EM algorithm

Thibault Randrianarisoa

University of Toronto, Winter 2026

February 25, 2026

Mixture of Gaussians

We combine simple models into a complex model by taking a mixture of K multivariate Gaussian densities of the form:

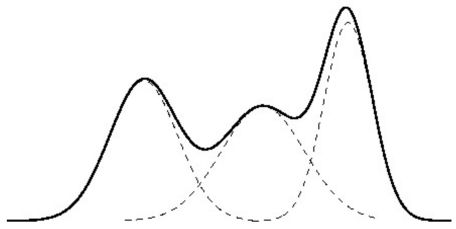
$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}_m(x|\mu_k, \Sigma_k),$$

where $\pi_k \geqslant 0$, $\sum_{k=1}^{K} \pi_k = 1$, and $\mathcal{N}_m(x|\mu_k, \Sigma_k)$ is the m-dim Gaussian density.

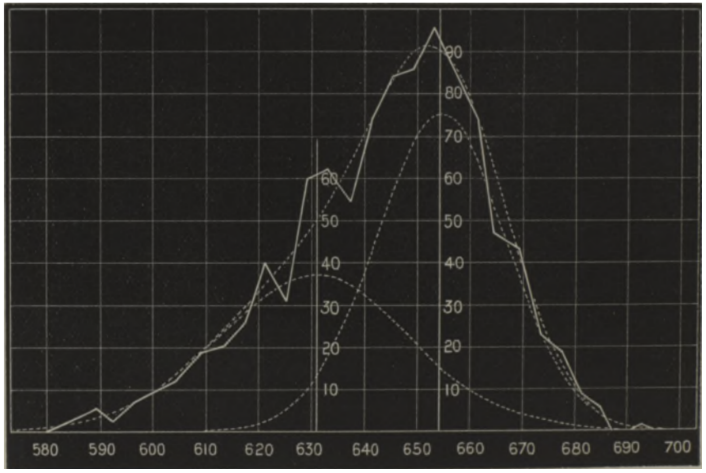
- Each Gaussian component has its own mean vector μ_k and covariance matrix Σ_k.
- The parameters π_k are called the mixing coefficients.

Example:

- $K = 3$ (three Gaussian components)
- $m = 1$ (univariate Gaussians)



The crabs from Naples bay



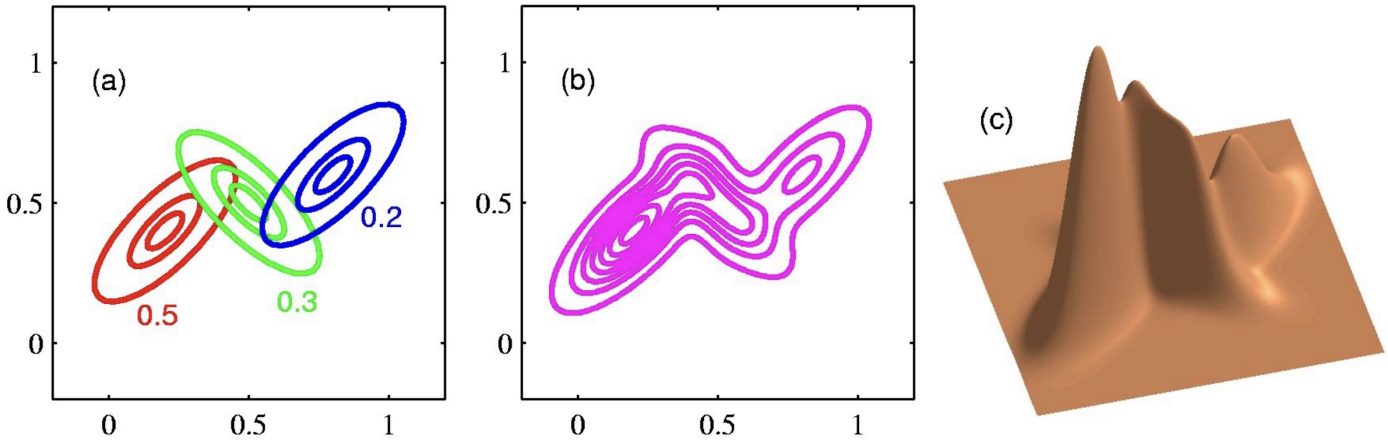
In 1892, scientists collected data on crab populations and observed that the ratio of forehead width to the body length actually showed a highly skewed distribution.

Source: *On Certain Correlated Variations in Carcinus maenas* (1893) W. F. Weldon.

They wondered whether this distribution could be the result of the population being a mix of two different normal distributions (two sub-species).

In **1894**, Karl Pearson proposed a method to fit this model (read here), whose modern version is the "method of moments". The method involved solving a higher order polynomial.

Illustration of a mixture of 3 Gaussians in a 2-dimensional space



(a) Contours of constant density of each of the mixture components, along with the mixing coefficients

(b) Contours of marginal probability density $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$

(c) A surface plot of the distribution $p(\mathbf{x})$.

Mixture of Gaussians as a latent variable model

Recall: $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}_m(x|\mu_k, \Sigma_k)$.

- Consider a latent variable z with K states $z \in \{1, \ldots, K\}$.
- The distribution of z given by the mixing coefficients:

$$p(z = k) = \pi_k.$$

- Specify the conditional as $p(x|z = k) = \mathcal{N}_m(x|\mu_k, \Sigma_k)$ with joint:

$$p(x, z = k) = p(z = k)p(x|z = k) = \pi_k \mathcal{N}_m(x|\mu_k, \Sigma_k).$$

- Then the marginal $p(x)$ satisfies

$$p(x) = \sum_{k=1}^{K} p(x, z = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}_m(x|\mu_k, \Sigma_k).$$

Mixture of Gaussians: inference

- If we have several observations $x_1, \ldots, x_N$, for every observed data point x_n there is a corresponding latent z_n.

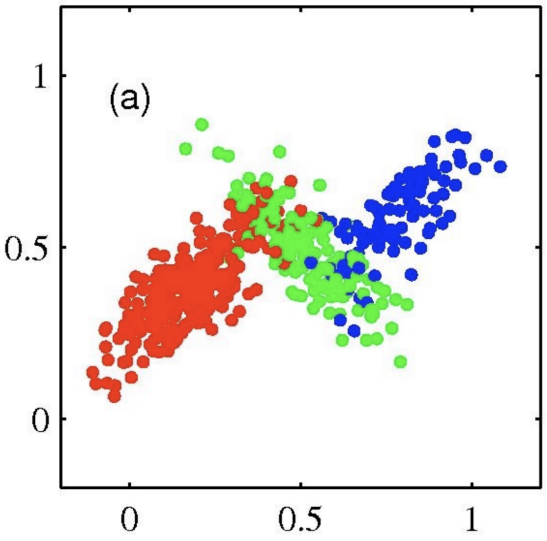
- Consider the conditional $p(z|x)$:

$$p(z = k|x) = \frac{p(z = k)p(x|z = k)}{\sum_{j=1}^{K} p(z = j)p(x|z = j)} \quad \text{(Bayes' rule)}$$

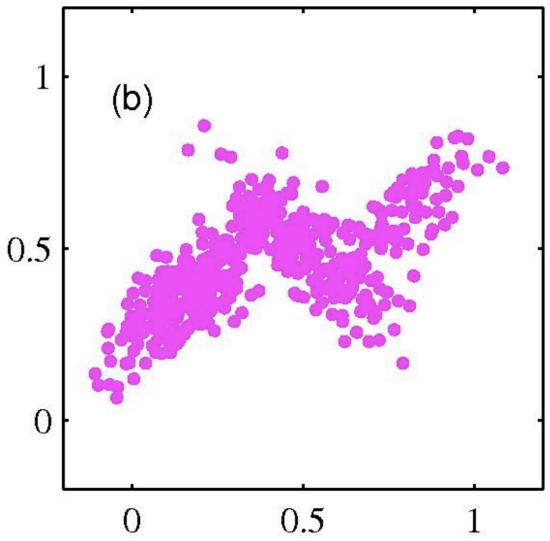
$$= \frac{\pi_k \mathcal{N}_m(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}_m(x|\mu_j, \Sigma_j)}$$

- We view π_k as prior probability that $z = k$, and $p(z = k|x)$ is the corresponding posterior once we have observed the data.

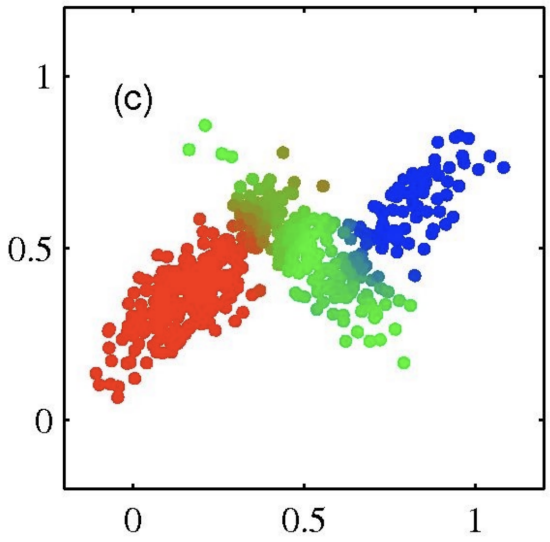
- 500 points drawn from a mixture of 3 Gaussians.



Samples from the joint distribution $p(x, z)$.



Samples from the marginal distribution $p(x)$.



Same samples where colors represent the value of responsibilities $p(z = k|x)$.

The Likelihood function

Parameters: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_K)$.

Recall: $p(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k N_m(x|\mu_k, \Sigma_k)$

- Represent the dataset $\{x_1, \ldots, x_N\}$ as $\boldsymbol{X} \in \mathbb{R}^{N \times m}$.
- The latent variable is represented by a vector $\boldsymbol{z} \in \mathbb{R}^N$.
- The log-likelihood takes the form

$$\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$$

Maximum Likelihood (μ)

Recall: $\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$.

- Differentiating wrt μ_k and setting to zero gives:

$$
\begin{aligned}
0 &= \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} \Sigma_k^{-1}(x_n - \mu_k) = \sum_{n=1}^{N} p(z_n = k|x_n) \Sigma_k^{-1}(x_n - \mu_k) \\
&= \Sigma_k^{-1} \left(\sum_{n=1}^{N} p(z_n = k|x_n) x_n - \mu_k \sum_{n=1}^{N} p(z_n = k|x_n) \right).
\end{aligned}
$$

- Equivalently (as Σ_k is positive definite)

$$
\mu_k = \sum_n \frac{p(z = k|x_n)}{N_k} x_n, \qquad N_k = \sum_n p(z = k|x_n).
$$

- Simple interpretation: the MLE for μ_k is given by the weighted mean of the data weighted by the posterior $p(z = k|x_n)$.

Maximum Likelihood (Σ, π)

Recall: $\log p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_m(x_n|\mu_k, \Sigma_k) \right)$.

- Differentiating wrt Σ_k and setting to zero gives:

$$\Sigma_k = \sum_n \frac{p(z = k|x_n)}{N_k} (x_n - \mu_k)(x_n - \mu_k)^\top.$$

- Again data points weighted by posterior probabilities.
- Finally, for the weights π_k the MLE is

$$\sum_k \pi_k = 1$$

$$\pi_k = \frac{N_k}{\sum_{j=1}^{K} N_j} = \frac{N_k}{N}, \qquad N_k = \sum_n p(z = k|x_n).$$

Motivating the EM algorithm

- The MLE does not have a closed form solution.
- The estimates depend on the posterior probabilities $p(z = k|x_n)$, which themselves depend on those parameters.
- Indeed, recall that

$$p(z = k|x_n) = \frac{\pi_k N_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N_m(x_n|\mu_j, \Sigma_j)}.$$

- Iterative solution (EM algorithm):
  - Initialize the parameters to some values.
  - E-step Update the posteriors $p(z = k|x_n)$.
  - M-step Update model parameters π, μ, Σ.
  - Repeat.

EM algorithm for Gaussian mixtures

- Initialize π, μ, Σ.
- E-step: for each k, n compute the posterior probabilities

$$p(z = k|x_n) = \frac{\pi_k N_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N_m(x_n|\mu_j, \Sigma_j)}.$$

- M-step: Re-estimate model parameters

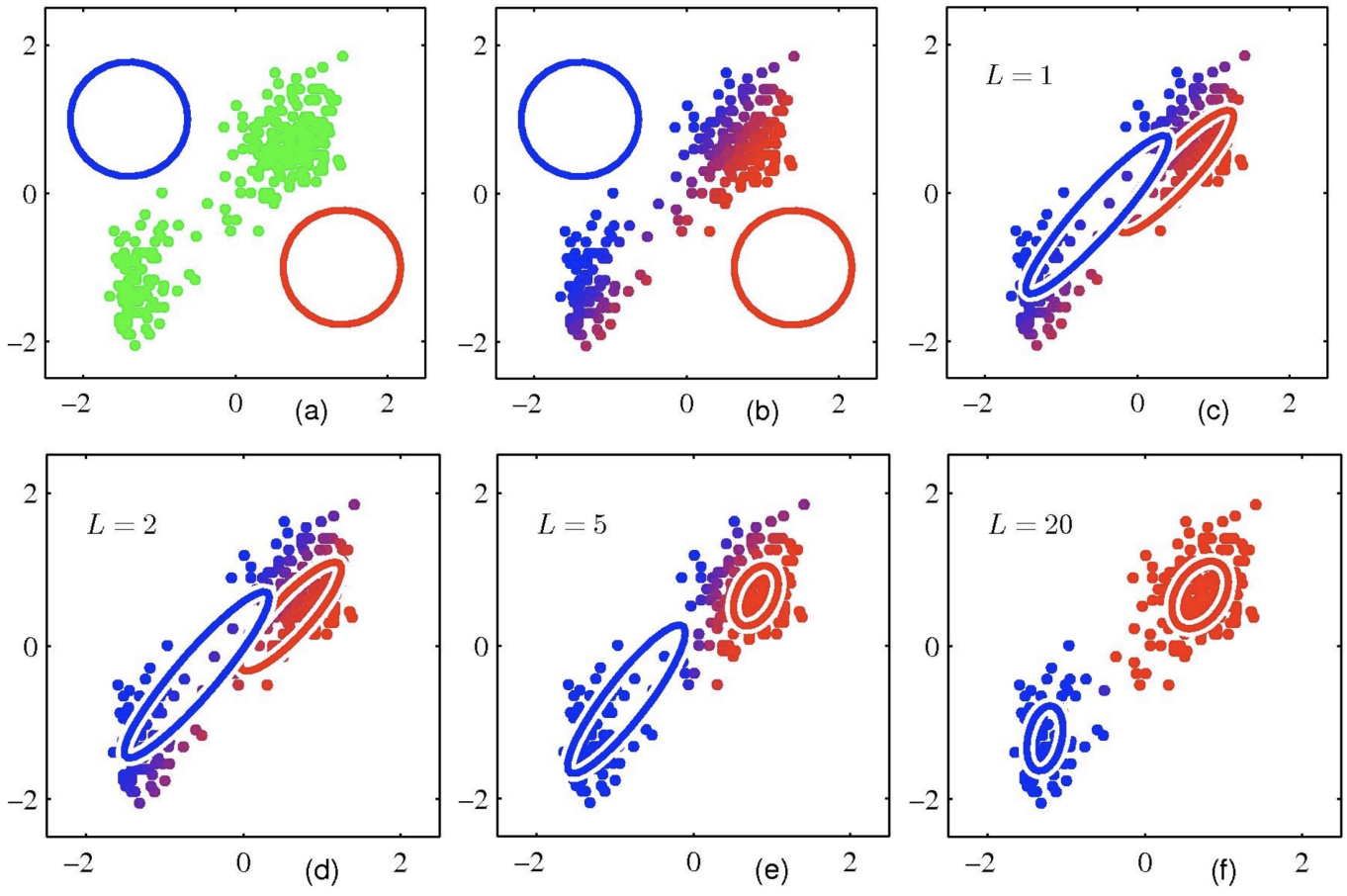
$$\mu_k^{\text{new}} = \sum_{n=1}^{N} \frac{p(z = k|x_n)}{N_k} x_n, \quad N_k = \sum_{n=1}^{N} p(z = k|x_n),$$

$$\Sigma_k^{\text{new}} = \sum_{n=1}^{N} \frac{p(z = k|x_n)}{N_k} (x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^\top,$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}.$$

- Evaluate the log-likelihood and check for convergence.

- Illustration of the EM algorithm (much slower convergence compared to K-means)



The General EM algorithm

Consider a general setting with latent variables.

- Observed dataset $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, latent variables $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$.

Maximize the log-likelihood $\log p(\boldsymbol{X}|\theta) = \log\left(\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\theta)\right)$.

- Initialize parameters θ^{old}.
- **E-step**: use θ^{old} to compute the posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{\text{old}})$.
- **M-step**: $\theta^{\text{new}} = \arg\max_\theta Q(\theta, \theta^{\text{old}})$, where

$$Q(\theta, \theta^{\text{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \theta^{\text{old}}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) = \mathbb{E}\Big(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \Big| \boldsymbol{X}, \theta^{\text{old}} \Big)$$

  which is tractable in many applications.

- Replace $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$. Repeat until convergence.

$$\frac{p(z|\theta^{\text{old}})\, p(x|z, \theta^{\text{old}})}{\sum_z p(z|\theta^{\text{old}})\, p(x|z, \theta^{\text{old}})}$$

Example: Gaussian mixture

- If z was observed, the MLE would be trivial

$$\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) = \sum_{n=1}^{N} \log p(x_n, z_n|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}(z_n = k) \log(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)).$$

For the E-step: $p(\boldsymbol{Z}|\boldsymbol{X}, \theta) = \prod_{n=1}^{N} p(z_n|\boldsymbol{X}, \theta)$ we have

$$p(z_n = k|\boldsymbol{X}, \theta) = p(z_n = k|x_n, \theta) = \frac{\pi_k \mathcal{N}_m(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}_m(x_n|\mu_j, \Sigma_j)}.$$

For the M-step: $\mathbb{E}(\mathbb{1}(z_n = k)|\boldsymbol{X}, \theta^{\text{old}}) = p(z_n = k|\boldsymbol{X}, \theta^{\text{old}})$ and so

$$\mathbb{E}\Big(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta) \Big| \boldsymbol{X}, \theta^{\text{old}} \Big) = \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_n = k|\boldsymbol{X}, \theta^{\text{old}}) \log(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)).$$

Maximizing gives the formulas on Slide 12.

Relationship to K-Means (STA 314?)

- Consider a Gaussian mixture, s.t. $\Sigma_k = \epsilon I$ for all $k = 1, \ldots, K$.
- We have

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{m/2}} \exp\left(-\frac{1}{2\epsilon}\|x - \mu_k\|^2\right).$$

- Consider the EM algorithm in this special case, $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu})$.
- The posterior probabilities take the form

$$p(z_n = k|\boldsymbol{X}, \theta) = \frac{\pi_k \exp(-\|x_n - \mu_k\|^2/2\epsilon)}{\sum_{j=1}^{K} \pi_j \exp(-\|x_n - \mu_j\|^2/2\epsilon)}.$$

- If $\epsilon \to 0$, the term with smallest $\|x_n - \mu_j\|$ tends to zero most slowly.
- Thus

$$p(z_n = k|\boldsymbol{X}, \theta) \to r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\| \\ 0 & \text{otherwise} \end{cases}$$

$\varepsilon \to 0$

Relationship to K-Means

Recall: $\mathbb{E}\left(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta)|\boldsymbol{X}, \theta^{\text{old}}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|\boldsymbol{X}, \theta^{\text{old}}) \log(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k))$.

As $\epsilon \to 0$, we have

$$p(z_n = k|\boldsymbol{X}, \theta) \to r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\| \\ 0 & \text{otherwise} \end{cases}$$

which gives

$$\mathbb{E}\Big(\log p(\boldsymbol{X}, \boldsymbol{Z}|\theta)\Big|\boldsymbol{X}, \theta^{\text{old}}\Big) \to -\frac{1}{2\epsilon^2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 + \text{const.}$$

- In the limit, maximizing the expected log-likelihood is equivalent to minimizing the distortion measure in the K-means algorithm.
- The EM-algorithm is slower but more flexible and accurate.

EM as Bound Optimization (ELBO)

The goal of EM is to maximize the log marginal likelihood of the observed data:

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) = \sum_{n=1}^{N} \log \left(\sum_{z_n} p(x_n, z_n|\theta) \right)$$

Introduce an arbitrary distribution $q_n(z_n)$ for each hidden variable. Using Jensen's inequality:

$$\begin{aligned}
\ell(\theta) &= \sum_{n=1}^{N} \log \left(\sum_{z_n} q_n(z_n) \frac{p(x_n, z_n|\theta)}{q_n(z_n)} \right) \\
&\geqslant \sum_{n=1}^{N} \sum_{z_n} q_n(z_n) \log \frac{p(x_n, z_n|\theta)}{q_n(z_n)} \\
&= \sum_{n=1}^{N} \Big(\mathbf{E}_{q_n}[\log p(x_n, z_n|\theta)] + \mathbb{H}(q_n) \Big) \triangleq \sum_{n=1}^{N} \mathcal{L}(\theta, q_n|x_n)
\end{aligned}$$

- This $\mathcal{L}(\theta, q_n|x_n)$ is exactly the **Evidence Lower Bound (ELBO)**!

E-step, M-step, and Variational EM

We can rewrite the ELBO for a single data point using the KL divergence:

$$\mathcal{L}(\theta, q_n|x_n) = \log p(x_n|\theta) - \mathrm{KL}(q_n(z_n) \;||\; p(z_n|x_n, \theta))$$

- **E-step**: Maximize the lower bound wrt q_n. This requires minimizing $\mathrm{KL}(q_n||p)$.
  - Setting $q_n^*(z_n) = p(z_n|x_n, \theta)$ gives $\mathrm{KL} = 0$. This ensures the ELBO is a **tight** lower bound.

- **M-step**: Maximize the bound wrt θ, using the fixed q_n^* computed in the E-step.
  - Since the entropy $\mathbb{H}(q_n^*)$ is constant wrt θ, we are left maximizing:

$$\theta^{\text{new}} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbf{E}_{q_n^*}[\log p(x_n, z_n|\theta)]$$

  - This is exactly the $Q(\theta, \theta^{\text{old}})$ function we saw earlier!

Connection to Variational Inference

If the exact posterior $p(z_n|x_n, \theta)$ is intractable, we restrict q_n to a tractable approximate family. This yields a non-tight lower bound. This generalized version of EM is known as **Variational EM**.

Summary

- EM algorithm is a classical method in statistics.
- It can be used in the presence of latent variables.
- When applied to Gaussian mixtures, compared to k-means, it captures the covariance structure of the data.
- Variational inference can be used in the E-step.