Lecture 6|13 : Generative Models so far

train $P(X)$ = density / curve [param] fit to data $\times$ (sep for each Y)

predict $P(Y|X) = \dfrac{P(X|Y) \cdot P(Y)}{P(X)}$

- For each Y $P(X|Y) = N(X|\mu, \Sigma)$ (D-dim) $\Longrightarrow$ Gauss DA
- For each Y $P(X|Y) = P(x^1 x^2 \ldots x^D|Y)$
  $= P(x^1|Y) \cdot P(x^2|Y) \ldots P(x^D|Y) \Rightarrow$ Naive Bayes
- Not-So-Naive Bayes: → Belief Network: group features that are very dependent

For each Y: $P(X|Y) = P(x^1 x^2 \ldots x^D|Y) =$ indep

$= P(x^1, x^2|Y) \cdot P(x^3|Y) \cdot P(x^4, x^5|Y, x^1) \cdot P(x^6, x^7|Y) \ldots$

2dim — 1-dim — 2dim fit for Y, x^1

MIXT OF GAUSSIANS (GMM)

Today / Next : $P(x) = \boxed{w_1} \cdot N_1(x \mid \underbrace{\mu_1 \Sigma_1}_{D\text{-dim}}) + \boxed{w_2} N_2(x \mid \underbrace{\mu_2, \Sigma_2}_{D\text{-dim}}) + \boxed{w_3} N_3(x \mid \underbrace{\mu_3 \Sigma_3}_{D\text{-dim}})$

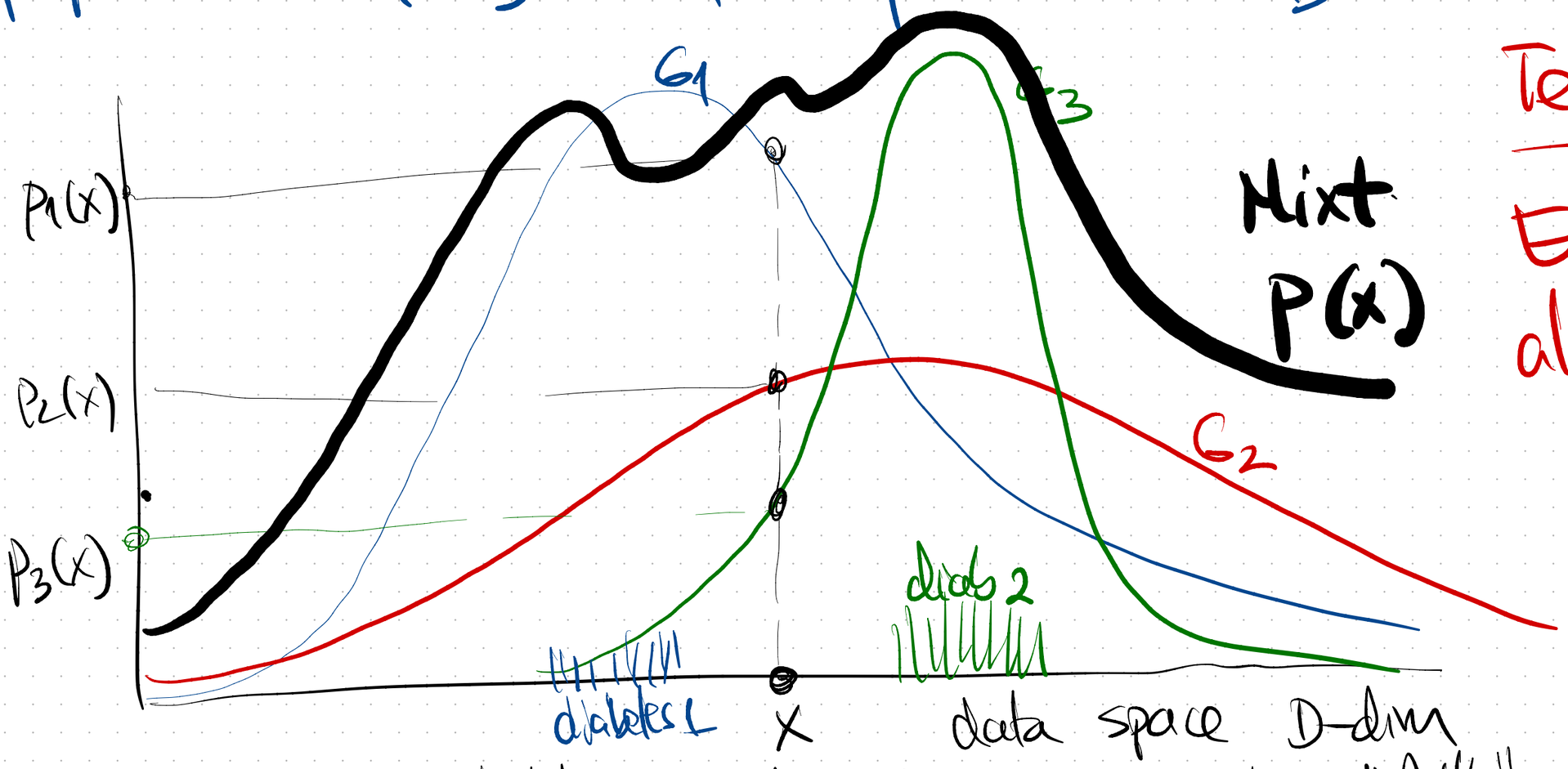
w_1 fixed, w_2 fixed, w_3 fixed

$w_1 + w_2 + w_3 = 1$

proportions

$K = 3$ # of "components"

D = # data dim



Mixt P(x)

Technique

Exp(p) Max) algorithm

Mixture adv: ability to fit data with multiple "hills" "humps"

GMM $= w_1 \cdot \mathcal{N}_1(\,) + w_2 \cdot \mathcal{N}_2(\,) + w_3 \cdot \mathcal{N}_3(\,)$ $\quad K=3$

- "generative process": assume that datapoint x_i is generated indep of all other datapoints, in 2 steps.
  1) select one comp (gauss), with fix probab $w_1 | w_2 | w_3$
     generator $\mathcal{N}_k(\mu_k, \Sigma_k)$
  2) sample x_i from $\mathcal{N}_k$ density (D-dim)

- Think of K-means (hard) and soft-kmeans mechanics

- π_{ik} = membership
  RV. { 1 if x_i generated by $\mathcal{N}_k(\,)$
      { 0 if not

  prob π_{ik} = prob that $\mathcal{N}_k(\,)$ generated x_i

$\pi_{ik} = \langle \pi_{ik} \rangle = E[\pi_{ik}] = \text{prob}(\pi_{ik} = 1)$

fixed i: $\sum_{k=1}^{K} \langle \pi_{ik} \rangle = 1$

fixed k (cluster/generator)

$\sum_{i=1}^{N} \langle \pi_{ik} \rangle = E[\#\text{datapoints generated by } \mathcal{N}_k] \simeq w_k$

Recap for intuition: Kmeans clustering algorithm.

X D-dim data $\Rightarrow$ cluster into $K=6$ groups

$i=1:N$ $\quad$ $k=1:6$

- π_{ik} = membership
  - 1 if $x_i \rightarrow$ cluster k
  - 0 if not
- μ_k = centroid for cluster k — param

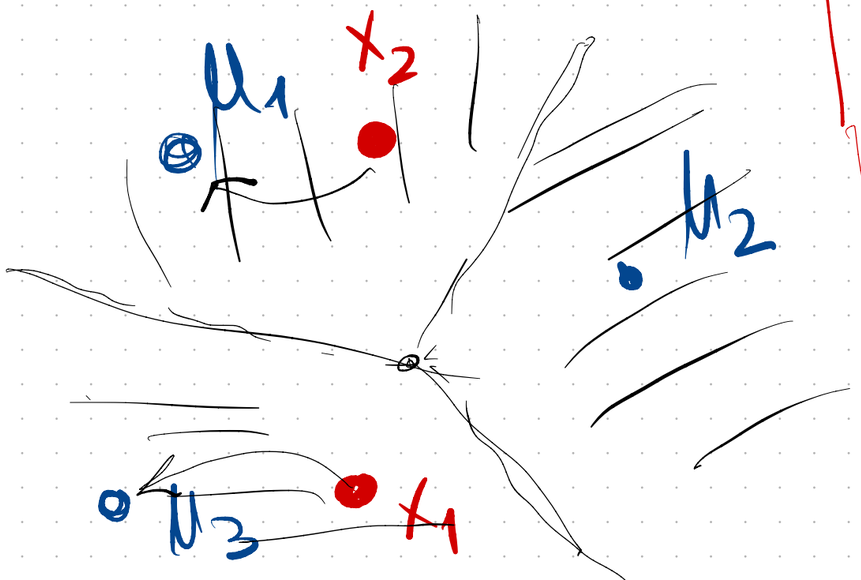
E step

calculate π_{ik}, given μ_k

$\pi_{ik} = 1$ for closest μ_k

$= \underset{k}{\operatorname{argmin}} \|x_i - \mu_k\|^2$

HARD: $\pi_{ik} \in \{0, 1\}$



M-Step

calculate μ_k, given π_{ik}

μ_k = avg of (x points) in cluster k

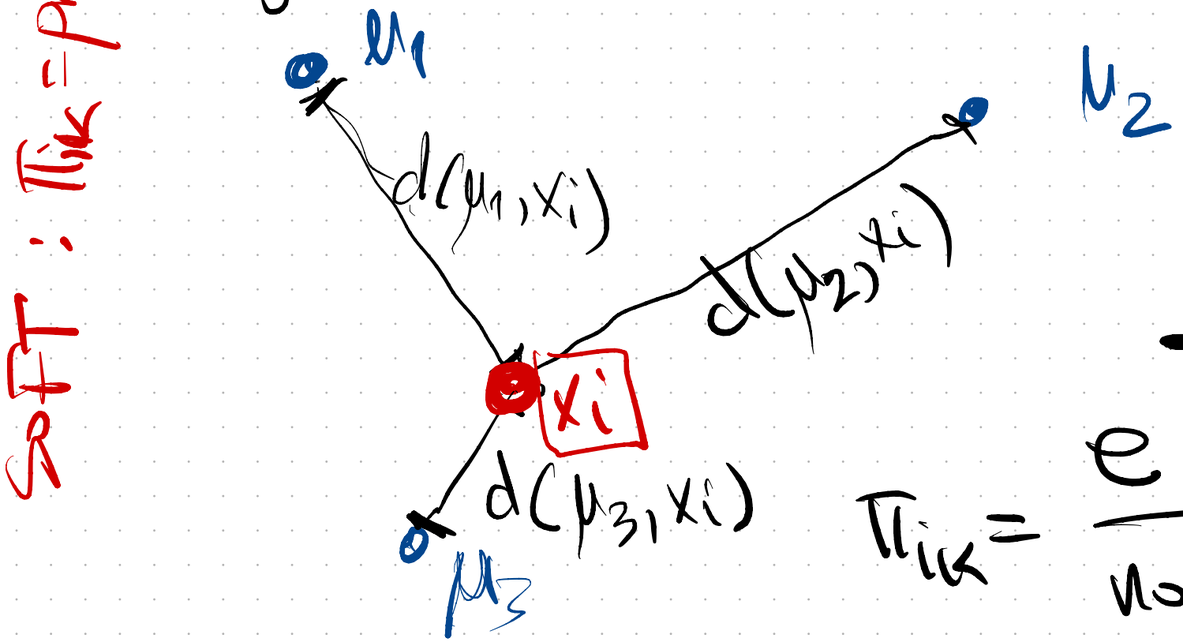
$$= \frac{\sum_{i=1}^{N} x_i \cdot \pi_{ik}}{\sum_{i=1}^{N} \pi_{ik}}$$

filter only points in cluster k

$\sum_{i=1}^{N} \pi_{ik} = N_k$ = # of points in cluster k

E step

- score dist rep output ⟹ probability



π_{ik} : $\pi_{ik} = \text{prob}(x_i \leftrightarrow \mu_k)$

RBF

fix datapoint x_i

$[\pi_{i1}\ \pi_{i2} \ldots \pi_{iK}]$ distribution over clusters

- parts of x_i distributed → $x_i \cdot \pi_{ik}$
- which cluster/group x_i is in ⟹ which generator μ_k produced datap x_i

$$\pi_{ik} = \frac{e^{-1/\beta \overbrace{\|x_i - \mu_k\|^2}^{d(x_i, \mu_k)}}}{\text{normalized over all } k}$$

small dist $\|x_i - \mu_k\|^2 \equiv$ large π_{ik}

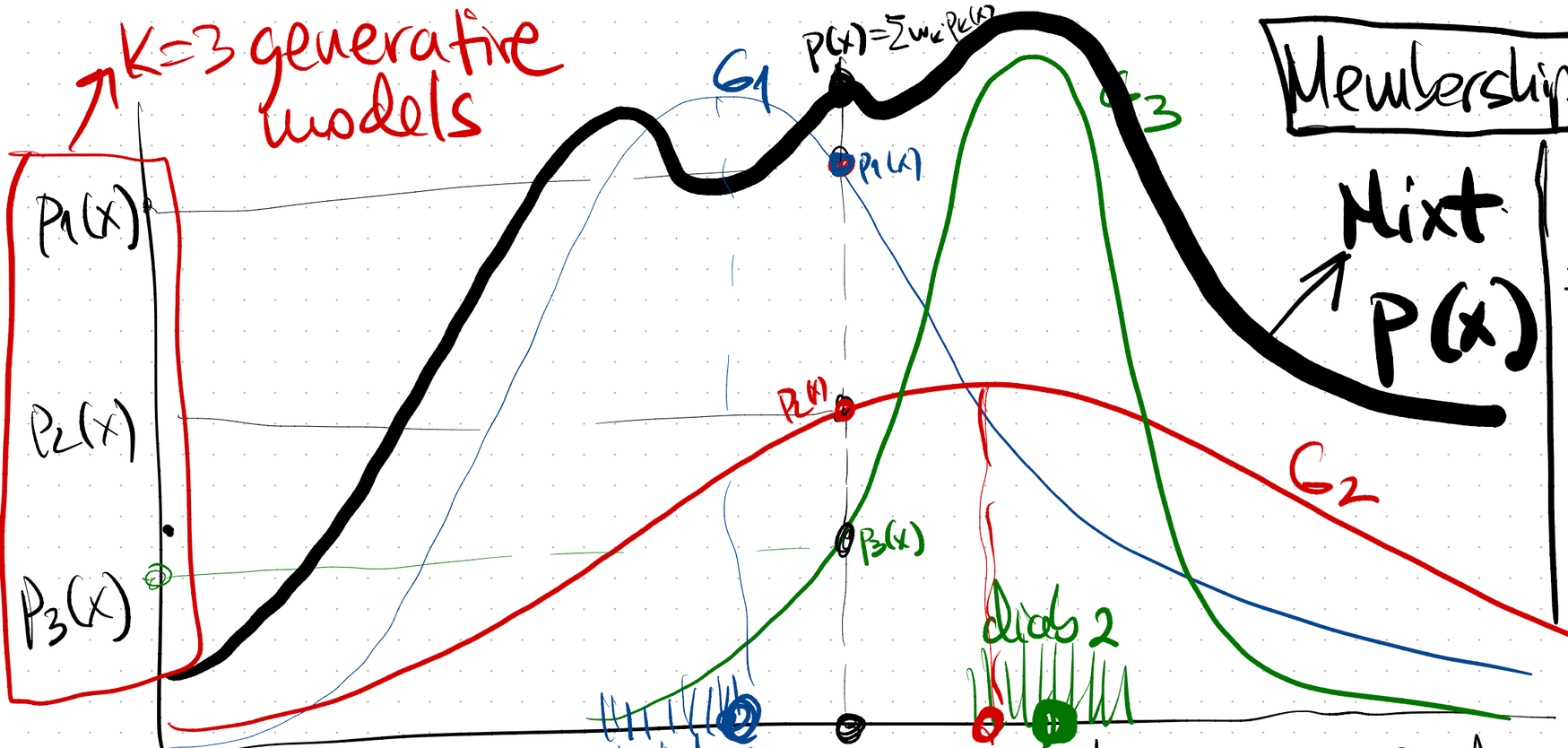
M step (same)

$$\mu_k = \frac{\sum_{i=1}^{N} x_i \cdot \boxed{\pi_{ik}}}{\sum_{i=1}^{N} \pi_{ik}}$$

proportion (weight)

→ weighted avg

$N_k = E[\#\text{points in cluster } k]$



Membership π_{ik} = prob that x_i was generated by Gauss k

$= pr(k\,gen\,|\,x_i) = w_k \cdot pr(x_i \mid \mu_k \Sigma_k)$

param k

normalized over all k

const: w_k = proportion size of cluster/source k

$E[\#datap]$ from k $= \sum_{i=1}^{N} \pi_{ik} / N$

Mixture adv: ability to fit data with multiple "hills" "humps"

$$p(x) = \sum_{k=1}^{3} w_k \cdot \underbrace{W_k(x \mid \mu_k \Sigma_k)}_{D\text{-dim}}$$

↓ prob to observe patient/email/car x_i overall sources $k=1, k=2, \ldots k=K$

π_{ik} = probab of source/generator k for observed datap x_i (model the evidence)

w_k = prob of source/gen k in general (prior) → given x_i

E step

calculate π_{ik} given params $(\mu_k, \Sigma_{ik}, w_k)_{k=1:K}$

prob of x_i generated by comp-k

$\langle \pi_{ik} \rangle =$

$\pi_{ik} = pr(k \text{ source generated } x_i)$

$= pr(k \mid x_i) = \dfrac{Pr(x_i \mid k) \cdot pr(k)}{Pr(x_i)}$

$= \dfrac{N_k(x_i \mid \mu_k, \Sigma_{ik}) \cdot w_k}{\text{normalize over } k \text{ generators}}$

M step

calculate param $(\mu_k, \Sigma_k, w_k)_{k=1:K}$ given memberships π_{ik}

- Log likelihood (data) = LL(x)
- take expectation w.r.t π_{ik}

  $E[LL(x)] \simeq LL(x)$
- take differentials

$\dfrac{\partial LL(x)}{\partial \mu_k} = 0 \quad \left| \quad \dfrac{\partial LL(x)}{\partial \Sigma_k^{-1}} = 0 \right.$

Constrain $w_1 + w_2 + w_3 = 1$

Lagrangian

$\dfrac{\partial LL(x)}{\partial w_k} \Rightarrow ?$

$$\text{Likelihood}(\text{data } X) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} \left[p(x_i \mid \pi_{ik}) \cdot p(\pi_{ik}) \right]$$

π_{ik} RV — $p(\pi_{ik})$ RV — π_{ik} RV

indep of i — w_k — filter 1/0

$$\text{Log Likelihood}(X) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\pi_{ik} \log\left(p(x_i \mid \pi_{ik}) \right) + \pi_{ik} \log(w_k) \right]$$

TOO HARD TO OPTIMIZE! $\Rightarrow$ $E[\text{loglikelihood}]$ w.r.t. π_{ik}

$$ELL(X) = \sum_{i=1}^{N} \sum_{k=1}^{K} \langle \pi_{ik} \rangle \log N_k(x_i \mid \mu_k \Sigma_k) + \langle \pi_{ik} \rangle \log(w_k)$$

(avg over π_{ik}) — prob — pros

$$N_k(x_i \mid \mu_k \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (x_i - \mu_k) \Sigma_k^{-1} (x_i - \mu_k)^T \right\}$$

D-dim gauss

① optimize for μ_k : $\log(N_k(x_i \mid \mu_k, \Sigma_k)) =$

$$= \log \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} - \frac{1}{2} (x_i - \mu_k) \Sigma_k^{-1} (x_i - \mu_k)^T$$

$$\frac{\partial \log(N_k(x_i))}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\left[\log \frac{1}{(2\pi)^{p/2}(\Sigma_k)^{1/2}}\right] - \frac{\partial}{\partial \mu_k}\left[\frac{1}{2}(x_i-\mu_k)\Sigma_k^{-1}(x_i-\mu_k)^T\right]$$

one point

$$= -\frac{1}{2}\frac{\partial}{\partial \mu_k}(x_i-\mu_k)\Sigma_k^{-1}(x_i-\mu_k)$$

$$= \boxed{(x_i-\mu_k)\Sigma_k^{-1}}$$

Vector diff: $\frac{\partial}{\partial x}[xAx^T] = x(A+A^T)$

$\Sigma_k^{-1} + (\Sigma_k^{-1})^T = 2\Sigma_k^{-1}$

$$\frac{\partial ELL(X^{\text{all data}})}{\partial \mu_k} = \sum_{i=1}^{N} \langle \pi_{ik} \rangle \left[(x_i-\mu_k)\Sigma_k^{-1}\right]$$

$\sum_{k=1}^{K} \ldots\ldots$ only k term relevant

want $= 0$

$\times \Sigma_k$

$$\sum_{i=1}^{N} \langle \pi_{ik} \rangle x_i = \sum_{i=1}^{N} \mu_k \cdot \langle \pi_{ik} \rangle$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle x_i}{\sum_{i=1}^{N} \langle \pi_k \rangle}$$

weighted Avg

② Diff w.r.t $(\Sigma_k)^{-1}$ = COVAR MATRIX of component k

$$\frac{\partial \log N_k(x_i \mid \mu_k \Sigma_k)}{\partial \Sigma_k^{-1}} = \frac{\partial}{\partial \Sigma_k^{-1}} \left\{ \frac{1}{2} \log |\Sigma_k^{-1}| - \frac{1}{2} (x_i - \mu_k) \Sigma_k^{-1} (x_i - \mu_k)^T \right\}$$

$$= \frac{1}{2} \underset{D \times D}{\Sigma_k} - \frac{1}{2} \underset{1 \times D}{(x_i - \mu_k)^T} \underset{D \times 1}{(x_i - \mu_k)}$$

matrix diff

- $\frac{\partial \log |M|}{\partial M} = (M^{-1})^T$
- $\frac{\partial}{\partial M} x M x^T = x^T x$

All datapoints:

$$\sum_{i=1}^{N} \langle \pi_{ik} \rangle \left[\frac{1}{2} \Sigma_k - \frac{1}{2} (x_i - \mu_k)^T (x_i - \mu_k) \right] \overset{\text{want}}{=} 0$$

$$\sum_{i=1}^{N} \langle \pi_{ik} \rangle \Sigma_k = \sum_{i=1}^{N} \langle \pi_{ik} \rangle (x_i - \mu_k)^T (x_i - \mu_k)$$

$$\Rightarrow \Sigma_k = \frac{\sum_{i=1}^{N} \langle \pi_{ik} \rangle \cdot (x_i - \mu_k)^T (x_i - \mu_k)}{\sum_{i=1}^{N} \langle \pi_{ik} \rangle}$$

EMPIRICAL COVAR (comp k)

weighted avg

(3) diff w.r.t w_k = mixture coef constraint $w_1+w_2+w_3=1$

Lagrangian $ELL(x_i) - \lambda\cdot[w_1+w_2+w_3-1]$

one point

$$\frac{\partial L_{ag}}{\partial w_k} = \frac{\partial}{\partial w_k}\left[\langle \pi_{ik}\rangle \cdot \log(w_k) - \lambda[w_1+w_2+w_3-1]\right]$$

$$= \langle \pi_{ik}\rangle \cdot \frac{1}{w_k} - \lambda$$

All datapoints: $\frac{\partial Lag}{\partial w_k} = \sum_{i=1}^{N}\langle \pi_{ik}\rangle \cdot \frac{1}{w_k} - \lambda \overset{\text{want}}{=} 0$

$$\sum_{i=1}^{N}\langle \pi_{ik}\rangle = \lambda w_k \Rightarrow w_k = \frac{\sum_{i=1}^{N}\langle \pi_{ik}\rangle}{\lambda} = \frac{\sum_{i=1}^{N}\langle \pi_{ik}\rangle}{N}$$

constraint (or equiv $\frac{\partial}{\partial \lambda}$): $1 = w_1+w_2+w_3 \Rightarrow$

$$N \Big/ \sum_{i=1}^{N}\sum_{k=1}^{K}\langle \pi_{ik}\rangle \Big/ \lambda = 1$$

$$\Rightarrow \boxed{\lambda = N}$$