \log(2\pi\sigma^2) - \log \sum_k \exp\left(-\frac{d_{nk}}{2\sigma^2}\right).$$

As $\sigma^2 \to 0$, the scaled loss $2\sigma^2 \ell_n$ tends to $\min_k d_{nk}$. Consequently, $2N\sigma^2 \mathcal{L}$ tends to the k-means loss after choosing each point's best assignment. This is a limit with centres fixed, not equality of the objectives at a finite variance.

GMMs and k-means: discussion notes

Advantages of a GMM. A GMM represents uncertainty about component labels and allows groups with different variability. Full covariances also model correlations and orientation. The model gives a density and can generate observations. These features are useful when groups overlap or differ in shape, as in the example in Lecture 9.

Costs and limitations. K-means is simpler and usually cheaper. It learns KD centre coordinates. A full-covariance GMM also learns $K-1$ independent mixing weights and $D(D+1)/2$ covariance parameters per component. Estimating these parameters requires more data and computation. Diagonal or shared covariances reduce this cost. Both methods depend on K, features and initialisation. A GMM's component probabilities depend on the chosen model; they do not guarantee that its components match the groups we intended.

Covariance collapse. For a nontrivial mixture, one component can centre itself on one observation and shrink its covariance towards zero, driving the likelihood arbitrarily high while other components cover the remaining observations. A covariance constraint or regularisation prevents this collapse. The positive variance floor in the optional exercise provides one such constraint.

Training and EM: discussion notes

Explain gradient training using the loss from Problem 2: parameterise valid weights and covariances, evaluate a minibatch loss, differentiate, and update the parameters. The optional Problem 3 gives the mean gradient. A minibatch samples observations; the mixture density still sums over all components.

For EM, connect the E-step to k-means assignments and the M-step to the mean update. The E-step computes responsibilities from the old parameters. The M-step holds them fixed and maximises

$$Q(\theta, \theta^{\text{old}}) = \sum_n \sum_k \gamma_{nk}^{\text{old}} \left[\log \pi_k + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right].$$

This is the expected log-likelihood of the observations and their latent labels, using the old posterior probabilities. The mean update, for $N_k = \sum_n \gamma_{nk}^{\text{old}} > 0$, is

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_n \gamma_{nk}^{\text{old}} \mathbf{x}_n.$$

K-means uses hard assignments in the same weighted average. Full EM also updates weights and covariances, unless the model holds them fixed.

The responsibilities depend on the current parameters, so the weighted-mean equation is not a direct solution for all unknowns. EM repeatedly recomputes the responsibilities and solves the weighted fitting problem. Well-defined exact EM steps do not decrease the observed-data log-likelihood; the proof is in PRML Section 9.4. EM does not guarantee a global optimum.

Problem 3(a) (optional). Only the k-th Gaussian in s_n depends on $\boldsymbol{\mu}_k$. Differentiate the original negative log-likelihood:

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}_k} \ell_n &= -\frac{1}{s_n} \nabla_{\boldsymbol{\mu}_k} s_n \\
&= -\frac{\pi_k}{s_n} \nabla_{\boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \sigma^2 \mathbf{I}) \\
&= -\underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \sigma^2 \mathbf{I})}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \sigma^2 \mathbf{I})}}_{\gamma_{nk}} \frac{\mathbf{x}_n - \boldsymbol{\mu}_k}{\sigma^2} \\
&= \frac{\gamma_{nk}}{\sigma^2} (\boldsymbol{\mu}_k - \mathbf{x}_n).
\end{aligned}$$

Averaging over observations gives

$$\boxed{\nabla_{\boldsymbol{\mu}_k} \mathcal{L} = \frac{1}{N\sigma^2} \sum_n \gamma_{nk} (\boldsymbol{\mu}_k - \mathbf{x}_n).}$$

The responsibility appears when we differentiate the logarithm of the mixture sum. This is the exact gradient; the responsibility depends on the current means and must be recomputed after an update.

For general *fixed* positive-definite covariances, the same calculation gives

$$\nabla_{\boldsymbol{\mu}_k} \mathcal{L} = \frac{1}{N} \sum_n \gamma_{nk} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_k - \mathbf{x}_n).$$

Problem 3(b) (optional). Substituting part (a) into the gradient descent rule gives

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \boldsymbol{\mu}_k^{\text{old}} - \eta \, \nabla_{\boldsymbol{\mu}_k} \mathcal{L} \big|_{\theta = \theta^{\text{old}}} \\
&= \boxed{\boldsymbol{\mu}_k^{\text{old}} + \frac{\eta}{N\sigma^2} \sum_n \gamma_{nk}^{\text{old}} (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}).}
\end{aligned}$$

Each point contributes a vector pointing from the current mean to that point. Its responsibility scales this contribution. In k-means, r_{nk} makes the contribution either present or absent. Up to the constant factor, the gradients have the same weighted-displacement form.

K-means moves directly to the mean of its assigned points. Gradient descent takes a step whose length depends on η. A large step can overshoot and increase the loss.

Problem 3(c) (optional). Differentiate the average minibatch loss, using the per-observation gradient from part (a):

$$\begin{aligned}\widehat{\mathbf{g}}_k = \nabla_{\boldsymbol{\mu}_k} \mathcal{L}_{\mathcal{B}} &= \frac{1}{B} \sum_{n \in \mathcal{B}} \nabla_{\boldsymbol{\mu}_k} \ell_n \\ &= \frac{1}{B\sigma^2} \sum_{n \in \mathcal{B}} \gamma_{nk} (\boldsymbol{\mu}_k - \mathbf{x}_n).\end{aligned}$$

Evaluate this gradient at the old parameters and substitute into the update:

$$\begin{aligned}\boldsymbol{\mu}_k^{\text{new}} &= \boldsymbol{\mu}_k^{\text{old}} - \eta \widehat{\mathbf{g}}_k^{\text{old}} \\ &= \boxed{\boldsymbol{\mu}_k^{\text{old}} + \frac{\eta}{B\sigma^2} \sum_{n \in \mathcal{B}} \gamma_{nk}^{\text{old}} (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}).}\end{aligned}$$

The average of this estimate over the sampling process equals the full-data gradient. If $B = 1$, this is a single-observation SGD update; using all N observations gives full-batch gradient descent. Each observation's mixture density and responsibilities still sum over all K components. There is no need to sample z_n. Stochastic updates need not decrease the full-data loss on every step.

Learning weights and variances. For the parameterisation on the question sheet,

$$\pi_k = \frac{e^{a_k}}{\sum_j e^{a_j}}, \qquad v_{kd} = \varepsilon + \text{softplus}(b_{kd}), \qquad \boldsymbol{\Sigma}_k = \text{diag}(v_{k1}, \ldots, v_{kD}),$$

automatic differentiation applies the chain rule to the *original* negative log-likelihood. There is no need to derive covariance gradients in class.

For an optional check, the per-observation derivatives are

$$\begin{aligned}\frac{\partial \ell_n}{\partial a_k} &= \pi_k - \gamma_{nk}, \\ \frac{\partial \ell_n}{\partial b_{kd}} &= \frac{\gamma_{nk}}{2} \left(\frac{1}{v_{kd}} - \frac{(x_{nd} - \mu_{kd})^2}{v_{kd}^2} \right) \text{sigmoid}(b_{kd}).\end{aligned}$$

The first formula follows from $\partial \pi_j / \partial a_k = \pi_j (\mathbb{I}[j = k] - \pi_k)$. The second follows by differentiating the diagonal Gaussian log-density with respect to v_{kd}, then using $dv_{kd}/db_{kd} = \text{sigmoid}(b_{kd})$. Average these derivatives over the selected observations before updating.

Implementation and limitations. Use log-softmax for $\log \pi_k$, stable softplus for the variances, and log-sum-exp for the mixture density. Initialise distinct means; perfectly symmetric components can remain identical under gradient updates. K-means centres provide one possible starting point.

The objective is nonconvex. Without a covariance constraint or regularisation, one component in a nontrivial mixture can put its mean on one observation and shrink its covariance towards zero, making the likelihood unbounded while other components cover the remaining observations. The positive variance floor prevents this collapse for the diagonal model. Different starts can still give different fitted models.