1921 UN SECOLO DI STORIA DAVANTI A NOI 2021

UNIVERSITÀ CATTOLICA del Sacro Cuore

MIXTURE MODELS AND EXPECTATION MAXIMIZATION ALGORITHM

K-MEANS CLUSTERING

K-MEANS CLUSTERING

We start with a recap of the K-means algorithm for clustering. Assume that we observe a D-variate data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^D \; \forall i$, with $D \geq 1$.

Goal. Partition the dat ainto K clusters (with K known) so that data points inside the same cluster have smaller distances with respect to data points in different clusters.
Formally, let $\boldsymbol{\mu}_k$ identify the center of cluster k. We want to identify $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and find a cluster assignment for each data point in order to minimize:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

K-MEANS CLUSTERING

We start with a recap of the K-means algorithm for clustering. Assume that we observe a D-variate data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^D \; \forall i$, with $D \geq 1$.

Goal. Partition the dat ainto K clusters (with K known) so that data points inside the same cluster have smaller distances with respect to data points in different clusters.
Formally, let $\boldsymbol{\mu}_k$ identify the center of cluster k. We want to identify $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and find a cluster assignment for each data point in order to minimize:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

Distortion measure:
Sum of square distances between data and the assigned centers.

Cluster membership of ith data point:

- r_{ik}=1 if $\boldsymbol{x}_i$ is assigned to cluster k
- r_{ik}=0 otherwise

K-MEANS CLUSTERING

We start with a recap of the K-means algorithm for clustering. Assume that we observe a D-variate data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^D \; \forall i$, with $D \geq 1$.

Goal. Partition the dat ainto K clusters (with K known) so that data points inside the same cluster have smaller distances with respect to data points in different clusters.
Formally, let $\boldsymbol{\mu}_k$ identify the center of cluster k. We want to identify $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and find a cluster assignment for each data point in order to minimize:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

To solve the minimization problem we need to find values for r_{ik} and $\boldsymbol{\mu}_k$, jointly.

K-MEANS CLUSTERING

- If we know $\boldsymbol{\mu}_k$, r_{ik} can be chosen to be one for the closest center to data point $\mathbf{x}_i$. Indeed, J is linear in r_{ik}. The terms involving different i are independent so we can directly minimize $\forall i$:

$$\sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

which gives:

$$r_{ik} = \begin{cases} 1, \text{ if } k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ 0, \text{ otherwise} \end{cases}$$

K-MEANS CLUSTERING

- If we know $\boldsymbol{\mu}_k$, r_{ik} can be chosen to be one for the closest center to data point $\mathbf{x}_i$. Indeed, J is linear in r_{ik}. The terms involving different i are independent so we can directly minimize $\forall i$:

$$\sum_{k=1}^{K} r_{ik}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

which gives:

$$r_{ik} = \begin{cases} 1, \text{ if } k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ 0, \text{ otherwise} \end{cases}$$

- If we know r_{ik}, $\boldsymbol{\mu}_k$ are the cluster means. Indeed minumization of J gives:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2\sum_{i=1}^{n} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k) = \mathbf{0}$$

$$\Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

K-MEANS CLUSTERING

- If we know $\boldsymbol{\mu}_k$, r_{ik} can be chosen to be one for the closest center to data point $\mathbf{x}_i$. Indeed, J is linear in r_{ik}. The terms involving different i are independent so we can directly minimize $\forall i$:

$$\sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

which gives:

$$r_{ik} = \begin{cases} 1, \text{ if } k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ 0, \text{ otherwise} \end{cases}$$

1. Start with some initial values of $\boldsymbol{\mu}_k$, and find the r_{ik}.

K-MEANS CLUSTERING

2. Fix r_{ik} as computed from last iteration and re-compute the $\boldsymbol{\mu}_k$.

- If we know r_{ik}, $\boldsymbol{\mu}_k$ are the cluster means. Indeed minumization of J gives:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2 \sum_{i=1}^{n} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k) = \mathbf{0}$$

$$\Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

K-MEANS CLUSTERING

Iterative method

- If we know $\boldsymbol{\mu}_k$, r_{ik} can be chosen to be one for the closest center to data point $\mathbf{x}_i$. Indeed, J is linear in r_{ik}. The terms involving different i are independent so we can directly minimize $\forall i$:

$$\sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

which gives:

$$r_{ik} = \begin{cases} 1, \text{ if } k = \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ 0, \text{ otherwise} \end{cases}$$

- If we know r_{ik}, $\boldsymbol{\mu}_k$ are the cluster means. Indeed minumization of J gives:

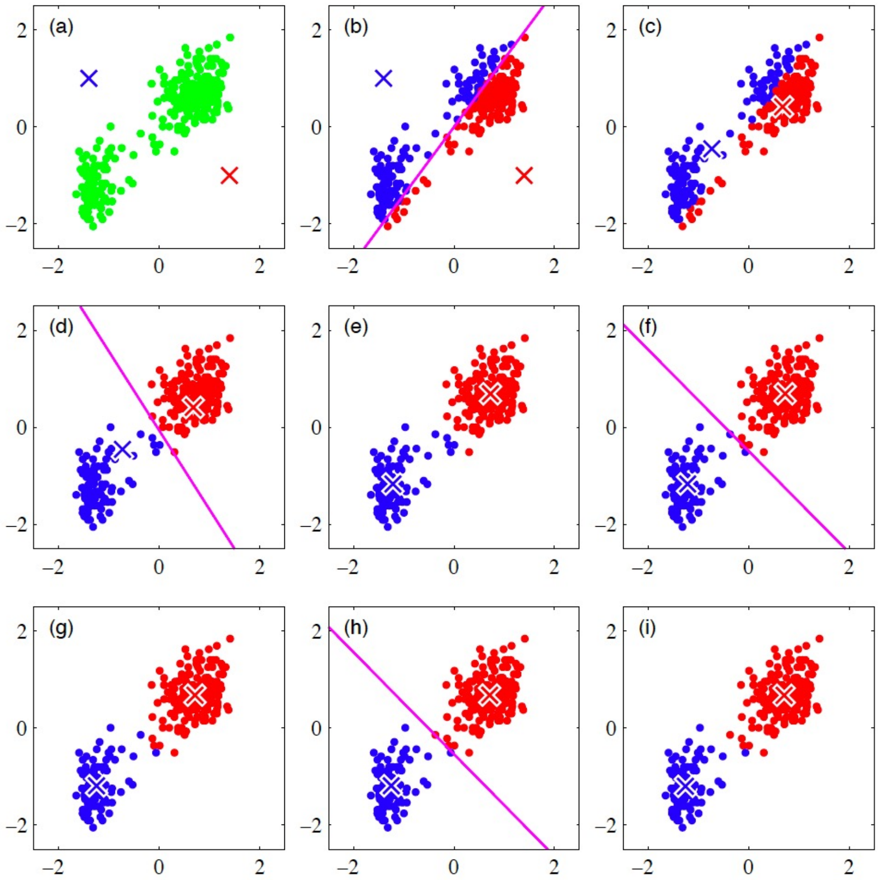
$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2 \sum_{i=1}^{n} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k) = \mathbf{0}$$

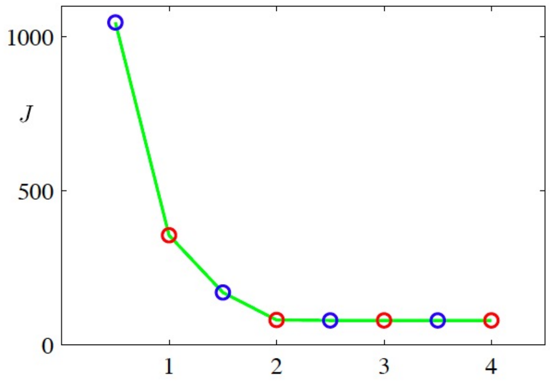
$$\Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

K-MEANS CLUSTERING

- Each iteration reduces the value of J. So the iterative algorithm converges.
- The algorithm might converge to a local minimum of J instead of the global one.
- The point of convergence (global or local minimum) depends on the initialization.

K-MEANS CLUSTERING





MIXTURE OF GAUSSIANS

MIXTURE OF GAUSSIANS

A mixture of Gaussian distributions can be written as a linear superposition of Gaussian pdfs:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

Mixing probabilities, $\sum_{k=1}^{K} \pi_k = 1$

Gaussian pdf with mean and covariance matrix $\boldsymbol{\mu}_k, \Sigma_k$

MIXTURE OF GAUSSIANS

Alternative way to define a mixture.
Introduce a binary variable $\mathbf{z} \in \mathbb{R}^K$ having only one element equal to one and all other elements equal to zero:

$$\forall k \in \{1, \dots, K\} \; z_k \in \{0, 1\}, \quad \text{and} \quad \sum_{k=1}^{K} z_k = 1.$$

The vector $\mathbf{z}$ has K possible states, according to which element is non-zero. Assume:

$$\mathbb{P}(z_k = 1) = \pi_k, \; \text{with } \pi_k \in [0, 1], \sum_{k=1}^{K} \pi_k = 1.$$

This is equivalent to:

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

MIXTURE OF GAUSSIANS

Let $\mathbf{X}$ be a random variable in $\mathbb{R}^D$ with conditional distribution

$$p(\mathbf{x}|z_k = 1) = N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k).$$

This is also equivalent to

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k}.$$

The joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is of course

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

and the marginal distribution of $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

MIXTURE OF GAUSSIANS

Let $\mathbf{X}$ be a random variable in $\mathbb{R}^D$ with conditional distribution

$$p(\mathbf{x}|z_k = 1) = N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k).$$

This is also equivalent to

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)^{z_k}.$$

The joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is of course

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

and the marginal distribution of $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

Equivalent formulation of a Gaussian mixture involving the latent variables ***z***.

MIXTURE OF GAUSSIANS

We define an additional useful quantity:

$$\gamma(z_k) = \mathbb{P}(z_k = 1|\mathbf{x}) = \frac{p(\mathbf{x}|z_k = 1)\mathbb{P}(z_k = 1)}{\sum_{j=1}^{K} p(\mathbf{x}|z_k = 1)\mathbb{P}(z_j = 1)}$$

$$= \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \Sigma_j)}$$

Posterior probability of component k

Prior probability of component k

MIXTURE OF GAUSSIANS

We define an additional useful quantity:

$$\gamma(z_k) = \mathbb{P}(z_k = 1|\mathbf{x}) = \frac{p(\mathbf{x}|z_k = 1)\mathbb{P}(z_k = 1)}{\sum_{j=1}^{K} p(\mathbf{x}|z_k = 1)\mathbb{P}(z_j = 1)}$$

$$= \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}|\boldsymbol{\mu}_j, \Sigma_j)}$$

Posterior probability of component k

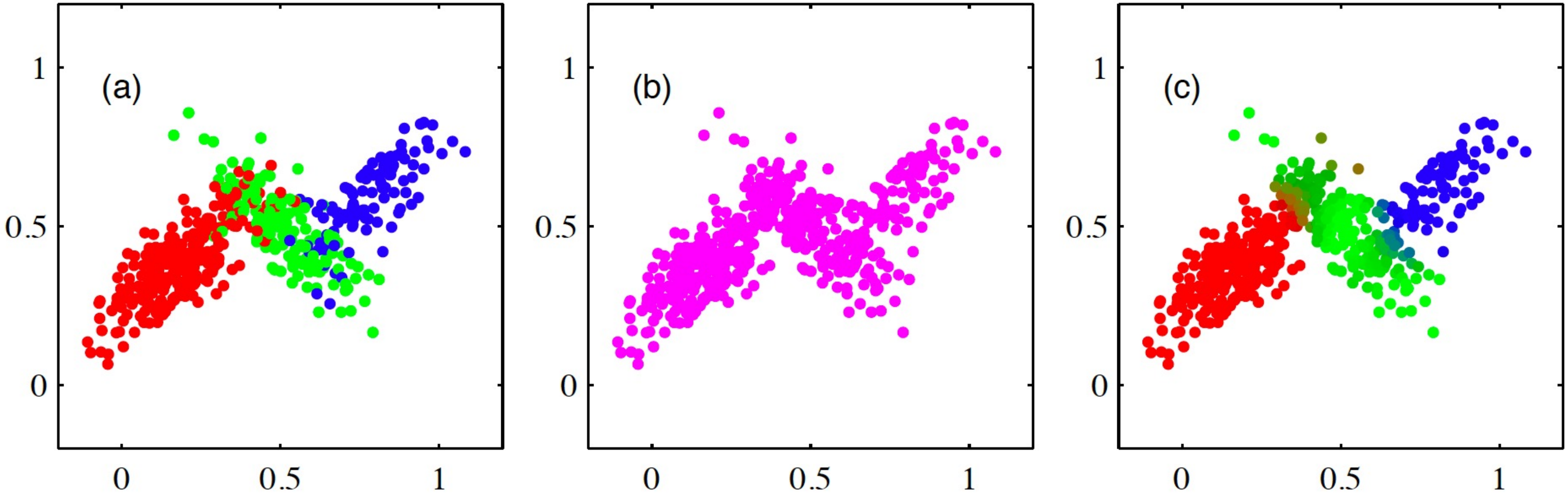
Prior probability of component k

The quantity $\gamma(z_k)$ is also called responsibility: responsibility that component k takes for explaining the observed $\mathbf{x}_k$.

MIXTURE OF GAUSSIANS

Random sample ($K = 3$):

- First generate $\mathbf{z}$.
- Second generate $\mathbf{x}|\mathbf{z}$.



Joint distribution | Marginal of x | Responsibilities

MAXIMUM LIKELIHOOD ESTIMATION

Assume now that we observe $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and that we want to model them as a mixture. Let X be the $(n \times D)$ data matrix. Similarly, denote as Z the (unobserved) $(n \times D)$ matrix of the latent variables. The log-likelihood is the following:

$$\ell(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \ln p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

The parameters of the model $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ can be obtained by maximizing the log-likelihood.

MAXIMUM LIKELIHOOD ESTIMATION

For simplicity, assume $D = 1$ and $K = 2$. Let $\mu_1 = \overline{x}$, $\sigma_1 = s_x$ for the first component and $\mu_2 = x_j$, $\sigma_2 \to 0$ for some $j = 1, \dots, n$. Assume $\pi_1, \pi_2 > 0$.
In this example, the log likelihood is

$$\sum_{i=1}^{n} \ln\left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right)$$

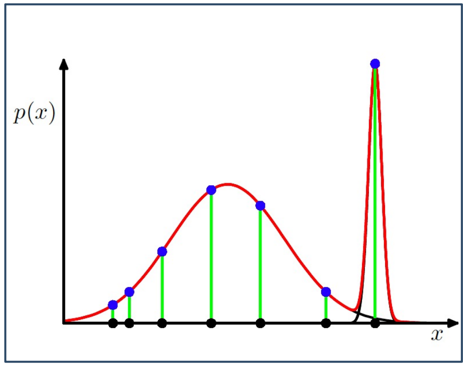
$$= \sum_{i \neq j} \ln\left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right) +$$

$$+ \ln\left(\pi_1 N(x_j|\overline{x}, s_x) + \pi_2 N(x_j|x_j, \sigma_2^2)\right)$$

MAXIMUM LIKELIHOOD ESTIMATION

For simplicity, assume $D = 1$ and $K = 2$. Let $\mu_1 = \overline{x}$, $\sigma_1 = s_x$ for the first component and $\mu_2 = x_j$, $\sigma_2 \to 0$ for some $j = 1, \dots, n$. Assume $\pi_1, \pi_2 > 0$.
In this example, the log likelihood is

$$\sum_{i=1}^{n} \ln\left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right)$$
$$= \sum_{i \neq j} \ln\left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right) +$$
$$+ \ln\left(\pi_1 N(x_j|\overline{x}, s_x) + \pi_2 N(x_j|x_j, \sigma_2^2)\right)$$



The second term is infinite for $\sigma_2 \to 0$. So, the max of the log likelihood is infinite, and it correspond to a singular solution!

MAXIMUM LIKELIHOOD ESTIMATION

For simplicity, assume $D = 1$ and $K = 2$. Let $\mu_1 = \overline{x}$, $\sigma_1 = s_x$ for the first component and $\mu_2 = x_j$, $\sigma_2 \to 0$ for some $j = 1, \dots, n$. Assume $\pi_1, \pi_2 > 0$.
In this example, the log likelihood is

$$\sum_{i=1}^{n} \ln \left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right)$$

$$= \sum_{i \neq j} \ln \left(\pi_1 N(x_i|\overline{x}, s_x) + \pi_2 N(x_i|x_i, \sigma_2^2)\right) +$$

$$+ \ln \left(\pi_1 N(x_j|\overline{x}, s_x) + \pi_2 N(x_j|x_j, \sigma_2^2)\right)$$

The second term is infinite for $\sigma_2 \to 0$. So, the max of the log likelihood is infinite, and it correspond to a singular solution!

This holds in general, for $D \geq$ and $K \geq 2$: if we have at least two components in the mixture, the likelihood cannot be directly maximised. We should instead seek for non-singular local maxima.

EXPECTATION MAXIMIZATION ALGORITHM FOR GAUSSIAN MIXTURES

EM FOR GAUSSIAN MIXTURES

Maximizing the log-likelihood for finding $\boldsymbol{\mu}_k$.

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_k} = -\sum_{i=1}^{n} \frac{\pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k N(\mathbf{x}_i|\boldsymbol{\mu}_j, \Sigma_j)} \sum_{k=1}^{K} (\mathbf{x}_i - \boldsymbol{\mu}_k)$$

$$= \gamma(z_{ik}) \sum_{k=1}^{K} (\mathbf{x}_i - \boldsymbol{\mu}_k) = 0$$

$$\Rightarrow \boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i \qquad (1)$$

$$n_k = \sum_{i=1}^{n} \gamma(z_{ik})$$

The mean of component k is a weighted mean of all data points where the weights are the responsibilities.

EM FOR GAUSSIAN MIXTURES

Maximizing the log-likelihood for finding Σ_k.

$$\frac{\partial \ell}{\partial \Sigma_k} = 0$$

$$\Rightarrow \Sigma_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)' \quad (2)$$

The variance of component k is a weighted mean of the contributions to variance of all data points where the weights are the responsibilities.

EM FOR GAUSSIAN MIXTURES

Maximizing the log-likelihood for finding π_k.
In the case of π_k, we cannot directly look for a stationary point of the log-likelihood, since we also have the constraint $\sum_{k=1}^{K} \pi_k = 1$. We use Lagrange multipliers: we need to maximize

$$\ell(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\begin{cases} \sum_{k=1}^{K} \pi_k - 1 = 0 \\ \frac{\partial \ell}{\partial \pi_k} + \lambda = 0 \end{cases}$$

$$\Rightarrow \boxed{\begin{cases} \lambda = -n \\ \pi_k = \frac{n_k}{n} \end{cases}} \tag{3}$$

The mixing probabilities are the effective number of data points contributing to each component divided by n.

EM FOR GAUSSIAN MIXTURES

$$\boldsymbol{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n}\gamma(z_{ik})\mathbf{x}_i$$

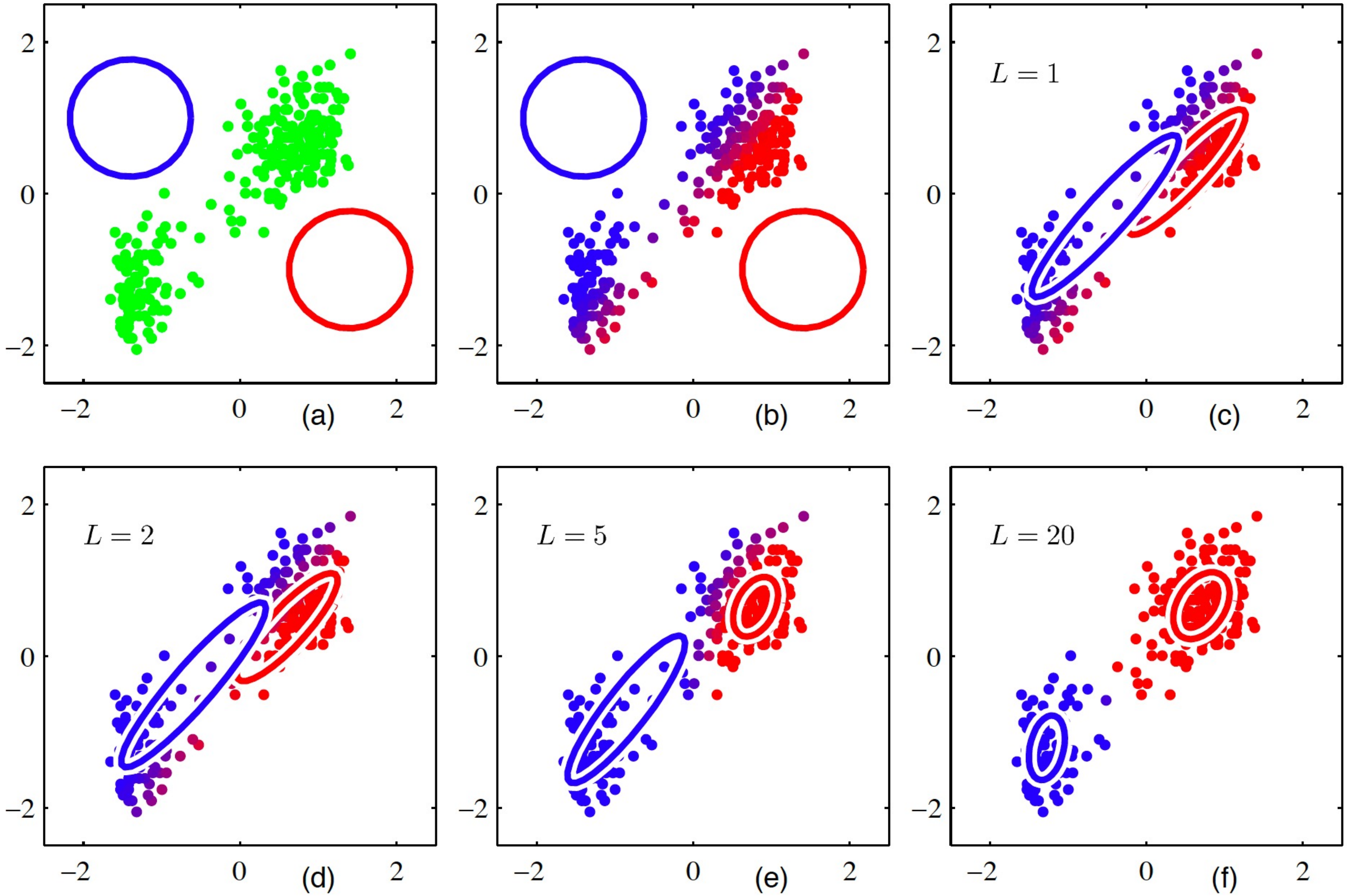
$$\Sigma_k = \frac{1}{n_k}\sum_{i=1}^{n}\gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)'$$

$$\begin{cases}\lambda = -n \\ \pi_k = \frac{n_k}{n}\end{cases}$$

Observe that (1) + (2) + (3) do not give a closed-form solution, since all terms are expressed as a function of the responsibilities $\gamma(z_{ik})$, which in turn depend on $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1,\dots,n}$ in a complex way.
However, they suggest an iterative way for finding a solution:

1. Start choosing values for $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1,\dots,n}$.
2. **E step:** use the current values of $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1,\dots,n}$ to evaluate $\gamma(z_{ik})$.
3. **M step:** re-estimate all parameters using (1), (2), (3), and the current value of $\gamma(z_{ik})$.
4. Check for convergence. If convergence is not met, return to step 2.

EM FOR GAUSSIAN MIXTURES



EM VS K-MEANS

k-means. It is based on a hard assignment: each data point only belongs to one cluster.

EM. It is based on a soft assignment: at the end of the algorithm, we have posterior probabilities of belonging to one mixture component, that can be viewed as posterior probabilities of belonging to one cluster.

Relation between the two methods.
Assume that $\Sigma_k = \epsilon I$, where ϵ assume the same value for all mixture components. Assume also that ϵ is a fixed known constant (we don't want to estimate it). The Gaussian pdf becomes

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}$$

EM VS K-MEANS

E-step:
The responsibilities are:

$$\gamma(z_{ik}) = \frac{\pi_k \exp(-\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2/2\epsilon)}{\sum_{j=1}^{K} \pi_j \exp(-\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2/2\epsilon)}$$

Consider now the limit of $\gamma(z_{ik})$ for $\epsilon \to 0$. In the denominator, the term $\tilde{j}$ for which $\|\mathbf{x}_i - \boldsymbol{\mu}_j\|$ is the smallest goes to 0 most slowly. So:

$$\lim_{\epsilon \to 0} \gamma(z_{ik}) = \begin{cases} 0 & \forall i \neq \tilde{j} \\ 1 & i = \tilde{j} \end{cases}$$

So, each data point is assigned to the cluster with the closest mean!
Note that this is independent on the π_k, as long as they all are strictly positive.

EM VS K-MEANS

M-step:

We only need to find $\boldsymbol{\mu}_k$, since the E-step does not depend on the π_k, and ε is fixed. In this case, we have trivially from (1):

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i$$

Where $n_k = \sum_{i=1}^{n} \gamma(z_{ik})$ is the number of points assigned to cluster k, since $\gamma(z_{ik})$ are either zero or one. Hence $\boldsymbol{\mu}_k$ are exactly the cluster means (as in K-means).

EM VS K-MEANS

M-step:
We only need to find $\boldsymbol{\mu}_k$, since the E-step does not depend on the π_k, and ε is fixed. In this case, we have trivially from (1):

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma(z_{ik}) \mathbf{x}_i$$

Where $n_k = \sum_{i=1}^{n} \gamma(z_{ik})$ is the number of points assigned to cluster k, since $\gamma(z_{ik})$ are either zero or one. Hence $\boldsymbol{\mu}_k$ are exactly the cluster means (as in K-means).

K-means is the limit of an EM algorithm obtained when the variance is constant for each component and goes to zero (in order to induce hard assignment).

EXPECTATION MAXIMIZATION ALGORITHM IN THE GENERAL CASE

GENERAL EM ALGORITHM

Given a joint distribution $p(X, Z|\boldsymbol{\theta})$ over observed variables X, latent variables Z and parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood with respect to θ. The general EM algorithm work as follows.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{old}$.

Expectation

2. **E-step.** Evaluate the posterior probabilities $p(Z|X, \boldsymbol{\theta}^{old})$. Use it to find the expectation of the log-likelihood, that is

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{Z} p(Z|X, \boldsymbol{\theta}^{old}) \ln p(X, Z|\boldsymbol{\theta})$$

3. **M-step.** Maximize the expected log-likelihood finding a new set of parameters $\boldsymbol{\theta}^{new}$:

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

4. Check for convergence of the method (in either the log-likelihood or in the parameter values). If the convergence criterion is met, stop. Otherwise, set

$$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

and return to step 2.

GENERAL EM ALGORITHM

Given a joint distribution $p(X, Z|\boldsymbol{\theta})$ over observed variables X, latent variables Z and parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood with respect to θ. The general EM algorithm work as follows.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{old}$.

2. **E-step.** Evaluate the posterior probabilities $p(Z|X, \boldsymbol{\theta}^{old})$. Use it to find the expectation of the log-likelihood, that is

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{Z} p(Z|X, \boldsymbol{\theta}^{old}) \ln p(X, Z|\boldsymbol{\theta})$$

Maximization

3. **M-step.** Maximize the expected log-likelihood finding a new set of parameters $\boldsymbol{\theta}^{new}$:

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

4. Check for convergence of the method (in either the log-likelihood or in the parameter values). If the convergence criterion is met, stop. Otherwise, set

$$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

and return to step 2.

CONVERGENCE

Remind that our goal is to maximize

$$p(X|\boldsymbol{\theta}) = \sum_Z p(X, Z|\boldsymbol{\theta})$$

where X collects all the observed variables, and Z collects all the latent variables. Assume that the direct maximization of $p(X|\boldsymbol{\theta})$ is difficult (and might lead to singular solutions), while maximization of $p(X, Z|\boldsymbol{\theta})$ is significantly easier. Introduce an arbitrary distribution $q(Z)$ on the latent variables. First, we note that for arbitrary $q(Z)$ we have the decomposition:

$$\ln p(X|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

with:

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_Z q(z) \ln \left(\frac{p(X, Z|\boldsymbol{\theta})}{q(Z)} \right)$$

Expected likelihood under q(Z)

$$\mathrm{KL}(q\|p) = -\sum_Z q(Z) \ln \left(\frac{p(Z|X, \boldsymbol{\theta})}{q(Z)} \right)$$

Kullback-Leibler divergence between q(Z) and p(Z|X,θ)

CONVERGENCE

To prove the decomposition, observe that:

$$\ln p(X, Z|\boldsymbol{\theta}) = \ln p(Z|\boldsymbol{\theta}) + \ln p(X|\boldsymbol{\theta})$$

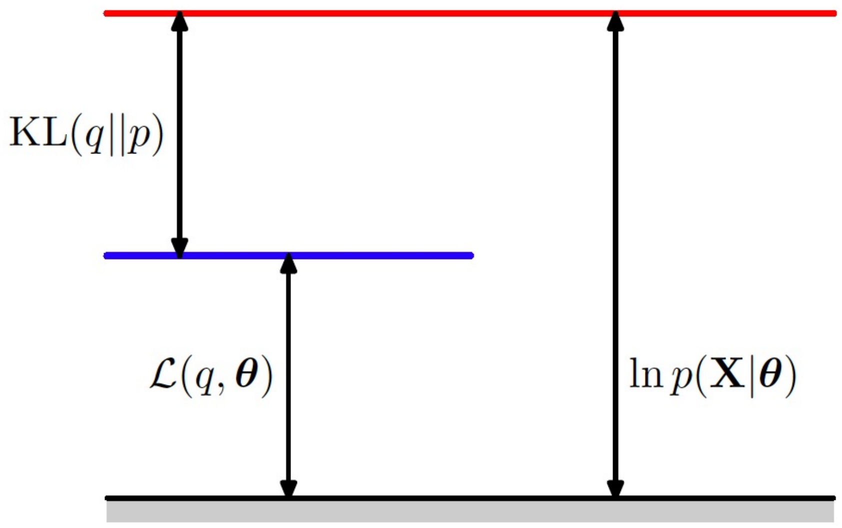
$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_Z q(Z) \left[\ln p(Z|\boldsymbol{\theta}) + \ln p(X|\boldsymbol{\theta}) - \ln q(Z)\right] \\
&= \sum_Z q(Z) \ln p(X|\boldsymbol{\theta}) + \sum_Z q(Z) \ln \left(\frac{p(Z|\boldsymbol{\theta})}{q(Z)}\right) \\
&= \ln p(X|\boldsymbol{\theta}) \sum_Z q(Z) - \mathrm{KL}(q\|p) \\
&= \ln p(X|\boldsymbol{\theta}) - \mathrm{KL}(q\|p)
\end{aligned}$$

CONVERGENCE

$$\ln p(X|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q \| p)$$

Remember that the Kullback-Leibler divergence between two probability distributions is a measure of distance between the two distributions. In particular, $\mathrm{KL}(q\|p) \geq 0$, and

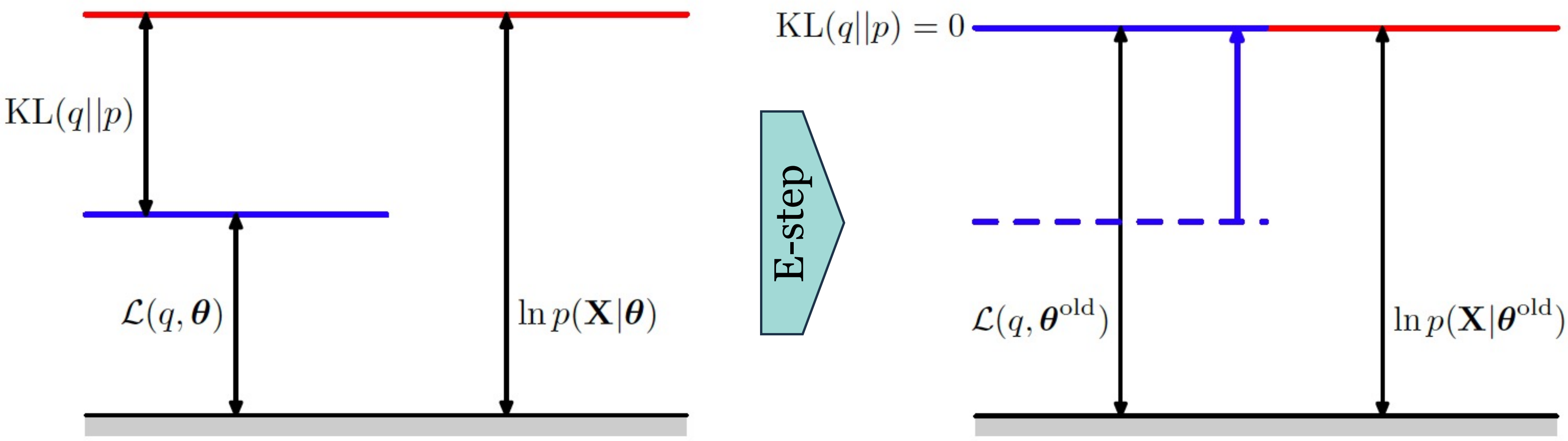
$$\mathrm{KL}(q\|p) = 0 \iff q(Z) = p(Z|X, \boldsymbol{\theta})$$



CONVERGENCE

Now, come back to the EM algorithm. Suppose that the current value of the parameters is $\boldsymbol{\theta}^{old}$. The distribution $q(Z)$ will be our estimate of the posterior probabilities $p(Z|X, \boldsymbol{\theta})$.

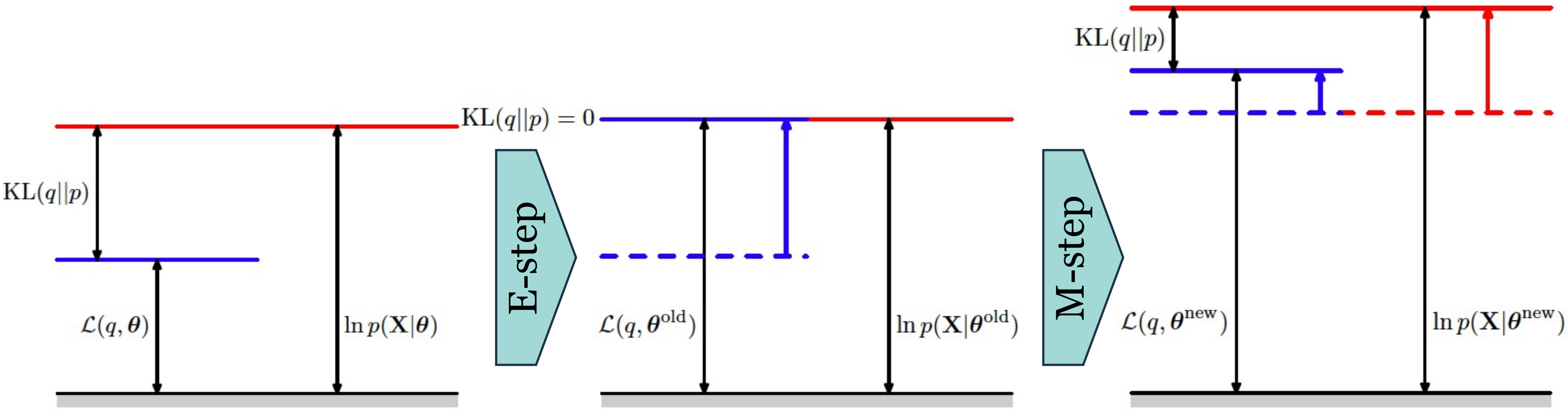
E-step.
In the E-step, $q(Z) = p(Z|X, \boldsymbol{\theta}^{old})$. This is equivalent to maximizing $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ with respect to $q(Z)$. Indeed, $\ln p(X|\boldsymbol{\theta}^{old})$ does not depend on $q(Z)$, so $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ is maximized when $\mathrm{KL}(q||p) = 0$.



CONVERGENCE

M-step.
Now, $q(Z)$ is held fixed, and $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$, obtaining new parameters $\boldsymbol{\theta}^{new}$. This will cause $\mathcal{L}(q, \boldsymbol{\theta})$ to increase, and in particular $\mathcal{L}(q, \boldsymbol{\theta}^{neq}) \geq \mathcal{L}(q, \boldsymbol{\theta}^{old})$. In addition, we will also have a non-zero K-L divergence, since $q(Z) = p(Z|X, \boldsymbol{\theta}^{old}) \neq p(Z|X, \boldsymbol{\theta}^{new})$. So:

$$\ln p(X|\boldsymbol{\theta}^{new}) \geq \ln p(X|\boldsymbol{\theta}^{old})$$



EM ALGORITHM: TAKE HOME MESSAGE

- Algorithm defined in general to find parameters of a model where we have both observed variables and unobserved latent variables.
  - Mixtures of Gaussians
  - Bernoulli mixtures
  - Missing data
  - Hidden Markov Models
  - …
- In the case of Mixtures of Gaussians, it has an easy formulation. In such a case, we can show that it is closely relater to *k*-means clustering.
- We can prove convergence in the general case.