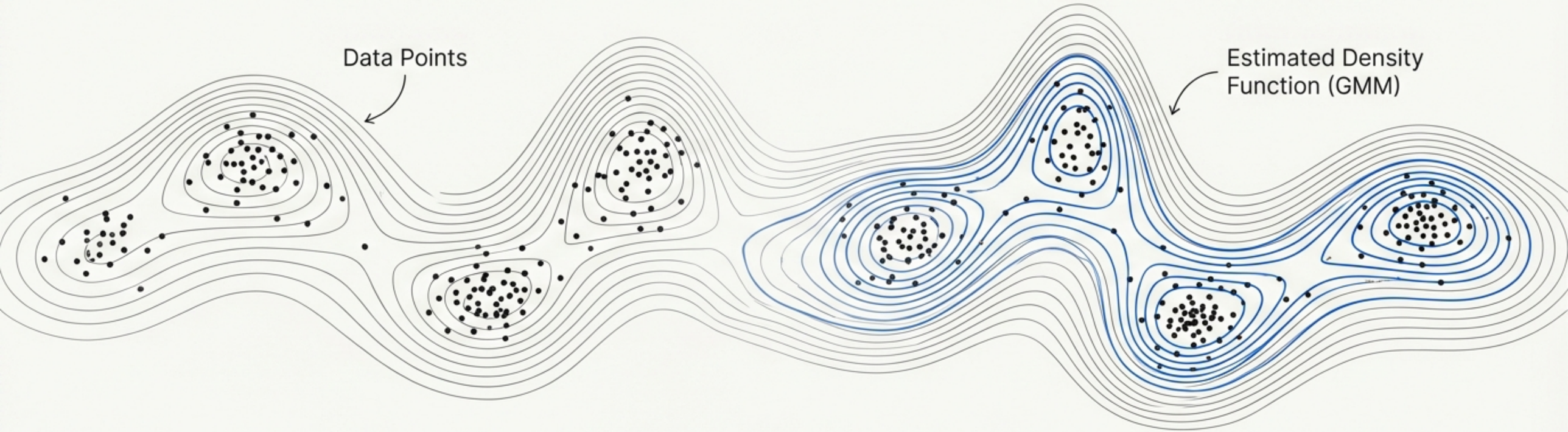
Clustering & The EM Algorithm

Derivations, Variations, and Theoretical Connections

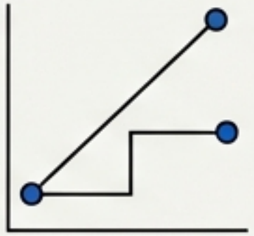
Based on ML-IN2064 Exercise 12 Solutions



Robust Clustering: The Derivation of K-Medians

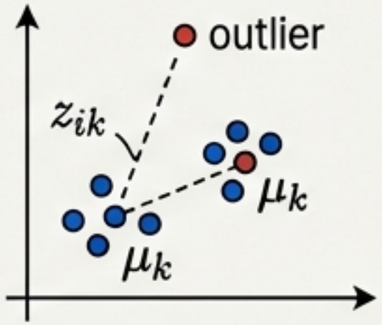
The Objective

Standard K-Means relies on the L2 distance (squared Euclidean), which is sensitive to outliers. To improve robustness, we substitute the L1 distance (Manhattan).



$$J(X, Z, \mu) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|_1$$

This is the "K-Medians" objective.



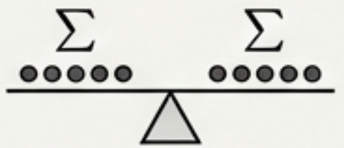
The Derivation

We minimize the objective w.r.t the centroid μ_k. Dimensions d are independent.

$$\rightarrow \frac{\partial}{\partial \mu_{kd}} \sum_{i=1}^{N} z_{ik} |x_{id} - \mu_{kd}| = 0$$

The derivative of the absolute value is the sign function:

$$\sum_{i=1}^{N} z_{ik} \cdot \text{sign}(\mu_{kd} - x_{id}) = 0$$



$$\Rightarrow \sum \mathbb{I}(\mu_{kd} > x_{id}) = \sum \mathbb{I}(\mu_{kd} < x_{id})$$

We need an equal number of points to the left and right.

Conclusion: The optimal centroid is the component-wise median.

$$\mu_{kd} = \text{median}\{x_{id} \mid z_{ik} = 1\}$$

Gaussian Mixture Models: Statistical Moments

Deriving properties via the Law of Iterated Expectations

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

1. Expected Value $E[x]$

Use iterated expectation:

$$E[x] = E_z[E_x[x|z]]$$

Substitute conditional mean:

$$E[x|z=k] = \mu_k$$

$$\boxed{E[x] = \sum_{k=1}^{K} \pi_k \mu_k}$$

2. Covariance $Cov[x]$

Identity: $Cov[x] = E[xx^T] - E[x]E[x]^T$

Compute second moment conditioning on z:

$$E[xx^T] = \sum_{k=1}^{K} \pi_k E[xx^T|z=k] = \sum_{k=1}^{K} \pi_k (\Sigma_k + \mu_k \mu_k^T)$$

$$\boxed{Cov[x] = \sum_{k=1}^{K} \pi_k (\Sigma_k + \mu_k \mu_k^T) - E[x]E[x]^T}$$

Total Covariance = Within-cluster spread + Between-cluster spread

The Expectation Step: Inferring Latent Variables

Calculating the posterior distribution (Responsibilties)

The Goal

We need to evaluate the probability that a specific data point x_i belongs to cluster k, given our current parameter estimates.

$$\gamma_t(z_i = k) := p(z_i = k \,|\, x_i, \pi^{(t)}, \mu^{(t)}, \Sigma^{(t)})$$

The Derivation (Bayes' Theorem)

Step 1: Apply Bayes' Rule:

$$\gamma_t(z_i = k) = \frac{p(x_i \,|\, z_i = k, \theta^{(t)}) \cdot p(z_i = k \,|\, \theta^{(t)})}{p(x_i \,|\, \theta^{(t)})}$$

Step 2: Expand terms using GMM definitions:

- Likelihood: $p(x_i \,|\, z_i = k) = \mathcal{N}(x_i \,|\, \mu_k^{(t)}, \Sigma_k^{(t)})$
- Prior: $p(z_i = k) = \pi_k^{(t)}$
- Evidence (marginal): $\sum_j \pi_j^{(t)} \mathcal{N}(x_i \,|\, \mu_j^{(t)}, \Sigma_j^{(t)})$

The E-Step Update Rule

$$\gamma_t(z_i = k) = \frac{\pi_k^{(t)} \mathcal{N}(x_i \,|\, \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(x_i \,|\, \mu_j^{(t)}, \Sigma_j^{(t)})}$$

The Maximization Step: Updating Priors (π)

Constrained Optimization via Lagrangian Multipliers

The Objective

$$\text{Maximize } L_z = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i = k) \log \pi_k \text{ subject to } \sum \pi_k = 1.$$

The Lagrangian

$$f(\pi, \lambda) = \sum_{k=1}^{K} N_k \log \pi_k + \lambda \left(1 - \sum_{k=1}^{K} \pi_k \right)$$

Where $N_k = \sum_{i=1}^{N} \gamma_t(z_i = k)$ is the effective cluster size.

The Solution

Derivative w.r.t π_k

$$\frac{\partial f}{\partial \pi_k} = \frac{N_k}{\pi_k} - \lambda = 0 \Rightarrow \pi_k = \frac{N_k}{\lambda}.$$

Solving for λ

Summing over k: $\sum N_k = \lambda \sum \pi_k \Rightarrow N = \lambda \cdot 1$

$$\pi_k^{(t+1)} = \frac{N_k}{N}$$

The new prior is simply the fraction of total responsibility assigned to cluster k.

The Maximization Step: Updating Parameters (μ, Σ)

Likelihood Maximization via Matrix Calculus

$$\text{Maximize } L_x = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i = k) \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

Flow 1: The Mean (μ_k)

$$\frac{\partial L_x}{\partial \mu_k} = \sum_{i=1}^{N} \gamma_t(z_i = k) \Sigma_k^{-1} (x_i - \mu_k) = 0$$

Multiply by Σ_k and solve:

$$\mu_k^{(t+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_t(z_i = k) x_i$$

Weighted Sample Mean

Flow 2: The Covariance (Σ_k)

$\frac{\partial L_x}{\partial \Sigma_k} \Rightarrow$ Derivative of log-determinant and quadratic form.

Solution yields the weighted scatter matrix:

$$\Sigma_k^{(t+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_t(z_i = k)(x_i - \mu_k)(x_i - \mu_k)^T$$

Weighted Sample Covariance

Q.E.D

Unification: K-Means as a Special Case of GMM

Consider a GMM with isotropic covariances: $\Sigma_k = \sigma^2 I$.

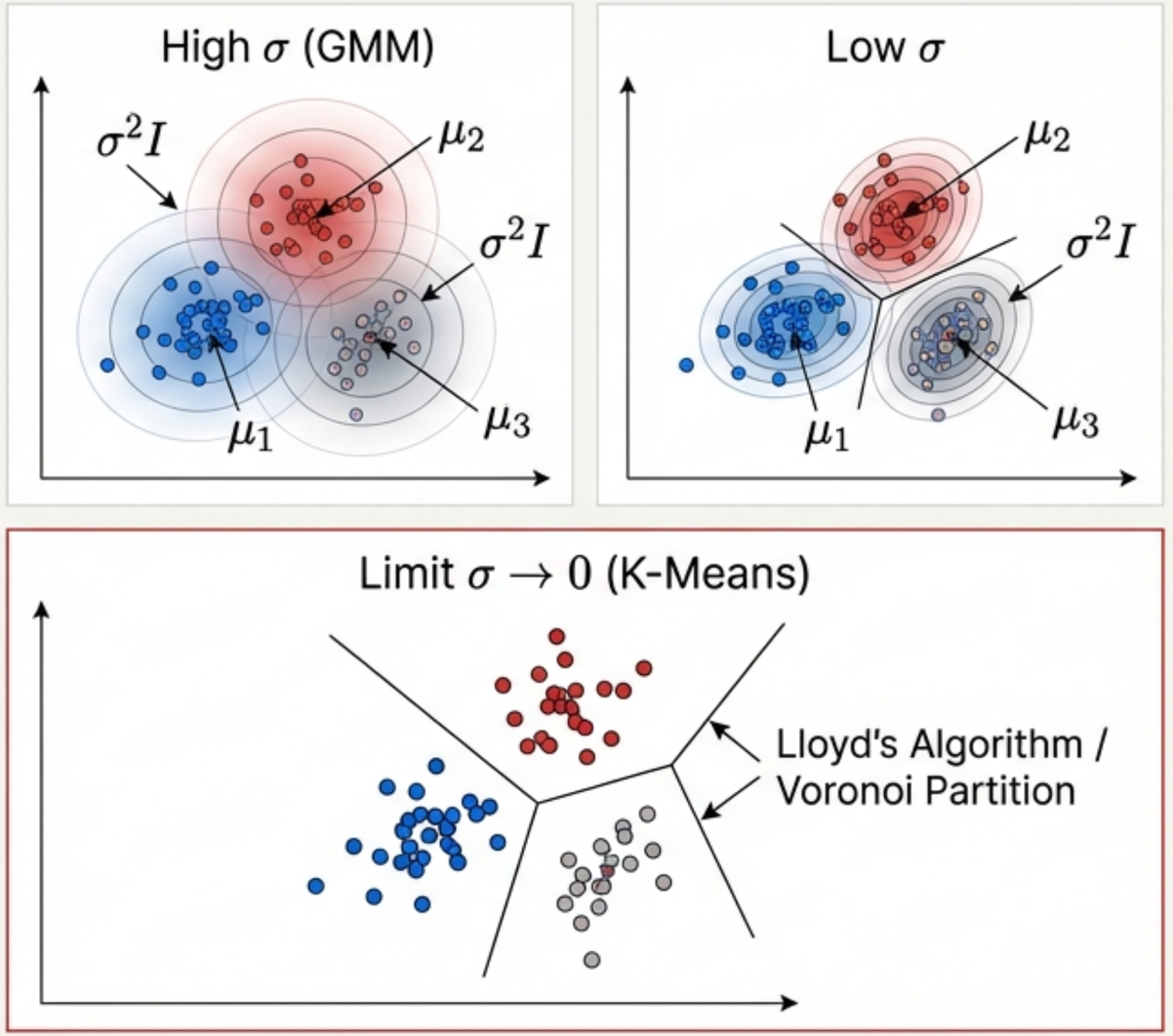
We investigate the limit as $\sigma^2 \to 0$.

Responsibility Analysis:

$$\gamma_t(z_i = k) \propto \exp\left(-\frac{||x_i - \mu_k||^2}{2\sigma^2}\right).$$

As $\sigma \to 0$, the exponential term for the closest centroid dominates significantly over all others.

Result: γ_t becomes a 'one-hot' vector (Hard Assignment).



Q.E.D

Generalizing EM: Mixture of Bernoullis

Modeling Binary Data (e.g., Black & White Pixels)

The Model:
Instead of Gaussian, use independent Bernoulli variables for each dimension d.

$$p(x|z=k) = \prod_{d=1}^{D} \theta_{kd}^{x_{id}} (1-\theta_{kd})^{1-x_{id}}$$

The E-Step:
Calculated via standard Bernoulli likelihoods. Structure remains identical to GMM, only the probability density function changes.

$$\gamma_t(z_i=k) \propto \pi_k \prod_{d=1}^{D} \theta_{kd}^{x_{id}} (1-\theta_{kd})^{1-x_{id}}$$

Q.E.D

Mixture of Bernoullis: The M-Step

Deriving the parameter update rule

Goal:

Maximize Log-Likelihood w.r.t parameters θ_{kd}.

Derivative:

$$\frac{\partial L}{\partial \theta_{kd}} = \sum_{i=1}^{N} \gamma_t(z_i = k) \left(\frac{x_{id}}{\theta_{kd}} - \frac{1 - x_{id}}{1 - \theta_{kd}} \right) = 0$$

Note: Parameters θ_{kd} do not interact across dimensions.

The Solution:

Solving for θ_{kd} yields a familiar form:

$$\theta_{kd}^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_t(z_i = k) x_{id}}{\sum_{i=1}^{N} \gamma_t(z_i = k)}$$

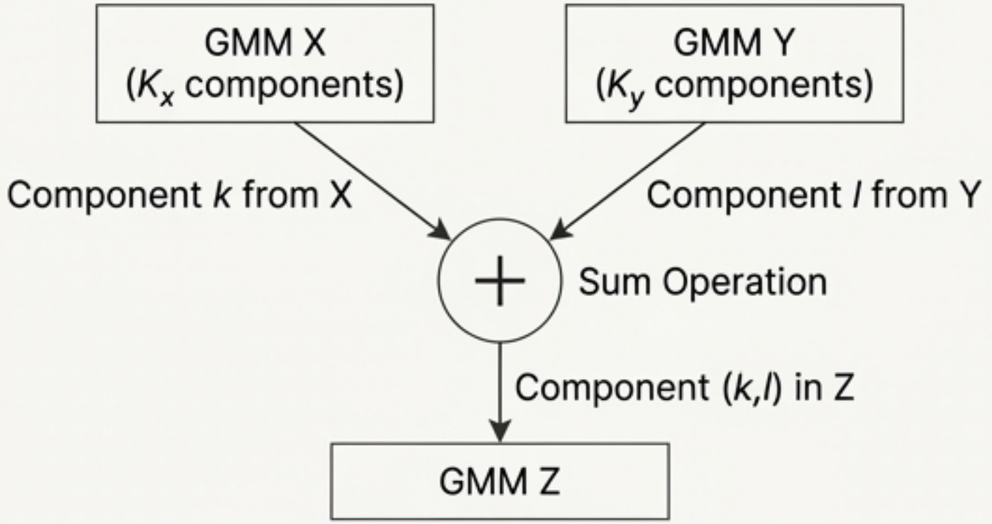
Intuition: This is a “Soft” version of the standard Maximum Likelihood Estimator (average pixel intensity weighted by responsibility).

Q.E.D

Theoretical Operations on GMMs

What is the distribution of the sum of two GMMs?

Given independent $x \sim \text{GMM}(\theta_x)$ and $y \sim \text{GMM}(\theta_y)$, find distribution of $z = x + y$.



The resulting PDF is a mixture of $K_x \times K_y$ Gaussians.

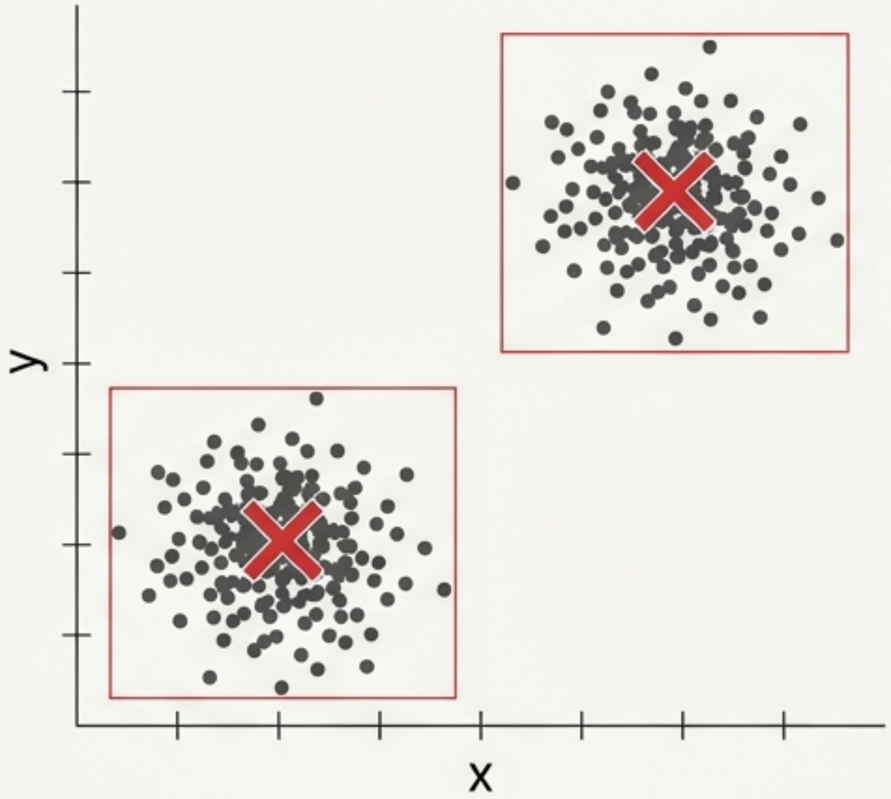
$$p(z) = \sum_{k=1}^{K_x} \sum_{l=1}^{K_y} \pi_{xk} \pi_{yl} \mathcal{N}(z \mid \mu_{xk} + \mu_{yl}, \Sigma_{xk} + \Sigma_{yl})$$

Means add. Covariances add. Weights multiply.

Q.E.D

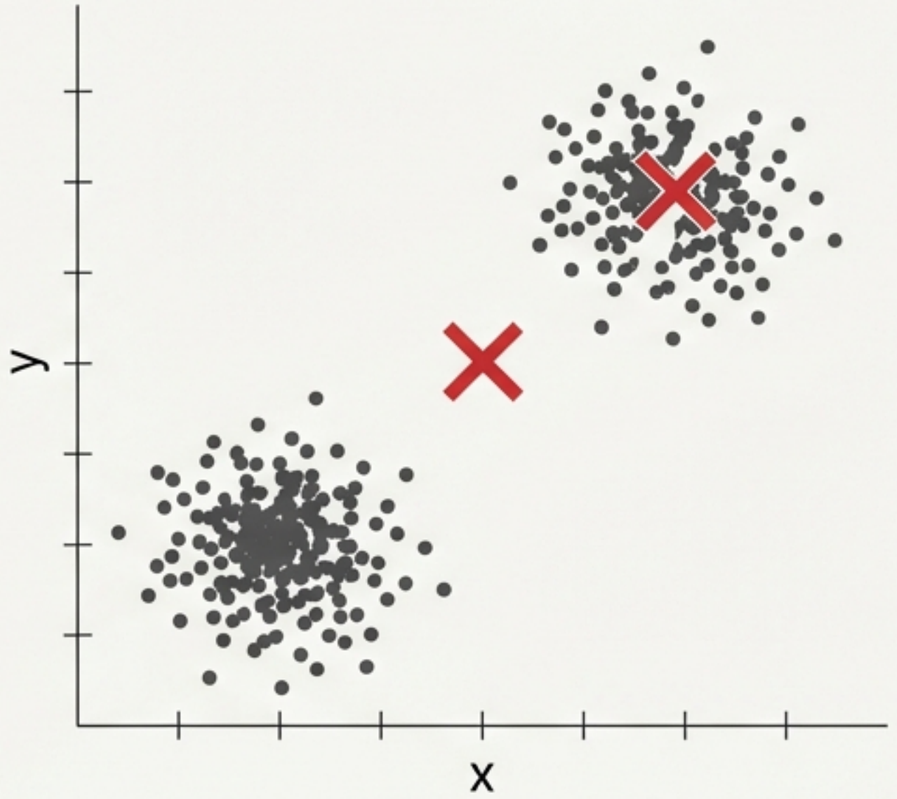
The Impact of Initialization on Convergence

A Optimal Initialization



Converges in 1 iteration.

B Poor Initialization



Requires multiple iterations to migrate centroids.

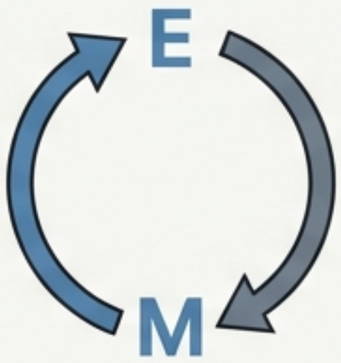
Initialization strategy (e.g., K-Means++) critically affects performance.

Summary & Key Takeaways

Clustering Spectrum

We moved from Hard assignments (K-Means/Medians) to Soft probabilistic assignments (GMM).

K-Means is simply GMM with $\sigma \to 0$.

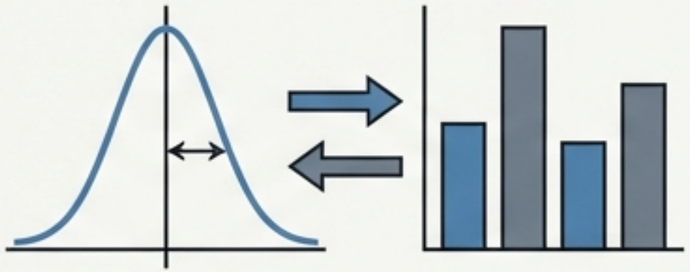


The EM Algorithm

The universal solver for mixture models.

E-Step: Infer latent variables (Responsibilities).

M-Step: Update parameters (MLE with weights).



Model Flexibility

The framework adapts to the data distribution.

Gaussian → Continuous data.
Bernoulli → Binary data.
L1 Norm → Robust K-Medians.

Q.E.D