15. Clustering

STA3142 Statistical Machine Learning

Kibok Lee

Assistant Professor of

Applied Statistics / Statistics and Data Science

Apr 30, 2024

Announcement

- No class @ week 10 (May 7, 9)

- No class & no final exam @ week 16
  - Assignment 5 is the replacement
  - You should submit A5 for your attendance @ week 16

Midterm Grading

- Ongoing; we are trying to release it this week

- If you don’t agree with the Honor code – your midterm score is 0.
  - If you didn’t write the pledge and your name on the first page properly, you receive 0 point.
  - If you did so, your submission will be graded after you complete it.
  - Your academic career is built on academic honesty.

Post-Midterm

- Let’s solve some questions that you felt difficult.
- A survey will be out together with midterm results.
  - To determine questions we are going to solve together

- If you feel you didn’t do well,
  - You are not alone; other students would too.
  - Problem-solving skills can be improved by practice.
    - E.g., Derive ML algorithms we have learned from scratch
    - Don’t just memorize them

Assignment 3

- Due **Friday 5/3, 11:59pm**
- Topics
  - (Programming) K-Nearest Neighbors
  - (Math) MLE vs. MAP
  - (Math) Kernel Methods
  - (Math/Programming) SVM Primal
- Please read the instruction carefully!
  - Submit one pdf and one zip file separately
  - Write your code only in the designated spaces
  - Do not import additional libraries
  - …
- If you feel difficult, consider to take **option 2**.

Recap: Machine Learning Tasks

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Density estimation
  - Embedding / Dimensionality reduction
- Reinforcement Learning
  - Learning to act
    (e.g., robot control, decision making, etc.)

Recap: Supervised Learning

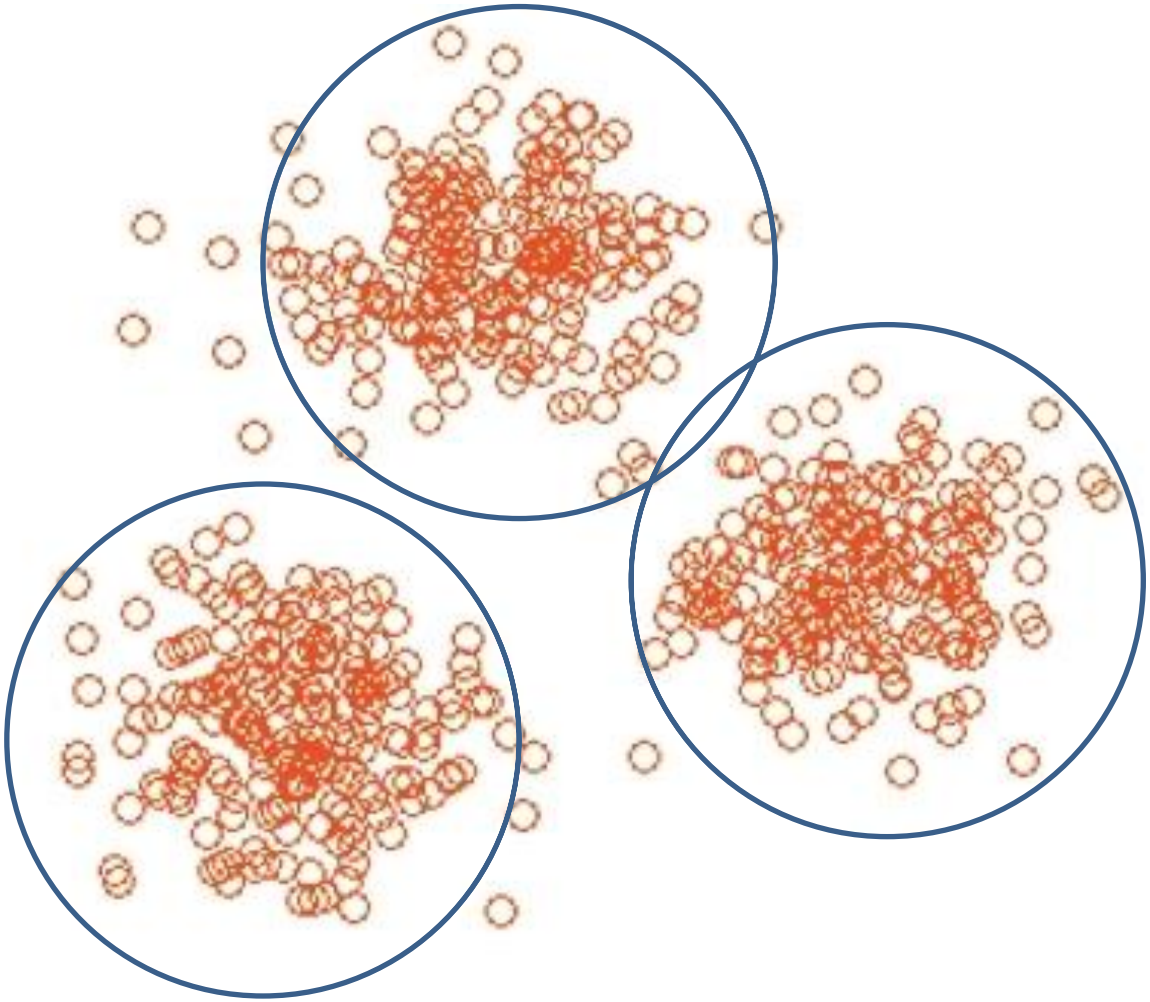
- Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where
  - $x_i \in \mathcal{X}$: input (feature)
  - $y_i \in \mathcal{Y}$: output (label)
- A black box ML algorithm produces a prediction function $h: \mathcal{X} \to \mathcal{Y}$, such that $h(x)$ can predict the y values for all x
  - Not only for all training data $x_i \in D$,
    but also for unseen test data $x^* \in \mathcal{X}$.
- Labels could be discrete or continuous
  - Discrete labels: **classification**
  - Continuous labels: **regression**

Unsupervised Learning

- Given a dataset $D = \{x_1, \ldots, x_n\}$ without any labels, learning the underlying **structure** or **distribution** of the data
  - Clustering
  - Probability distribution (density)
  - Generating data
  - Embedding & neighborhood relations
- "Learning without teacher (supervision)"

Unsupervised Learning: Clustering

- Grouping into similar examples



Unsupervised Learning: Clustering

- Grouping into similar examples

Cat image is CC0 public domain
Dog image is CC0 Public Domain
Monkey image is CC0 Public Domain

Cat image is CC0 public domain
Monkey Image is CC0 public domain

Cat image is CC0 public domain
Dog image is CC0 public domain
Dog image is CC0 public domain

Slide Credit: Justin Johnson

Outline

- Expectation Maximization (EM)
  - K-Means
  - Gaussian Mixture Models (GMM)
- General View of EM

Expectation Maximization (EM)

- Iteratively learning parameters when data is not fully observed
- Suppose we have observed variables X and latent (hidden) variables Z
  - e.g., clustering: X: data, Z: cluster labels
- Iterate **E-steps** and **M-steps** until converged:
  - **E-step**: Inference about Z given X: $Q = P(Z|X)$
  - **M-step**: Update parameters with Q found at E-step
- EM algorithms for clustering:
  - K-Means (a special case of GMM)
  - Gaussian Mixture Models (GMM)

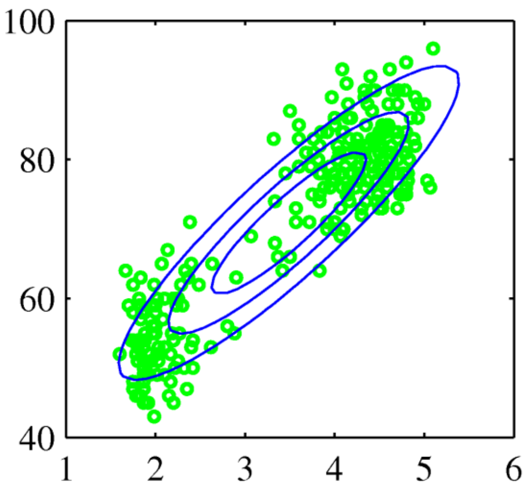
K-Means

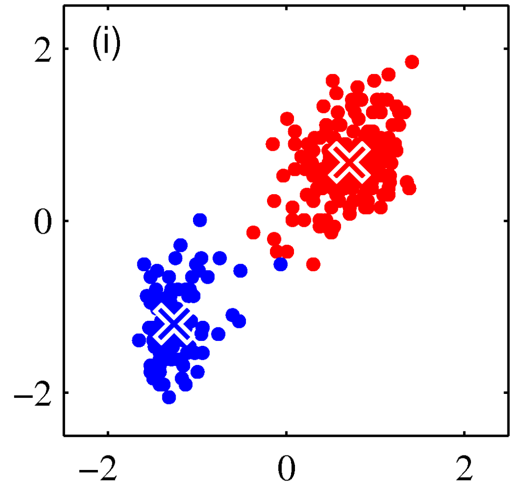
K-Means

- Given unlabeled data $x^{(n)}$ for $n = 1, \ldots, N$,

- Assume that each data belongs to one of the K clusters,

- How do we find the cluster labels?





K-Means: Formulation

- Cluster centers: $\boldsymbol{\mu}_k, k = 1, \dots, K$
- Indicator variables: $r_{nk} \in \{0,1\}, n = 1, \dots, N$
  - $r_{nk} = 1$ if $\mathbf{x}^{(n)}$ is in cluster k.
  - $r_{nj} = 0$ for all $j \neq k$.

- Minimize $J(r, \boldsymbol{\mu})$: sum of squared distances of points from the center of its assigned cluster.

$$J(r, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \left\| \mathbf{x}^{(n)} - \boldsymbol{\mu}_k \right\|^2$$

K-Means Algorithm

- Initialize the cluster centers arbitrarily.
- Repeat the following updates until convergence:
  1. **E-Step**:
  $$r := \operatorname*{argmin}_{r} J(r, \boldsymbol{\mu})$$
  2. **M-Step**:
  $$\boldsymbol{\mu} := \operatorname*{argmin}_{\boldsymbol{\mu}} J(r, \boldsymbol{\mu})$$

K-Means Algorithm

- Initialize the cluster centers arbitrarily.
- Repeat the following updates until convergence:
  1. **E-Step**: Cluster assignment
     - Assign each point to the closest center.

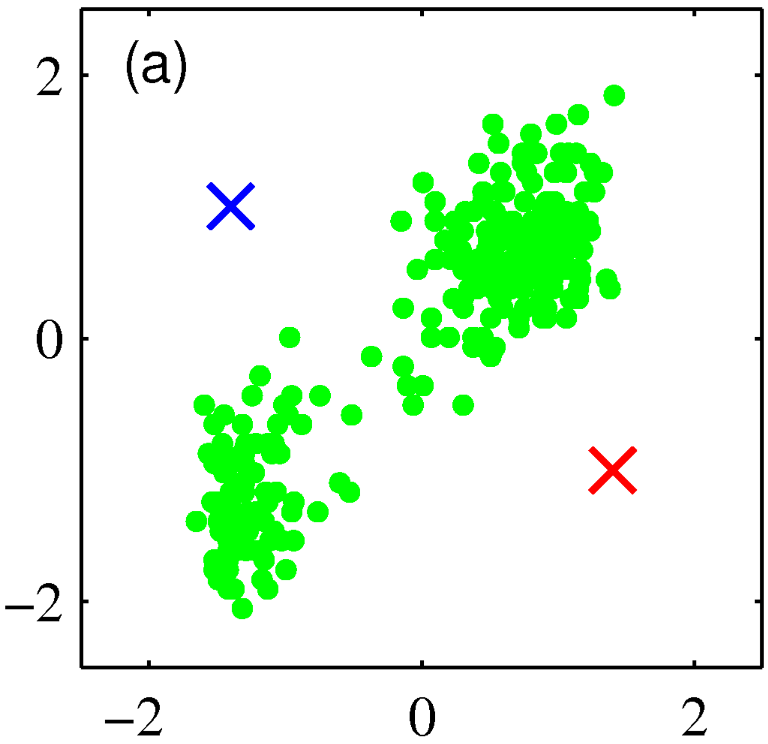
$$r_{nk} := \begin{cases} 1 & \text{if } k = \operatorname*{argmin}_j \left\| \mathbf{x}^{(n)} - \boldsymbol{\mu}_k \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

  2. **M-Step**: Parameter update
     - Update cluster centers

$$\boldsymbol{\mu} := \frac{\sum_{n=1}^{N} r_{nk} \mathbf{x}^{(n)}}{\sum_{n=1}^{N} r_{nk}}$$

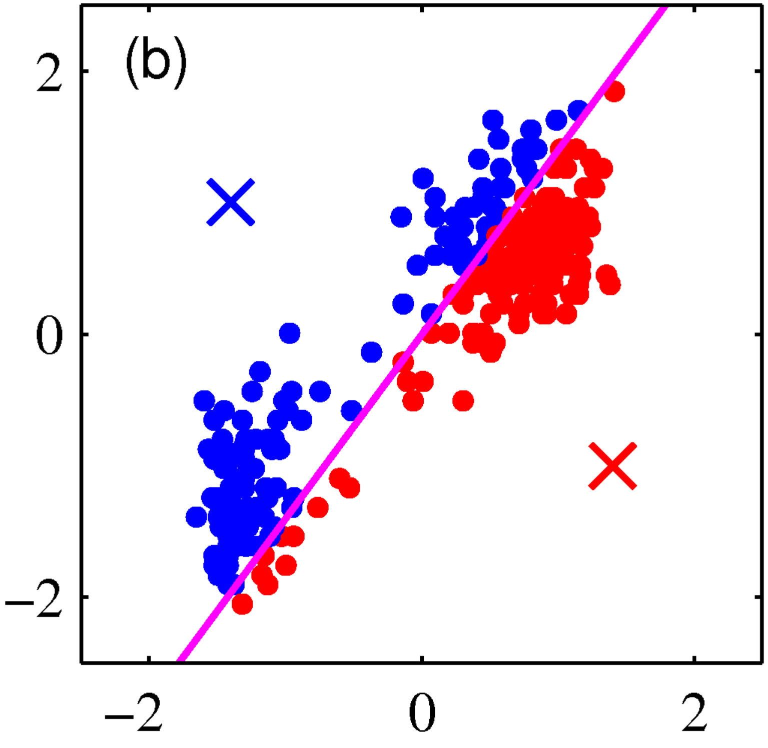
K-Means Example: Initialization

- Choose K and pick random means.
- In this example, $K = 2$.



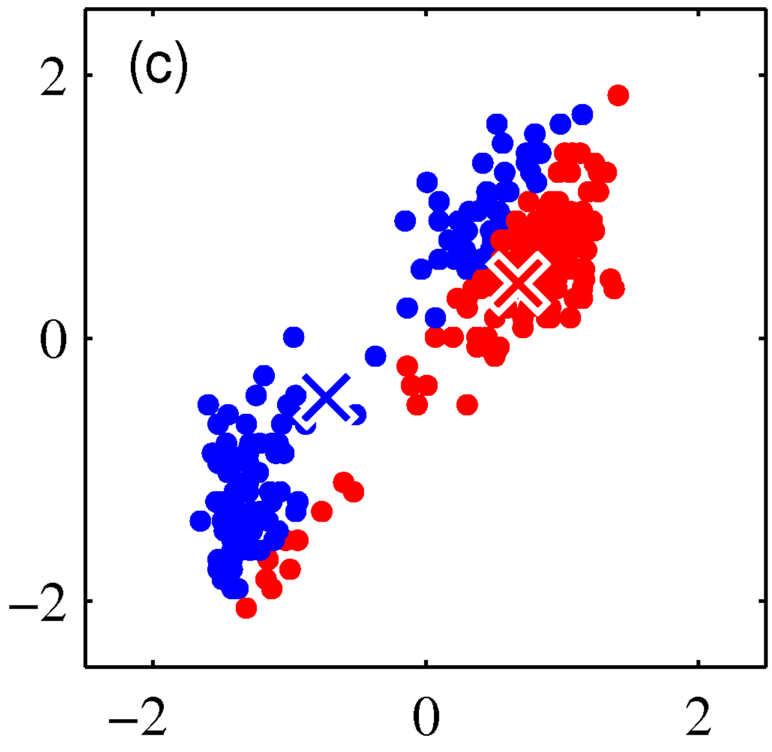
K-Means Example: 1st E-Step

- Assign each data to the nearest center.



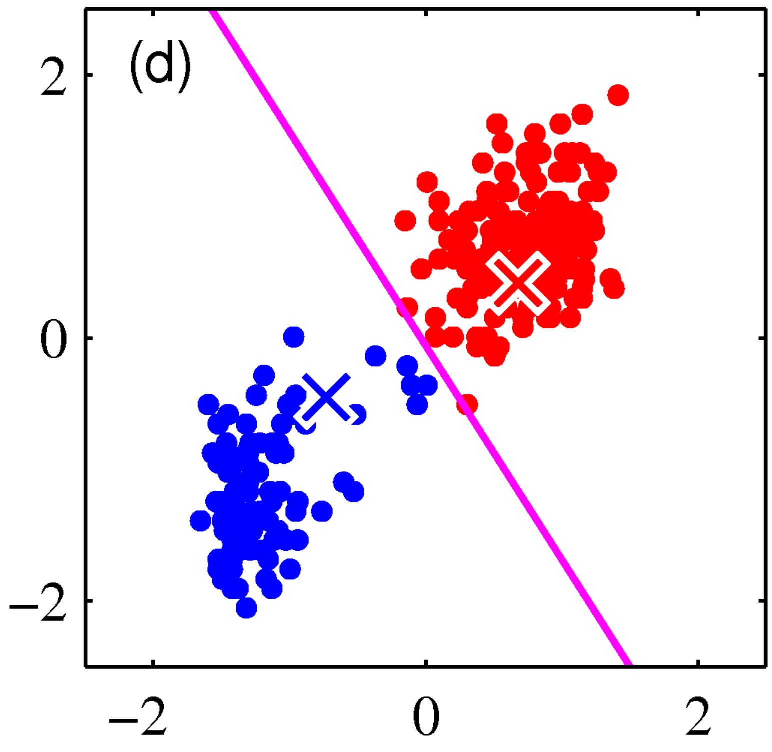
K-Means Example: 1st M-Step

- Compute new centers for each cluster.



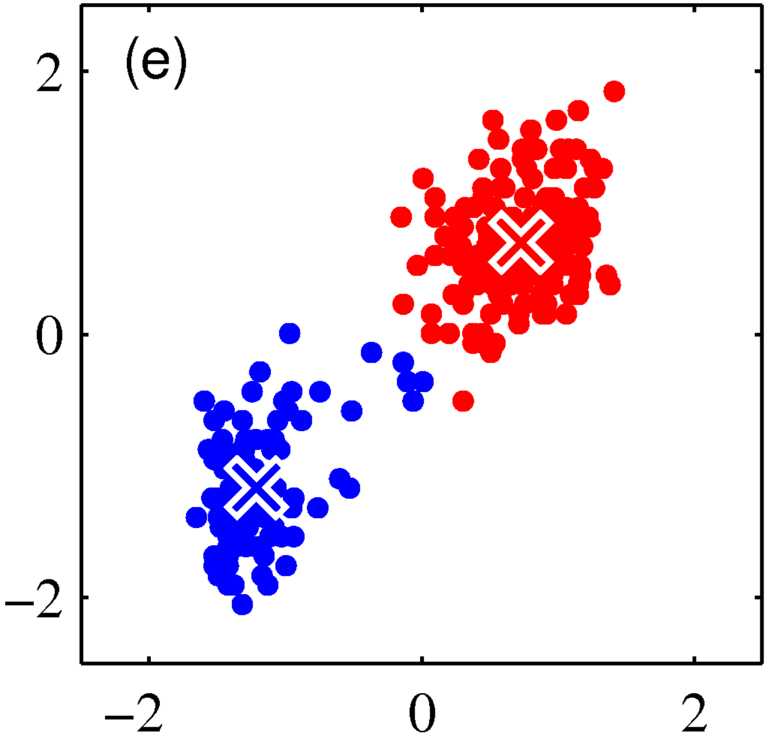
K-Means Example: 2nd E-Step

- Reassign data to the nearest center.



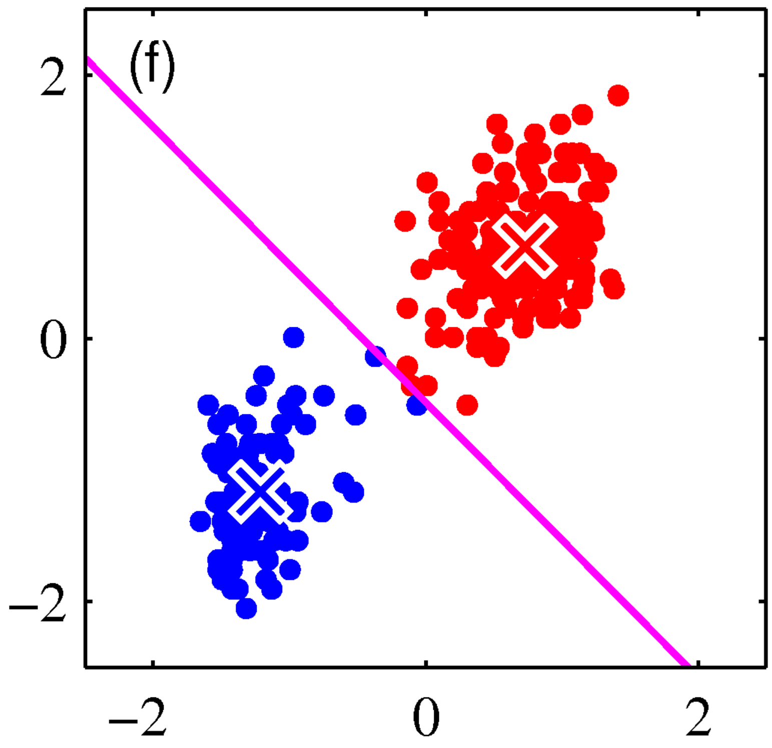
K-Means Example: 2nd M-Step

- Compute new centers for each cluster.



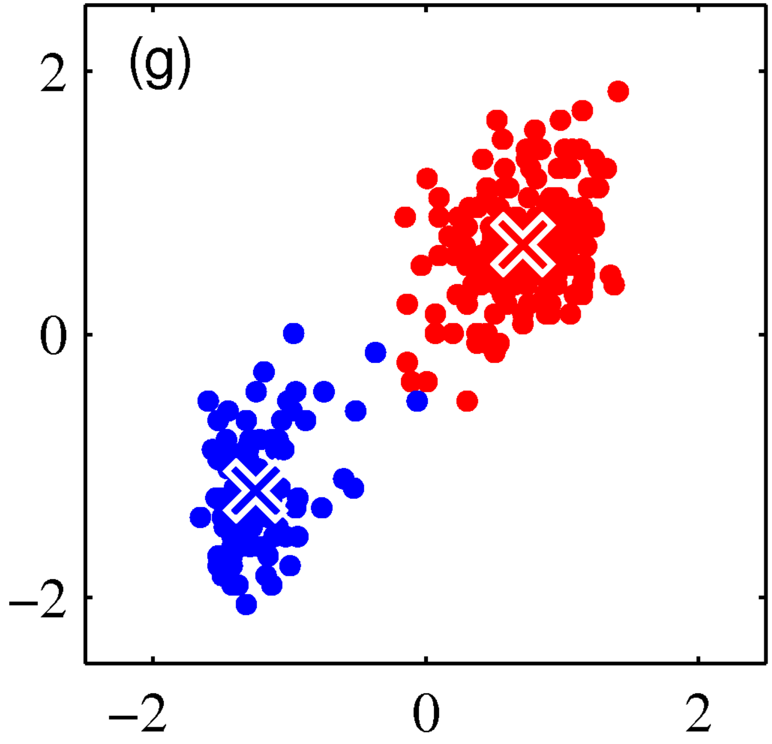
K-Means Example: 3rd E-Step

- Reassign data to the nearest center.



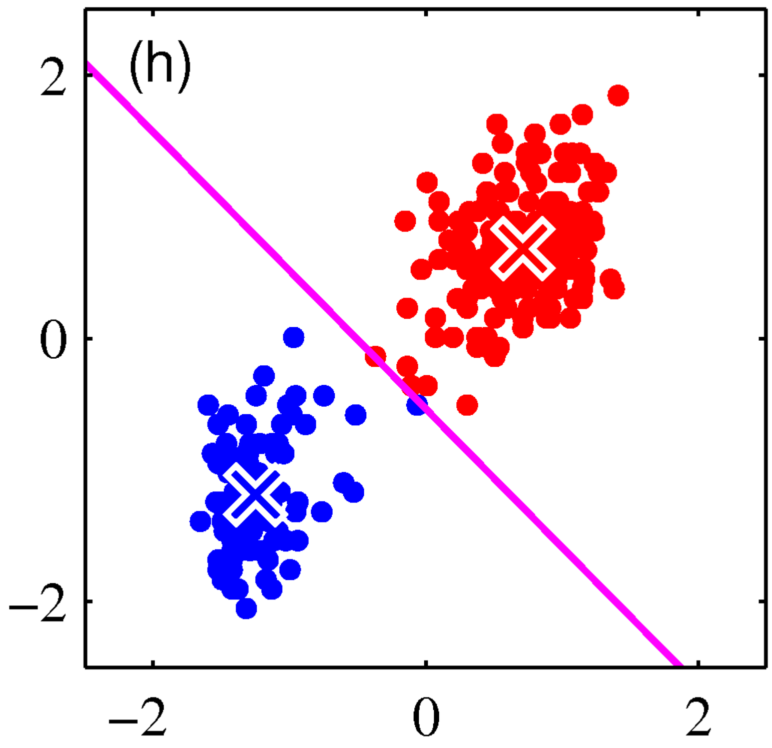
K-Means Example: 3rd M-Step

- Compute new centers for each cluster.



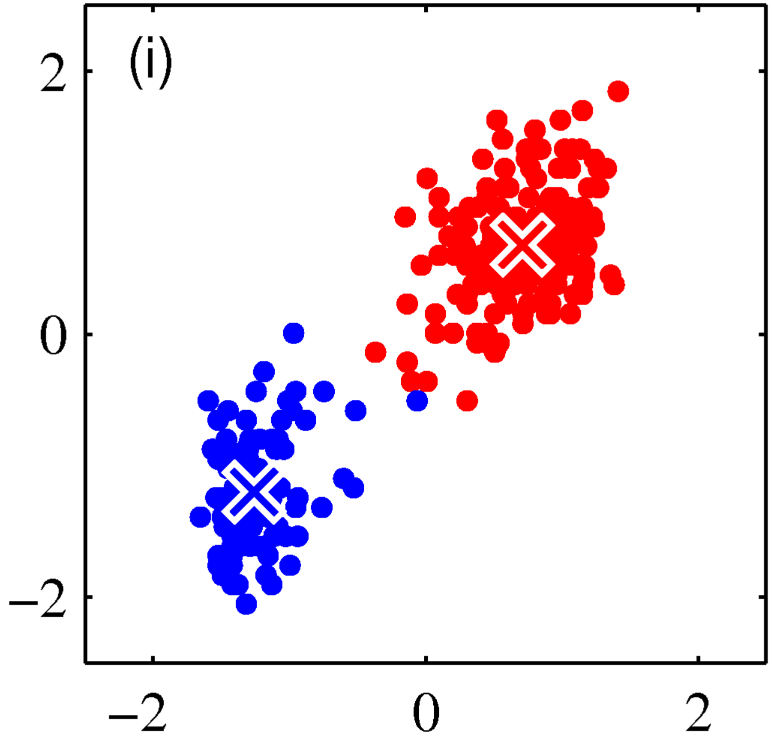
K-Means Example: 4th E-Step

- Reassign data to the nearest center.



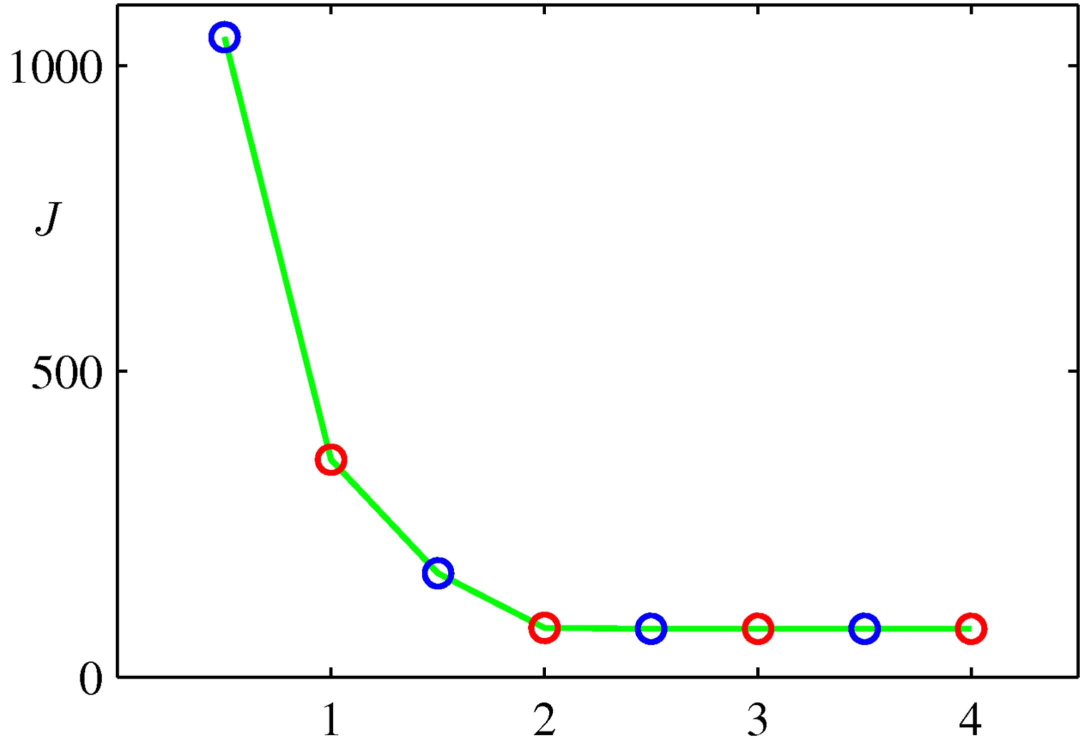
K-Means Example: 4th M-Step

- Compute new centers for each cluster.
- Stop here; cluster centers have stopped changing.



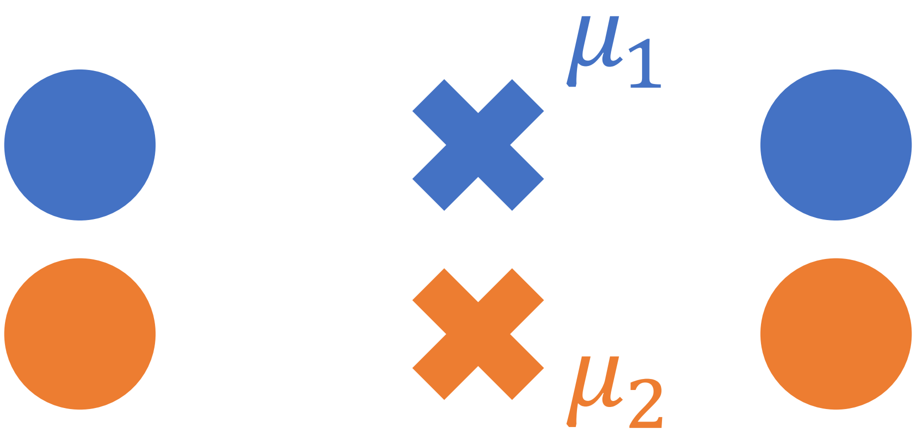
K-Means: Convergence

- Convergence is relatively quick, in # of steps.
  - Blue circles after **E-step**: Assign each point to a cluster
  - Red circles after **M-step**: Recompute the cluster centers
  - However, all those distance computations are expensive.



K-Means: Properties

- The objective function $J(r, \boldsymbol{\mu})$ monotonically decreases over time.
  - It is a general property of the EM algorithm.
- No guarantee to find the **global optimum**.
  - Guaranteed to converge to **local optimum**.
  - Clustering result depends on the initial values.
  - e.g., the following clustering is a stable local optimum



Gaussian Mixture Models

Hard vs. Soft Clusters

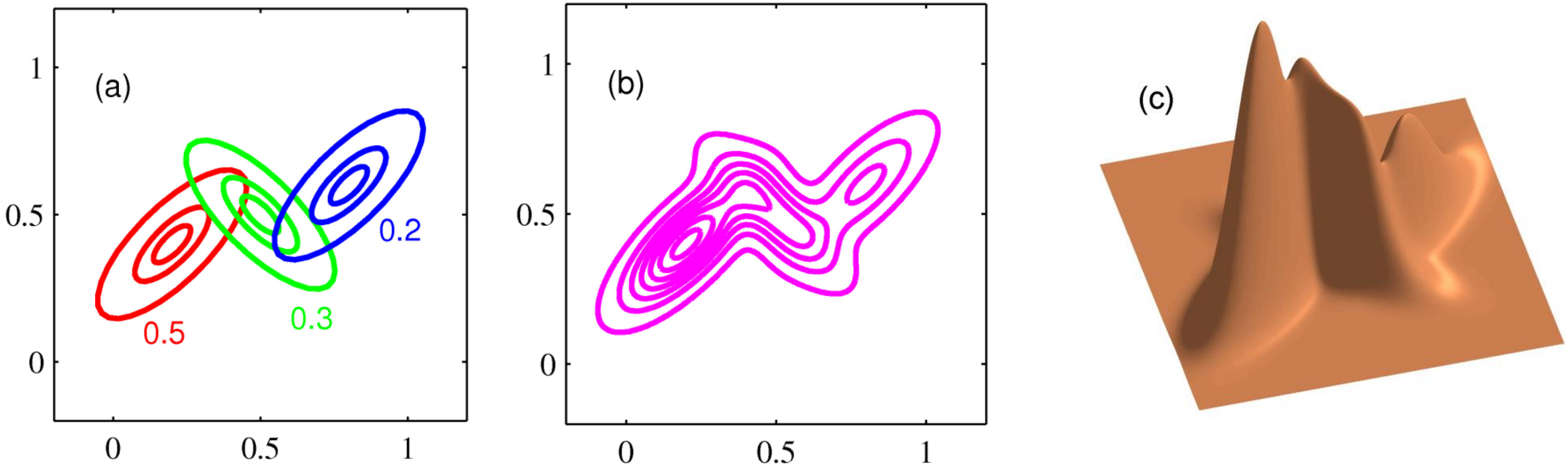
- K-means uses **hard** clustering assignment.
  - Each data belongs to exactly one cluster.

- Gaussian mixture model (GMM) for **soft** clustering
  - Each data is assigned to more than one cluster.
  - Different clusters take different levels of responsibility (posterior probability) for each point.
    - Each data was generated by only one cluster, but we don’t know which one.
  - Note that GMM itself is a probabilistic model, not a clustering method; **EM for GMM** is a clustering method on top of the probabilistic model.

Gaussian Mixture Models

- GMMs make it possible to describe much richer distributions.

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



GMM: Formulation

- Mixing coefficients: π_k, where $\sum_{k=1}^{K} \pi_k = 1$
- Cluster assignments: $\mathbf{z} \in \{0,1\}^K$ (1-of-K)
- Marginal distribution of $\mathbf{z}$:

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

- Conditional distribution of $\mathbf{x}$:

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

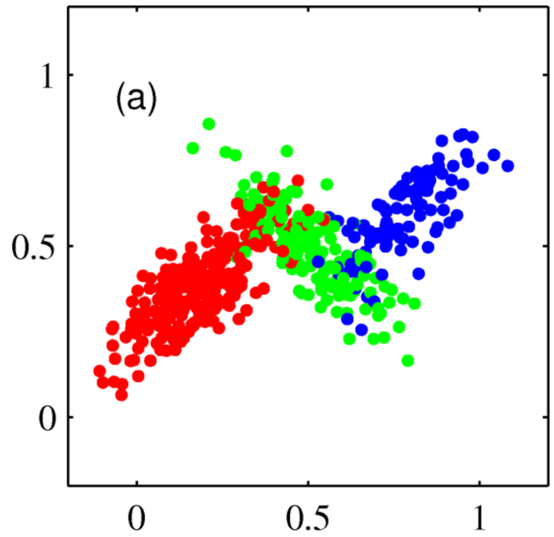
- Marginal distribution of $\mathbf{x}$:

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

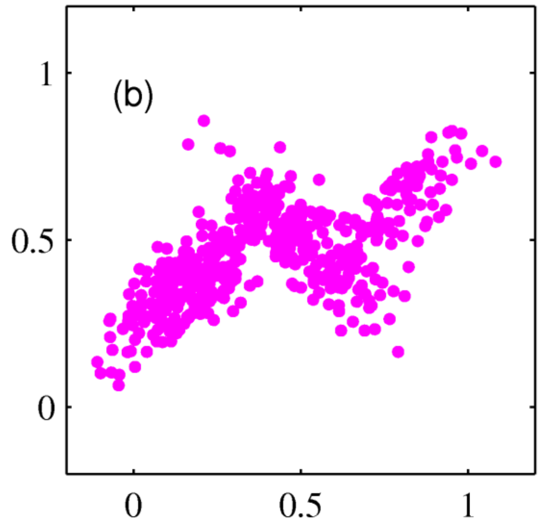
GMM: Formulation

- To generate samples from a Gaussian mixture distribution $p(\mathbf{x})$, use $p(\mathbf{x}, \mathbf{z})$:
  - Select a value $\mathbf{z}$ from the marginal $p(\mathbf{z})$;
  - Then select a value $\mathbf{x}$ from $p(\mathbf{x}|\mathbf{z})$ for that $\mathbf{z}$.

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



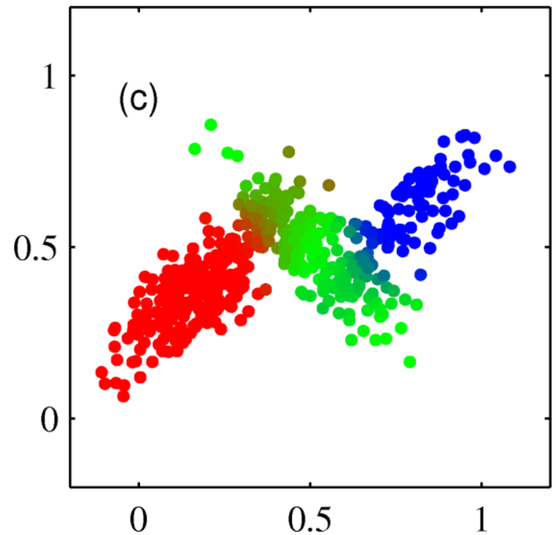
$p(\mathbf{x})$



EM for GMM: E-Step

- Responsibility $\gamma(z_k)$: The degree (posterior prob.) to which each Gaussian explains an observation $\mathbf{x}$.

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



EM for GMM: M-Step Formulation

- Log-likelihood of observing the data $\mathbf{x}$

$$
\begin{aligned}
\log p(\mathbf{x}) &= \sum_{k=1}^{K} \gamma(z_k) \log p(\mathbf{x}) \\
&= \sum_{k=1}^{K} \gamma(z_k) \log \frac{p(\mathbf{x}, z_k = 1)}{p(z_k = 1 | \mathbf{x})} \\
&= \sum_{k=1}^{K} \gamma(z_k) \log p(\mathbf{x}, z_k = 1) - \sum_{k=1}^{K} \gamma(z_k) \log \gamma(z_k)
\end{aligned}
$$

- Assume $\gamma(z_k)$ is a constant

$$
\log p(\mathbf{x}) = \sum_{k=1}^{K} \gamma(z_k) \log\big(\pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \Sigma_k)\big) + C
$$

EM for GMM: M-Step Formulation

$$
\begin{aligned}
\log p(\mathbf{x}) &= \sum_{k=1}^{K} \gamma(z_k) \log\left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)\right) + C \\
&= \sum_{k=1}^{K} \gamma(z_k) \log \pi_k + \sum_{k=1}^{K} \gamma(z_k) \log \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) + C \\
&= \sum_{k=1}^{K} \gamma(z_k) \log \pi_k + \frac{1}{2} \sum_{k=1}^{K} \gamma(z_k) \log\left|\Sigma_k^{-1}\right| \\
&\quad - \frac{1}{2} \sum_{k=1}^{K} \gamma(z_k) (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + C
\end{aligned}
$$

EM for GMM: M-Step Formulation

- Log-likelihood:

$$L = \log p\left(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\right) = \sum_{n=1}^{N} \log p\left(\mathbf{x}^{(n)}\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log\left|\Sigma_k^{-1}\right|$$

$$- \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)^T \Sigma_k^{-1} \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right) + C$$

EM for GMM: M-Step Formulation

- Learning objective: maximum log-likelihood

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

$$\text{subject to} \sum_{k=1}^{K} \pi_k = 1$$

  - where

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log\left|\Sigma_k^{-1}\right|$$
$$- \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)^T \Sigma_k^{-1} \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)$$

EM for GMM: M-Step Derivation

- MLE with respect to $\boldsymbol{\mu}_k$:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log\pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log\left|\Sigma_k^{-1}\right|$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)^T\Sigma_k^{-1}\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)$$

$$\frac{\partial L}{\partial\boldsymbol{\mu}_k} = \frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\Sigma_k^{-1}\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right) = 0$$
$$\therefore\ \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}^{(n)}}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

EM for GMM: M-Step Derivation

- MLE with respect to $M = \Sigma_k^{-1}$:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log\pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log|M|$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)^T M\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)$$

$$\frac{\partial\log|X|}{\partial X} = (X^{-1})^T \qquad \frac{\partial\mathbf{a}^T X\mathbf{b}}{\partial X} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial L}{\partial M} = \frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})M^{-1} - \frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)^T = 0$$

$$\therefore M^{-1} = \Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)\left(\mathbf{x}^{(n)}-\boldsymbol{\mu}_k\right)^T}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

EM for GMM: M-Step Derivation

- MLE with respect to π_k:

$$\max_{\{\pi_k\}_{k=1}^K} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k$$

$$\text{subject to } \sum_{k=1}^{K} \pi_k = 1$$

- Lagrangian function:

$$\mathcal{L}(\pi_1, \dots, \pi_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

EM for GMM: M-Step Derivation

- Lagrangian function:

$$\mathcal{L}(\pi_1, \ldots, \pi_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} - \alpha = 0$$

$$\Rightarrow \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{\alpha}$$

- From the constraint:

$$\sum_{k=1}^{K} \pi_k = \frac{1}{\alpha} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) = \frac{N}{\alpha} = 1$$

$$\therefore \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

EM for GMM: M-Step

- The mean of a cluster is the weighted mean, weighted by the responsibilities $\gamma(z_{nk})$.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}^{(n)}$$

  - where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$ is the effective number of data in cluster k

- Likewise for covariance:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)\left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)^T$$

- Mixing coefficients: $\pi_k = \frac{N_k}{N}$

EM for GMM: Summary

- Initialize means $\boldsymbol{\mu}_k$, covariances Σ_k, and mixing coefficients π_k for K Gaussians.
- **E-Step**: Given the parameters $\{\boldsymbol{\mu}_k, \Sigma_k, \pi_k\}$, evaluate the responsibilities $\gamma(z_{nk})$.

$$\gamma(z_{nk}) = p\left(z_k = 1 | \mathbf{x}^{(n)}\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}^{(n)} \middle| \boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}^{(n)} \middle| \boldsymbol{\mu}_j, \Sigma_j\right)}$$

- **M-Step**: Given the responsibilities $\gamma(z_{nk})$, estimate the parameters $\{\boldsymbol{\mu}_k, \Sigma_k, \pi_k\}$.

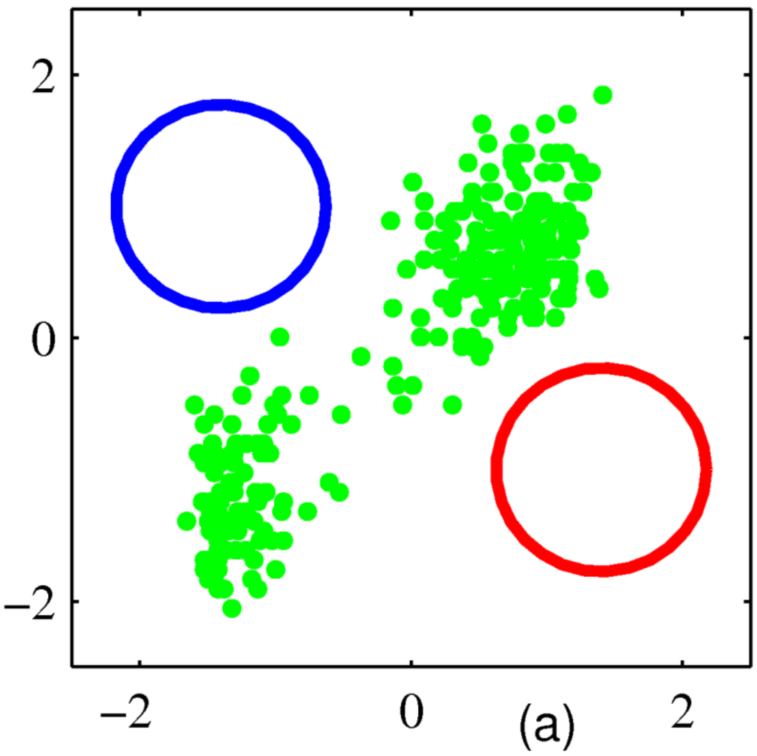
$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}^{(n)}, \pi_k^{\text{new}} = \frac{N_k}{N}, \text{where } N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k^{\text{new}}\right)^T$$

- Stop when the parameters or likelihood converges.

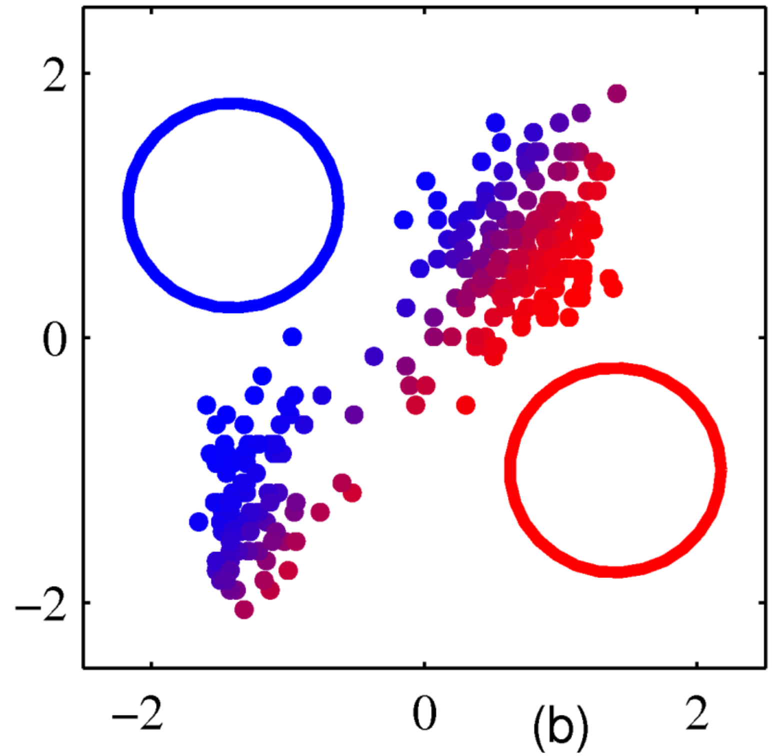
EM for GMM Example

- Initialize parameters: means, covariances, and mixing coefficients.



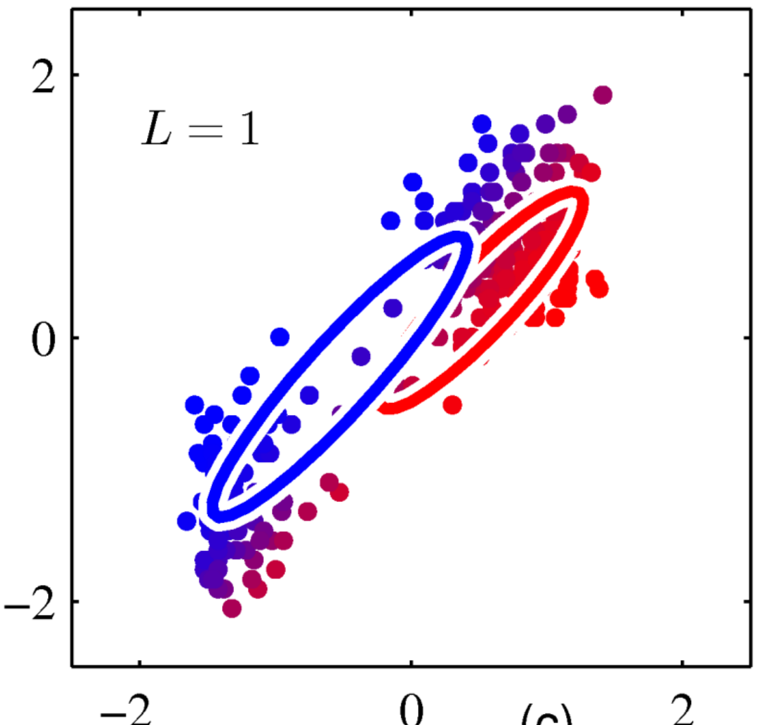
EM for GMM Example: 1st E-Step

- Evaluate the responsibilities.



EM for GMM Example: 1st M-Step

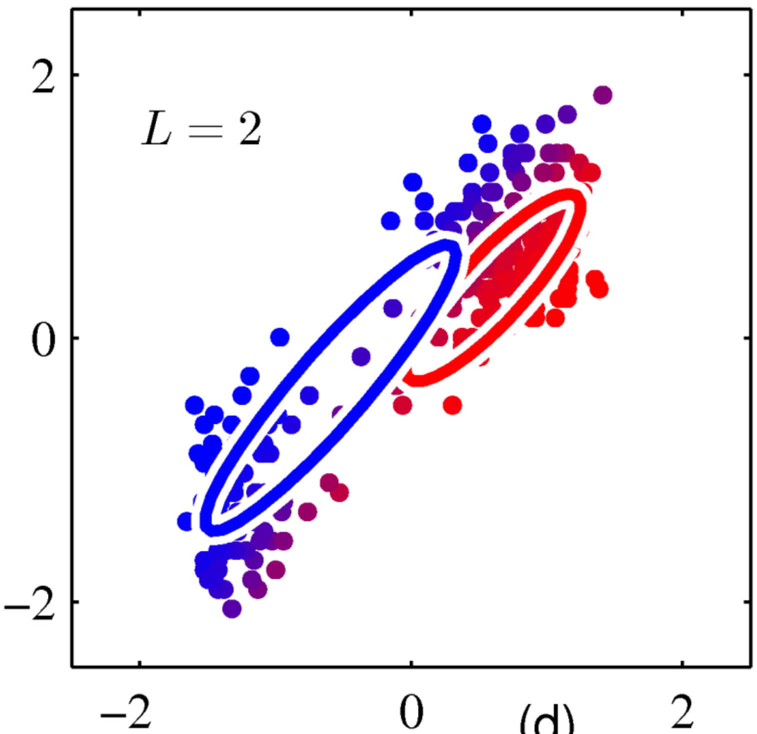
- Estimate the parameters.



(c)

EM for GMM Example: 2nd Steps

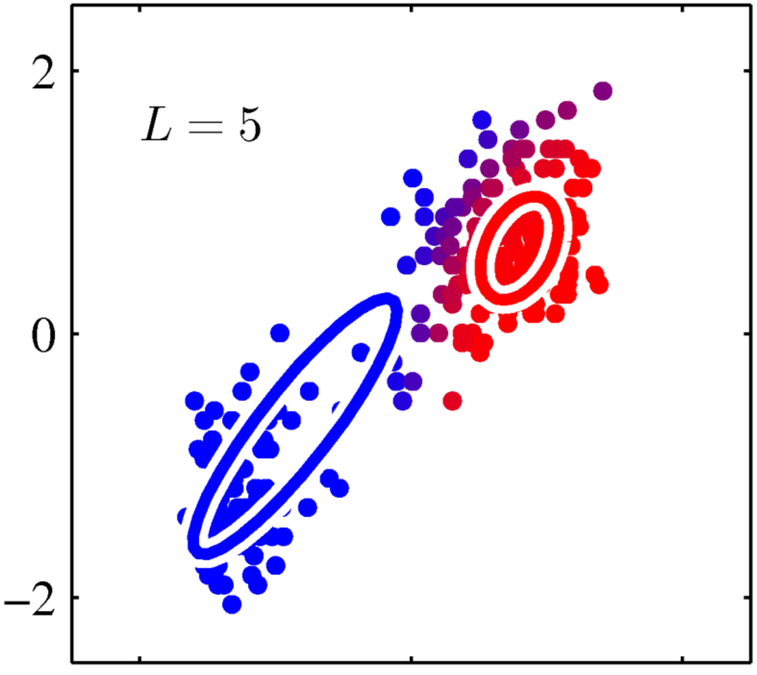
- Evaluate the responsibilities first, then estimate the parameters.



(d)

EM for GMM Example: 5^{th} Steps

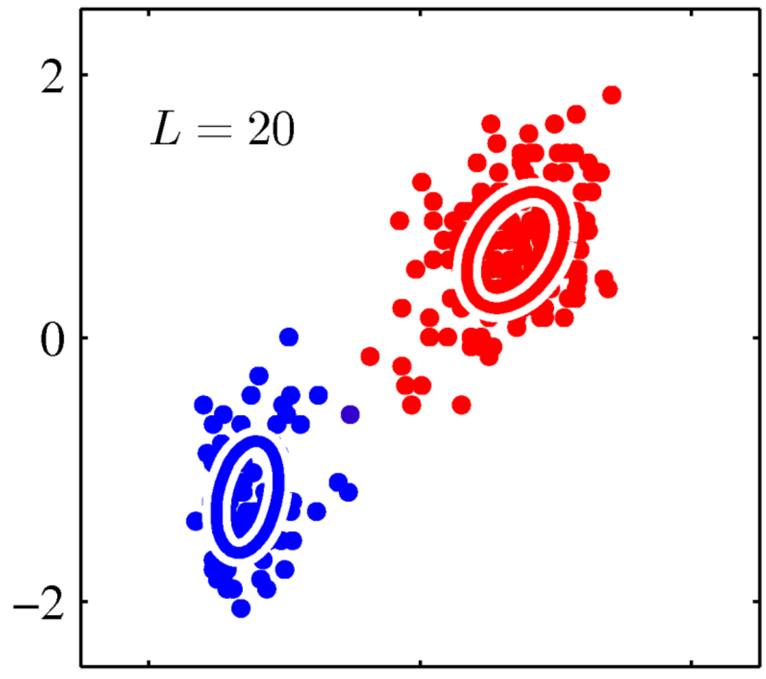
- Evaluate the responsibilities first, then estimate the parameters.



(e)

EM for GMM Example: 20^{th} Steps

- Evaluate the responsibilities first, then estimate the parameters.



(f)

EM for GMM vs. K-Means

- Fix the covariance matrix for each cluster as a diagonal matrix:

$$\Sigma_k = \sigma^2 I$$

- If we take $\sigma^2 \rightarrow 0$, then the update rules converge to K-means clustering.

General View of EM

EM: Motivation

- Suppose a system with observed variables X.
- It may be easier to understand with additional latent variables Z, which are not observed.

- E.g., in GMM, the latent variable Z specifies which Gaussian generated the sample X.
  - The responsibility is the posterior $p(Z|X)$.

EM: Motivation

- In ML, we usually find model parameters θ by maximizing log-likelihood of observed data.
- If we had complete data $\{X, Z\}$, we could easily maximize likelihood $p(X, Z|\theta)$.
- However, when not all variables are observed, we can marginalize over the unobserved variables:

$$\log p(X|\theta) = \log\left\{\sum_{Z} p(X, Z|\theta)\right\}$$

  - If Z is continuous, replace the sum with integral
- The summation over the latent variables is inside the logarithm, resulting in complicated expressions.

EM: Formulation

- EM finds the local maximum likelihood of $\log p(X)$ by alternating:
- **E-Step**: Given current parameters θ^{old}, find the posterior distribution of Z given X: $p(Z|X, \theta^{\text{old}})$
- Then, we find the expectation of the complete-data log-likelihood using the posterior:

$$Q(\theta, \theta^{\text{old}}) = \sum_{Z} \underbrace{p(Z|X, \theta^{\text{old}})}_{\text{Constant w.r.t. } \theta} \log p(X, Z|\theta)$$

- **M-Step**: Maximize the expectation:

$$\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{\text{old}})$$

EM: Derivation

- Goal: Maximize $p(X|\theta) = \sum_Z p(X, Z|\theta)$
- For any distribution $q(Z)$, $(q(Z) \geq 0, \sum_Z q(Z) = 1)$

$$\log p(X|\theta) = \sum_Z q(Z) \log p(X|\theta)$$

  - $\log p(X|\theta)$ is independent to Z

EM: Derivation

- Goal: Maximize $p(X|\theta) = \sum_Z p(X, Z|\theta)$
- For any distribution $q(Z)$, $(q(Z) \geq 0, \sum_Z q(Z) = 1)$

$$\log p(X|\theta) = \sum_Z q(Z) \log p(X|\theta)$$
$$= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)}$$

  - Conditional probability: $p(X, Z|\theta) = p(X|\theta)p(Z|X, \theta)$

EM: Derivation

- Goal: Maximize $p(X|\theta) = \sum_Z p(X, Z|\theta)$
- For any distribution $q(Z)$, ($q(Z) \geq 0, \sum_Z q(Z) = 1$)

$$\log p(X|\theta) = \sum_Z q(Z) \log p(X|\theta)$$
$$= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)}$$
$$= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} \frac{q(Z)}{p(Z|X, \theta)}$$

  - Introduce auxiliary $q(Z)$

EM: Derivation

- Goal: Maximize $p(X|\theta) = \sum_Z p(X,Z|\theta)$
- For any distribution $q(Z)$, $(q(Z) \geq 0, \sum_Z q(Z) = 1)$

$$
\begin{aligned}
\log p(X|\theta) &= \sum_Z q(Z) \log p(X|\theta) \\
&= \sum_Z q(Z) \log \frac{p(X,Z|\theta)}{p(Z|X,\theta)} \\
&= \sum_Z q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} \frac{q(Z)}{p(Z|X,\theta)} \\
&= \sum_Z q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{p(Z|X,\theta)} \\
&= \mathcal{L}(q,\theta) + KL\big(q(Z)||p(Z|X)\big)
\end{aligned}
$$

- $\mathcal{L}(q,\theta)$: The lower bound we maximize
- $KL\big(q(Z)||p(Z|X)\big)$: Gap between $q(Z)$ and $p(Z|X)$

EM: Derivation

- Goal: Maximize $p(X|\theta) = \sum_Z p(X, Z|\theta)$
- For any distribution $q(Z)$, $(q(Z) \geq 0, \sum_Z q(Z) = 1)$

$$
\begin{aligned}
\log p(X|\theta) &= \sum_Z q(Z) \log p(X|\theta) \\
&= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} \\
&= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} \frac{q(Z)}{p(Z|X, \theta)} \\
&= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{p(Z|X, \theta)} \\
&= \mathcal{L}(q, \theta) + KL\big(q(Z) || p(Z|X)\big) \\
&\geq \mathcal{L}(q, \theta)
\end{aligned}
$$

- KL divergence is nonnegative: $KL\big(q(Z) || p(Z|X)\big) \geq 0$

EM: Another Derivation

- Given the observed variables X, latent variables Z, and parameters θ:

$$\log p(X|\theta) = \log \sum_{Z} p(X, Z|\theta)$$

  - Introduce Z

EM: Another Derivation

- Given the observed variables X, latent variables Z, and parameters θ:

$$\log p(X|\theta) = \log \sum_{Z} p(X, Z|\theta)$$

$$= \log \sum_{Z} q(Z) \frac{p(X, Z|\theta)}{q(Z)}$$

  - Introduce auxiliary $q(Z)$ where $q(Z) \geq 0, \sum_Z q(Z) = 1$

EM: Another Derivation

- Given the observed variables X, latent variables Z, and parameters θ:

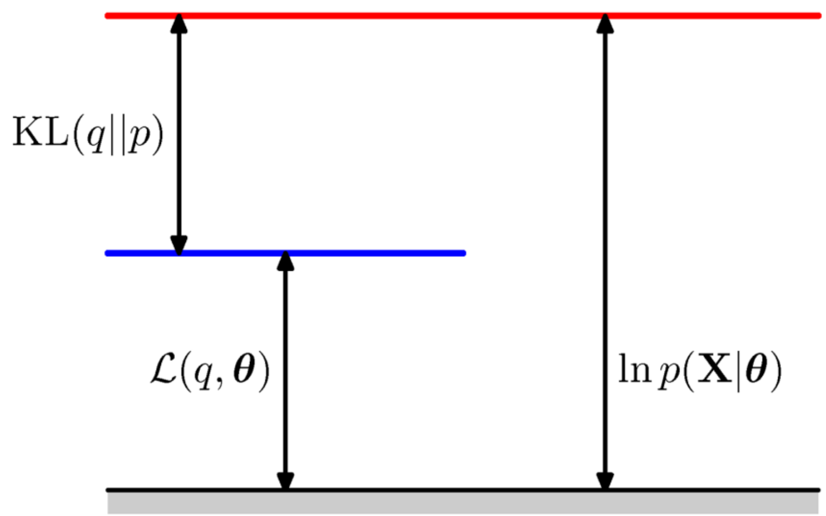
$$\log p(X|\theta) = \log \sum_{Z} p(X,Z|\theta)$$
$$= \log \sum_{Z} q(Z) \frac{p(X,Z|\theta)}{q(Z)}$$
$$\geq \sum_{Z} q(Z) \log \frac{p(X,Z|\theta)}{q(Z)} = \mathcal{L}(q,\theta)$$

  - Jensen's inequality
  - Equality holds when $p(X,Z|\theta)/q(Z)$ is constant.

EM: Visualizing Decompositions

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + KL\big(q(Z)||p(Z|X)\big)$$

- $KL(q||p) \geq 0$; equality holds only when $q = p$.
- Thus, $\mathcal{L}(q, \theta)$ is the lower bound of $\log p(X|\theta)$, which EM maximizes.

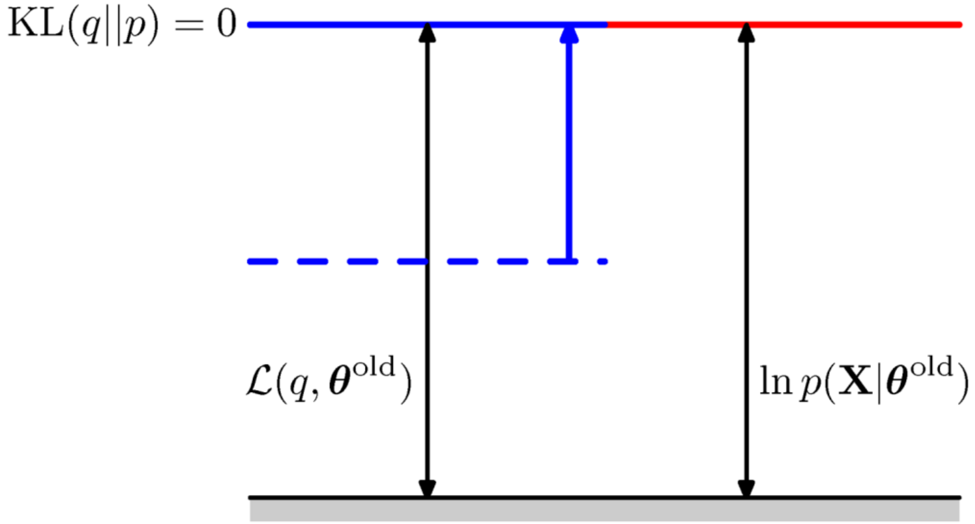


EM: Visualizing Decompositions

$$\log p(X|\theta) = \mathcal{L}(q,\theta) + KL\big(q(Z)||p(Z|X)\big)$$

- **E-Step** updates $q(Z)$ to maximize $\mathcal{L}(q,\theta)$.
- $q(Z)$ is maximized when $KL\big(q(Z)||p(Z|X)\big) = 0$

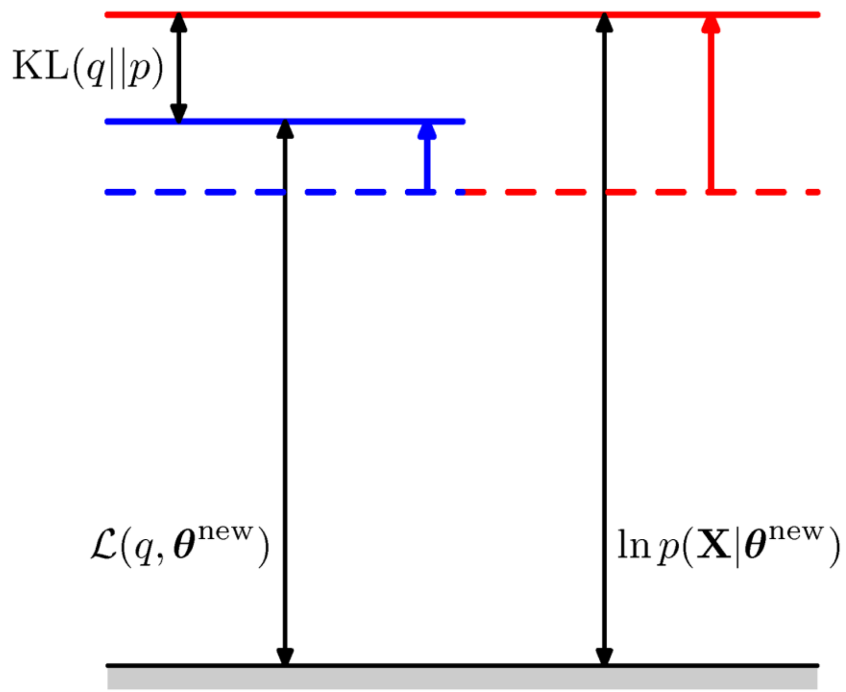
$$q(Z) = p(Z|X,\theta)$$



EM: Visualizing Decompositions

$$\log p(X|\theta) = \mathcal{L}(q,\theta) + KL\big(q(Z)||p(Z|X)\big)$$

- **M-Step**: Increase $\mathcal{L}(q,\theta)$ with $q(Z)$ as constant
  - This increases $\log p(X|\theta)$, but then $q(Z) \neq p(Z|X,\theta)$
- Iterate EM until converged



Next: Dimensionality Reduction