18-661 Recitation

Ching-Yi Lin

Outline

- EM
- GMM + (hw6 coding)
- PCA

EM – Setup

Find $\boldsymbol{\theta}$ to maximize $p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$

- Given incomplete data $\mathcal{D} = \{\boldsymbol{x}_n\}$ our goal is to compute MLE of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \arg\max \ell(\boldsymbol{\theta}) = \arg\max \log p(\mathcal{D}) = \arg\max \sum_n \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

$$= \arg\max \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$$

log-sum (log(∑x)) is hard to compute. We prefer sum-log (∑(logx)) (why?)

① Maximize $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$

log Σ

EM - Jensen's inequality $f(\mathbb{E}X) \leq \mathbb{E}f(X)$

1. Rewrite $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$ into $\sum_n \log \sum_{z_n} q(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)}$

2. If $q(z_n)$ is a distribution, $\ell(\boldsymbol{\theta}) = \sum_n \log \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)}\right]$

3. Applying Jensen's ineq., $\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} q(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)}$

Σ log

Maximize $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} q(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)}$$

EM – Get tight lower bound

Jensen's inequality tells you $\ell(\boldsymbol{\theta}) \geq \sum_n \underbrace{\sum_{z_n} q(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q(\boldsymbol{z}_n)}}_{\mathbb{E}_{q(z_n)}\left[f\left(\frac{p(x_n, z_n|\theta)}{q(z_n)}\right)\right]}$

The equality holds when $q_t(\boldsymbol{z}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$

- We use the latest $\boldsymbol{\theta}^t$ to replace $\boldsymbol{\theta}$
- $q_t(\boldsymbol{z}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$ implies $q_t(\boldsymbol{z}_n) = p(\boldsymbol{z}_n = k|\boldsymbol{x}_n; \boldsymbol{\theta}^t)$
- What if your q_t is not $\propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$?

EM – Get tight lower bound

Maximize $\ell(\theta) = \sum_n \log \sum_{z_n} p(x_n, z_n|\theta)$

$$\ell(\theta) \geq \sum_n \sum_{z_n} q(z_n) \log \frac{p(x_n, z_n|\theta)}{q(z_n)}$$

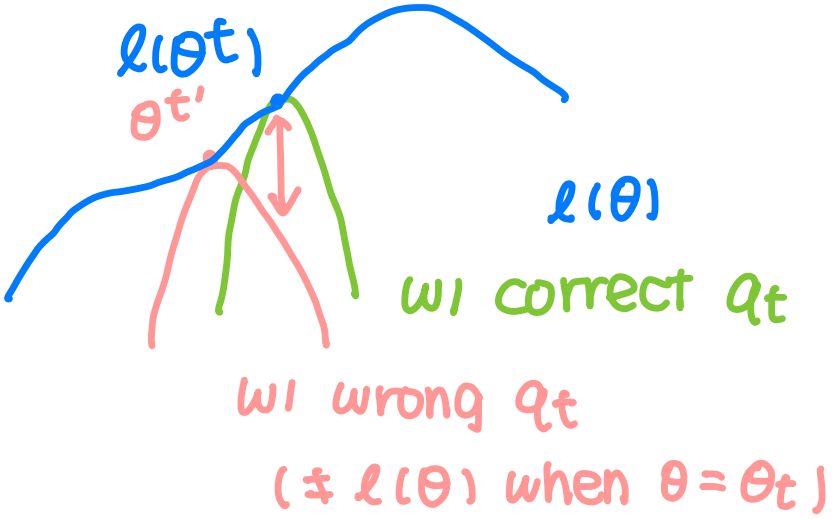
The equality holds when $q_t(z_n) = p(z_n = k|x_n; \theta^t)$

What if your q_t is not $\propto p(x_n, z_n|\theta^t)$?

- Still a lower bound? Yes
- Strictly lower bound?

If $q_t' \propto p(x_n, z_n|\theta^{t'})$

$$\ell(\theta^{t'}) = \sum\sum q_t' \log \frac{p(x, z|\theta^{t'})}{q_t'}$$



Substituting $q_t(z_n) = p(z_n = k|x_n; \theta^t)$ gives us $\ell(\theta) \geq \sum_n \sum_{z_n} p(z_n|x_n; \theta^t) \log \frac{p(x_n, z_n|\theta)}{p(z_n|x_n; \theta^t)}$

EM – E-step

Maximize $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$$

We can decompose $\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$ into

- $g(\boldsymbol{\theta}^t)$: Independent with $\boldsymbol{\theta}$
- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$: To-be-maximized

$Q(\theta|\theta^t)$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \mathbb{E}_{q_t} \sum_n \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$$

- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ is more easily optimized (why?) sum-log is easier than log-sum
- When $\boldsymbol{\theta} = \boldsymbol{\theta}^t$, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^t)$
- Maximize $\mathbb{E}_{q_t} \sum_n \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$ instead of maximizing $\ell(\boldsymbol{\theta})$

E for E-step

EM – M-step

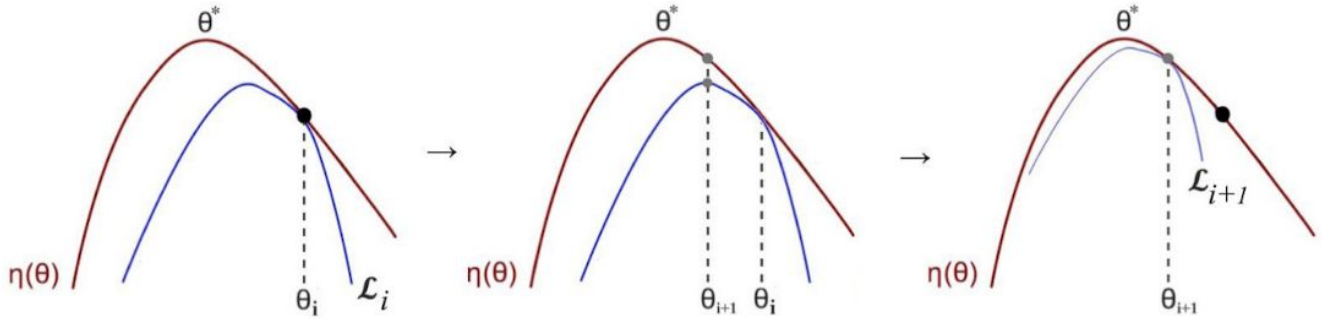
From E-step, $\ell(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) - g(\boldsymbol{\theta}^t)$

- $\boldsymbol{\theta}^{t+1}$ = argmax $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ also maximizes $\ell(\boldsymbol{\theta})$

3 ~~Maximize $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(x_n, z_n|\boldsymbol{\theta})$~~

$$\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(z_n|x_n;\boldsymbol{\theta}^t) \log \frac{p(x_n, z_n|\boldsymbol{\theta})}{p(z_n|x_n;\boldsymbol{\theta}^t)}$$

4 Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$



Adapted from Jonathan Hui's Medium tutorial.

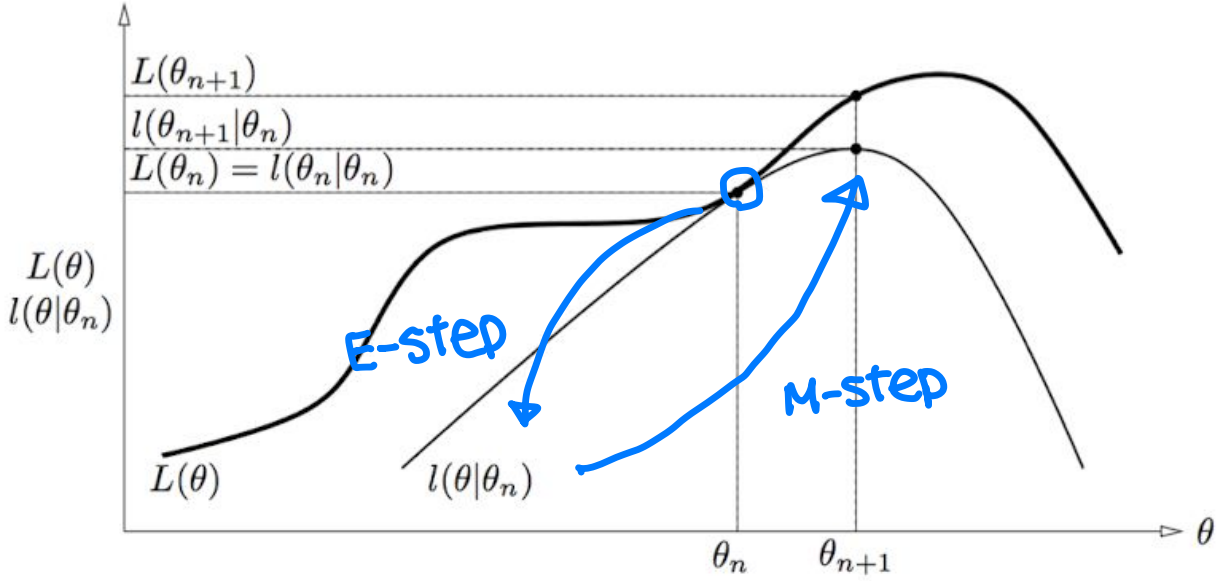
EM – Several exercises

Maximize $\ell(\theta) = \sum_n \log \sum_{z_n} p(x_n, z_n|\theta)$

$$\ell(\theta) \geq \sum_n \sum_{z_n} p(z_n|x_n;\theta^t) \log \frac{p(x_n, z_n|\theta)}{p(z_n|x_n;\theta^t)}$$

Where are E- and M- steps

Maximizing $Q(\theta|\theta^t)$



(Figure from tutorial by Sean Borman)

EM – Several exercises

Maximize $\ell(\theta) = \sum_n \log \sum_{z_n} p(x_n, z_n|\theta)$

$$\ell(\theta) \geq \sum_n \sum_{z_n} p(z_n|x_n;\theta^t) \log \frac{p(x_n, z_n|\theta)}{p(z_n|x_n;\theta^t)}$$

Maximizing $Q(\theta|\theta^t)$

Where are E- and M- steps

$$\ell(\theta^{t+1}) \geq Q(\theta^{t+1}|\theta^t) = \max_\theta Q(\theta|\theta^t) \geq Q(\theta^t|\theta^t) = \ell(\theta^t)$$

E M E

EM – What if z is known?

Maximize $\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n; \boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n; \boldsymbol{\theta}^t)}$$

Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$

$p(z_n)$ and never changes

No hidden variable -> Complete data

Define $\mathcal{D}' = \{\boldsymbol{x}_n, z_n\}_{n=1}^N$, $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$

- $\mathcal{D}'$ is the complete data
- $\mathcal{D}$ the incomplete data

In EM,

- $q_t(\boldsymbol{z}_n) = p(\boldsymbol{z}_n = k|\boldsymbol{x}_n; \boldsymbol{\theta}^t)$ doesn't depend on $\boldsymbol{\theta}^t$ anymore
- $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = Q(\boldsymbol{\theta})$, pure optimization problem!

EM Overview

E-step: Find a tight lower bound $\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n,\boldsymbol{z}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$

Derive $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \underbrace{\log p(\boldsymbol{x}_n,\boldsymbol{z}_n|\boldsymbol{\theta})}_{\text{complete-data likelihood}} = \mathbb{E}_{q_t} \sum_n \log p(\boldsymbol{x}_n,\boldsymbol{z}_n|\boldsymbol{\theta})$

M-step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$, i.e., $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$

GMM setup

Parameters: $\boldsymbol{\theta} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

Hidden variable: r_{nk} (contribution from each cluster)

$$\sum_n \log p(\boldsymbol{x}_n, z_n) = \sum_n \log p(z_n)p(\boldsymbol{x}_n|z_n) = \sum_k \sum_{n:z_n=k} \log p(z_n)p(\boldsymbol{x}_n|z_n)$$

$$= \sum_k \sum_n r_{nk} \log p(z=k)p(\boldsymbol{x}_n|z=k)$$

conditional prob.

$$= \sum_k \sum_n r_{nk} [\log \omega_k + \log N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]$$

hidden variable

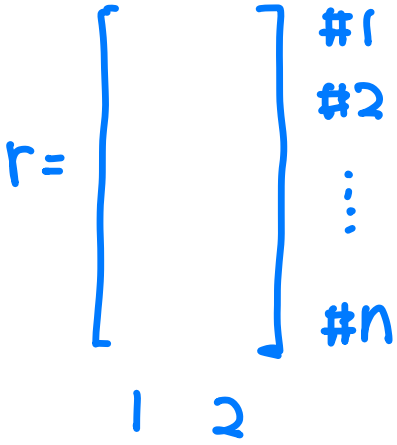
prior knowledge

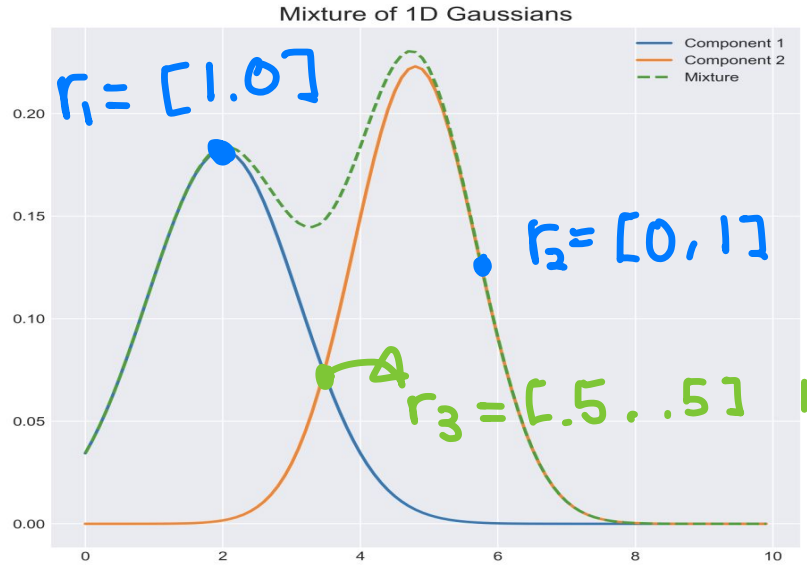
$$\sum_n \log p(\boldsymbol{x}_n, z_n) = \sum_k \sum_n r_{nk} \log p(z = k) p(\boldsymbol{x}_n | z = k)$$

GMM setup

Parameters: $\boldsymbol{\theta} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$

Hidden variable: r_{nk} (contribution from each cluster)





mixed contribution for each cluster

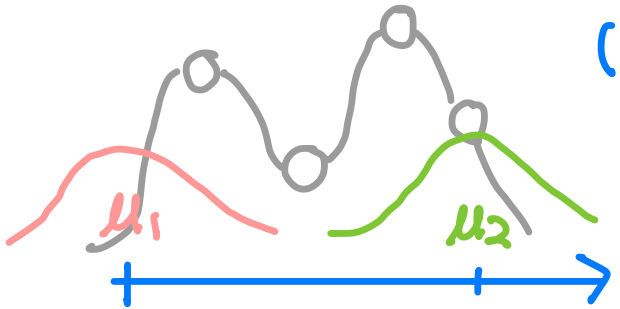
GMM from EM view

Parameters: $\boldsymbol{\theta} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

E-step: Find a tight lower bound $\ell(\boldsymbol{\theta}) \geq \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$

Derive $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \underbrace{\log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}_{} = \mathbb{E}_{q_t} \sum_n \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$

M-step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$, i.e., $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$



(Assume init is perfect)

4 data (N=4)

2 clusters (K=2)

GMM – E-step

Goal: Derive $Q(\theta|\theta^t) = \sum_n \sum_{z_n} p(z_n|x_n;\theta^t) \underbrace{\log p(x_n, z_n|\theta)}_{} = \mathbb{E}_{q_t} \sum_n \log p(x_n, z_n|\theta)$

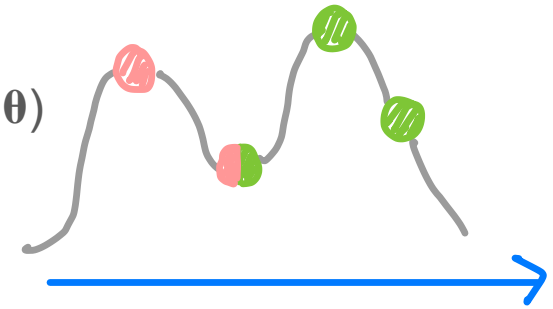
- $q_t(z_n) = p(z_n = k|x_n;\theta^t)$: Posterior probability

$$p(z_n = k|x_n) = \frac{p(x_n|z_n = k)p(z_n = k)}{p(x_n)} = \frac{p(x_n|z_n = k)p(z_n = k)}{\sum_{k'=1}^{K} p(x_n|z_n = k')p(z_n = k')}$$

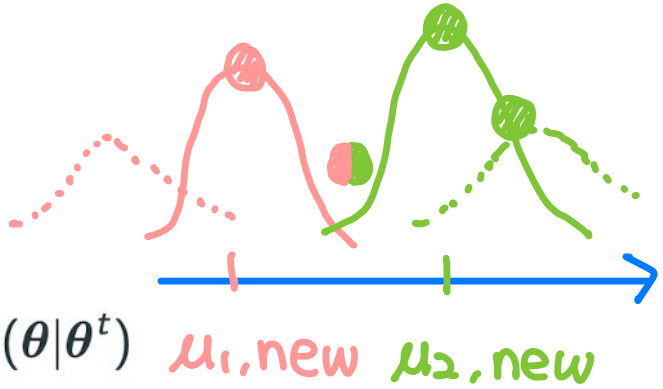
$$= \frac{N(x_n|\mu_k, \Sigma_k) \times \omega_k}{\sum_{k'=1}^{K} N(x_n|\mu_{k'}, \Sigma_{k'}) \times \omega_{k'}}$$

$$r = \begin{bmatrix} 1 & 0 \\ .5 & .5 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

- $\log p(x_n, z_n|\theta)$: pdf from Gaussian model (with parameter θ)



GMM – M-step



Parameters: $\boldsymbol{\theta} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

Goal: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$, i.e., $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$

$$\omega_k = \frac{\sum_n r_{nk}}{\sum_k \sum_n r_{nk}}, \quad \boldsymbol{\mu}_k = \frac{1}{\sum_n r_{nk}} \sum_n r_{nk} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_n r_{nk}} \sum_n r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$$

GMM – Algorithm

Alternate between estimating r_{nk} and computing parameters

- Step 0: initialize $\boldsymbol{\theta}$ with some values (random or otherwise)
- Step 1 (E-Step): set $r_{nk} = p(z_n = k|\boldsymbol{x}_n)$ for current $\boldsymbol{\theta}$ using Bayes Rule
- Step 2 (M-Step): update $\boldsymbol{\theta}$ using these r_{nk}s using MLE
- Step 3: go back to Step 1

Before PCA – Basis

We have several choices of basis

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \sqrt{2}\begin{bmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{bmatrix} + 1\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 1\begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

basis #1 #2 basis #3

They are some special basis

- Orthogonal: Perpendicular to each other #1, #3
- Normal: Each has unit norm #1, #2

Before PCA – Eigen-decomposition

Form: $C_X = Q^T \Lambda Q$

- Λ is a diagonal matrix
- Each component in Λ is in order and called eigenvalue
- Each basis is called eigenvector
- Orthonormal basis

eigenvalue #1

$$C_X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

eigenvector #1

PCA – With 2-d

Eigen-decomposition on covariance matrix $C_X = \frac{1}{n} \begin{bmatrix} -- & \mathbf{x}_1^T & -- \\ -- & \mathbf{x}_2^T & -- \\ & \vdots & \\ -- & \mathbf{x}_d^T & -- \end{bmatrix} \begin{bmatrix} | & | & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_d \\ | & | & | \end{bmatrix}$

Choose P as the first k columns of Q

$$\mathbf{\Lambda} = \mathbf{Q}^T \mathbf{C_X} \mathbf{Q}$$

$$\mathbf{C_Z} = \mathbf{P}^T \mathbf{C_X} \mathbf{P}$$

$$\begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & \dots & \\ & & \lambda_k \end{bmatrix} = \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ & \vdots & \\ - & v_k & - \end{bmatrix} C_X \begin{bmatrix} | & | & | \\ v_1 & v_2 & v_k \\ | & | & | \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix} = \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \end{bmatrix} C_X \begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix}$$

PCA – Example

If $X = \begin{bmatrix} 3 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 3 & -1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}^T$

- $\sigma_1^2 = \frac{1}{8} (3^2 + 1^2 + 1^2 + \ldots + (-1)^2) = 2$
- $\sigma_2^2 = \frac{1}{8} (1^2 + 3^2 + (-1)^2 + \ldots + (-1)^2) = 2$
- $\sigma_{12} = \frac{1}{8} ((3x1) + (1x3) + (1x-1) + \ldots + ((-1)x(-1)) = 1$

$$C_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

PCA – Example with K=2

If $X = \begin{bmatrix} 3 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 3 & -1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}^T$

$$C_X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} - & p_1 & - \\ - & p_2 & - \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} | & | \\ p_1 & p_2 \\ | & | \end{bmatrix} \Rightarrow \begin{cases} p_1 = \begin{bmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{bmatrix} \\ p_2 = \begin{bmatrix} \frac{1}{2}\sqrt{2} \\ \frac{-1}{2}\sqrt{2} \end{bmatrix} \end{cases}$$

$$x_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \left(\begin{bmatrix} 3 \\ 1 \end{bmatrix}^T p_1 \right) p_1 + \left(\begin{bmatrix} 3 \\ 1 \end{bmatrix}^T p_2 \right) p_2 = 2\sqrt{2}\, p_1 + \sqrt{2}\, p_2$$

When $k = \dim(x)$, PCA fully recovers

PCA – Example with K=1

If $X = \begin{bmatrix} 3 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 3 & -1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}^T$

Only $P_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ (Remove $P_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$)

$x_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ is represented as $\left(\begin{bmatrix} 3 \\ 1 \end{bmatrix}^T P_1\right) P_1 = 2\sqrt{2} P_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$