Mathematics for Machine Learning
— Gaussian Mixture Models

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,
National Taiwan Ocean University

Fall 2025

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Arnold J. Insel, Lawrence E. Spence, Stephen H. Friedberg: Linear Algebra, 4th Edition. Prentice Hall. 2013.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra, 12th Edition. Wiley. 2019.*
- We could partially refer to the monograph: *Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213*

Outline

1. Introduction & Gaussian Mixture Model (GMM)
2. Parameter Learning via Maximum Likelihood
   - Updating the Means
   - Updating the Covariances
   - Updating the Mixture Weights

Outline

Introduction

Focus

- **Goal:** Density Estimation.
- Covering two important concepts:
  - Expectation maximization (EM).
  - Latent variable perspective.

Motivation

- A straightforward way to represent data: Let them present themselves directly.
- **Issue:** The data might be *dirty* or too huge to show all of them.

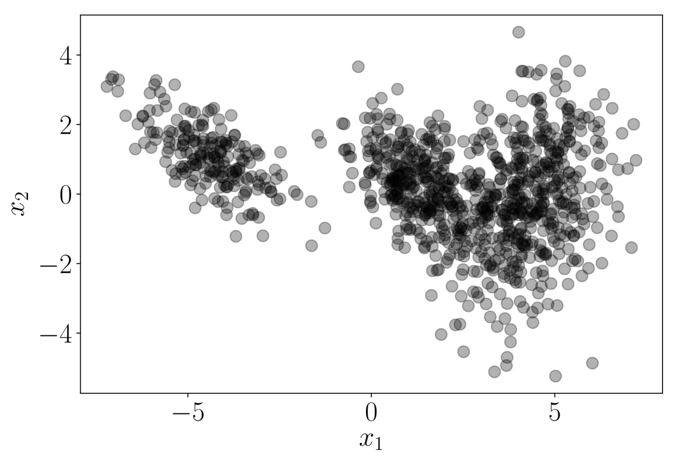
Motivation

- A straightforward way to represent data: Let them present themselves directly.
- **Issue:** The data might be *dirty* or too huge to show all of them.

We want to represent the data compactly using a density from a parametric family, such as Gaussian or Beta distribution.

- Mean & variance.

One Gaussian representation might not be meaningful.



A Solution

- Consider mixture models:
  - A convex combination of K simple base distributions.
  - A distribution $p(\mathbf{x})$:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}),$$

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1.$$

    - π_k: *mixture weights*.
  - More expressive than a base distribution.

A Solution

- Consider mixture models:
  - A convex combination of K simple base distributions.
  - A distribution $p(\mathbf{x})$:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}),$$

$$0 \leq \pi_k \leq 1, \ \sum_{k=1}^{K} \pi_k = 1.$$

    - π_k: *mixture weights*.
  - More expressive than a base distribution.
  - Gaussian mixture modesl (GMMs): the base distributions are Gaussians.

Gaussian Mixture Model

Gaussian Mixture Model

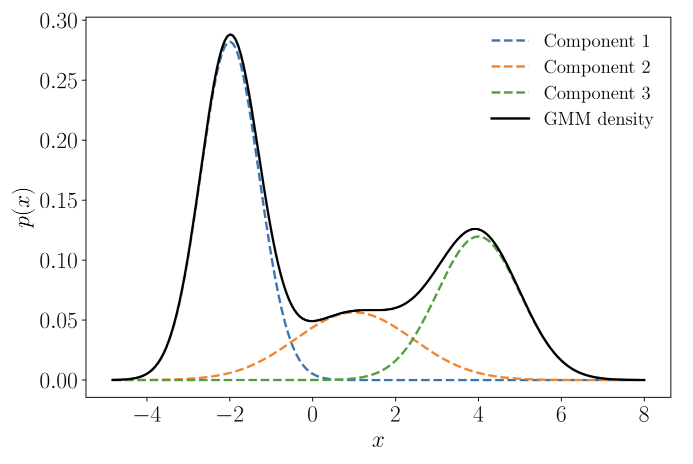
A Gaussian mixture model is a density model where we combine a finite number of K Gaussian distributions $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ such that

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

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1,$$

where $\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k \mid k = 1, \ldots, K\}$.

GMMs



$$p(x \mid \boldsymbol{\theta}) = 0.5\mathcal{N}(x \mid -2, 0.5) + 0.2\mathcal{N}(x \mid 1, 2) + 0.3\mathcal{N}(x \mid 4, 1).$$

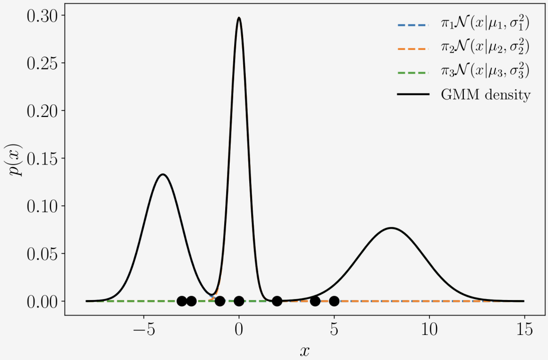
Outline

The Setting

- A dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where each $\mathbf{x}_i$ is drawn i.i.d. from an unknown distribution $p(\mathbf{x})$.
- Parameters: $\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k \mid k = 1, \ldots, K\}$.

Example of an Initial Setting

- $\mathcal{X} = \{-3, -2.5, -1, 0, 2, 4, 5\}$.
- $K = 3$.
- $p_1(x) = \mathcal{N}(x \mid -4, 1)$, $p_2(x) = \mathcal{N}(x \mid 0, 0.2)$, $p_3(x) = \mathcal{N}(x \mid 8, 3)$.
- $\pi_1 = \pi_2 = \pi_3 = 1/3$.



The Likelihood

By the i.i.d. assumption, we have the factorized likelihood

$$p(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta}), \quad p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Then the log-likelihood is

$$\mathcal{L} := \log p(\mathcal{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The Likelihood

By the i.i.d. assumption, we have the factorized likelihood

$$p(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta}), \quad p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Then the log-likelihood is

$$\mathcal{L} := \log p(\mathcal{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- **Goal:** Find parameters $\boldsymbol{\theta}^*_{ML}$.

MLE

- We cannot obtain a closed-form solution here (except for $K = 1$, i.e., single Gaussian).

MLE

- We cannot obtain a closed-form solution here (except for $K = 1$, i.e., single Gaussian).
- We exploit an iterative scheme to find $\boldsymbol{\theta}^*_{ML}$:

MLE

- We cannot obtain a closed-form solution here (except for $K = 1$, i.e., single Gaussian).
- We exploit an iterative scheme to find $\boldsymbol{\theta}^*_{ML}$: the EM algorithm.

MLE

- We cannot obtain a closed-form solution here (except for $K = 1$, i.e., single Gaussian).
- We exploit an iterative scheme to find $\boldsymbol{\theta}^*_{ML}$: the EM algorithm.
- **The key idea:** Update one model parameter at a time while keeping the others fixed.

Necessary conditions for a local optimum of $\mathcal{L}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top \iff \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0}^\top \iff \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0}^\top$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \pi_k} = 0.$$

Applying the chain rule:

$$\frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = {\color{orange}\frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})}} {\color{blue}\frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}$$

and

$${\color{orange}\frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})}} = \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

Responsibilities: Facilitating our discussions

Responsibility of the *k*th mixture conponent for *n*th data point

$$r_{ik} := \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Note that

$$p(\mathbf{x}_i \mid \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

which is proportional to the likelihood.

Responsibilities: Facilitating our discussions

Responsibility of the *k*th mixture conponent for *n*th data point

$$r_{ik} := \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Note that

  $$p(\mathbf{x}_i \mid \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  which is proportional to the likelihood.
- High responsibility $\Longrightarrow$ The data point is plausible sample from that mixture component.

Remark

$\mathbf{r}_i := [r_{i1}, \ldots, r_{iK}]^\top \in \mathbb{R}^K$ is a normalized probability vector.

Remark

$\mathbf{r}_i := [r_{i1}, \ldots, r_{iK}]^\top \in \mathbb{R}^K$ is a normalized probability vector.

- A *soft assignment* of $\mathbf{x}_i$ to the K mixture component.

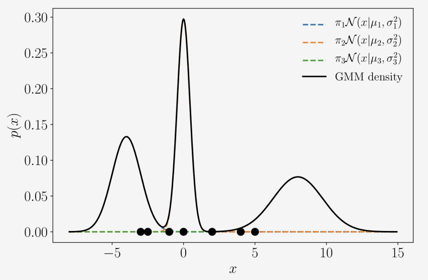
Remark

$\mathbf{r}_i := [r_{i1}, \ldots, r_{iK}]^\top \in \mathbb{R}^K$ is a normalized probability vector.

- A *soft assignment* of $\mathbf{x}_i$ to the K mixture component.
    - Similar idea: softmax functions.

Example (responsibilities of the previous example)

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 0.057 & 0.943 & 0.0 \\ 0.001 & 0.999 & 0.0 \\ 0.0 & 0.066 & 0.934 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \in \mathbb{R}^{N \times K}.$$



- Try to compute it by yourselves.

Update of the GMM Means

Theorem [Update of the Means]

The update of the mean parameters $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, of the GMM is given by

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{i=1}^{N} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{N} r_{ik}}.$$

$$
\begin{aligned}
\frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} &= \sum_{j=1}^{K} \pi_j \frac{\partial \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\partial \boldsymbol{\mu}_k} = \pi_k \frac{\partial \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \pi_k (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).
\end{aligned}
$$

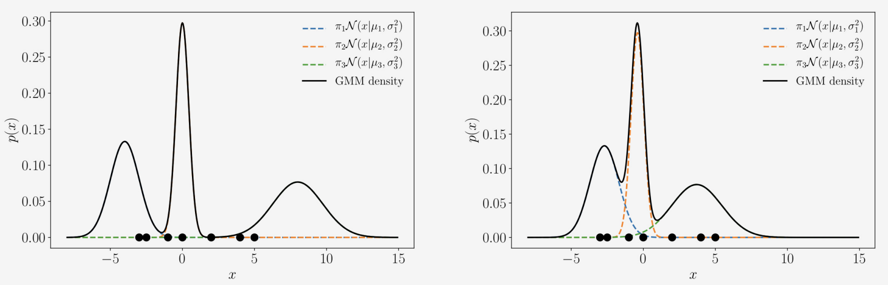
$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} &= \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\
&= \sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}.
\end{aligned}
$$

Solving $\dfrac{\partial \mathcal{L}(\boldsymbol{\mu}_k^{new})}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{new})^\top \boldsymbol{\Sigma}_k^{-1} = \mathbf{0}^\top$:

$$\sum_{i=1}^{N} r_{ik}\mathbf{x}_i = \sum_{i=1}^{N} r_{ik}\boldsymbol{\mu}_k^{new}$$

$$\iff \quad \boldsymbol{\mu}_k^{new} = \frac{\sum_{i=1}^{N} r_{ik}\mathbf{x}_i}{\sum_{i=1}^{N} r_{ik}} = \frac{1}{N_k}\sum_{i=1}^{N} r_{ik}\mathbf{x}_i,$$

where $N_k := \sum_{i=1}^{N} r_{ik}$.



- $\mu_1 : -4 \rightarrow -2.7.$
- $\mu_2 : 0 \rightarrow -0.4.$
- $\mu_3 : 8 \rightarrow 3.7.$

Remark

- r_{ik} is a function of $\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ for all $j = 1, \ldots, K$.
- Hence the updates depend on all parameters of the GMM.

Update of the GMM Covariances

Theorem [Update of the Covariances]

The update of the covariance parameters $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$, of the GMM is given by

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top,$$

where

$$r_{ik} := \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

and $N_k := \sum_{i=1}^{N} r_{ik}$.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}$$

$$\begin{aligned}
\frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(\pi_k (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \right) \\
&= \pi_k (2\pi)^{-\frac{D}{2}} \left[\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \right. \\
&\quad \left. + \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \right]
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} = -\frac{1}{2} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \boldsymbol{\Sigma}_k^{-1}, \\
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) = -\boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}
\end{aligned}$$

$$\frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right]$$

Thus,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} &= \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{i=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\
&\quad \cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right] \\
&= -\frac{1}{2} \sum_{i=1}^{N} r_{ik} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k))(\mathbf{x}_i - \boldsymbol{\mu}_k))^\top \boldsymbol{\Sigma}_k^{-1} \\
&= -\frac{1}{2} \left(\sum_{i=1}^{N} r_{ik} \right) \boldsymbol{\Sigma}_k^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1}.
\end{aligned}$$

$$\frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right]$$

Thus,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} &= \sum_{i=1}^{N} \frac{\partial \log p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^{N} \frac{1}{p(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{i=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\
&\quad \cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right] \\
&= -\frac{1}{2} \sum_{i=1}^{N} r_{ik} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k))(\mathbf{x}_i - \boldsymbol{\mu}_k))^\top \boldsymbol{\Sigma}_k^{-1} \\
&= -\frac{1}{2} N_k \boldsymbol{\Sigma}_k^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1}.
\end{aligned}$$

Setting $\dfrac{\partial \mathcal{L}}{\partial \mathbf{\Sigma}_k} = \mathbf{0}^\top$, we have

$$N_k \mathbf{\Sigma}_k^{-1} = \mathbf{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right) \mathbf{\Sigma}_k^{-1}$$

Setting $\dfrac{\partial \mathcal{L}}{\partial \mathbf{\Sigma}_k} = \mathbf{0}^\top$, we have

$$N_k \mathbf{\Sigma}_k^{-1} = \mathbf{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right) \mathbf{\Sigma}_k^{-1}$$

Then,

$$N_k \boldsymbol{I} = \mathbf{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right)$$

Setting $\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0}^\top$, we have

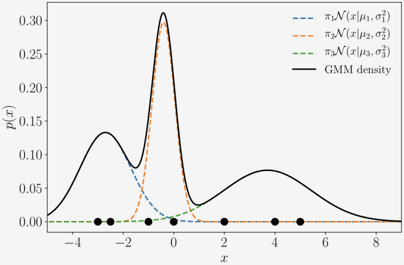
$$N_k \boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1}$$

Then,

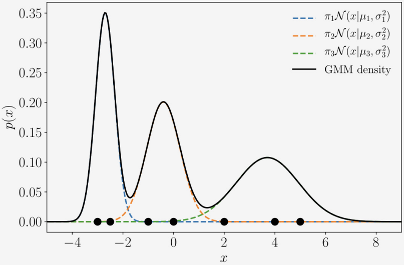
$$N_k \boldsymbol{I} = \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \right)$$

Hence,

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top .$$



(a) GMM density and individual components prior to updating the variances.



(b) GMM density and individual components after updating the variances.

- $\sigma_1^2 : 1 \to 0.14$.
- $\sigma_2^2 : 0.2 \to 0.44$.
- $\sigma_3^2 : 3 \to 1.53$.

Update of the GMM Mixture Weights

Theorem [Update of the Mixture Weights]

The update of the mixture weights of the GMM is given by

$$\pi_k^{new} = \frac{N_k}{N}, \quad k = 1, \ldots, K.$$

- N: the number of data points.
- $N_k := \sum_{i=1}^{N} r_{ik}$.

- We account for the constraint $\sum_k \pi_k = 1$.
  - Using Lagrange multipliers.

- We account for the constraint $\sum_k \pi_k = 1$.
  - Using Lagrange multipliers.
- The Lagrangian:

$$
\begin{aligned}
\mathfrak{L} &= \mathcal{L} + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \\
&= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).
\end{aligned}
$$

Obtain the partial derivative of $\mathfrak{L}$ w.r.t. π_k:

$$
\begin{aligned}
\frac{\partial \mathfrak{L}}{\partial \pi_k} &= \sum_{i=1}^{N} \frac{\mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \\
&= \frac{1}{\pi_k} \sum_{i=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \\
&= \frac{N_k}{\pi_k} + \lambda,
\end{aligned}
$$

and the partial derivative w.r.t. λ is

$$
\frac{\partial \mathfrak{L}}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1.
$$

Now we have

$$\frac{\partial \mathfrak{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda$$

$$\frac{\partial \mathfrak{L}}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1$$

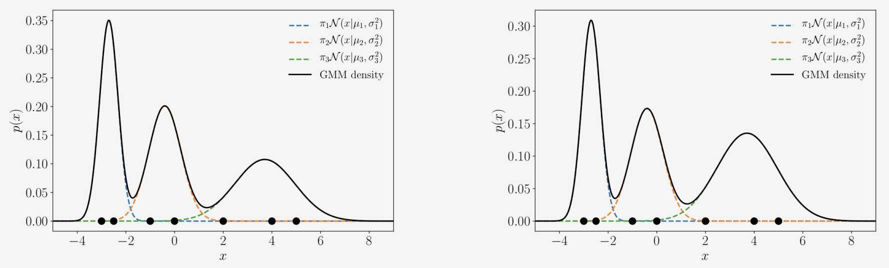
Now we have

$$\frac{\partial \mathfrak{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda$$
$$\frac{\partial \mathfrak{L}}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1$$

Setting both to $\mathbf{0}^\top$ we have

$$\pi_k = -\frac{N_k}{\lambda}$$
$$1 = \sum_{k=1}^{K} \pi_k = -\sum_{k=1}^{K} \frac{N_k}{\lambda} = -\frac{N}{\lambda}$$

So $\lambda = -N \implies \pi_k^{new} = \frac{N_k}{N}$.



- $\pi_1 : \frac{1}{3} \to 0.29.$
- $\pi_2 : \frac{1}{3} \to 0.29.$
- $\pi_3 : \frac{1}{3} \to 0.42.$

Discussions