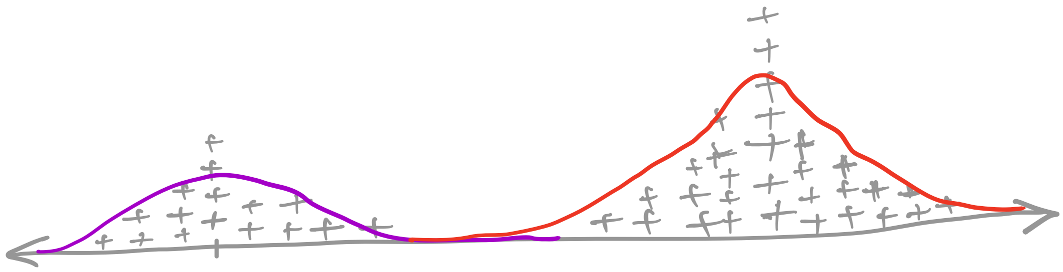


$x \in \mathbb{R}$

data $x_1, x_2, \ldots \ldots x_n$; $x_i \in \mathbb{R}$

mixture models

* model the probability of observing a data point x_i as a mixture of k- pure types

$\rightarrow$ $p(x|\theta_j)$, $j = 1, \ldots \ldots k$

θ_j ?

Assumption: 'k' is known

$$* \quad p(x) = \sum_{j=1}^{k} \pi_j \, p(x \mid \theta_j)$$

(mixture density) (pure type density)

Were we can see that

- $0 \leq \pi_j \leq 1$
- $\sum_{j=1}^{k} \pi_j = 1$

(mixture propotions)

Gaussian mixture model

* Assuming Gaussian pure types, we write

$$\rightarrow p(x) = \sum_{j=1}^{k} \pi_j \, N(x \mid \mu_j, \sigma_j^2) \quad \cdots\cdots \quad (1)$$

$$\boxed{N(x \mid \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp\left\{ \frac{(x-\mu_j)^2}{-2\sigma_j^2} \right\}}$$

Unknown parameters of the mixture

* π_i ; $0 \le \pi_{ij} \le 1$, $\sum_{j=1}^{k} \pi_{ij} = 1$
* $\mu_1, \mu_2, \ldots \mu_k \leftarrow$ mean
* $\sigma_1, \sigma_2, \ldots \sigma_k \leftarrow$ variance

An equivalent formulation — "missing data" representation

Dempster et al. (1977) ← Introduced EM algorithm

* we introduce a binary membership variable

$$z_{ij} \in \{0,1\} \; ; \quad \vec{z}_i \in \{0,1\}^k , \quad \sum_{j=1}^{k} z_j = 1$$

for every data point x_i

E.g. if x_i belongs to cluster j

$$z_{ij} = 1 \quad \text{and} \quad z_{ij'\neq j} = 0 \quad , j' = 1, \cdots k$$

* we use $\boxed{p(x) = \sum_{z} p(x, \vec{z}) = p(x|\vec{z})\, p(\vec{z})}$ ←

- the marginal $p(\vec{z})$ is modeled as a multinomial

$$p(z_j == 1) = \pi_j$$

$$p(\vec{z}) = \cancel{\theta} \prod_{j=1}^{k} \pi_j^{z_j} \quad \cdots\cdots \quad (2)$$

- Assuming normality, we write

$$p(x \mid z_j == 1) = N(x \mid \mu_j, \sigma_j^2) \leftarrow$$

$$p(x \mid \vec{z}) = \prod_{j=1}^{k} \underbrace{N(x \mid \mu_j, \sigma_j)}^{z_j} \quad \cdots\cdots \quad (3)$$

membership

- Combining (2) & (3), we have

$$p(x, z) = \underbrace{\prod_{j=1}^{k} \left[\pi_j \, \mathcal{N}(x \mid \mu_j, \sigma_j^2)\right]^{z_j}}_{\text{the complete data likelihood}}$$

$z_i \in \{0, 1\}^k$

$z_i \in (0, 1)^k$

- $\underline{p(x)}$

$$= \sum_{z} p(x, z)$$

$$= \sum_{z} \underbrace{\prod_{j=1}^{k} \left[\pi_j \, \mathcal{N}(x \mid \mu_j, \sigma_j^2)\right]^{z_j}}_{\text{other densities}}$$

$$= \sum_{j=1}^{k} \pi_j \, \mathcal{N}(x \mid \mu_j, \sigma_j^2)$$

(a Gaussian mixture)

__Note:__ Summation over z consists of 'k' terms;

(a) for j^{th} term, $z_j == 1$ and

(b) the product becomes $\pi_j \, \mathcal{N}(x \mid \mu_j, \sigma_j^2)$

for $j' \neq j^{th}$ term, $z_{j'} == 0$

Adv.

We have a joint density $p(x, z)$ with a hidden variable z — missing data. It helps

- MLE using EM (Dempster et al. 1977)
- and Bayesian posterior inference

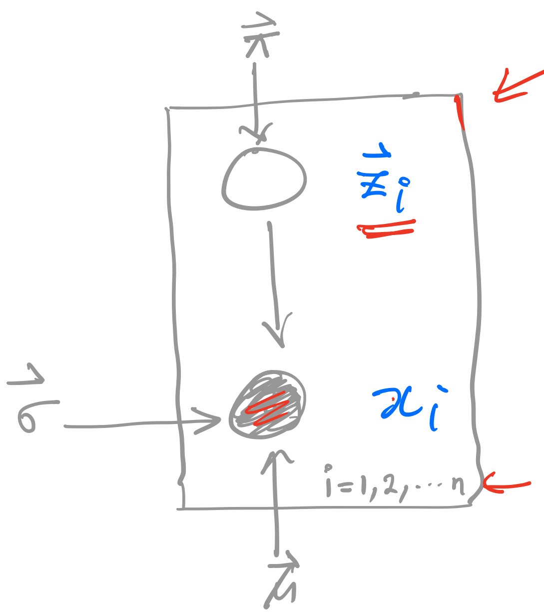
The hierarchical model (induced by the missing data formulation)

We assume

- there are k clusters in the data $\leftarrow$ known
- $\underbrace{\vec{\pi}}_{\text{mixture proportion}}, \underbrace{\mu_1, \dots \mu_k}_{\text{mean}}, \underbrace{\sigma_1, \dots \sigma_k}_{\text{vari}}$ are parameters

For every data point x_i, $i = 1, 2, \dots n$ $\qquad \vec{\pi} \in (0,1)^k$

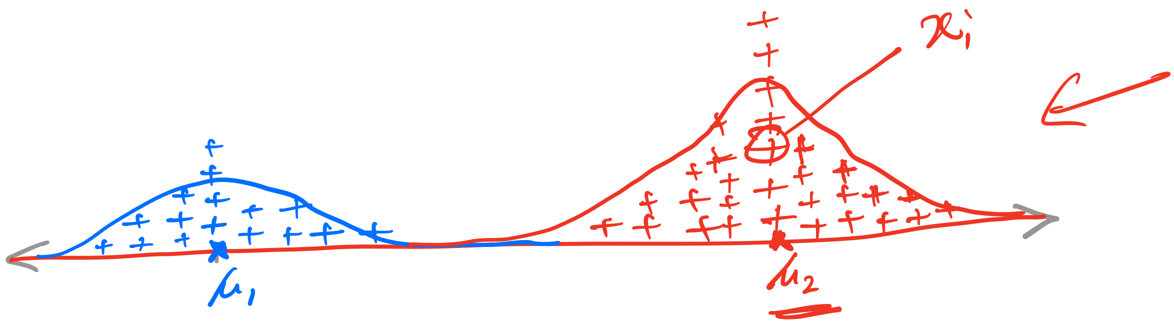
plate



$\sim \underline{p(\vec{z} \mid \vec{\pi})}$ "missing data"

$\sim p(x \mid \vec{z}_i, \mu_1, \dots \mu_k, \sigma_1, \dots \sigma_k)$

Example data generated ($k=2$, $x_i \in \mathbb{R}$)

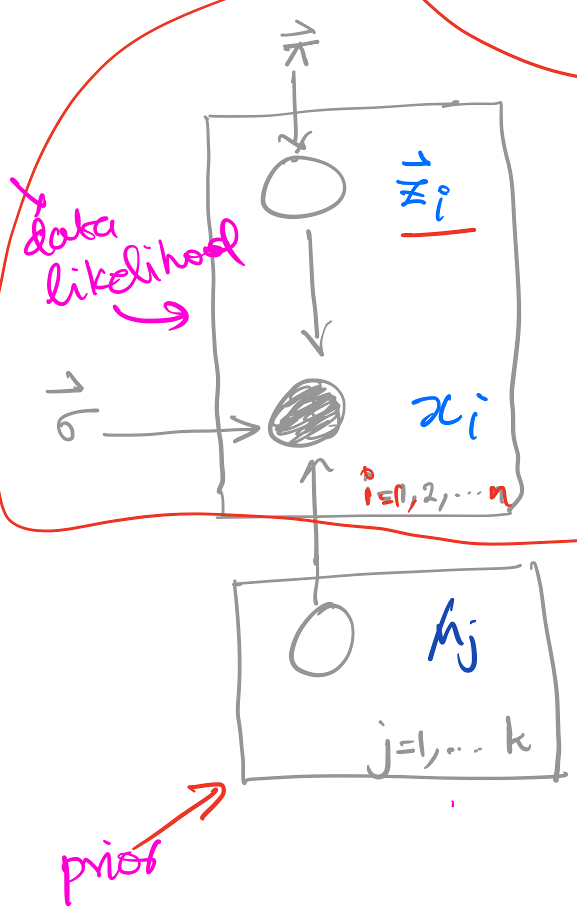


Bayesian formulation

We assume that the parameters, for example, π, μ, σ are random and distributed according to some prior

To keep it simple, we assume

* μ_j are normally distributed
* σ_j are constant ; π follows a uniform distribution



$\vec{z}_i \sim Cat(\vec{\pi})$ — 2 $\pi = (1/k, \dots, 1/k)$

$x_i \sim N(x \mid \mu_{z_i}, \sigma^2)$ — 3

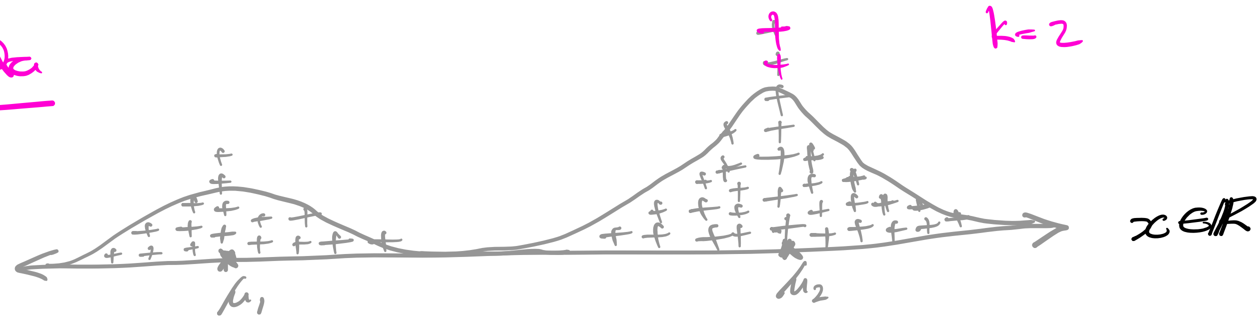
$\mu_j \sim N(\mu \mid 0, \tau_j^2)$ ← 1

Posterior a distribution over hidden variables

- z — missing data
- μ — prior

given observed data — $\underbrace{p(z, \mu \mid x)}$

our data



$p(z, \mu \mid x)$ is intractable

For a given choice of prior, and the likelihood $\rightarrow x_i$

$$\boxed{\prod_{i=1}^{n}} \sum_{j=1}^{k} \pi_j \, N(x_i \mid \mu_j, \sigma_j^2),$$

the posterior is a mixture with $\underbrace{k^n \text{ terms}}$

Gibbs Sampler

* we have a posterior density $p(z, \mu | x)$
* Our interests: the characteristics of a marginal

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

- intractable

* Gibbs sampler allows to sample
  - $\underbrace{z_1, z_2, \ldots\ldots z_M \sim p(z)}$ [without requiring $p(z)$]

  - Once we have a large sample, to calculate the mean of $p(z)$ we can use the sample mean

  - Gibbs sampler generates a sample from $p(z)$ by sampling from
    (1) $p(z | \mu)$ ←
    (2) $p(\mu | z)$
    iteratively

$z \sim p_x(z | \mu)$ ← Multinomi[al]

$\mu \sim p_x(\mu | z)$ ← Normal sample