Gaussian mixture model (GMM)

If $\Sigma > 0$ then the density for $x \sim N_p(\mu, \Sigma)$ is

$$f(x;\mu,\Sigma) = \left(\frac{1}{2\pi}\right)^{p/2} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

GMM assumes that $x_1 \dots x_n$ are drawn i.i.d. from density:

$$p(x) = \sum_{k=1}^{K} \pi_k f(x; \mu_k, \Sigma_k),$$ where

mixing coefficients $\pi_k \geqslant 0$ and $\sum_{k=1}^{K} \pi_k = 1$

that is a mixture of K multivariate Gaussian distributions.

Unknown parameters: $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$

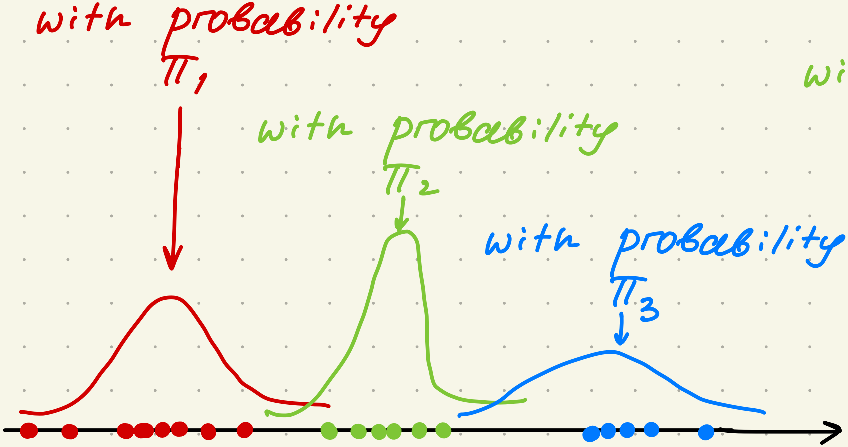
generative model for GMM:

$$z_i = \begin{cases} 1 & \text{with probability } \pi_1 \\ \ldots & \\ K & \text{with probability } \pi_K \end{cases}$$

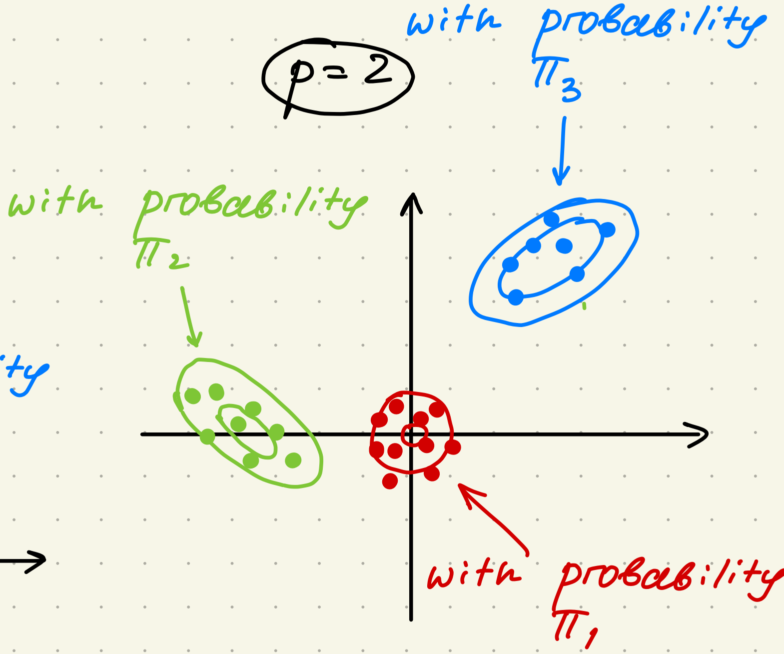
latent variable

$$x_i \mid z_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$$

$p = 1$



$p = 2$



Maximize log-likelihood:

$$\sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(x_i; \mu_k, \Sigma_k)\right)$$

- If $K=1$ the solution is easy:

$$\pi_1 = 1 \qquad \mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \Sigma_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

- In general, $\log\left(\sum_{k=1}^{K} \dots\right)$ is the problem.

1 If we know $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$

we could compute $p(z_i = k \mid x_i)$

$$p(z_i = k \mid x_i) = \frac{p(z_i = k) \cdot p(x_i \mid z_i = k)}{p(x_i)} =$$

$$= \frac{\pi_k \, f(x_i ; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, f(x_i ; \mu_j, \Sigma_j)}$$

x_i belongs to $\to$ cluster 1 with $p(z_i = 1 \mid x_i)$

....

$\to$ cluster K with $p(z_i = K \mid x_i)$

We can pick $z_i = \underset{k=1\dots K}{\operatorname{argmax}} \, p(z_i = k \mid x_i)$

② If we knew z_i then

- $\pi_k = \frac{1}{n} \sum_{i=1}^{n} I(z_i = k)$ - proportion of observations in cluster k
- $\mu_k = \dfrac{\sum_{i=1}^{n} I(z_i = k)\, x_i}{\sum_{i=1}^{n} I(z_i = k)}$ - sample mean of cluster k
- $\Sigma_k = \dfrac{\sum_{i=1}^{n} I(z_i = k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} I(z_i = k)}$ - sample co-variance of cluster k

$$\sum_{i=1}^{n} \log p(x_i, z_i) = \sum_{i=1}^{n} \left[\log p(x_i \mid z_i) + \log p(z_i)\right]$$

$$= \sum_{i=1}^{n} \left[\log f(x_i; \mu_{z_i}, \Sigma_{z_i}) + \log \pi_{z_i}\right] =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} I(z_i = k)\left(\log f(x_i; \mu_k, z_k) + \log \pi_k\right)$$

Estimating π_k:

$$\sum_{k=1}^{K} \log \pi_k \cdot \underbrace{\sum_{i=1}^{n} I(z_i = k)}_{n_k} = \sum_{k=1}^{K} n_k \log \pi_k \longrightarrow \max$$

(*)

$$\sum_{k=2}^{K} n_k \log \pi_k + n_1 \log \left(1 - \sum_{k=2}^{K} \pi_k\right)$$

$$\nabla_{\pi_k} = \frac{n_k}{\pi_k} - \frac{n_1}{1 - \sum_{k=2}^{K} \pi_k} = 0 \implies n_k \underbrace{\left(1 - \sum_{k=2}^{K} \pi_k\right)}_{\pi_1} = n_1 \pi_k$$

$$n_k \pi_1 = n_1 \pi_k \implies \pi_k = n_k \cdot \alpha$$

$$\pi_1 + \dots + \pi_K = 1 \implies \alpha = \frac{1}{n_1 + \dots + n_K} = \frac{1}{n} \implies \pi_k = \frac{n_k}{n}$$

Estimating μ_k, Σ_k:

$$\sum_{i=1}^{n} I(z_i = k) \log f(x_i; \mu_k, \Sigma_k) = \sum_{i \in C_k} \log f(x_i; \mu_k, \Sigma_k)$$

Thus $\mu_k = \frac{1}{n_k} \sum_{i \in C_k} x_i$

$$\Sigma_k = \frac{1}{n_k} \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

① Given $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ compute

$$p(z_i = k \mid x_i) = \frac{\pi_k f(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f(x_i; \mu_j, \Sigma_j)}$$

$$z_i = \underset{k=1\ldots K}{\operatorname{argmax}}\ p(z_i = k \mid x_i), \quad w_{ik} = I(z_i = k)$$

each x_i is assigned to cluster $\underset{\substack{\uparrow \\ w_{i1}}}{1} \ldots \underset{\substack{\uparrow \\ w_{iK}}}{K}$

② Given z_i compute

- $$\pi_k = \frac{1}{n} \sum_{i=1}^{n} I(z_i = k) = \frac{\sum_{i=1}^{n} w_{ik}}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}}$$
- $$\mu_k = \frac{\sum_{i=1}^{n} I(z_i = k)\, x_i}{\sum_{i=1}^{n} I(z_i = k)} = \frac{\sum_{i=1}^{n} w_{ik}\, x_i}{\sum_{i=1}^{n} w_{ik}}$$
- $$\Sigma_k = \frac{\sum_{i=1}^{n} I(z_i = k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} I(z_i = k)} = \frac{\sum_{i=1}^{n} w_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} w_{ik}}$$

EM algorithm

Expectation (E) step: Given $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$

$$w_{ik} = p(z_i = k \mid x_i) = \frac{\pi_k \, f(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, f(x_i; \mu_j, \Sigma_j)}$$

each x_i is partially assigned to cluster $1 \ldots K$

$$\underset{w_{i1}}{1} \; \ldots \; \underset{w_{iK}}{K}$$

Maximization (M) step: Given w_{ik}

- $$\pi_k = \frac{\sum_{i=1}^{n} w_{ik}}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}}$$
- $$\mu_k = \frac{\sum_{i=1}^{n} w_{ik} x_i}{\sum_{i=1}^{n} w_{ik}}$$
- $$\Sigma_k = \frac{\sum_{i=1}^{n} w_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} w_{ik}}$$

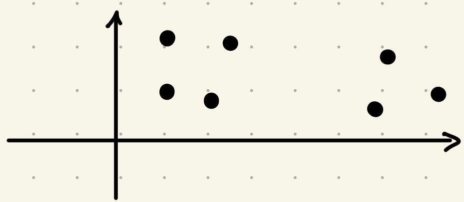
<u>E-step</u>: which of $N_p(\mu_1, \Sigma_1) \ldots N_p(\mu_K, \Sigma_K)$ generated each x_i?

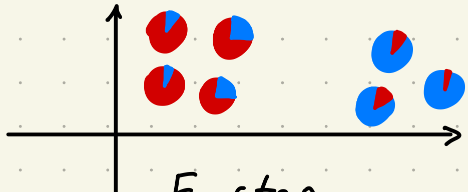
We are not sure, so we assign probabilities

x_i came from $N(\mu_1, \Sigma_1)$ with probability w_{i1}
$\vdots$ $\vdots$
$N(\mu_K, \Sigma_K)$ with probability w_{iK}

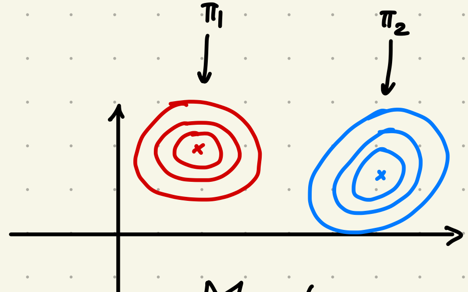
<u>M-step</u>: we want to estimate $\mu_1 \ldots \mu_K$ and $\Sigma_1 \ldots \Sigma_K$

For μ_k and Σ_k each observation x_i will have weight w_{ik}. We use weighted sample mean and sample variance.





E-step



M-step

EM algorithm : motivation

Denote by θ the set of parameters and by X the set of observations.

$$\ell(X;\theta)=\sum_{i=1}^{n}\log p(x_i;\theta)=\sum_{i=1}^{n}\log\Big(\sum_{k=1}^{K}p(x_i,z_i=k;\theta)\Big)$$

any observation

- $$\log p(x;\theta)\geqslant\sum_{k=1}^{K}\omega_k\cdot\log\Big(\frac{p(x,z=k;\theta)}{\omega_k}\Big)$$

  for $\omega_1\ldots\omega_K$ such that $\omega_1+\ldots+\omega_K=1$ and $\omega_k\geqslant 0$

log is concave function thus

$$\log\Big(\sum_{k=1}^{K}a_k\Big)=\log\Big(\sum_{k=1}^{K}\omega_k\cdot\frac{a_k}{\omega_k}\Big)\geqslant\sum_{k=1}^{K}\omega_k\log\Big(\frac{a_k}{\omega_k}\Big)$$

- If $\omega_k = p(z=k \mid x;\theta)$ then inequaily becomes equality

$$\frac{p(x, z=k; \theta)}{\omega_k} = \frac{p(x, z=k; \theta)}{p(z=k \mid x; \theta)} = p(x;\theta)$$

$$\sum_{k=1}^{k} \omega_k \cdot \log\left(\frac{p(x, z=k; \theta)}{\omega_k}\right) = \sum_{k=1}^{k} \omega_k \cdot p(x;\theta) = p(x;\theta)$$

- Denote weights for observation x_i by $w_{i1} \dots w_{ik}$

<u>E-step</u>: at iteration t compute

$$w_{ik}^{(t)} = p(z_i = k \mid x_i; \theta^{(t-1)}).$$

for each i, $w_{ik}^{(t)} \geq 0$ and $\sum_{k=1}^{K} w_{ik}^{(t)} = 1$

$$\ell(X;\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} p(x_i, z_i = k; \theta)\right) \geq$$

$$\sum_{i=1}^{n} \sum_{k=1}^{k} w_{ik}^{(t)} \log\left(\frac{p(x_i, z_i = k; \theta)}{w_{ik}^{(t)}}\right)$$

- M-step: at iteration t maximize the lower-bound for $\ell(x;\theta)$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log\left(\frac{p(x_i, z_i = k; \theta)}{w_{ik}^{(t)}}\right) =$$

ignore

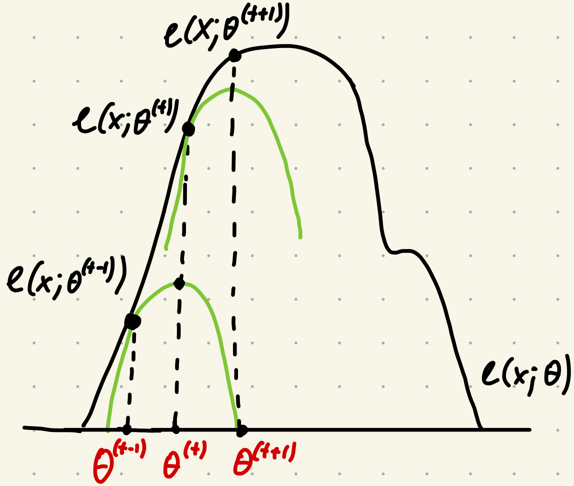
$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \left[\log\left(p(x_i, z_i = k; \theta)\right) - \log w_{ik}^{(t)}\right] = Q(\theta) + \ldots$$

M-step: $\max_{\theta}$imize $Q(\theta)$

- $\ell(x;\theta^{(t-1)}) \leq \ell(x;\theta^{(t)})$

$$\ell(x;\theta^{(t-1)}) = Q(\theta^{(t-1)}) + \ldots \leq$$
$$\leq Q(\theta^{(t)}) + \ldots \leq \ell(x;\theta^{(t)})$$

- log-likelihood converges to local maximum.



EM for GMM

E-step Given $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$

$$w_{ik} = p(z_i = k \mid x_i) = \frac{\pi_k f(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f(x_i; \mu_j, \Sigma_j)}$$

M-step maximize$_\theta$

$$Q(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log\big(p(x_i, z_i = k; \theta)\big) =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \big(\log f(x_i; \mu_k, \Sigma_k) + \log \pi_k\big)$$

Estimating π_k:

$$\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log \pi_k = \sum_{i=1}^{n} \Big(\sum_{k=2}^{K} w_{ik} \log \pi_k + w_{i1} \log\big(1 - \sum_{k=2}^{K} \pi_k\big)\Big)$$

$$\nabla_{\pi_k} = \sum_{i=1}^{n} \Big[w_{ik}/\pi_k - w_{i1}/\big(1 - \sum_{k=2}^{K} \pi_k\big)\Big] = \frac{\sum_{i=1}^{n} w_{ik}}{\pi_k} - \overset{1/\alpha}{\frac{\sum_{i=1}^{n} w_{i1}}{\pi_1}} = 0 \Rightarrow \pi_k = \alpha \sum_{i=1}^{n} w_{ik}$$

as $\sum_{k=1}^{K} \pi_k = 1$ then $\alpha = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Rightarrow \pi_k = \frac{\sum_{i=1}^{n} w_{ik}}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}}$

Estimating μ_k:

$$\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \ (x_i - \mu_k)^T \Sigma_k (x_i - \mu_k)$$

$$\nabla_{\mu_k} = -2 \sum_{i=1}^{n} w_{ik} \Sigma_k (x_i - \mu_k) = -2 \Sigma_k \sum_{i=1}^{n} w_{ik} (x_i - \mu_k) = 0$$

$$\sum_{i=1}^{n} w_{ik} x_i = \mu_k \cdot \sum_{i=1}^{n} w_{ik} \Rightarrow \mu_k = \frac{\sum_{i=1}^{n} w_{ik} x_i}{\sum_{i=1}^{n} w_{ik}}$$