Mixture Distribution

Statistical Machine Learning (ENGG*6600*02)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh
Summer 2023

Point Estimation

Maximum Likelihood Estimation

- Assume we have a sample with size n, i.e., $\{x_1, \ldots, x_n\}$. Also assume that we know the distribution from which this sample has been randomly drawn but we do not know the parameters of that distribution.
- For example, we know it is drawn from a normal distribution but the mean and variance of this distribution are unknown. The goal is to estimate the parameters of the distribution using the sample $\{x_1, \ldots, x_n\}$ available from it.
- This estimation of parameters from the available sample is called "**point estimation**". One of the approaches for point estimation is **Maximum Likelihood Estimation (MLE)**.
- As it is obvious from its name, MLE deals with the likelihood of data.
- We postulate that the values of sample, i.e., $x_1, \ldots, x_n$, are independent random variates of data having the sample distribution. In other words, the data has a joint distribution $f_X(x_1, \ldots, x_n|\Theta)$ with parameter Θ and we assume the variates are independent and identically distributed (*iid*) variates, i.e., $x_i \overset{iid}{\sim} f_X(x_i|\Theta)$ with the same parameter Θ.
- Considering the Bayes rule, we have:

$$\underbrace{f(\Theta|x_1, \ldots, x_n)}_{\text{posterior}} = \frac{\overbrace{f_X(x_1, \ldots, x_n|\Theta)}^{\text{likelihood}}\overbrace{\pi(\Theta)}^{\text{prior}}}{\underbrace{f_X(x_1, \ldots, x_n)}_{\text{prior}}}. \tag{1}$$

- The MLE aims to find parameter Θ which maximizes the likelihood:

$$\widehat{\Theta} = \arg\max_{\Theta} f_X(x_1, \ldots, x_n|\Theta). \tag{2}$$

Maximum Likelihood Estimation

- As the points are i.i.d., the likelihood can be written as:

$$L(x_1, \ldots, x_n|\Theta) = f(x_1, \ldots, x_n; \Theta) = \prod_{i=1}^{n} f(x_i; \Theta). \tag{3}$$

- Note that in literature, the $L(x_1, \ldots, x_n|\Theta)$ is also denoted by $L(\Theta)$ for simplicity.
- Usually, for more convenience, we use log-likelihood rather than likelihood:

$$\ell(\Theta) := \log L(\Theta) \tag{4}$$

$$= \log \prod_{i=1}^{n} f(x_i, \Theta) = \sum_{i=1}^{n} \log f(x_i, \Theta). \tag{5}$$

- So, MLE is:

$$\widehat{\Theta} = \arg\max_{\Theta} \ell(\Theta). \tag{6}$$

- Often, the logarithm is a natural logarithm for the sake of compatibility with the exponential in the well-known normal density function.
- Notice that as logarithm function is monotonic, it does not change the location of maximization of the likelihood.
- In **Maximum A Posterior (MAP)**, we maximize the posterior rather than the likelihood:

$$\widehat{\Theta} = \arg\max_{\Theta} f(\Theta|x_1, \ldots, x_n). \tag{7}$$

Expectation Maximization

- Sometimes, the data are not fully observable. For example, the data are known to be whether zero or greater than zero.
- As an illustration, assume the data are collected for a particular disease but for convenience of the patients participated in the survey, the severity of the disease is not recorded but only the existence or non-existence of the disease is reported. So, the data are not giving us complete information as $X_i > 0$ is not obvious whether is $X_i = 2$ or $X_i = 1000$.
- In this case, MLE cannot be directly applied as **we do not have access to complete information and some data are missing**.
- In this case, **Expectation Maximization (EM)** is useful.
- The main idea of EM can be summarized in this short friendly conversation:
  – What shall we do? The data is missing! The log-likelihood is not known completely so MLE cannot be used.
  – Mmm, probably we can replace the missing data with something...
  – Aha! Let us replace it with its mean.
  – You are right! We can take the mean of log-likelihood over the possible values of the missing data. Then everything in the log-likelihood will be known, and then...
  – And then we can do MLE!

Expectation Maximization

- Assume $D^{(obs)}$ and $D^{(miss)}$ denote the observed data (X_i's $= 0$ in the above example) and the missing data (X_i's > 0 in the above example).
- The EM algorithm includes two main steps, i.e., E-step and M-step.
- In the **E-step (Expectation Step)**, the log-likelihood (equation (4)), is taken expectation with respect to the missing data $D^{(miss)}$ in order to have a mean estimation of it. Let $Q(\Theta)$ denote the expectation of the likelihood with respect to $D^{(miss)}$:

$$Q(\Theta) := \mathbb{E}_{D^{(miss)}|D^{(obs)},\Theta}[\ell(\Theta)]. \tag{8}$$

- Note that in the above expectation, the $D^{(obs)}$ and Θ are conditioned on, so they are treated as constants and not random variables.
- In the **M-step (Maximization Step)**, the MLE approach is used where the log-likelihood is replaced with its expectation, i.e., $Q(\Theta)$; therefore:

$$\widehat{\Theta} = \arg\max_{\Theta} Q(\Theta). \tag{9}$$

- These two steps are iteratively repeated until convergence of the estimated parameters $\widehat{\Theta}$.

Introduction to Mixture Distribution

Introduction to Mixture Distribution

- Every random variable can be considered as a sample from a distribution, whether a well-known distribution or a not very well-known (or "ugly") distribution. Some random variables are drawn from one single distribution, such as a normal distribution. But life is not always so easy! Most of real-life random variables might have been generated from a mixture of several distributions and not a single distribution.
- The mixture distribution is a weighted summation of K distributions $\{g_1(x;\Theta_1),\ldots,g_K(x;\Theta_K)\}$ where the weights $\{w_1,\ldots,w_K\}$ sum to one. As is obvious, every distribution in the mixture has its own parameter Θ_k.
- The mixture distribution is formulated as:

$$f(x;\Theta_1,\ldots,\Theta_K)=\sum_{k=1}^{K} w_k g_k(x;\Theta_k), \quad \text{subject to} \quad \sum_{k=1}^{K} w_k = 1. \tag{10}$$

- The distributions can be from different families, for example from beta and normal distributions. However, this makes the problem very complex and sometimes useless; therefore, mostly the distributions in a mixture are from one family (e.g., all normal distributions) but with different parameters.
- We aim to find the parameters of the distributions in the mixture distribution $f(x;\Theta)$ as well as the weights (also called "mixing probabilities") w_k.

EM for Mixture Distribution

EM for Mixture Distribution

- We want to fit a mixture of K distributions $g_1(x;\Theta_1),\ldots,g_K(x;\Theta_K)$ to the data. Again, in theory, these K distributions are not necessarily from the same distribution family. For more convenience of reader, equation (10) is repeated here:

$$\begin{cases} f(x;\Theta_1,\ldots,\Theta_K) = \sum_{k=1}^{K} w_k\, g_k(x;\Theta_k), \\ \text{subject to} \quad \sum_{k=1}^{K} w_k = 1. \end{cases}$$

The likelihood and log-likelihood for this mixture are:

$$L(\Theta_1,\ldots,\Theta_K) = f(x_1,\ldots,x_n;\Theta_1,\ldots,\Theta_K) \overset{(a)}{=} \prod_{i=1}^{n} f(x_i;\Theta_1,\ldots,\Theta_K)$$

$$= \prod_{i=1}^{n}\sum_{k=1}^{K} w_k g_k(x_i;\Theta_k)$$

$$\ell(\Theta_1,\ldots,\Theta_K) = \sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} w_k g_k(x_i;\Theta_k)\Big],$$

where (a) is because of assumption that $x_1,\ldots,x_n$ are *iid*.

EM for Mixture Distribution

- We had:

$$\ell(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} w_k g_k(x_i; \Theta_k) \right].$$

- Optimizing this log-likelihood is difficult because of the summation within the logarithm. We use a trick:

$$\Delta_{i,k} := \begin{cases} 1 & \text{if } x_i \text{ belongs to } g_k(x; \Theta_k), \\ 0 & \text{otherwise,} \end{cases}$$

and its probability is:

$$\begin{cases} \mathbb{P}(\Delta_{i,k} = 1) = w_k, \\ \mathbb{P}(\Delta_{i,k} = 0) = 1 - w_k. \end{cases}$$

Therefore, the log-likelihood can be written as:

$$\ell(\Theta_1, \ldots, \Theta_K) = \begin{cases} \sum_{i=1}^{n} \log \left[w_1 \, g_1(x_i; \Theta_1) \right] & \text{if } \Delta_{i,1} = 1 \text{ and } \Delta_{i,k} = 0 \ \ \forall k \neq 1 \\ \sum_{i=1}^{n} \log \left[w_2 \, g_2(x_i; \Theta_2) \right] & \text{if } \Delta_{i,2} = 1 \text{ and } \Delta_{i,k} = 0 \ \ \forall k \neq 2 \\ \vdots & \\ \sum_{i=1}^{n} \log \left[w_K \, g_K(x_i; \Theta_K) \right] & \text{if } \Delta_{i,K} = 1 \text{ and } \Delta_{i,k} = 0 \ \ \forall k \neq K \end{cases}$$

EM for Mixture Distribution

- We had:

$$\ell(\Theta_1, \ldots, \Theta_K) = \begin{cases} \sum_{i=1}^{n} \log \left[w_1 \, g_1(x_i; \Theta_1) \right] & \text{if } \Delta_{i,1} = 1 \text{ and } \Delta_{i,k} = 0 \;\; \forall k \neq 1 \\ \sum_{i=1}^{n} \log \left[w_2 \, g_2(x_i; \Theta_2) \right] & \text{if } \Delta_{i,2} = 1 \text{ and } \Delta_{i,k} = 0 \;\; \forall k \neq 2 \\ \vdots & \\ \sum_{i=1}^{n} \log \left[w_K \, g_K(x_i; \Theta_K) \right] & \text{if } \Delta_{i,K} = 1 \text{ and } \Delta_{i,k} = 0 \;\; \forall k \neq K \end{cases}$$

- The above expression can be restated as:

$$\ell(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \Delta_{i,k} \log \left(w_k g_k(x_i; \Theta_k) \right) \right].$$

- The $\Delta_{i,k}$ here is the incomplete (missing) datum because we do not know whether it is $\Delta_{i,k} = 0$ or $\Delta_{i,k} = 1$ for x_i and a specific k. Therefore, using the EM algorithm, we try to estimate it by its expectation.

EM for Mixture Distribution

- We found:

$$\ell(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \Delta_{i,k} \log \left(w_k g_k(x_i; \Theta_k) \right) \right].$$

- The E-step in EM:

$$Q(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \mathbb{E}[\Delta_{i,k} | X, \Theta_1, \ldots, \Theta_K] \times \log \left(w_k g_k(x_i; \Theta_k) \right) \right].$$

- The $\Delta_{i,k}$ is either 0 or 1; therefore:

$$\begin{aligned} \mathbb{E}[\Delta_{i,k} | X, \Theta_1, \ldots, \Theta_K] &= 0 \times \mathbb{P}(\Delta_{i,k} = 0 | X, \Theta_1, \ldots, \Theta_K) + 1 \times \mathbb{P}(\Delta_{i,k} = 1 | X, \Theta_1, \ldots, \Theta_K) \\ &= \mathbb{P}(\Delta_{i,k} = 1 | X, \Theta_1, \ldots, \Theta_K). \end{aligned}$$

- According to Bayes rule, we have:

$$\begin{aligned} \mathbb{P}(\Delta_{i,k} = 1 | X, \Theta_1, \ldots, \Theta_K) &= \frac{\mathbb{P}(X, \Theta_1, \ldots, \Theta_K, \Delta_{i,k} = 1)}{\mathbb{P}(X; \Theta_1, \ldots, \Theta_K)} \\ &= \frac{\mathbb{P}(X, \Theta_1, \ldots, \Theta_K | \Delta_{i,k} = 1) \, \mathbb{P}(\Delta_{i,k} = 1)}{\sum_{k'=1}^{K} \mathbb{P}(X, \Theta_1, \ldots, \Theta_K | \Delta_{i,k'} = 1) \mathbb{P}(\Delta_{i,k'} = 1)}. \end{aligned}$$

EM for Mixture Distribution

- In summary, we had:

$$Q(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \mathbb{E}[\Delta_{i,k} | X, \Theta_1, \ldots, \Theta_K] \times \log \left(w_k g_k(x_i; \Theta_k) \right) \right],$$

$$\mathbb{E}[\Delta_{i,k} | X, \Theta_1, \ldots, \Theta_K] = \mathbb{P}(\Delta_{i,k} = 1 | X, \Theta_1, \ldots, \Theta_K),$$

$$\mathbb{P}(\Delta_{i,k} = 1 | X, \Theta_1, \ldots, \Theta_K) = \frac{\mathbb{P}(X, \Theta_1, \ldots, \Theta_K | \Delta_{i,k} = 1)\, \mathbb{P}(\Delta_{i,k} = 1)}{\sum_{k'=1}^{K} \mathbb{P}(X, \Theta_1, \ldots, \Theta_K | \Delta_{i,k'} = 1) \mathbb{P}(\Delta_{i,k'} = 1)}$$

- The marginal probability in the denominator is:

$$\mathbb{P}(X; \Theta_1, \ldots, \Theta_K) = \sum_{k'=1}^{K} w_{k'}\, g_{k'}(x_i; \Theta_{k'}).$$

- Assuming that $\widehat{\gamma}_{i,k} := \mathbb{E}[\Delta_{i,k} | X, \Theta_1, \ldots, \Theta_K]$ (called responsibility of x_i), we have:

$$\widehat{\gamma}_{i,k} = \frac{\widehat{w}_k\, g_k(x_i; \Theta_k)}{\sum_{k'=1}^{K} \widehat{w}_{k'}\, g_{k'}(x_i; \Theta_{k'})}, \tag{11}$$

and

$$Q(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\gamma}_{i,k} \log \left(w_k g_k(x_i; \Theta_k) \right). \tag{12}$$

EM for Mixture Distribution

- We found:

$$Q(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\gamma}_{i,k} \log \left(w_k g_k(x_i; \Theta_k) \right). \quad (*)$$

- Some simplification of $Q(\Theta_1, \ldots, \Theta_K)$ will help in next step:

$$Q(\Theta_1, \ldots, \Theta_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\widehat{\gamma}_{i,k} \log w_k + \widehat{\gamma}_{i,k} \log g_k(x_i; \Theta_k) \right].$$

- The M-step in EM:

$$\widehat{\Theta}_k, \widehat{w}_k = \arg \max_{\Theta_k, w_k} Q(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K),$$

$$\text{subject to } \sum_{k=1}^{K} w_k = 1.$$

- Note that the function $Q(\Theta_1, \ldots, \Theta_K)$ is also a function of $w_1, \ldots, w_K$ and that is why we wrote it as $Q(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K)$.

EM for Mixture Distribution

- We had:

$$\begin{cases} \widehat{\Theta}_k, \widehat{w}_k = \arg\max\limits_{\Theta_k, w_k} Q(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K), \\ \text{subject to} \quad \sum\limits_{k=1}^{K} w_k = 1. \end{cases}$$

- The above problem is a constrained optimization problem:

$$\begin{aligned} &\underset{\Theta_k, w_k}{\text{maximize}} \quad Q(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K), \\ &\text{subject to} \quad \sum_{k=1}^{K} w_k = 1, \end{aligned}$$

which can be solved using Lagrange multiplier:

$$\begin{aligned} \mathcal{L}(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K, \alpha) &= Q(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K) - \alpha\Big(\sum_{k=1}^{K} w_k - 1\Big) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\widehat{\gamma}_{i,k} \log w_k + \widehat{\gamma}_{i,k} \log g_k(x_i; \Theta_k)\Big] - \alpha\Big(\sum_{k=1}^{K} w_k - 1\Big) \end{aligned}$$

EM for Mixture Distribution

- We had:

$$\mathcal{L}(\Theta_1, \ldots, \Theta_K, w_1, \ldots, w_K, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\widehat{\gamma}_{i,k} \log w_k + \widehat{\gamma}_{i,k} \log g_k(x_i; \Theta_k) \right] - \alpha \left(\sum_{k=1}^{K} w_k - 1 \right)$$

$$\frac{\partial \log(g(\theta))}{\partial \theta} = \frac{1}{g(\theta)} \frac{\partial g(\theta)}{\partial \theta}$$

$$\frac{\partial \mathcal{L}}{\partial \Theta_k} = \sum_{i=1}^{n} \frac{\widehat{\gamma}_{i,k}}{g_k(x_i; \Theta_k)} \frac{\partial g_k(x_i; \Theta_k)}{\partial \Theta_k} \overset{\text{set}}{=} 0 \tag{13}$$

$$\frac{\partial \mathcal{L}}{\partial w_k} = \sum_{i=1}^{n} \frac{\widehat{\gamma}_{i,k}}{w_k} - \alpha \overset{\text{set}}{=} 0 \implies w_k = \frac{1}{\alpha} \sum_{i=1}^{n} \gamma_{i,k}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{k=1}^{K} w_k - 1 \overset{\text{set}}{=} 0 \implies \sum_{k=1}^{K} w_k = 1$$

$$\therefore \quad \sum_{k=1}^{K} \frac{1}{\alpha} \sum_{i=1}^{n} \gamma_{i,k} = 1 \implies \alpha = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k}$$

$$\therefore \quad \widehat{w}_k = \frac{\sum_{i=1}^{n} \gamma_{i,k}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} \gamma_{i,k'}} \tag{14}$$

- Solving equations (13) and (14) gives us the estimations $\widehat{\Theta}_k$ and $\widehat{w}_k$ (for $k \in \{1, \ldots, K\}$) in every iteration.

Algorithm for Mixture Distribution

- We had:

$$\widehat{\gamma}_{i,k} = \frac{\widehat{w}_k \, g_k(x_i; \Theta_k)}{\sum_{k'=1}^{K} \widehat{w}_{k'} \, g_{k'}(x_i; \Theta_{k'})},$$

$$\frac{\partial \mathcal{L}}{\partial \Theta_k} = \sum_{i=1}^{n} \frac{\widehat{\gamma}_{i,k}}{g_k(x_i; \Theta_k)} \frac{\partial g_k(x_i; \Theta_k)}{\partial \Theta_k} \stackrel{\text{set}}{=} 0,$$

$$\widehat{w}_k = \frac{\sum_{i=1}^{n} \gamma_{i,k}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} \gamma_{i,k'}}.$$

START: Initialize $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$, $\widehat{w}_1, \ldots, \widehat{w}_K$
while *not converged* **do**
  // E-step in EM:
  for *i from 1 to n* **do**
    for *k from 1 to K* **do**
      $\widehat{\gamma}_{i,k} \leftarrow$ equation (11)
  // M-step in EM:
  for *k from 1 to K* **do**
    $\widehat{\Theta}_k \leftarrow$ equation (13)
    $\widehat{w}_k \leftarrow$ equation (14)
  // Check convergence:
  Compare $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$, and $\widehat{w}_1, \ldots, \widehat{w}_K$ with their values in previous iteration

Algorithm 2: Fitting A Mixture of Several Distributions

Gaussian Mixture Model (GMM)

Gaussian Mixture Model

- Here, we consider a mixture of K one-dimensional Gaussian distributions as an example for mixture of several continuous distributions. In this case, we have:

$$g_k(x;\mu_k,\sigma_k^2)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)=\phi\left(\frac{x-\mu_k}{\sigma_k}\right),\quad \forall k\in\{1,\dots,K\}$$

- Therefore, equation (10) becomes:

$$f(x;\mu_1,\dots,\mu_K,\sigma_1^2,\dots,\sigma_K^2)=\sum_{k=1}^{K} w_k\,\phi\left(\frac{x-\mu_k}{\sigma_k}\right). \tag{15}$$

- The equation (11) becomes:

$$\widehat{\gamma}_{i,k}=\frac{\widehat{w}_k\,\phi\left(\frac{x_i-\mu_k}{\sigma_k}\right)}{\sum_{k'=1}^{K}\widehat{w}_{k'}\,\phi\left(\frac{x_i-\mu_{k'}}{\sigma_{k'}}\right)}. \tag{16}$$

- The $Q(\mu_1,\dots,\mu_K,\sigma_1^2,\dots,\sigma_K^2)$ is:

$$\begin{aligned}&Q(\mu_1,\dots,\mu_K,\sigma_1^2,\dots,\sigma_K^2)\\&=\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\widehat{\gamma}_{i,k}\log w_k+\widehat{\gamma}_{i,k}\left(-\frac{1}{2}\log(2\pi)-\log\sigma_k-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right)\right].\end{aligned}$$

Gaussian Mixture Model

- We found:

$$Q(\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Big[\widehat{\gamma}_{i,k} \log w_k + \widehat{\gamma}_{i,k} \big(-\frac{1}{2} \log(2\pi) - \log \sigma_k - \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \big) \Big].$$

- The Lagrangian is:

$$\mathcal{L}(\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, w_1, \ldots, w_K, \alpha) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Big[\widehat{\gamma}_{i,k} \log w_k + \widehat{\gamma}_{i,k} \big(-\frac{1}{2} \log(2\pi) - \log \sigma_k - \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \big) \Big] - \alpha \big(\sum_{k=1}^{K} w_k - 1 \big).$$

Therefore:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{i=1}^{n} \Big[\widehat{\gamma}_{i,k} \big(\frac{x_i - \mu_k}{\sigma_k^2} \big) \Big] \overset{\text{set}}{=} 0 \implies \widehat{\mu}_k = \frac{\sum_{i=1}^{n} \widehat{\gamma}_{i,k} \, x_i}{\sum_{i=1}^{n} \widehat{\gamma}_{i,k}}, \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_k} = \sum_{i=1}^{n} \Big[\widehat{\gamma}_{i,k} \big(-\frac{1}{\sigma_k} + \frac{(x_i - \mu_k)^2}{\sigma_k^3} \big) \Big] \overset{\text{set}}{=} 0 \implies \widehat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \widehat{\gamma}_{i,k} \, (x_i - \widehat{\mu}_k)^2}{\sum_{i=1}^{n} \widehat{\gamma}_{i,k}}, \tag{18}$$

and $\widehat{w}_k$ is the same as equation (14).

Algorithm for Mixture Distribution

- We had:

$$\widehat{\gamma}_{i,k} = \frac{\widehat{w}_k \, \phi\left(\frac{x_i - \mu_k}{\sigma_k}\right)}{\sum_{k'=1}^{K} \widehat{w}_{k'} \, \phi\left(\frac{x_i - \mu_{k'}}{\sigma_{k'}}\right)}, \quad \widehat{w}_k = \frac{\sum_{i=1}^{n} \gamma_{i,k}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} \gamma_{i,k'}}.$$

The Θ patameters:

$$\widehat{\mu}_k = \frac{\sum_{i=1}^{n} \widehat{\gamma}_{i,k} \, x_i}{\sum_{i=1}^{n} \widehat{\gamma}_{i,k}}, \quad \widehat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \widehat{\gamma}_{i,k} \, (x_i - \widehat{\mu}_k)^2}{\sum_{i=1}^{n} \widehat{\gamma}_{i,k}}.$$

START: Initialize $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K, \widehat{w}_1, \ldots, \widehat{w}_K$
while *not converged* **do**
  // E-step in EM:
  for i *from* 1 *to* n **do**
    for k *from* 1 *to* K **do**
      $\widehat{\gamma}_{i,k} \leftarrow$ equation (11)
  // M-step in EM:
  for k *from* 1 *to* K **do**
    $\widehat{\Theta}_k \leftarrow$ equation (13)
    $\widehat{w}_k \leftarrow$ equation (14)
  // Check convergence:
  Compare $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$, and $\widehat{w}_1, \ldots, \widehat{w}_K$ with their values in previous iteration

Algorithm 2: Fitting A Mixture of Several Distributions

Gaussian Mixture Model: Simulation

Gaussian Mixture Model: Simulation

- A sample with size $n = 2200$ from three distributions is randomly generated for this experiment:

$$\phi\left(\frac{x-\mu_1}{\sigma_1}\right) = \phi\left(\frac{x+10}{1.2}\right),$$
$$\phi\left(\frac{x-\mu_2}{\sigma_2}\right) = \phi\left(\frac{x-0}{2}\right),$$
$$\phi\left(\frac{x-\mu_3}{\sigma_3}\right) = \phi\left(\frac{x-5}{5}\right).$$

- For having generality, the size of subset of sample generated from the three densities are different, i.e., 700, 1000, and 500.
- Applying the EM algorithm and using equations (16), (17), (18), and (14) for mixture of $K = 3$ Gaussians gives us the estimated values for the parameters:

$$\begin{cases} \mu_1 = -9.99, & \sigma_1 = 1.17, & w_1 = 0.317 \\ \mu_2 = -0.05, & \sigma_2 = 1.93, & w_2 = 0.445 \\ \mu_3 = 4.64, & \sigma_3 = 4.86, & w_3 = 0.237 \end{cases}$$

- Comparing the estimations for μ_1, μ_2, μ_3 and $\sigma_1, \sigma_2, \sigma_3$ with those in original densities from which data were generated verifies the correctness of the estimations.

Gaussian Mixture Model: Simulation

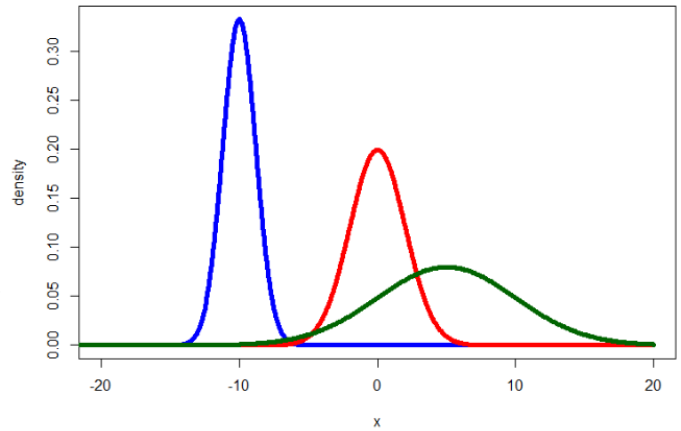

Figure 1. The original probability density functions from which the sample is drawn. The mixture includes three different Gaussians showed in blue, red, and green colors.

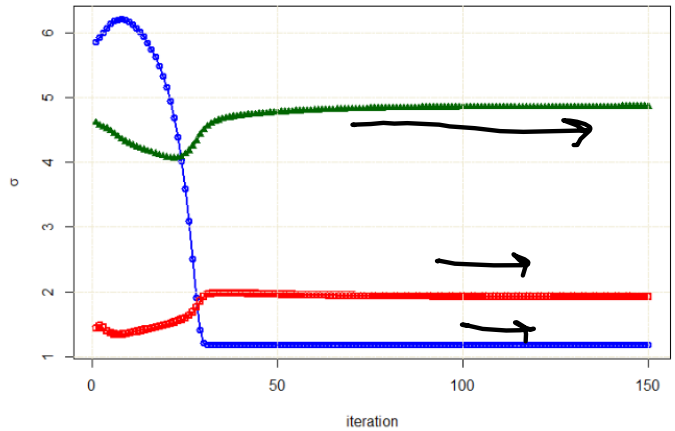

Figure 3. The change and convergence of σ_1 (shown in blue), σ_2 (shown in red), and σ_3 (shown in green) over the iterations.

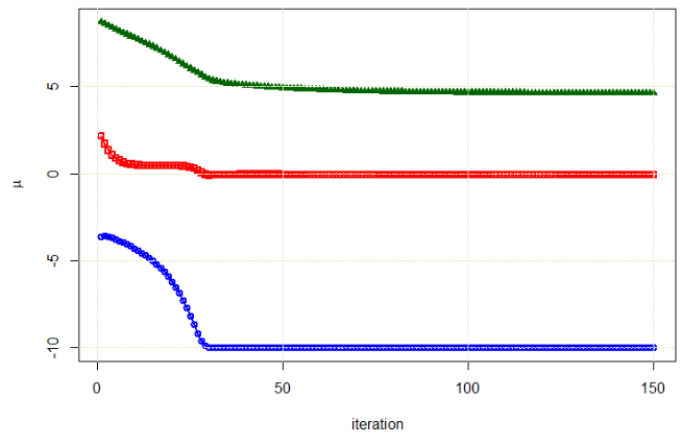

Figure 2. The change and convergence of μ_1 (shown in blue), μ_2 (shown in red), and μ_3 (shown in green) over the iterations.

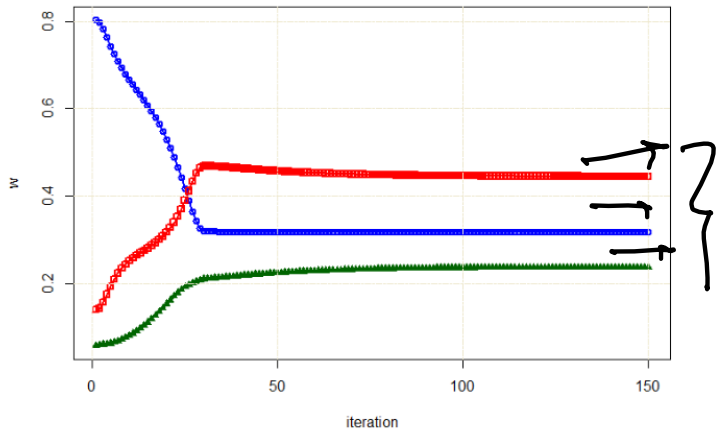

Figure 4. The change and convergence of w_1 (shown in blue), w_2 (shown in red), and w_3 (shown in green) over the iterations.

Gaussian Mixture Model: Simulation

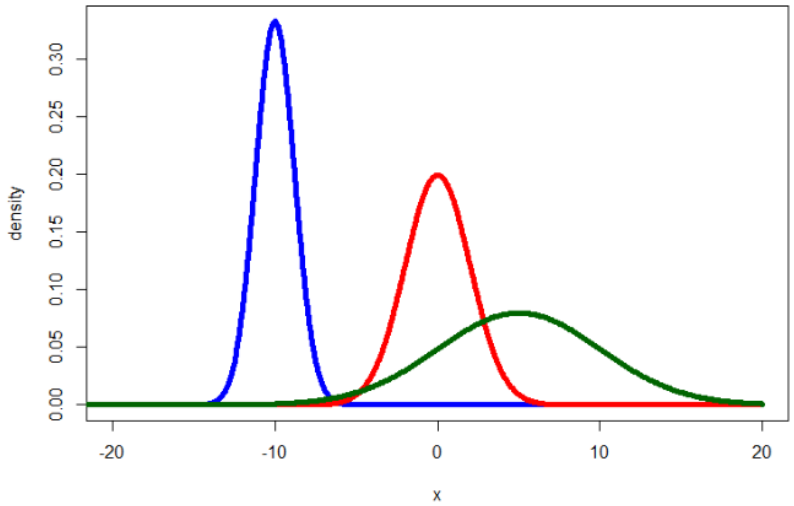

Figure 1. The original probability density functions from which the sample is drawn. The mixture includes three different Gaussians showed in blue, red, and green colors.

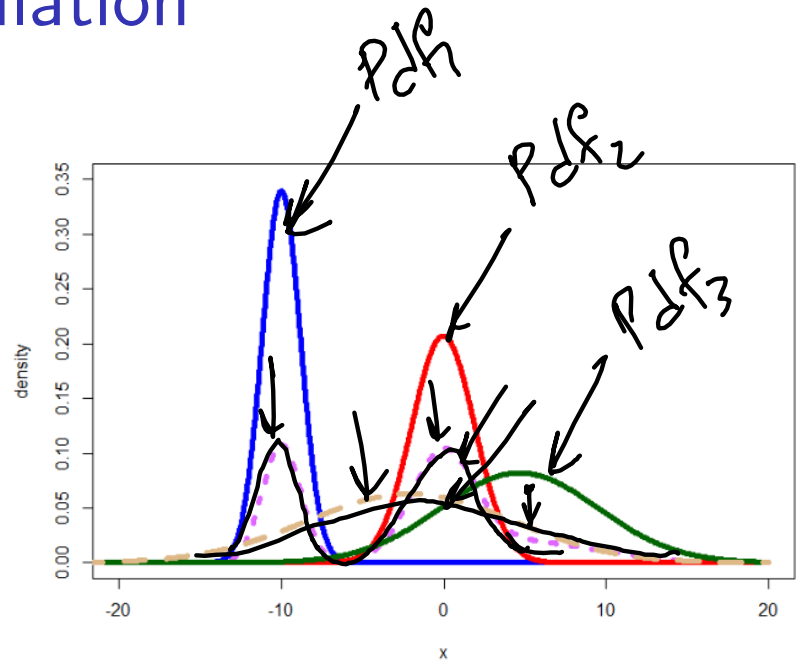

Figure 5. The estimated probability density functions. The estimated mixture includes three different Gaussians showed in blue, red, and green colors. The dashed purple density is the weighted summation of these three densities, i.e., $\sum_{k=1}^{3} w_k \phi(\frac{x-\mu_k}{\sigma_k})$. The dashed brown density is the fitted density whose parameters are estimated by MLE.

The code of this simulation in my GitHub page (in R language):
`https://github.com/bghojogh/Fitting-Mixture-Distribution`

Acknowledgment

- Some slides are based on our tutorial paper: "Fitting a mixture distribution to data: tutorial" [1]
- See our tutorial [1] for **mixture of discrete distributions**, such as mixture of **Poisson** distributions.
- Some slides of this slide deck are inspired by teachings of Prof. Mu Zhu at University of Waterloo, Department of Statistics.
- For more information on mixture distributions in machine learning, see the book: Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, Jerome H. Friedman. "The elements of statistical learning: data mining, inference, and prediction". Vol. 2. New York: springer, 2009 [2].
- The code of fitting mixture distribution in my GitHub page (in R language): `https://github.com/bghojogh/Fitting-Mixture-Distribution`
- Mixture distribution in sklearn: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html`

References

[1] B. Ghojogh, A. Ghojogh, M. Crowley, and F. Karray, "Fitting a mixture distribution to data: tutorial," *arXiv preprint arXiv:1901.06708*, 2019.

[2] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2.
Springer, 2009.