Cluster Analysis

- K-means
- Agglomerative clustering
- **EM Algorithm and Gaussian Mixture Model Clustering**

MLE

(likelihood function) Let $X_1, X_2, ..., X_n$ be a random sample from a distribution characterized by $p(x|\theta)$,[4] where θ is an unknown parameter[5]. Let $x_1, \ldots, x_n$ be the observed values. The likelihood function is defined by

$$L(\theta) = L(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i|\theta),$$

and the log-likelihood function is given by

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta).$$

(maximum likelihood estimator, **MLE**) The MLE $\hat{\theta}_{\text{mle}} = \hat{\theta}_{\text{mle}}(X_1, \ldots, X_n)$ of θ is

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} l(\theta).$$

[4] p is a density for a continuous r.v. and probability for a discrete r.v..
[5] θ can be a vector.

How to find MLE (1/2)

Analytical approach.

Verify $\frac{d^2}{d\theta^2} l(\theta) < 0$ and find the unique θ such that $\frac{d}{d\theta} l(\theta) = 0$.

Let $X_1, \ldots, X_n$ be independently sampled from $\mathcal{N}(\theta, 1)$. Find the MLE of θ.

Solution. The log-likelihood function is

$$l(\theta) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \theta)^2}{2}\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\theta - x_i)^2.$$

Then $\frac{dl(\theta)}{d\theta} = -\sum_{i=1}^{n}(\theta - x_i)$. Let $\frac{dl(\theta)}{d\theta} = 0$ to give $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$. We verify that this is really the maximizer by looking at the second order derivative $\frac{d^2 l(\theta)}{d\theta^2} = -n < 0$. We replace the observation by the random sample to give the maximum likelihood estimator (MLE),

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

□

How to find MLE (2/2)

Numerical approach for $\hat{\theta} = \arg\max_\theta l(\theta)$.

- Coordinate ascent. For example, if $\theta = (\theta_1, \theta_2)'$,

$$\theta_1^{(k+1)} = \arg\max_{\theta_1} l(\theta_1, \theta_2^{(k)} | \boldsymbol{x}),$$
$$\theta_2^{(k+1)} = \arg\max_{\theta_2} l(\theta_1^{(k+1)}, \theta_2 | \boldsymbol{x}).$$

- Newton-Raphson. When θ is one-dimensional,

$$\theta^{(k+1)} = \theta^{(k)} - \frac{l'}{l''}.$$

  Generally,

$$\theta^{(k+1)} = \theta^{(k)} - J_F(\theta^{(k)})^{-1} F(\theta^{(k)}),$$

  where F is the derivative of f, and J_F is the Jacobi matrix of F.

- **Expectation-maximization, the EM algorithm**

EM algorithm (1/3)

• EM algorithm is an iterative method to find MLE, when there is latent variable (hidden variable).

• Let $\boldsymbol{x} = \{x_i\}_{i=1}^m$ be the observed data, and $\boldsymbol{z} = \{z_i\}_{i=1}^n$ be the unobserved data.[6] One has (by Bayes rule)

$$p(\boldsymbol{x}, \boldsymbol{z}|\theta) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)p(\boldsymbol{x}|\theta) \quad \Longleftrightarrow \quad p(\boldsymbol{x}|\theta) = \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{p(\boldsymbol{z}|\boldsymbol{x}, \theta)}.$$

One takes logarithms to get[7]

$$\log p(\boldsymbol{x}|\theta) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) - \log p(\boldsymbol{z}|\boldsymbol{x}, \theta).$$

Let $l(\theta|\boldsymbol{x}) := \log p(\boldsymbol{x}|\theta)$.

Goal: find some iteration, such that for any given θ^{old}, an updated θ^{new} in each round of iteration, so that

$$l(\theta^{\text{new}}|\boldsymbol{x}) \geq l(\theta^{\text{old}}|\boldsymbol{x}).$$

Equivalently, $\theta^{(0)} \quad \to \quad \theta^{(1)} \quad \to \quad \theta^{(2)} \quad \to \cdots$

$$l(\theta^{(0)}|\boldsymbol{x}) \leq l(\theta^{(1)}|\boldsymbol{x}) \leq l(\theta^{(2)}|\boldsymbol{x}) \leq \cdots$$

[6] $(\boldsymbol{x}, \boldsymbol{z})$ together is called the complete data.

[7] If the marginal density of $p(x|\theta)$ can be obtained, we can use MLE directly.

EM Algorithm (2/3)

Recall that for any θ,

$$l(\theta|\boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta) - \log p(\boldsymbol{z}|\boldsymbol{x}, \theta).$$

Given $\boldsymbol{x}$ and θ^{old}, take conditional expectations $\int \cdots p(\boldsymbol{z}|\boldsymbol{x}, \theta^{\text{old}})\, d\boldsymbol{z}$ to get[8]

$$\begin{aligned} l(\theta|\boldsymbol{x}) &= \int [\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]\, p(\boldsymbol{z}|\boldsymbol{x}, \theta^{\text{old}})\, d\boldsymbol{z} - \int [\log p(\boldsymbol{z}|\boldsymbol{x}, \theta)]\, p(\boldsymbol{z}|\boldsymbol{x}, \theta^{\text{old}})\, d\boldsymbol{z} \\ &=: Q(\theta, \theta^{\text{old}}) - H(\theta, \theta^{\text{old}}). \end{aligned}$$

EM algorithm.

Step 1. Initialize unknown parameter $\theta^{(0)}$.

Step 2. (**E**xpectation.) Evaluate $Q(\theta, \theta^{(0)})$.

Step 3. (**M**aximization.) $\theta^{(1)} = \arg\max_\theta Q(\theta, \theta^{(0)})$.

Step 4. Repeat **E** step and **M** step until convergence.

EM algorithm maximizes log-likelihood in the sense that $l(\theta^{\text{new}}|\boldsymbol{x}) \geq l(\theta^{\text{old}}|\boldsymbol{x})$.

[8] For discrete r.v.s., change densities to probabilities and integral to sum.

EM Algorithm (3/3)

$$l(\theta^{\text{new}}|\boldsymbol{x}) \geq l(\theta^{\text{old}}|\boldsymbol{x}).$$

Multinomial distribution

- Binomial distribution.
  - Do a test n times independently. There are only two outcomes of each test: success and fail (e.g. flip a coin).
  - The probability of success is p.
  - Let B be the r.v. denoting the number of success. Then $\mathbb{P}(B=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^k$.
  - $\mathbb{E}[B]=np$ and $\mathrm{Var}(B)=np(1-p)$.
- Multi-nomial distribution.
  - Do a test n times independently.
  - Each trial has $m\geq 2$ outcomes (e.g. roll a dice).
  - The probability of the occurrence of outcome i is p_i: $\sum_{i=1}^m p_i=1$.
  - Let B_i be the r.v. for the number of occurrence of outcome i.
  - $\mathbb{P}(B_1=k_1,\cdots,B_m=k_m)=\frac{(k_1+\cdots+k_m)!}{k_1!\cdots k_m!}p_1^{k_1}\cdots p_k^{k_m}$.
  - $\mathbb{E}[B_i]=np_i$ and $\mathrm{Var}(B_i)=np_i(1-p_i)$.

Example *without* hidden variables (1/2)

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$ follow a multinomial distribution with probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right), \quad \theta \in (0,1) \text{ unknown.}$$

Use 197 samples of $\boldsymbol{Y}$ with $\boldsymbol{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ and **MLE** to estimate θ.

Solution. Given θ, the vector $\boldsymbol{Y}$ has a multinomial distribution with

$$\begin{aligned} p(y_1, y_2, y_3, y_4|\theta) &:= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4|\theta) \\ &= \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}(1-\theta)\right)^{y_2+y_3}\left(\frac{\theta}{4}\right)^{y_4}. \end{aligned}$$

The log-likelihood function is (where C is a constant independent of θ)

$$\begin{aligned} l(\theta|\boldsymbol{y}) &= \log p(y_1, y_2, y_3, y_4|\theta) \\ &= C + y_1 \log(2+\theta) + (y_2+y_3)\log(1-\theta) + y_4 \log\theta. \end{aligned}$$

Example *without* hidden variables (2/2)

Therefore

$$\frac{\partial l(\theta|\boldsymbol{y})}{\partial \theta} = \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta} = 0,$$

which implies

$$(y_1+y_2+y_3+y_4)\theta^2 + (-y_1+2y_2+2y_3+y_4)\theta - 2y_4 = 0.$$

Since $\theta \in (0,1)$, we drop the negative root and obtain the MLE:

$$\hat{\theta}_{\text{mle}} = \frac{1}{2}r + \sqrt{\frac{1}{4}r^2 + \frac{2y_4}{y_1+y_2+y_3+y_4}},$$

where $r = \frac{y_1-2y_2-2y_3-y_4}{y_1+y_2+y_3+y_4}$.

Example *with* hidden variables

Let $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)$ follow a multinomial distribution with probabilities

$$\left(\frac{1}{2},\ \frac{\theta}{4},\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{\theta}{4}\right), \quad \theta \in (0,1) \text{ unknown},$$

We have 197 samples of $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)$:

$$x_1 + x_2 = y_1 = 125, \quad x_3 = y_2 = 18, \quad x_4 = y_3 = 20, \quad x_5 = y_4 = 34.$$

Here, X_1 and X_2 are not observable but $X_1 + X_2 := Y_1$ is observable. Estimate θ from the data by **EM**.

Solution.

- Define the Q function.[9] Recall $Q(\theta, \theta^{\text{old}}) = \int [\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]\, p(\boldsymbol{z}|\boldsymbol{x}, \theta^{\text{old}})\, d\boldsymbol{z}$.

[9] Here, $\boldsymbol{x} = (x_1, \cdots, x_5)$, $\boldsymbol{Y} = (Y_1, \cdots, Y_4)$ and $\boldsymbol{y} = (y_1, \cdots, y_4)$.

- E step: compute the Q function.

First,

$$\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{X}=\boldsymbol{x}|\theta)=\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|\theta)=\frac{(x_1+\cdots+x_5)!}{x_1!\cdots x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\theta}{4}\right)^{x_2+x_5}\left(\frac{1-\theta}{4}\right)^{x_3+x_4}.$$

Second,

$$\begin{aligned}
Q(\theta,\theta^{\text{old}}) &= \sum_{\boldsymbol{x}}\{C+x_2\log\theta+x_5\log\theta+(x_3+x_4)\log(1-\theta)\}\,\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y},\theta^{\text{old}}) \\
&= \sum_{\boldsymbol{x}}\{C+x_5\log\theta+(x_3+x_4)\log(1-\theta)\}\,\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y},\theta^{\text{old}}) \\
&\quad+\sum_{\boldsymbol{x}}x_2\log\theta\;\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y},\theta^{\text{old}}) \\
&= C+(y_2+y_3)\log(1-\theta)+y_4\log\theta+\frac{\theta^{\text{old}}y_1}{2+\theta^{\text{old}}}\log\theta.
\end{aligned}$$

- M step: update θ.

Take derivatives to get

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \theta} = -\frac{y_2 + y_3}{1-\theta} + \left(\frac{\theta^{\text{old}} y_1}{2 + \theta^{\text{old}}} + y_4 \right) \frac{1}{\theta}.$$

Letting $\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \theta} = 0$:

$$\theta^{\text{new}} = \frac{2y_4 + (y_1 + y_4)\theta^{\text{old}}}{2(y_2 + y_3 + y_4) + (\sum_{j=1}^{4} y_j)\theta^{\text{old}}},$$

i.e. the EM iteration takes the form $\theta^{\text{new}} = M(\theta^{\text{old}})$, where

$$M(\theta) = \frac{2y_4 + (y_1 + y_4)\theta}{2(y_2 + y_3 + y_4) + (\sum_{j=1}^{4} y_j)\theta}.$$

$\Rightarrow \theta^{(0)} \to \theta^{(1)} = M(\theta^{(0)}) \to \theta^{(2)} = M(\theta^{(1)}) \to \cdots$ until convergence.

- Convergence analysis.

Recall that the true solution θ^* satisfies

$$\theta^* = \frac{2y_4 + (y_1 + y_4)\theta^*}{2(y_2 + y_3 + y_4) + (\sum_{j=1}^4 y_j)\theta^*} = M(\theta^*).$$

Using the formula: for $\theta \in (0, 1)$,

$$\frac{d}{d\theta}\frac{A + B\theta}{C + D\theta} = \frac{B(C + D\theta) - D(A + B\theta)}{(C + D\theta)^2} = \frac{BC - DA}{(C + D\theta)^2},$$

we have

$$\begin{aligned}
|M'(\theta)| &= \left| \frac{2(y_1 + y_4)(y_2 + y_3 + y_4) - 2y_4(y_1 + y_2 + y_3 + y_4)}{\Big(2(y_2 + y_3 + y_4) + \theta(y_1 + y_2 + y_3 + y_4)\Big)^2} \right| \\
&\leq \left| \frac{2(y_1 + y_4)(y_2 + y_3 + y_4) - 2y_4(y_1 + y_2 + y_3 + y_4)}{4(y_2 + y_3 + y_4)^2} \right| \\
&\leq \frac{y_1(y_2 + y_3)}{2(y_2 + y_3 + y_4)^2} = \frac{125 \times 38}{2 \times 72^2} = 0.4581.
\end{aligned}$$

So

$$|M(\theta_1) - M(\theta_2)| = |M'(\theta)(\theta_1 - \theta_2)| \leq 0.4581|\theta_1 - \theta_2|.$$

Gaussian Mixture Distribution

The Gaussian mixture distribution is defined to be the distribution with the following density

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

Here

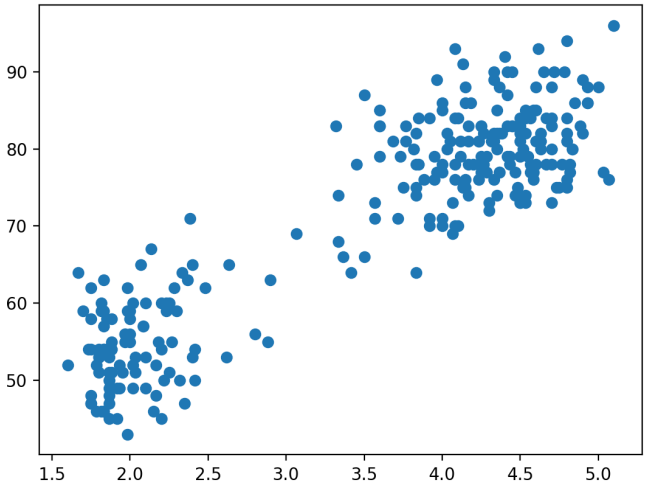
- $\boldsymbol{x} \in \mathbb{R}^n$, so the Gaussian mixture distribution is supported on $\mathbb{R}^n$.
- Mixture proportion[10]: $\pi_1, \ldots, \pi_K \in (0, 1)$ and $\pi_1 + \cdots + \pi_K = 1$.
- Mixture component: $\boldsymbol{\mu}_k \in \mathbb{R}^n$ is the mean, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{n \times n}$ is the covariance matrix, of the k'th Gaussian density function $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Parameters: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^\top$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)^\top$.

Why Gaussian mixture?
Gaussian mixture density is a universal approximator for smooth densities.

[10] The proportion of each cluster.

Gaussian mixture model (GMM) clustering

- Data.



- Assume each data point is generated from one mixture component:

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

  Unknown: which component each point is generated from, parameters of each component, mixture proportion.

- Goal: parameter estimation and *probabilistic* cluster prediction.
- Hidden variable $\Rightarrow$ EM algorithm.

EM for GMM clustering: general procedure

- Initialize the means $\boldsymbol{\mu}_k$'s, the covariance matrices $\boldsymbol{\Sigma}_k$'s, and the mixture proportion π_k's with $\pi_1 + \ldots + \pi_K = 1$.
- **E step**. For each $k = 1, \ldots, K$ and $n = 1, \ldots, N$, define

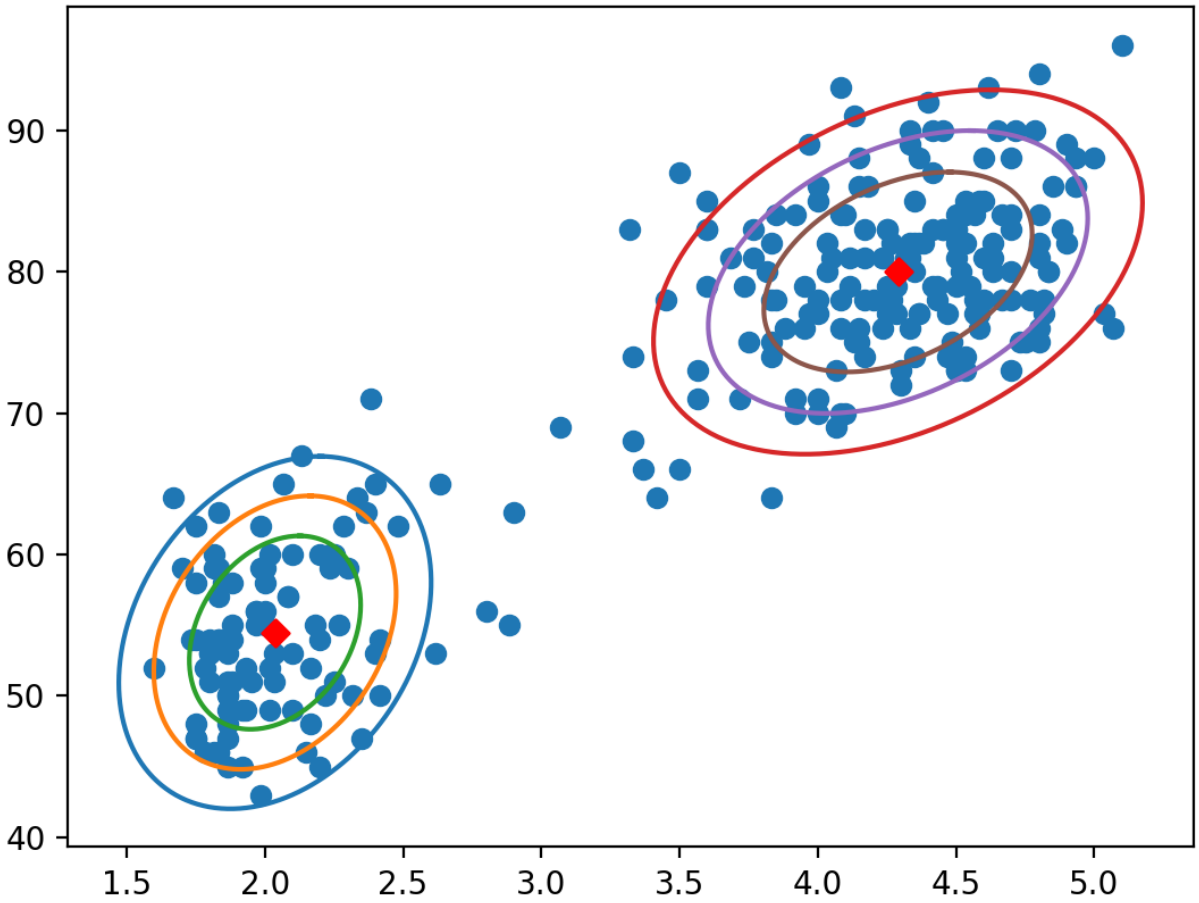
$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \qquad N_k = \sum_{i=1}^N \gamma_{ik}.$$

- **M step**. For each $k = 1, \ldots, K$, update:

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} \boldsymbol{x}_i, \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})^\top, \\
\pi_k^{\text{new}} &= \frac{N_k}{N}.
\end{aligned}$$

- Iterate.
- Clustering of the sample $\boldsymbol{x}_i$: $\arg\max_{k=1,\cdots,K} \gamma_{ik}$

Illustration: GMM with $K = 2$



- Ellipses are GMM.
- Refer to the tutorial for python code.

Details of EM in GMM clustering (1/6)

- **Identification r.v. for each sample.**
  - Let the $\mathbb{R}^K$-valued r.v. $\boldsymbol{\Delta}_i := (\Delta_{i1}, \cdots, \Delta_{iK})$ be the identification of sample $\boldsymbol{x}_i$, $i = 1, \cdots, N$, i.e.

$$\boldsymbol{\Delta}_i = (0, \cdots, 0, 1, 0, \cdots, 0) \text{ if sample } i \text{ belongs to cluster } k.$$

  - Only one coordinate of $\boldsymbol{\Delta}_i$ is 1; all others are 0.
  - $\{\boldsymbol{\Delta}_i\}_{i=1}^K$ are independent of each other.
  - For all i,

$$\mathbb{P}(\Delta_{ik} = 1) = \pi_k.$$

  - Define [11]

$$\delta_{ik} = \begin{cases} 1, & \text{if sample } i \text{ belongs to cluster } k, \\ 0, & \text{else.} \end{cases}$$

  Then,

$$p(\boldsymbol{\delta}_i) := \mathbb{P}(\boldsymbol{\Delta}_i = (\delta_{i1}, \cdots, \delta_{iK})) = \pi_1^{\delta_{i1}} \times \pi_2^{\delta_{i2}} \times \cdots \times \pi_K^{\delta_{iK}} = \Pi_{k=1}^K \pi_k^{\delta_{ik}}.$$

[11] Consider δ_{ik} as observations of Δ_{ik}.

Details of EM in GMM clustering (2/6)

- **Compute the likelihood function.** Given a set of parameters $\theta := (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi)$, the likelihood function with complete data is

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_N | \theta) = p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N | \delta_1, \cdots, \boldsymbol{\delta}_N, \theta) p(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_N | \theta),$$

where

$$p(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_N | \theta) = p(\boldsymbol{\delta}_1 | \theta) \cdots p(\boldsymbol{\delta}_N | \theta) = \Pi_{i=1}^N \Pi_{k=1}^K \pi_k^{\delta_{ik}},$$

and

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N | \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_N, \theta) = \Pi_{i=1}^N p(\boldsymbol{x}_i | \boldsymbol{\delta}_i, \theta) = \Pi_{i=1}^N \Pi_{k=1}^K \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\delta_{ik}}.$$

Then

$$\begin{aligned} & p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_N | \theta) \\ = & \Pi_{i=1}^N \Pi_{k=1}^K \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\delta_{ik}} \times \Pi_{i=1}^N \Pi_{k=1}^K \pi_k^{\delta_{ik}} \\ = & \Pi_{i=1}^N \Pi_{k=1}^K \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\delta_{ik}} \pi_k^{\delta_{ik}}. \qquad \text{(why?)} \end{aligned}$$

Details of EM in GMM clustering (3/6)

- **Evaluate** $Q(\theta, \theta^{\text{old}})$. First, given θ, the likelihood function is

$$L(\boldsymbol{x}, \boldsymbol{\delta}|\theta) = \Pi_{i=1}^{N}\Pi_{k=1}^{K}\left\{\pi_k \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}\right\}^{\delta_{ik}},$$

which yields the log-likelihood

$$l(\boldsymbol{x}, \boldsymbol{\delta}|\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}\log\left\{\pi_k \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)}\right\}.$$

Second, compute the conditional expectation of δ_{ik} given $\boldsymbol{x}$ and θ^{old}

$$\begin{aligned}
\mathbb{E}[\Delta_{ik}|\boldsymbol{X}=\boldsymbol{x}, \theta^{\text{old}}] &= \mathbb{P}(\Delta_{ik}=1|\boldsymbol{X}=\boldsymbol{x}, \theta^{\text{old}}) = \mathbb{P}(\Delta_{ik}=1|\boldsymbol{X}_i=\boldsymbol{x}_i, \theta^{\text{old}}) \\
&= \frac{\mathbb{P}(\boldsymbol{X}_i=\boldsymbol{x}_i|\Delta_{ik}=1, \theta^{\text{old}})\mathbb{P}(\Delta_{ik}=1|\theta^{\text{old}})}{\sum_{k=1}^{K}\mathbb{P}(\boldsymbol{X}_i=\boldsymbol{x}_i|\Delta_{ik}=1, \theta^{\text{old}})\mathbb{P}(\Delta_{ik}=1|\theta^{\text{old}})} \\
&= \frac{\pi_k^{\text{old}}\mathcal{N}(\boldsymbol{x}_i|\mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{k=1}^{K}\pi_k^{\text{old}}\mathcal{N}(\boldsymbol{x}_i|\mu_k^{\text{old}}, \Sigma_k^{\text{old}})} := \gamma_{ik}(\theta^{\text{old}}).
\end{aligned}$$

Details of EM in GMM clustering (4/6)

Third, evaluate Q (E step):

$$
\begin{aligned}
& Q(\theta, \theta^{\text{old}}) \\
&= \mathbb{E}[l(\boldsymbol{X}, \boldsymbol{\Delta}) | \boldsymbol{X} = \boldsymbol{x}, \theta^{\text{old}}] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}[\Delta_{ik} | \boldsymbol{X} = \boldsymbol{x}, \theta^{\text{old}}] \log \left\{ \pi_k \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)} \right\} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}(\theta^{\text{old}}) \left\{ \log \pi_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} + C
\end{aligned}
$$

- **M step-1**: update μ.

Details of EM in GMM clustering (5/6)

Recall

$$
\begin{aligned}
& Q(\theta, \theta^{\text{old}}) \\
= & \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}(\theta^{\text{old}}) \left\{ \log \pi_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} + C
\end{aligned}
$$

M step-2: update Σ. We will use two facts

$$
\frac{\partial |\Sigma|}{\partial \Sigma} = |\Sigma| (\Sigma^{-1})', \qquad \frac{\partial A' \Sigma^{-1} B}{\partial \Sigma} = -(\Sigma^{-1})' A B' (\Sigma^{-1})'.
$$

Details of EM in GMM clustering (6/6)

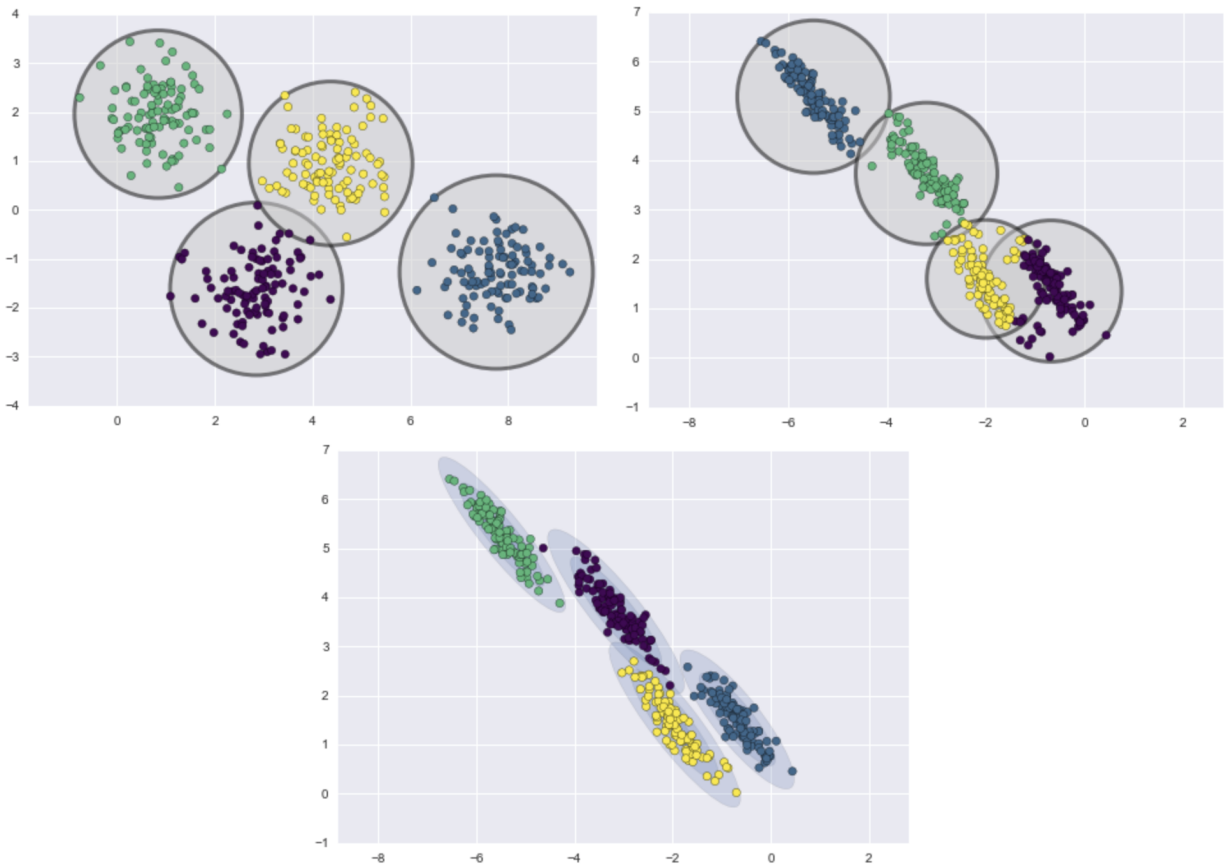
Recall

$$
\begin{aligned}
& Q(\theta, \theta^{\text{old}}) \\
= & \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}(\theta^{\text{old}}) \left\{ \log \pi_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} + C
\end{aligned}
$$

M step-3: update π.

Comparison of K-means and GMM

- K-means does not account for variance.
- K-means places a circle at the center of each cluster, with a radius defined by the farthest points in the cluster $\Rightarrow$ fine when data is circular.
- In contrast, Gaussian mixture models can handle even very oblong clusters.



Comparison of K-means and GMM, *cont'd*

- K-means performs hard classification whereas GMM performs soft classification
  - K-means tells each data point belongs to which cluster.
  - GMM tells the probability that each data point belongs to each of the possible clusters.
- Iteration times of EM are more than K-means. It is common to run K-means first to find a suitable initialization for EM.