GMM + EM 4/29/26 (1)

$p(x|\theta)$

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

k = cluster index

π_k = k selection prob → mixing coefficients

μ_k, Σ_k = single gaussian params.

θ = some params = $\{\pi_k, \mu_k \Sigma_k\}$

~~$2(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$~~

$$2(X|\theta) = \prod_{n=1}^{N} p(x_i|\theta)$$

$$\ell = \log 2(X|\theta) = \log \prod_{n=1}^{N} p(x_i|\theta)$$

$$\Rightarrow \sum_{n=1}^{N} \log p(x_i|\theta)$$

$$\Rightarrow \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \right]$$

~~instead of weight, assume we know which gaussian a datapoint comes from.~~

Log-likelihood of single datapoint.

$$\ln p(x|\theta) = \sum_z q(z) \ln p(x|\theta)$$ → Introduce additional variable for sum-sies..

Now rewrite $p(x|\theta)$ to include z.

$$p(x|\theta) = \frac{p(x,\theta)}{p(\theta)} = \frac{\sum_z p(x,\theta,z)}{\sum_z p(\theta,z)} = \frac{\sum_z p(x,z|\theta)\, p(\theta)}{\sum_z p(\theta,z)}$$??

see next page

By working backwards:

known: $p(x|\theta) = \dfrac{p(x,z|\theta)}{p(z|x,\theta)} = \dfrac{p(x,z,\theta)/p(\theta)}{p(x,z,\theta)/p(x,\theta)} = \dfrac{\frac{1}{p(\theta)}}{\frac{1}{p(x,\theta)}}$

$$\Rightarrow \frac{p(x,\theta)}{p(\theta)} = p(x|\theta)$$

Inverse:

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

$$\Rightarrow \frac{p(x,z,\theta)}{p(\theta)} \cdot \left[\frac{p(x,\theta)}{p(x,z,\theta)}\right]$$

$$\Rightarrow p(x,z|\theta) \cdot \left[\frac{p(x,z,\theta)}{p(x,\theta)}\right]^{-1}$$

$$\Rightarrow p(x,z|\theta) \cdot [p(z|x,\theta)]^{-1}$$

$$\Rightarrow \frac{p(x,z|\theta)}{p(z|x,\theta)}$$

$$\Rightarrow \ln p(x|\theta) = \sum_z q(z) \ln p(x|\theta) = \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{p(z|x,\theta)}\right]$$

$$\Rightarrow \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{p(z|x,\theta)} \cdot \frac{q(z)}{q(z)}\right]$$ → for funsies.

$$\Rightarrow \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)} \cdot \frac{q(z)}{p(z|x,\theta)}\right]$$

$$\Rightarrow \sum_z q(z) \left[\ln\left[\frac{p(x,z|\theta)}{q(z)}\right] + \ln\left[\frac{q(z)}{p(z|x,\theta)}\right]\right]$$

$$\Rightarrow \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right] + \sum_z q(z) \ln\left[\frac{q(z)}{p(z|x,\theta)}\right]$$

$$\ln p(x|\theta) = \underbrace{\sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right]}_{\mathcal{L}(q,\theta)\ \text{ELBO}} + \underbrace{\sum_z q(z) \ln\left[\frac{q(z)}{p(z|x,\theta)}\right]}_{KL(q\|p)}$$

$KL(q\|p) \geq 0$ due to Jensen's Inequality:

$$KL(q\|p) = \sum_x q(x) \ln\left[\frac{q(x)}{p(x)}\right] = -\sum_x q(x) \ln\left[\frac{p(x)}{q(x)}\right] \Rightarrow -E_q\left[\ln\left[\underbrace{\frac{p(x)}{q(x)}}_{u}\right]\right]$$

Jensen's Inequality for convex functions:

$$E(f(u)) \geq f(E(u))$$

Expectation

$f = -\ln$ only as $+\ln$ is not convex

$$E_q[-\ln(u)] \geq -\ln[E(u)]$$

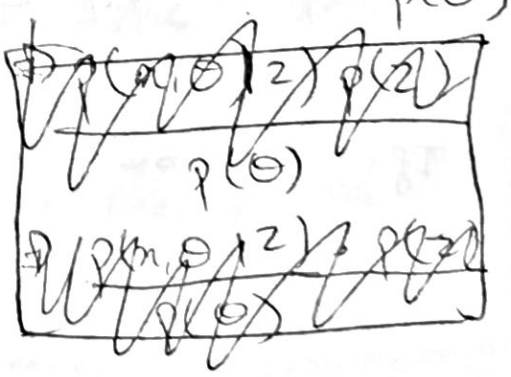
~~$E_q[-\ln(u)] \leq -\ln[E(u)] \Rightarrow E_q\left[\ln\left[\frac{p(x)}{q(x)}\right]\right] \leq -\ln\left[E_q\left[\frac{p(x)}{q(x)}\right]\right]$~~

$$-\sum_x q(x) \ln\left[\frac{p(x)}{q(x)}\right] \geq -\ln\left[\sum_x q(x) \frac{p(x)}{q(x)}\right]$$

$$\geq -\ln\left[\sum_x p(x)\right] \geq -\ln[1] \geq 0$$

$$\boxed{-\sum_x q(x) \ln\left[\frac{p(x)}{q(x)}\right] \geq 0}$$

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



$$\Rightarrow \frac{p(x|z,\theta)\,p(z,\theta)}{p(\theta)}$$

$$\Rightarrow p(x|z,\theta)\,\frac{p(z,\theta)}{p(\theta)}$$

$$\Rightarrow p(x|z,\theta)\,p(z|\theta)$$

$$p(x|\theta) = \sum_{k=1}^{k} p(x, z_k = 1|\theta)$$

cluster selector.

$$p(x,z|\theta) = p(z|\theta)\,p(x|z,\theta)$$

prior over latent cluster.

$$\ln p(x|\theta) = \log\left[\sum_k^K \pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k)\right] = \log \sum_k^K p(x,z|\theta)$$

~~Instead if we do marginalize k (z_k)~~ or

$$\Rightarrow \log \sum_k^K p(z_k|\theta)\, p(x|z_k,\theta) \Rightarrow \log \sum_k^K p(x,z_k|\theta)$$

Instead of marginalizing over z_k, we can introduce an alternative form for $p(x|\theta) = \dfrac{p(x,z|\theta)}{p(z|x,\theta)}$ which leads us to the ELBO + KL decomposition.

$$\Rightarrow \ln p(x|\theta) = \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right] + \sum_z q(z) \ln\left[\frac{q(z)}{p(z|x,\theta)}\right]$$

~~$\Rightarrow \sum_z q(z) \ln \left[\quad p(x,z|\theta) = \pi_k \mathcal{N}(x|\mu_k,\Sigma_k)\right]$~~

$$p(x,z|\theta) = \prod_k^K \left[\pi_k \mathcal{N}(x|\mu_k,\Sigma_k)\right]^{z_k} \quad \text{for } z_k = \delta(\text{cluster} = k)$$

EM Intuition: $\underbrace{\ln p(x|\theta)}_{\text{constant for some } \theta} = \underbrace{\mathcal{L}(q,\theta)}_{\text{ELBO}} + \underbrace{KL_{q\|p}}_{\geq 0}$

If we hold θ fixed: $\Delta \underbrace{\ln p(x|\theta)}_{0} = \Delta \mathcal{L}(q,\theta) + \Delta KL_{q\|p}$

$$\Delta \mathcal{L}(q,\theta) = -\Delta KL_{q\|p}$$

If we ~~minimize~~ decrease $KL_{q\|p}$ for θ fixed, the ELBO will go up.

$KL_{q\|p} = 0$ for $q = p$.

$\Rightarrow q(z) = p(z|x,\theta_{\text{fixed}})$

GMM - EM

w/ $KL_{q||p} = 0$, we have $\ln p(x|\theta) = \mathcal{L}(q, \theta)$

and because $KL_{q||p} = 0$ and can't go below 0, any changes we make to $\mathcal{L}(q, \theta)$ cannot reduce $\ln p(x|\theta)$.

E-step:

Start w/ find q such that $KL_{q||p} = 0$

$\Rightarrow q(z) = p(z|x, \theta) \Rightarrow p(z|x)$ as $\theta \rightarrow$ fixed

$$p(z|x)\, p(x) = p(x|z)\, p(z)$$

1 datapoint

$$\Rightarrow p(z_k|x) = \frac{p(x|z)\, p(z)}{p(x)} = \frac{\mathcal{N}(x|\mu_k, \Sigma_k)\, \pi_k}{\sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)} = \gamma_{nk}$$

Marginalized $p(x, z)$

all points

$$\Rightarrow q(Z) = \prod_{n=1}^{N} q(z_n)$$

$$\Rightarrow q(z) \Rightarrow \prod_{n=1}^{N} \prod_{z=1}^{k} [\gamma_{nk}]^{z_{nk}}$$

N datapoint $= \prod_{n=1}^{N} \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)\, \pi_k}{\sum_k \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$

over all z

$$p(z_n|x) = \prod_{z=1}^{k} [\gamma_{nk}]^{z_{nk}} = q(z_n)$$

i.e. responsibility for all clusters for a datapoint

M-Step

Insert $q(z)$ [such that $q(z) = p(z|x,\theta)$] into complete data log-likelihood eqn which makes $KL_{q||p} = 0$

$$\ln p(x|\theta) = \sum_z q(z) \ln\left[\frac{p(x,z|\theta)}{q(z)}\right]$$

$$\Rightarrow \sum_z q(z) \left[\ln p(x,z|\theta) - \ln q(z)\right]$$

$$\Rightarrow \underbrace{\sum_z q(z) \ln p(x,z|\theta)}_{\mathbb{E}_q[\ln p(x,z|\theta)]} - \underbrace{\sum_z q(z) \ln q(z)}_{\text{Entropy}}$$

determined using θ^{old} = constant

$$\therefore \frac{\partial}{\partial \theta} = 0$$

$$\frac{\partial}{\partial \theta} \ln p(x|\theta) = \frac{\partial}{\partial \theta} \sum_z q(z) \underbrace{\ln p(x,z|\theta)}_{\ln \prod_n^N p(x_n,z_n|\theta)}$$

$$= \sum_n^N \ln p(x_n,z_n|\theta)$$

$$= \sum_n^N \ln \prod_k^K \left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]^{z_{nk}}$$

$$= \sum_n^N \sum_k^K \ln\left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]^{z_{nk}}$$

$$\Rightarrow \mathbb{E}_q\left[\sum_n^N \sum_k^K \ln\left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]^{z_{nk}}\right]$$

$$\Rightarrow \mathbb{E}_q\left[\sum_n^N \sum_k^K z_{nk} \ln\left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]\right] \qquad \mathbb{E}(A+B) = \mathbb{E}(A) + \mathbb{E}(B)$$

$$\Rightarrow \sum_n^N \sum_k^K \mathbb{E}_q\left[z_{nk} \underbrace{\ln\left[\pi_k \mathcal{N}(x_k|\mu_k,\Sigma_k)\right]}_{\text{constant w.r.t } q}\right]$$

$$\Rightarrow \sum_n^N \sum_k^K \mathbb{E}_q(z_{nk}) \ln\left[\pi_k \mathcal{N}(x_k|\mu_k,\Sigma_k)\right]$$

GMM-EM

$$\Rightarrow \sum_{n}^{N}\sum_{k}^{K} \underbrace{E_q(z_{nk})}\ln\left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right]$$

$$\Rightarrow E_q(z_{nk}) = \sum_{z_n} q(z_n) z_{nk} \quad \text{(selector variable)}$$

$$= [1 \cdot q(z_{nk}=1)] + [0 \cdot q(z_{nk}=0)]$$

$$= q(z_{nk}=1) = \gamma_{nk}$$

$$\Rightarrow \sum_{n}^{N}\sum_{k}^{K} \gamma_{nk} \ln\left[\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right] = \ln p(x|\theta)$$

Derive parameter updates: (of a specific cluster so don't have to sum over k.)

μ_k

$$\Rightarrow \frac{\partial}{\partial \mu_k}\ln p(x|\theta) = \frac{\partial}{\partial \mu_k}\sum_{n}^{N}\gamma_{nk}\left[\underbrace{\ln \pi_k}_{\frac{\partial}{\partial \mu_k}=0} + \underbrace{\ln \mathcal{N}(x_n|\mu_k,\Sigma_k)}\right] = Q$$

$$\Rightarrow \frac{\partial}{\partial \mu_k}\sum_{n}^{N}\gamma_{nk}\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

$$\Rightarrow \frac{1}{2}\sum_{n}^{N}\gamma_{nk}\cdot\frac{\partial}{\partial \mu_k}\underbrace{\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right]}_{u \quad v}$$

$$\frac{\partial v}{\partial \mu_k} = -1, \quad \frac{\partial u}{\partial \mu_k} = -\Sigma^{-1}$$

$$\frac{\partial}{\partial \mu_k}(u\cdot v) = u'v + uv' \Rightarrow -\Sigma^{-1}(x-\mu) - (x-\mu)^T\Sigma^{-1}$$

since Σ^{-1} is symmetric $[a\Sigma^{-1}]^T = \Sigma^{-1}a^T$

$$\Rightarrow -2\Sigma^{-1}(x-\mu)$$

$$\frac{\partial Q}{\partial \mu_k} = -\frac{1}{2}\sum_{n}^{N}\gamma_{nk}\left[-2\Sigma^{-1}(x-\mu)\right] = \sum_{n}^{N}\gamma_{nk}\Sigma_k^{-1}(x_n-\mu_k)$$

$$\frac{\partial}{\partial \mu_k}\ln\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right]$$

$$\Rightarrow \ln\underbrace{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\partial/\partial\mu_k=0} - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

Find max

$$\frac{\partial Q}{\partial \mu_k} = 0 = \sum_n \gamma_{nk} \Sigma_k^{-1} (x_n - \mu_k) \Rightarrow \sum_n \gamma_{nk} \Sigma_k^{-1} x_n = \sum_n \gamma_{nk} \Sigma_k^{-1} \mu_k$$

$$\Rightarrow \frac{\sum_n \gamma_{nk} \Sigma_k^{-1} x_n}{\sum_n \gamma_{nk} \Sigma_k^{-1}} = \mu_k = \frac{\Sigma_k^{-1} \sum_n \gamma_{nk} x_n}{\Sigma_k^{-1} \sum_n \gamma_{nk}}$$

$$= \mu_k \Sigma_k \sum_n \gamma_{nk} \Sigma_k^{-1}$$

$$\boxed{\mu_k = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}}$$

π_k

$$\frac{\partial}{\partial \pi_k} Q = \sum_n \gamma_{nk} \ln \pi_k$$

But we want to optimize subject to constraint $\sum_k \pi_k = 1$

Constrained optimization:

$$\Rightarrow J = \sum_n \sum_k \gamma_{nk} \ln \pi_k + \lambda \left(\sum_k \pi_k - 1 \right)$$

$$\frac{\partial J}{\partial \pi_k} = \sum_n \sum_k \gamma_{nk} \cdot \frac{1}{\pi_k} + \frac{\partial}{\partial \pi_k} \left[\lambda \underbrace{\sum_k \pi_k}_{1} + \lambda \sum_k (-1) \right]$$

$\frac{\partial}{\partial \pi_k} = \lambda k$; $\frac{\partial}{\partial \pi_k} = 0$

$$\Rightarrow \sum_n \sum_k \frac{\gamma_{nk}}{\pi_k} + \sum_k \lambda = 0$$

or $\frac{\partial}{\partial \pi_k} = \sum_n \lambda$

for $k =$ sum things

$$\Rightarrow \sum_n \frac{\gamma_{nk}}{\pi_k} = \sum_n \left[\sum_k \frac{\gamma_{nk} + \lambda}{\pi_k} \right] \Rightarrow \sum_k \left[\sum_n \gamma_{nk} + \lambda \pi_k \right]$$

for one $k \Rightarrow \sum_n \gamma_{nk} + \lambda \pi_k = 0 \Rightarrow \sum_n \gamma_{nk} = -n \lambda \pi_k$

$$\Rightarrow \sum_n \sum_k \frac{\gamma_{nk}}{\pi_k} = \sum_k \sum_n \gamma_{nk}$$

$$\lambda = -\sum_n \gamma_{nk}$$

GMM - EM

$$\frac{\partial J}{\partial \pi_k} = \sum_n \sum_k \frac{\gamma_{nk}}{\pi_k} + \sum_k \lambda = 0$$

$$\left[\sum_k \frac{1}{\pi_k} \sum_n \gamma_{nk} \right] + \sum_k \lambda = 0 \Rightarrow \sum_k \left[\frac{1}{\pi_k} \sum_n \gamma_{nk} + \lambda \right] = 0$$

$\Rightarrow$ Can lose the $\sum_k$ b/c only the k-th term will have a non-zero derivative.

$$\Rightarrow \sum_n \frac{\gamma_{nk}}{\pi_k} + \lambda = 0 \Rightarrow \lambda = -\sum_n \frac{\gamma_{nk}}{\pi_k}$$

$$\Rightarrow \pi_k \lambda = -\sum_n \gamma_{nk} \Rightarrow \lambda \underbrace{\sum_k \pi_k}_{1} = -\sum_n \underbrace{\sum_k \gamma_{nk}}_{1}$$

All responsibility for any point $\Sigma = 1$

$$\Rightarrow \lambda = -\sum_n 1 \Rightarrow \boxed{\lambda = -N}$$

$$\sum_n \frac{\gamma_{nk}}{\pi_k} + (-N) = 0 \Rightarrow \sum_n \frac{\gamma_{nk}}{\pi_k} = N \Rightarrow \boxed{\frac{1}{N} \sum_n \gamma_{nk} = \pi_k}$$

Σ_k

$$\frac{\partial}{\partial \Sigma_k} \ln p(x|\theta) = \frac{\partial}{\partial \Sigma_k} \sum_n^N \gamma_{nk} \left[\ln \pi_k + \ln \mathcal{N}(x_n | \mu_k, \Sigma_k) \right]$$

$\ln \pi_k$: $\frac{\partial}{\partial \Sigma_k} = 0$

$$\Rightarrow \sum_n^N \gamma_{nk} \frac{\partial}{\partial \Sigma_k} \left[\ln \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} - \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right]$$

$$\Rightarrow \frac{\partial}{\partial \Sigma_k} \ln \left[(2\pi)^{d/2} |\Sigma_k|^{1/2} \right]^{-1} \Rightarrow \frac{\partial}{\partial \Sigma_k} \ln \left[(2\pi)^{d/2} |\Sigma_k| \right]^{-1/2} = \frac{\partial}{\partial \Sigma_k} -\frac{1}{2} \ln \left[(2\pi)^d |\Sigma_k| \right]$$

$$\Rightarrow -\frac{1}{2} \frac{\partial}{\partial \Sigma_k} \left[\underbrace{d \ln 2\pi}_{0} + \ln |\Sigma_k| \right] \Rightarrow -\frac{1}{2} \Sigma^{-1}$$

via magic

$$\frac{\partial}{\partial \Sigma_k} \underbrace{(x-\mu)^T \Sigma_k^{-1}}_{u} \underbrace{(x-\mu)}_{v} \Rightarrow \frac{\partial u}{\partial \Sigma_k} (x-\mu^T)\Sigma_k^{-1} \Rightarrow (x-\mu^T)\frac{\partial}{\partial \Sigma_k}\Sigma^{-1}$$

$$\frac{\partial}{\partial \Sigma_k} uv = u'v + uv' \qquad \frac{\partial v}{\partial \Sigma_k} = 0$$

$$\Rightarrow (x-\mu)^T \left[\Sigma^{-1}\frac{d\Sigma}{d\Sigma}\Sigma^{-1}\right]$$

$$\Rightarrow (x-\mu)^T \left[-\Sigma^{-1}\frac{d\Sigma}{d\Sigma}\Sigma^{-1}\right](x-\mu) \overset{\text{by magic}}{\Rightarrow} -\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1}$$

$$\frac{\partial}{\partial \Sigma_k} \ln p(x|\Theta) = \sum^n \gamma_{nk}\left[-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1}\right]$$

$$\Rightarrow \frac{1}{2}\sum^n \gamma_{nk}\left[\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1} - \Sigma_k^{-1}\right]$$

$$= \frac{\Sigma_k^{-1}}{2}\sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_k^{-1} - \mathbb{1}\right]$$

$$= \frac{\Sigma_k^{-1}}{2}\left[\sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T - \Sigma_k\right]\right]\Sigma_k^{-1} = 0$$

$$\Rightarrow \sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T - \Sigma_k\right] = 0$$

$$\Rightarrow \sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T\right] - \sum^n \gamma_{nk}\Sigma_k = 0$$

$$\sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T\right] = \sum^n \gamma_{nk}\Sigma_k = \Sigma_k \sum^n \gamma_{nk}$$

$$\boxed{\Sigma_k = \frac{\sum^n \gamma_{nk}\left[(x_n-\mu_k)(x_n-\mu_k)^T\right]}{\sum^n \gamma_{nk}}}$$