GAUSSIAN MIXTURE MODELS AND EXPECTATION MAXIMIZATION

Pei-Yuan Wu

National Taiwan University

OUTLINE

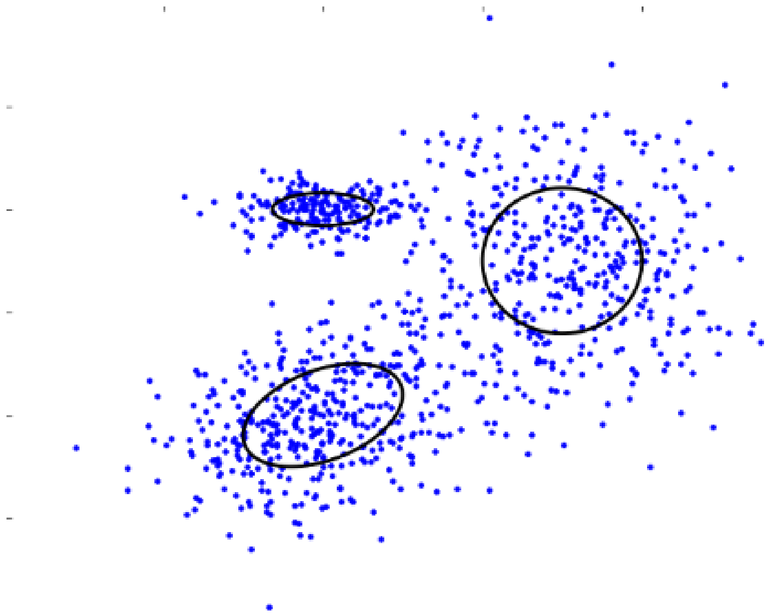
- Maximum-likelihood parameter estimation
- Gaussian Mixture Model (GMM)
- Expectation Maximization (EM) Algorithm

THE PROBLEM

- You have data that you believe is drawn from K populations
- You want to identify parameters for each population
- You don't know anything about the populations *a priori*
  - Except you believe that they're Gaussian…

GAUSSIAN MIXTURE MODELS (GMM)

- Fit a Set of K Gaussians that generates the data
- Maximum Likelihood over a mixture model



MAXIMUM-LIKELIHOOD ESTIMATION (MLE)

- We have data set $\mathcal{X} = \{x_1, \ldots, x_N\}$. We have a probability density function $p(x;\theta)$ that is governed by the parameters $\theta \in \Theta$.
- If we assume $x_1, \ldots, x_N$ are drawn independently from $p(x;\theta)$, then the joint probability density function for $\mathcal{X}$ is

$$p(\mathcal{X};\theta) = \prod_{i=1}^{N} p(x_i;\theta)$$

  - We call $p(\mathcal{X};\theta)$ the ***likelihood*** function (of θ given $\mathcal{X}$).
- Maximum likelihood parameter estimation:

$$\theta^* = \max_{\theta \in \Theta} p(\mathcal{X};\theta)$$

  - It is usually analytically easier to maximize the log-likelihood $\log p(\mathcal{X};\theta)$.
  - For some problems, θ^* can be analytically solved by setting the derivative of the log-likelihood to be zero.
  - For many problems, it is NOT possible to solve θ^* analytically.

MLE FOR SINGLE GAUSSIAN MODEL

- Suppose the data are drawn from a Gaussian distribution specified by $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^m$ is the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ is the (non-singular) covariance matrix.
  - Probability density function:

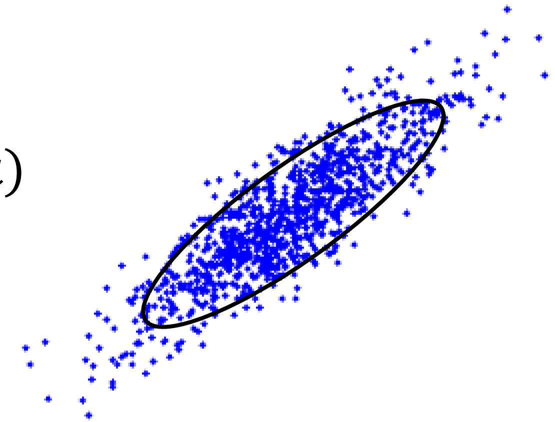
$$p(\boldsymbol{x};\theta) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Given data set $\mathcal{X} = \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}$, where $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N \in \mathbb{R}^m$. The likelihood function is

$$p(\mathcal{X};\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i;\theta) = \prod_{i=1}^{N}\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)$$

  Hence the log-likelihood is given by

$$\log p(\mathcal{X};\theta) = N\log\frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$



MLE FOR SINGLE GAUSSIAN MODEL (CONT'D)

For single Gaussian distribution, the optimal θ^* can be analytically solved by setting partial derivatives equal to zero.

- Partial derivative over $\boldsymbol{\mu}$

$$\nabla_{\boldsymbol{\mu}} \log p(\mathcal{X};\theta) = N\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$$

where $\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$ is the sample mean.

➢ Setting derivate equals zero $\Longrightarrow$ Take $\boldsymbol{\mu}^* = \overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$

- Partial derivative over $\boldsymbol{\Sigma}$: First rewrite

$$\log p(\mathcal{X};\theta) = \frac{N}{2}\left(-\log(2\pi)^m + \log|\boldsymbol{\Sigma}^{-1}| - \mathrm{Trace}(\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{x}})\right)$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$. Let $\boldsymbol{\Sigma}^{-1} = [a_{ij}]$, then

$$\frac{\partial}{\partial a_{ij}} \log p(\mathcal{X};\theta) = \frac{N}{2}\left(\frac{\partial \log|\boldsymbol{\Sigma}^{-1}|}{\partial a_{ij}} - \boldsymbol{e}_j^T \overline{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \boldsymbol{e}_i\right) = \frac{N}{2}\left(\boldsymbol{e}_j^T(\boldsymbol{\Sigma} - \overline{\boldsymbol{\Sigma}}_{\boldsymbol{x}})\boldsymbol{e}_i\right)$$

➢ Setting derivate equals zero $\Longrightarrow$ Take $\boldsymbol{\Sigma}^* = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$

$\boldsymbol{A} \in \mathbb{R}^{m\times m}$

$\boldsymbol{A}_{ij} \in \mathbb{R}^{(m-1)\times(m-1)}$

$$\begin{vmatrix} 1 & 4 & -5 \\ 6 & 9 & 2 \\ 2 & 3 & -6 \end{vmatrix} = 1 \times \begin{vmatrix} 9 & 2 \\ 3 & -6 \end{vmatrix} - 4 \times \begin{vmatrix} 6 & 2 \\ 2 & -6 \end{vmatrix} + (-5) \times \begin{vmatrix} 6 & 9 \\ 2 & 3 \end{vmatrix}$$

$$= 1 \times \begin{vmatrix} 9 & 2 \\ 3 & -6 \end{vmatrix} - 6 \times \begin{vmatrix} 4 & -5 \\ 3 & -6 \end{vmatrix} + 2 \times \begin{vmatrix} 4 & -5 \\ 9 & 2 \end{vmatrix}$$

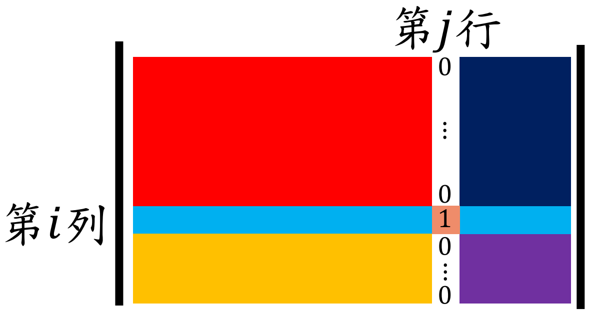
$$|\boldsymbol{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{ij}|$$

$$|\boldsymbol{A}| = \sum_{i=1}^{m} (-1)^{i+j} a_{ij} |A_{ij}|$$

$$\frac{\partial}{\partial a_{ij}} |\boldsymbol{A}| = (-1)^{i+j} |\boldsymbol{A}_{ij}|$$

Cramer rule

$\boldsymbol{Ax} = \boldsymbol{e}_i$ → $x^{(j)} =$



$|\boldsymbol{A}|$

→ $$\boldsymbol{e}_j^T A^{-1} \boldsymbol{e}_i = \frac{(-1)^{i+j} |A_{ij}|}{|\boldsymbol{A}|} = \frac{\partial log|\boldsymbol{A}|}{\partial a_{ij}}$$

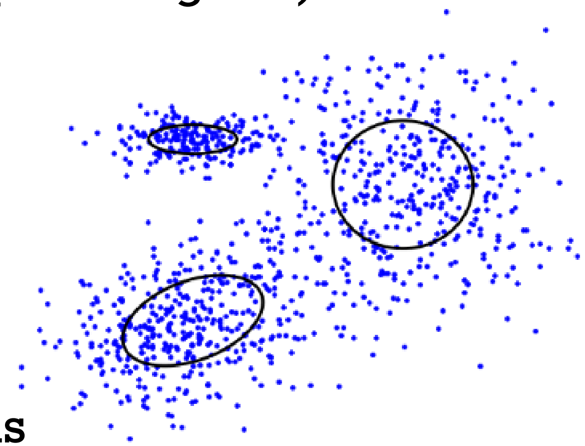
Let $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$, then

$$\boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_i = \frac{\partial log|\boldsymbol{\Sigma}^{-1}|}{\partial a_{ij}}$$

MLE FOR GMM

- Suppose the data are drawn from K Gaussian distributions specified by $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$, where $\pi_k \in \mathbb{R}$, $\boldsymbol{\mu}_k \in \mathbb{R}^m$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{m\times m}$ denotes the prior probability, mean, and (non-singular) covariance matrix of the k'th Gaussian distribution
  - Probability density function

$$p(\boldsymbol{x};\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



where $\sum_{k=1}^{K} \pi_k = 1$

- Given data set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^m$. The likelihood function is

$$p(\mathcal{X};\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i;\theta) = \prod_{i=1}^{N}\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)$$

Hence the log-likelihood is given by

$$\log p(\mathcal{X};\theta) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

For GMM, it is intractable to find optimal θ^* analytically.

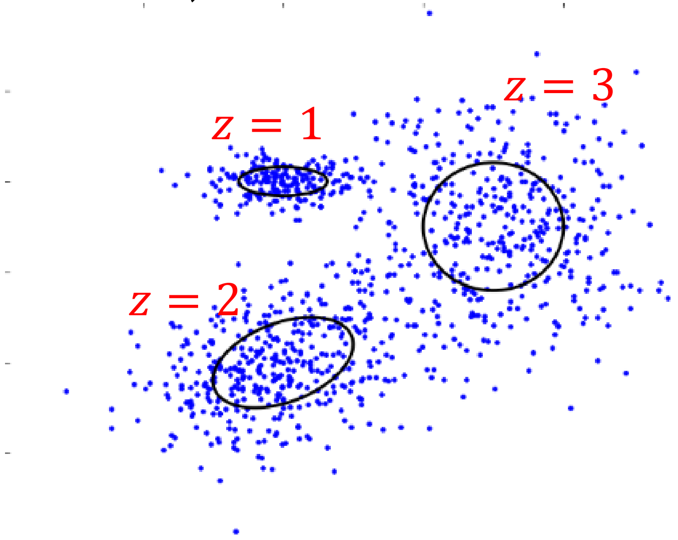
LATENT VARIABLES IN GMM

- Denote latent variable $z_i \in \{1, \dots, K\}$ indicating which Gaussian distribution $\boldsymbol{x}_i$ is drawn from.

$$p(\boldsymbol{x}, z = k; \theta) = \pi_k \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k{}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right)$$

- Denote $\mathcal{Z} = \{z_1, \dots, z_N\}$ as the collection of all latent variables.
- GMM can be applied to clustering problems.
  - Assign $\boldsymbol{x}$ to the k'th cluster if

$$k = \underset{k}{\operatorname{argmax}}\, p(z = k | \boldsymbol{x}; \theta)$$



EXPECTATION MAXIMIZATION THEORY

- Suppose we have parameters $\theta^{(t)}$. We want to find better parameters $\theta^{(t+1)}$ with larger log-likelihood

$$p\left(\mathcal{X};\theta^{(t+1)}\right) \geq p\left(\mathcal{X};\theta^{(t)}\right)$$

- Note that

$$\log p(\mathcal{X};\theta) = \sum_{\mathcal{Z}} p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right) \log p(\mathcal{X};\theta)$$

$$= \sum_{\mathcal{Z}} p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right)\left(\log p(\mathcal{X},\mathcal{Z};\theta) - \log p(\mathcal{Z}|\mathcal{X};\theta)\right)$$

$$= \sum_{\mathcal{Z}} p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right) \log p(\mathcal{X},\mathcal{Z};\theta) - \sum_{\mathcal{Z}} p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right) \log p(\mathcal{Z}|\mathcal{X};\theta)$$

$$= Q\left(\theta\middle|\theta^{(t)}\right) + H\left(\theta\middle|\theta^{(t)}\right)$$

- ✓ $p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right)$: Posterior prob. dist. of latent variables based on current parameters $\theta^{(t)}$
- ✓ $p(\mathcal{Z}|\mathcal{X};\theta)$: Posterior prob. dist. of latent variables based on parameters θ
- ✓ $\log p(\mathcal{X},\mathcal{Z};\theta)$: Log-likelihood of parameter θ, given data $\mathcal{X}$ and latent variable $\mathcal{Z}$
- ✓ $Q\left(\theta\middle|\theta^{(t)}\right)$: Expectation of log-likelihood function (of θ), assuming latent variables follows posterior prob. dist. based on current parameters $\theta^{(t)}$
- ✓ $H\left(\theta\middle|\theta^{(t)}\right)$: cross entropy between $p\left(\mathcal{Z}\middle|\mathcal{X};\theta^{(t)}\right)$ and $p(\mathcal{Z}|\mathcal{X};\theta)$

EXPECTATION MAXIMIZATION THEORY

- Let's look into the difference of log-likelihood functions between θ and θ_t

$$\log p(\mathcal{X};\theta) - \log p\left(\mathcal{X};\theta^{(t)}\right) =$$
$$= \left(Q\left(\theta|\theta^{(t)}\right) + H\left(\theta|\theta^{(t)}\right)\right) - \left(Q\left(\theta^{(t)}|\theta^{(t)}\right) + H\left(\theta^{(t)}|\theta^{(t)}\right)\right)$$

(Explanation in next slide)

Since $H\left(\theta|\theta^{(t)}\right) \geq H\left(\theta^{(t)}|\theta^{(t)}\right)$, we have

Lower bound

$$\log p(\mathcal{X};\theta) - \log p\left(\mathcal{X};\theta^{(t)}\right) \geq Q\left(\theta|\theta^{(t)}\right) - Q\left(\theta^{(t)}|\theta^{(t)}\right)$$

Idea: Find $\theta^{(t+1)}$ that maximizes the lower bound!

- Expectation Step (E-step): Compute

$$Q\left(\theta|\theta^{(t)}\right) = \sum_{\mathcal{Z}} p\left(\mathcal{Z}|\mathcal{X};\theta^{(t)}\right) \log p(\mathcal{X},\mathcal{Z};\theta)$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = arg\max_{\theta\in\Theta} Q\left(\theta|\theta^{(t)}\right)$$

WHY $H(\theta|\theta^{(t)}) \geq H(\theta^{(t)}|\theta^{(t)})$?

$$H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) = \sum_{\mathcal{Z}} p(\mathcal{Z}|\mathcal{X};\theta^{(t)}) \log \frac{p(\mathcal{Z}|\mathcal{X};\theta^{(t)})}{p(\mathcal{Z}|\mathcal{X};\theta)} \geq 0 \text{ ???}$$

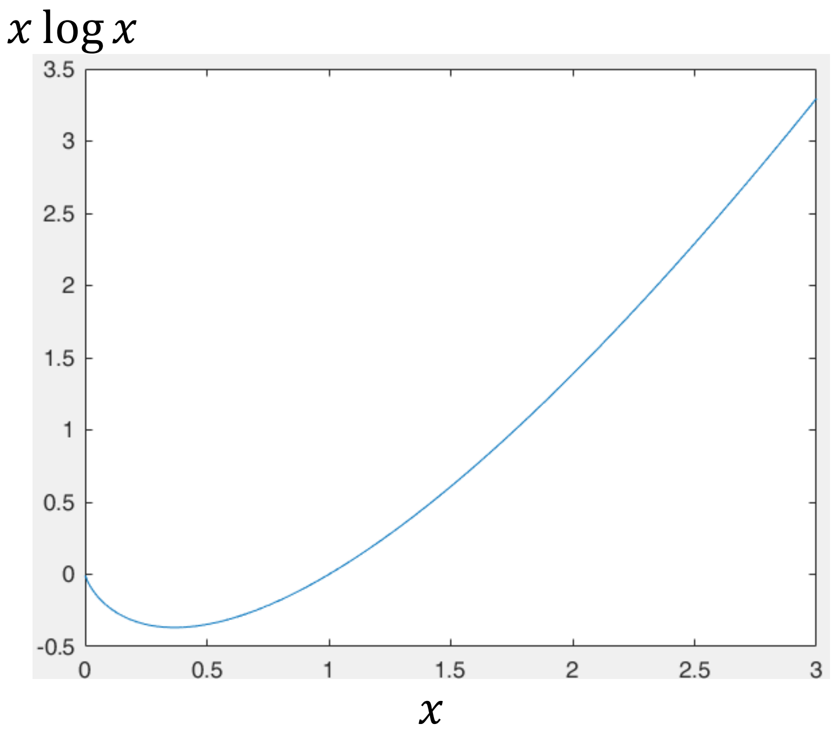
Theorem: Let p, q be two probability density functions on $\mathbb{R}^m$. If ~~$p(z) = 0$ whenever $q(z) = 0$~~, then

$$\int_{\mathbb{R}^m} p(z) \log \frac{p(z)}{q(z)} dz \geq 0$$

Proof: Let $f(x) = x \log x$, then f is convex and $f(1) = 0$, hence

$$\int_{\mathbb{R}^m} p(z) \log \frac{p(z)}{q(z)} dz = \int_{\mathbb{R}^m} q(z) f\left(\frac{p(z)}{q(z)}\right) dz = \mathbb{E}_{Z\sim q}\left[f\left(\frac{p(Z)}{q(Z)}\right)\right]$$

$$\geq f\left(\mathbb{E}_{Z\sim q}\left[\frac{p(Z)}{q(Z)}\right]\right) = f\left(\int_{\mathbb{R}^m} q(z) \frac{p(z)}{q(z)} dz\right) = f(1) = 0$$



Let $\Omega = \{z \in \mathbb{R}^m : p(z) > 0, q(z) = 0\}$. Let $r = \int_{\mathbb{R}^m \setminus \Omega} p(z) dz$ and take $\tilde{p}$ so that $p = r\tilde{p}$. Then

$$\int_{\mathbb{R}^m} p(z) \log \frac{p(z)}{q(z)} dz = \int_{\Omega} p(z) \log \frac{p(z)}{q(z)} dz + \int_{\mathbb{R}^m \setminus \Omega} p(z) \log \frac{p(z)}{q(z)} dz$$

$$= (1-r)\cdot\infty + \int_{\mathbb{R}^m \setminus \Omega} r\tilde{p}(z) \log \frac{r\tilde{p}(z)}{q(z)} dz \geq 0$$

EXPECTATION MAXIMIZATION ALGORITHM

- Randomly initialize parameters $\theta^{(1)}$.
- Iterate through step t=1,2,…
  - Expectation Step (E-step): Compute

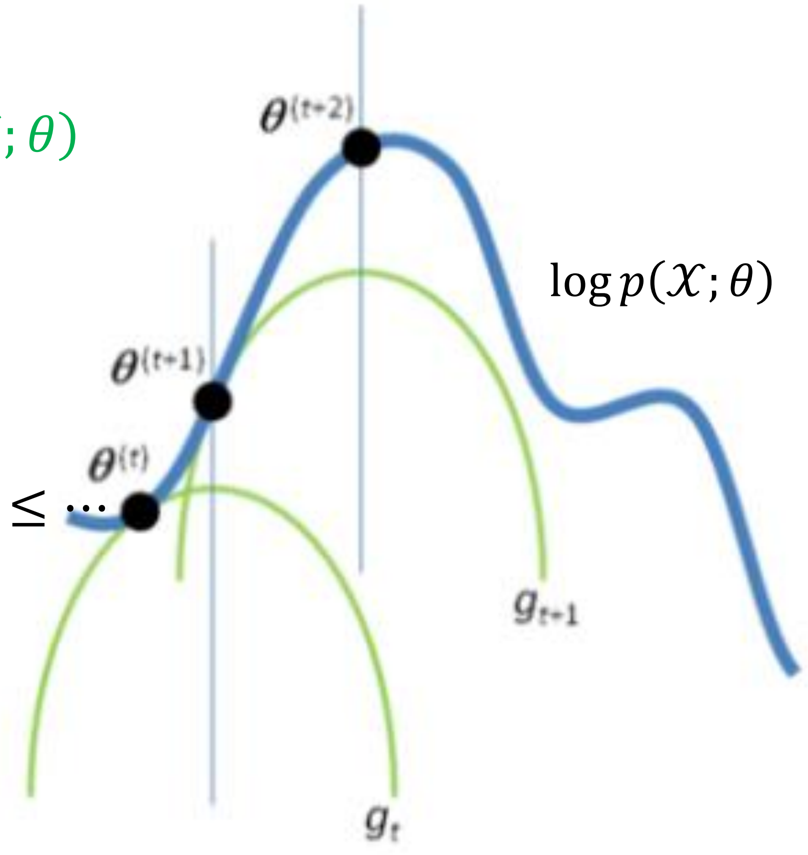
$$Q(\theta|\theta^{(t)}) = \sum_{\mathcal{Z}} p(\mathcal{Z}|\mathcal{X};\theta^{(t)}) \log p(\mathcal{X},\mathcal{Z};\theta)$$

  - Maximization Step (M-step): Choose

$$\theta^{(t+1)} = arg\max_{\theta\in\Theta} Q(\theta|\theta^{(t)})$$

- Log likelihood always non-decreasing

$$p(\mathcal{X};\theta^{(1)}) \le p(\mathcal{X};\theta^{(2)}) \le p(\mathcal{X};\theta^{(3)}) \le \cdots$$



$$g_t(\theta) = \log p(\mathcal{X};\theta^{(t)}) + Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$$

EM ALGORITHM FOR GMM – E STEP

- Current parameter estimates $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$
- Expectation Step (E-step): Compute

$$Q(\theta|\theta^{(t)}) = \sum_{\mathcal{Z}} p(\mathcal{Z}|\mathcal{X};\theta^{(t)}) \log p(\mathcal{X},\mathcal{Z};\theta) = \mathbb{E}_{\mathcal{Z}|\mathcal{X};\theta^{(t)}}[\log p(\mathcal{X},\mathcal{Z};\theta)]$$

$$= \sum_{i=1}^{N} \mathbb{E}_{\mathcal{Z}|\mathcal{X};\theta^{(t)}}[\log p(\boldsymbol{x}_i, z_i;\theta)] = \sum_{i=1}^{N} \mathbb{E}_{z_i|\boldsymbol{x}_i;\theta^{(t)}}[\log p(\boldsymbol{x}_i, z_i;\theta)]$$

➢ Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$

根據現有模型$\theta^{(t)}$,資料點$\boldsymbol{x}_i$有多少比例隸屬第k群

$$\mathbb{P}[z_i = k|\boldsymbol{x}_i;\theta^{(t)}] = \frac{p(\boldsymbol{x}_i, z_i = k;\theta^{(t)})}{\sum_{j=1}^{K} p(\boldsymbol{x}_i, z_i = j;\theta^{(t)})} = \frac{\pi_k^{(t)}\mathcal{N}\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K}\pi_j^{(t)}\mathcal{N}\left(\boldsymbol{x}_i;\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)}\right)} = \delta_{ik}^{(t)}$$

➢ Log-likelihood of parameter θ, given data $\boldsymbol{x}_i$ and latent variable z_i

$$\log p(\boldsymbol{x}_i, z_i = k;\theta) = \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$$

Hence

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}^{(t)}\left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$

EM ALGORITHM FOR GMM – M STEP

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{ \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\}$$

- Partial derivative over $\boldsymbol{\mu}_k$

$$\nabla_{\boldsymbol{\mu}_k} \log Q(\theta|\theta^{(t)}) = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{N} \delta_{ik}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$$

  - Setting derivate equals zero ⟹ Take

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

EM ALGORITHM FOR GMM – M STEP (CONT'D)

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \underset{\theta\in\Theta}{argmax}\, Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}$$

➢ Partial derivative over $\boldsymbol{\Sigma}_k$: First rewrite

$$Q(\theta|\theta^{(t)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \delta_{nk}^{(t)} \left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m}}\right) + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_k^{-1}| - \mathrm{Trace}\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\right)\right)\right\}$$

Let $\boldsymbol{\Sigma}_k^{-1} = [a_{ij}^k]$, then

$$\frac{\partial}{\partial a_{ij}^k}\log Q(\theta|\theta^{(t)}) = \frac{1}{2}\sum_{n=1}^{N}\delta_{nk}^{(t)}\left\{\frac{\partial \log|\boldsymbol{\Sigma}_k^{-1}|}{\partial a_{ij}^k} - \boldsymbol{e}_j^T(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{e}_i\right\} = \frac{1}{2}\sum_{n=1}^{N}\delta_{nk}^{(t)}\{\boldsymbol{e}_j^T(\boldsymbol{\Sigma}_k - (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T)\boldsymbol{e}_i\}$$

Setting derivate equals zero $\Longrightarrow$ Take

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N}\delta_{ik}^{(t)}(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^{N}\delta_{ik}^{(t)}}$$

EM ALGORITHM FOR GMM – M STEP (CONT'D)

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{ \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\}$$

  - Partial derivative over π_k :

    Note that we have constraint $\sum_{k=1}^{K} \pi_k = 1$. Hence we invoke Lagrange multiplier

$$\nabla_{\pi_k}\left(\log Q(\theta|\theta^{(t)}) - \lambda \sum_{k=1}^{K} \pi_k\right) = \sum_{i=1}^{N} \frac{\delta_{ik}^{(t)}}{\pi_k} - \lambda$$

    Setting derivate equals zero $\Longrightarrow$ Take $\pi_k^{(t+1)} = \lambda^{-1} \sum_{i=1}^{N} \delta_{ik}^{(t)}$

    The constraint $\sum_{k=1}^{K} \pi_k = 1$ implies $\lambda = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} = N$. Hence

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{ik}^{(t)}$$

EM ALGORITHM FOR GMM – SUMMARY

- Randomly initialize parameters $\theta^{(1)}$.
- Iterate through step t=1,2,…
  - Expectation Step (E-step): Compute

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{ \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}_k|}}\right) - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\}$$

$$\delta_{ik}^{(t)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}$$

Evaluate the "responsibilities" of each cluster with the current parameters

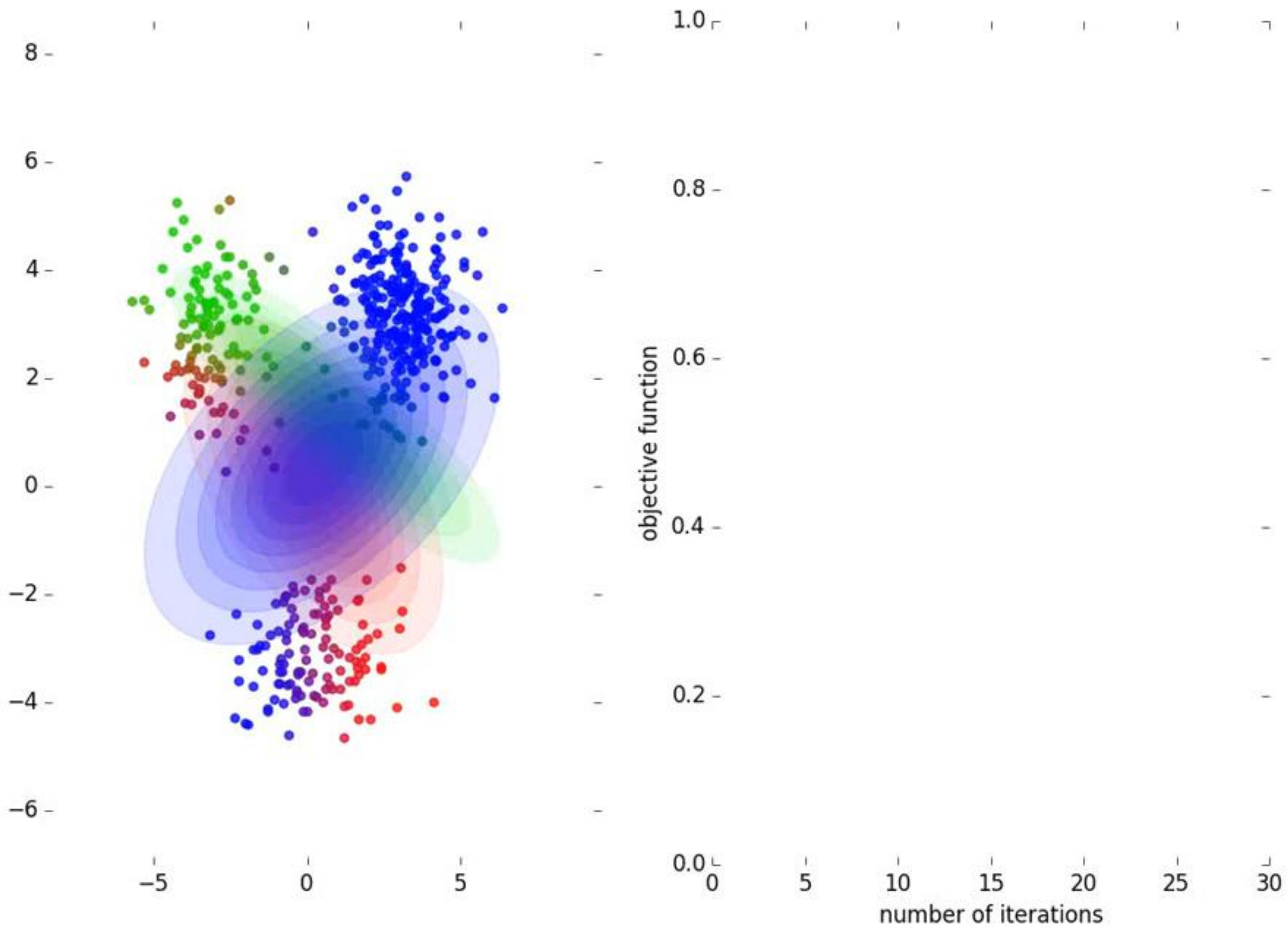
  - Maximization Step (M-step): Choose $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$, where

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{ik}^{(t)}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

Re-estimate parameters using the existing "responsibilities"



https://www.youtube.com/watch?v=uUtpiK5NEAM