Note on Clustering

This version: October 11, 2023

1 Introduction

Clustering is a type of unsupervised learning where the goal is to group or segment a set of data points into clusters or groups where data points in the same group are more similar to each other than those in other. groups.

Example 1. *Market segmentation in Netflix.*

Cluster 1: Users who like actions, thrillers
Cluster 2: Users who like comedies, musicals, romances
Cluster 3: Users who like fantasies, sci-fi

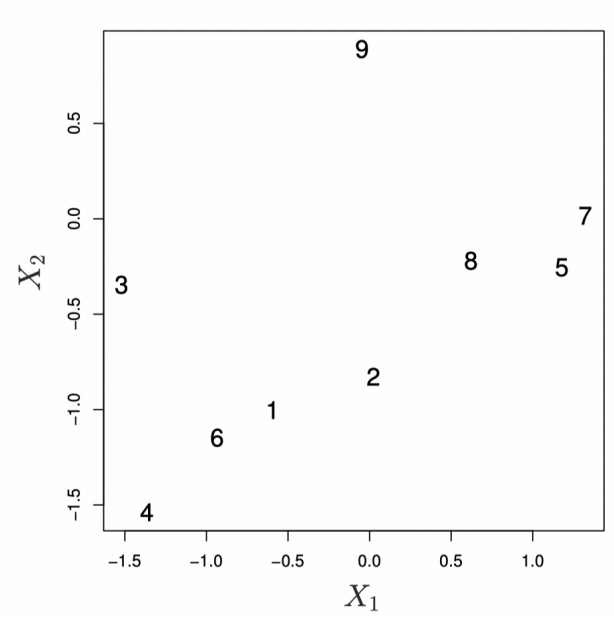
How can we find clusters?

2 Hierarchical Clustering

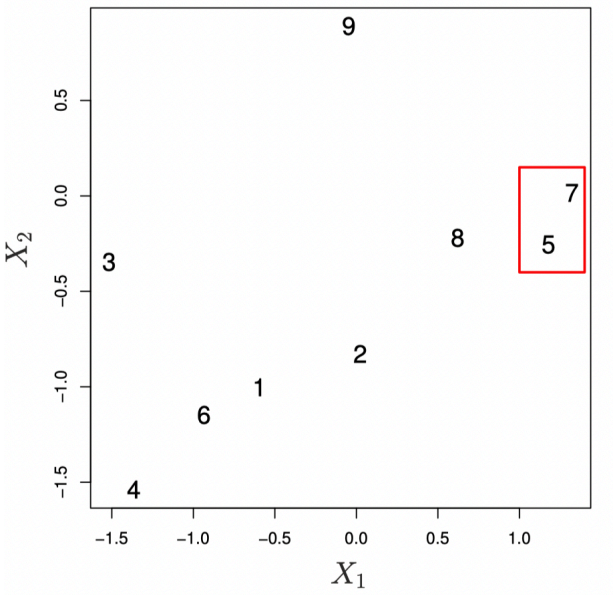
Hierarchical Clustering iteratively groups the closest observations. [1]

1. Initially, there are nine distinct clusters $\{1\}, \{2\}, \ldots, \{9\}$.

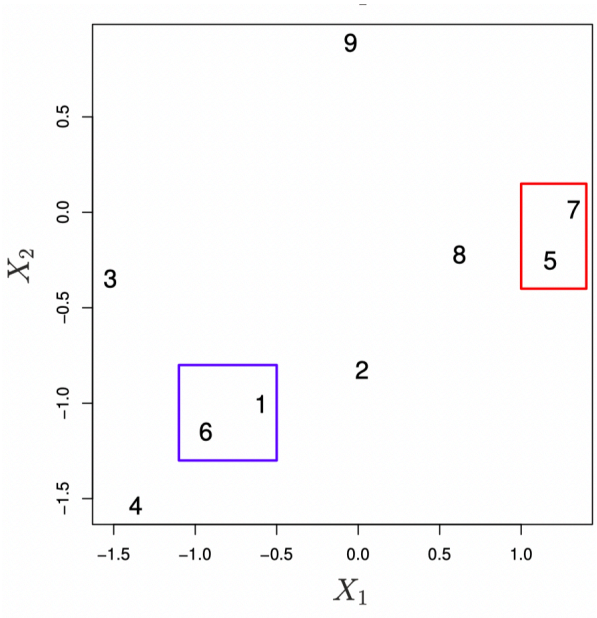
[1] Graphs in this subsection are copied from our textbook "An Introduction to Statistical Learning."



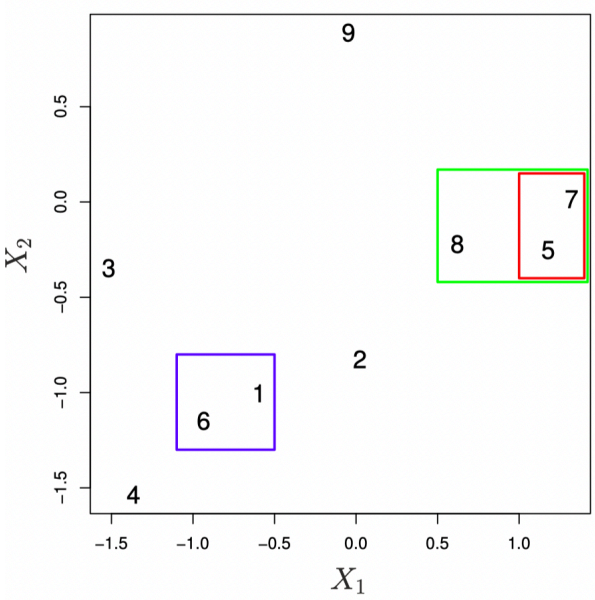
2. The two clusters that are closest together, {5} and {7}, are fused into a single cluster.



3. The two clusters that are closest together, {6} and {1}, are fused into a single cluster.



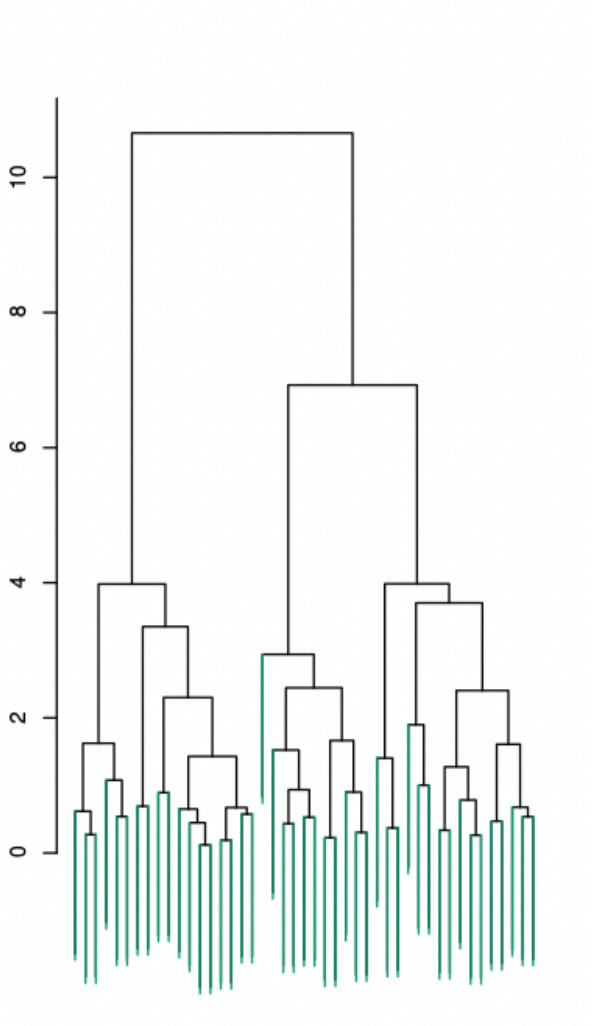
4. The two clusters that are closest together using complete linkage, $\{8\}$ and the cluster $\{5, 7\}$, are fused into a single cluster.



How to define distance between clusters?

Average Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.

Complete Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.



A Dendogram can visualize the hierarchical clustering. Its y-axis measures distance between two clusters being pooled. We can use it to decide the final clustering. The so-called 'Bottom up' approach.

Algorithm 1 (Hierarchical Clustering)**.**

1. *Begin with n observations and a measure of all the $\binom{n}{2}$ pairwise dissimilarities. Treat each observations as its own cluster.*
2. *For $i = n, n-1, \ldots, 2$:*

(a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are lest dissimilar. Fuse these two clusters. The dissimilarities between these two clusters indicates the height in the dendogram at which the fusion should be placed.

(b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

3 K-means Clustering

Suppose that we want to group observations into K clusters $C_1, C_2 \ldots, C_K$. These set must satisfy two properties:

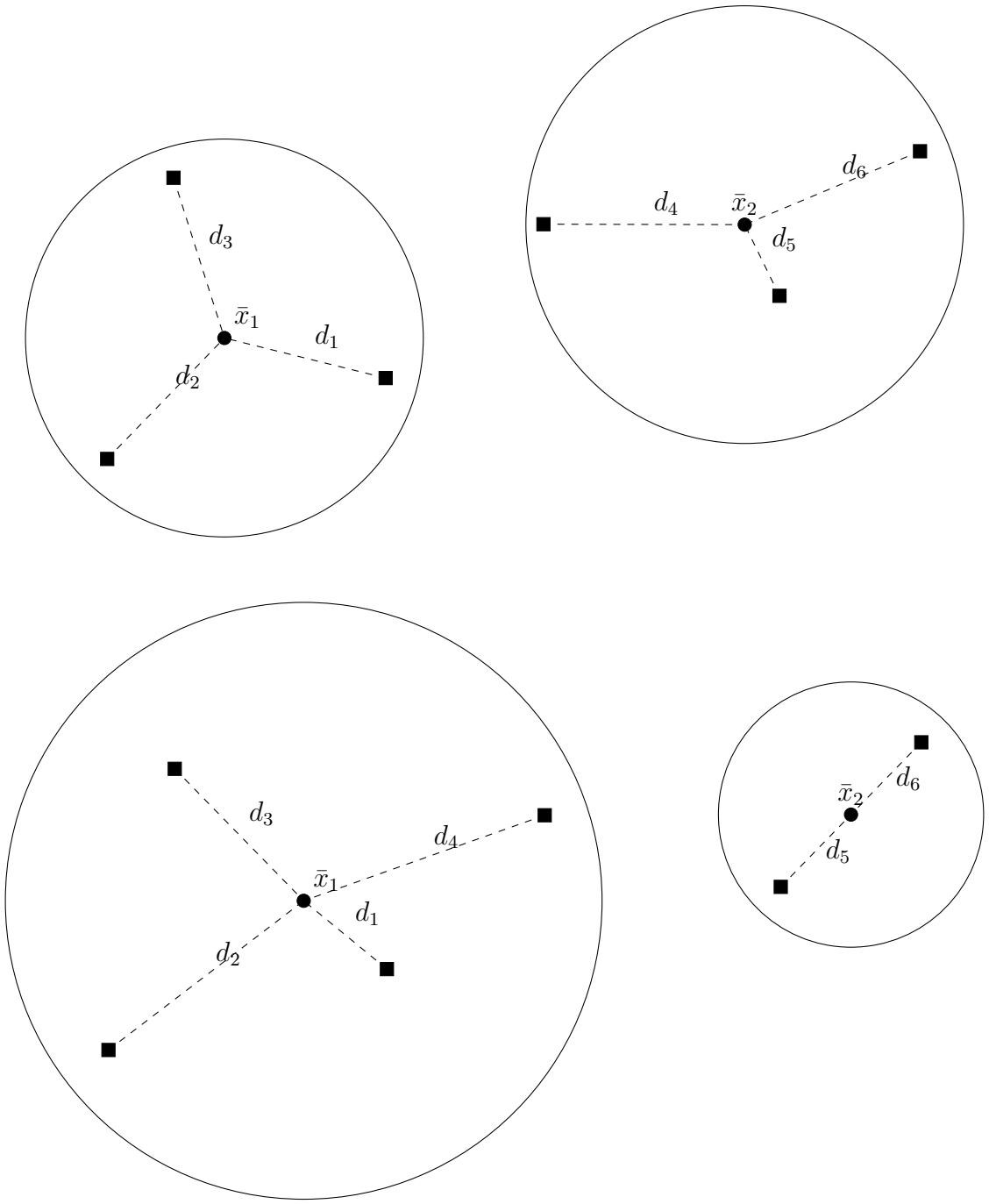
$$C_k \cap C_{k'} = \varnothing, \quad \forall k \neq k'$$
$$\bigcup_{k=1}^{K} C_k = \{1, 2, \ldots, n\}.$$

Given K, we would like to find the best clustering $C_1, C_2 \ldots, C_K$ that minimizes

$$\sum_{k=1}^{K} \overbrace{\sum_{i \in C_k} \underbrace{\sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2}_{\text{Distance to centroid}}}^{\text{Within cluster deviation}},$$

given that our data $\boldsymbol{x}$ is of p dimension.

Two possible clusterings:



The objection function is to minimize

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2.$$

How to find optimal $C_1, \ldots, C_K$?

1. Randomly assign group, e.g., randomly pick k observations as centroids.

2. Calculate $\bar{x}_k$

3. Reassign observation i to

$$\arg \min_{i=1,2,\ldots,k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

4. Repeat 2 and 3 until convergence.

3.1 Gaussian Mixture Models

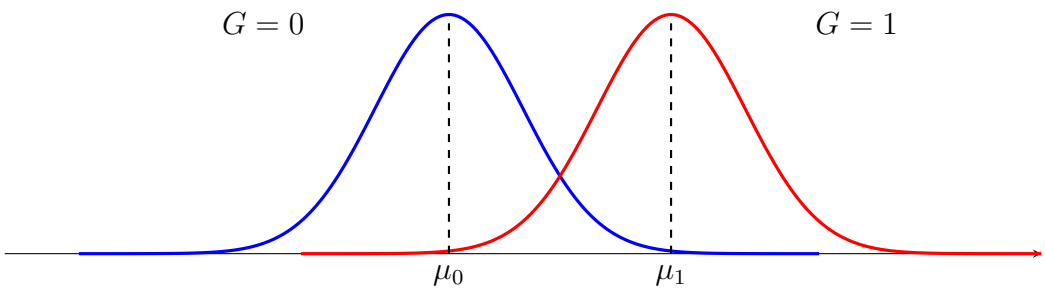
Let us perform clustering in a statistical fashion. Assume $\boldsymbol{X}$ comes from K distributions. For now, assume $K = 2$. Consider a Gaussian mixture model

$$\boldsymbol{X} = G\boldsymbol{Y}_1 + (1 - G)\boldsymbol{Y}_0,$$

where $G \sim Bernoulli(\pi)$, $\boldsymbol{Y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$, and $(G, \boldsymbol{Y}_0, \boldsymbol{Y}_1)$ are mutually independent, equivalently

$$\boldsymbol{X} = \begin{cases} \boldsymbol{Y}_1 & \text{if } G = 1 \\ \boldsymbol{Y}_0 & \text{if } G = 0 \end{cases}.$$

Example 2 ($p = 1$)**.** *The one dimension case can be visualized as the following plot*



Why imposing the model useful? If we know π, $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$,

$$\begin{aligned} f(G=1 \mid \boldsymbol{X}=\boldsymbol{x}) =& \frac{f_{\boldsymbol{X},G}(\boldsymbol{x},1)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ =& \frac{f_{\boldsymbol{X},G}(\boldsymbol{x},1)}{f_{\boldsymbol{X},G}(\boldsymbol{x},1)+f_{\boldsymbol{X},G}(\boldsymbol{x},0)} \\ =& \frac{f_{\boldsymbol{X}|G=1}(\boldsymbol{x})f_G(1)}{f_{\boldsymbol{X}|G=1}(\boldsymbol{x})f_G(1)+f_{\boldsymbol{X}|G=0}(\boldsymbol{x})f_G(0)} \end{aligned}$$

Using Bayes Rule

$$\begin{aligned} & f_G(1)=\pi \\ & \boldsymbol{X} \mid G=1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ & \boldsymbol{X} \mid G=0 \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \end{aligned}$$

How to calculate $\boldsymbol{\theta}: (\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$? We know the marginal distribution

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|G=1}(\boldsymbol{x})\pi + f_{\boldsymbol{X}|G=0}(\boldsymbol{x})(1-\pi).$$

So, given the data $x_1, x_2, ..., x_n$, the log likelihood function is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) =& \log f_{\boldsymbol{X}}(\boldsymbol{x}) \\ =& \log \prod_{i=1}^{n} f_{\boldsymbol{X}_1}(\boldsymbol{x}_i) \\ =& \sum_{i=1}^{n} \log f_{\boldsymbol{X}_1}(\boldsymbol{x}_i) \\ =& \sum_{i=1}^{n} \log \left(f_{\boldsymbol{X}_1|G=1}(\boldsymbol{x})\pi + f_{\boldsymbol{X}_1|G=0}(\boldsymbol{x})(1-\pi) \right). \end{aligned}$$

It is hard to use the first order condition to solve the maximum likelihood problem, since $\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}}$ involves $\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$, and is complicated.

By contrast, when G is observed

$$\begin{aligned} f_{\boldsymbol{X},G}(\boldsymbol{x}, g) =& f_{\boldsymbol{X}|G=g}(\boldsymbol{x}) f_G(g) \\ =& f_{\boldsymbol{X}|G=g}(\boldsymbol{x}) \pi^g (1-\pi)^{(1-g)} \end{aligned}$$

Now our data is $(x_1, z_1), (x_2, z_2), ..., (x_n, z_n)$, and the likelihood function becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{g}) =& \log f_{\boldsymbol{X},\boldsymbol{G}}(\boldsymbol{x}, \boldsymbol{g}) \\ =& \sum_{i=1}^{n} g_i \log \pi + (1-g_i) \log(1-\pi) + g_i f_{\boldsymbol{X}_1|G=1}(\boldsymbol{x}_i) + (1-g_i) f_{\boldsymbol{X}_1|G=0}(\boldsymbol{x}_i) \\ =& \sum_{i=1}^{n} g_i \log \pi + (1-g_i) \log(1-\pi) + g_i \log f_{\boldsymbol{X}_1|G=1}(\boldsymbol{x}_i) + (1-g_i) \log f_{\boldsymbol{X}_1|G=0}(\boldsymbol{x}_i) \end{aligned}$$

In our Gaussian setting

$$f_{\boldsymbol{X}_1|G=1}(\boldsymbol{x}_i) = (2\pi)^{-\frac{p}{2}} \det(\boldsymbol{\Sigma}_1)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_1) \right).$$

So

$$\frac{\partial \mathcal{L}}{\partial \pi} = \sum_{i=1}^{n} \frac{g_i}{\pi} - \frac{1-g_i}{1-\pi} = 0 \Rightarrow \hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} g_i.$$

That is, we estimate the proportion of group 1 by counting how many observations are in group 1, and $\hat{\pi}$ does not depend on other parameter estimators. For other parameters, MLE are also simple:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_1} = \sum_{i=1}^{n} g_i (\boldsymbol{x}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} = 0 \Rightarrow \hat{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^{n} g_i \boldsymbol{x}_i}{\sum_{i=1}^{n} g_i}$$

Similarly,

$$\hat{\boldsymbol{\mu}}_2 = \frac{\sum_{i=1}^n (1-g_i)\boldsymbol{x}_i}{\sum_{i=1}^n (1-g_i)}.$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_1} = \sum_{i=1}^n g_i \left(-\frac{1}{2}\frac{\partial \log\det(\boldsymbol{\Sigma}_1)}{\partial \boldsymbol{\Sigma}_1} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1) \right) = 0$$

Note that $\frac{\partial \log\det(\boldsymbol{\Sigma}_1)}{\partial \boldsymbol{\Sigma}_1} = \boldsymbol{\Sigma}_1^{-1}$, and

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1) =& \frac{\partial}{\partial \boldsymbol{\Sigma}_1} tr\left((\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)\right) \\
=& \frac{\partial}{\partial \boldsymbol{\Sigma}_1} tr\left(\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\right) \\
=& -\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

So,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_1} = \sum_{i=1}^n g_i \boldsymbol{\Sigma}_1^{-1} - \sum_{i=1}^n g_i \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1} = \mathbf{0}.$$

Hence

$$\widehat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{i=1}^n g_i(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)'}{\sum_{i=1}^n g_i}.$$

Similarly

$$\widehat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{i=1}^n (1-g_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_2)'}{\sum_{i=1}^n (1-g_i)}.$$

To solve the MLE without observing $(g_1, g_2, ..., g_n)$, we can start by pretending we do observe $g_i's$

1. Randomly assign $\boldsymbol{G}$
2. Calculate $\hat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Sigma}}$, $\hat{\pi}$

3. Update $\boldsymbol{G}$ by

$$P(G_i = 1 \mid \boldsymbol{X}) = \frac{\hat{f}_{\boldsymbol{X}|G=1}(\boldsymbol{X})\hat{\pi}}{\hat{f}_{\boldsymbol{X}}(\boldsymbol{X})}$$

4. Repeat 2 and 3 until convergence.

Example 3. *Suppose* $\pi = \frac{1}{2}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{I}$

$$P(G = 1 \mid \boldsymbol{X}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_1)'(\boldsymbol{X} - \boldsymbol{\mu}_1)\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_1)'(\boldsymbol{X} - \boldsymbol{\mu}_1)\right) + \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_0)'(\boldsymbol{X} - \boldsymbol{\mu}_0)\right)}$$
$$P(G = 0 \mid \boldsymbol{X}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_0)'(\boldsymbol{X} - \boldsymbol{\mu}_0)\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_1)'(\boldsymbol{X} - \boldsymbol{\mu}_1)\right) + \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_0)'(\boldsymbol{X} - \boldsymbol{\mu}_0)\right)}.$$

Now

$$f_{\boldsymbol{X}|G=1}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)'(\boldsymbol{x} - \boldsymbol{\mu}_1)\right)$$
$$f_{\boldsymbol{X}|G=0}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)'(\boldsymbol{x} - \boldsymbol{\mu}_0)\right),$$

so $f_{\boldsymbol{X}|G=1}(\boldsymbol{x}) > f_{\boldsymbol{X}|G=0}(\boldsymbol{x})$ when

$$(\boldsymbol{x} - \boldsymbol{\mu}_1)'(\boldsymbol{x} - \boldsymbol{\mu}_1) < (\boldsymbol{x} - \boldsymbol{\mu}_0)'(\boldsymbol{x} - \boldsymbol{\mu}_0).$$

That is, observation i is more likely coming from the distribution $\mathcal{N}_p(\mu_1, I)$ if it is closer to μ_1. Sounds like K-means? The example implies that we can view K-means is a version of Gaussian Mixture models with

1. $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{I}$
2. $\pi = 0.5$

and 'hard' assignment, in which group membership is deterministic rather probabilistic.