Recap

- We have been considering estimation of density functions given *iid* samples.

Recap

- We have been considering estimation of density functions given *iid* samples.
- We have studied maximum likelihood estimation and Bayesian estimation of density functions.

Recap

- We have been considering estimation of density functions given *iid* samples.
- We have studied maximum likelihood estimation and Bayesian estimation of density functions.
- Given estimated densities, we can implement Bayes classifier.

Recap

- We have been considering estimation of density functions given *iid* samples.
- We have studied maximum likelihood estimation and Bayesian estimation of density functions.
- Given estimated densities, we can implement Bayes classifier.
- We have also discussed the exponential family of densities and role of sufficient statistics in estimation.

Mixture densities

- The last topic we consider under parametric estimation is that of mixture densities.

Mixture densities

- The last topic we consider under parametric estimation is that of mixture densities.
- In many cases we may not be able to capture the class conditional density using any standard density model.

Mixture densities

- The last topic we consider under parametric estimation is that of mixture densities.
- In many cases we may not be able to capture the class conditional density using any standard density model.
- In such cases, often, modelling the class conditional density as a mixture of densities is helpful.

Mixture densities

- The last topic we consider under parametric estimation is that of mixture densities.
- In many cases we may not be able to capture the class conditional density using any standard density model.
- In such cases, often, modelling the class conditional density as a mixture of densities is helpful.
- We look at this and a special technique, called the EM algorithm, for ML estimation of mixture densities in this class.

Mixture density model

- Consider a density model

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x), \quad \lambda_k \geq 0, \text{ and } \sum_{k=1}^{K} \lambda_k = 1$$

where each f_k is a density function.

Mixture density model

- Consider a density model

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x), \quad \lambda_k \geq 0, \text{ and } \sum_{k=1}^{K} \lambda_k = 1$$

  where each f_k is a density function.
- Since each f_k is a density, given the conditions on λ_k, f is a convex combination of densities and hence is itself a density.

Mixture density model

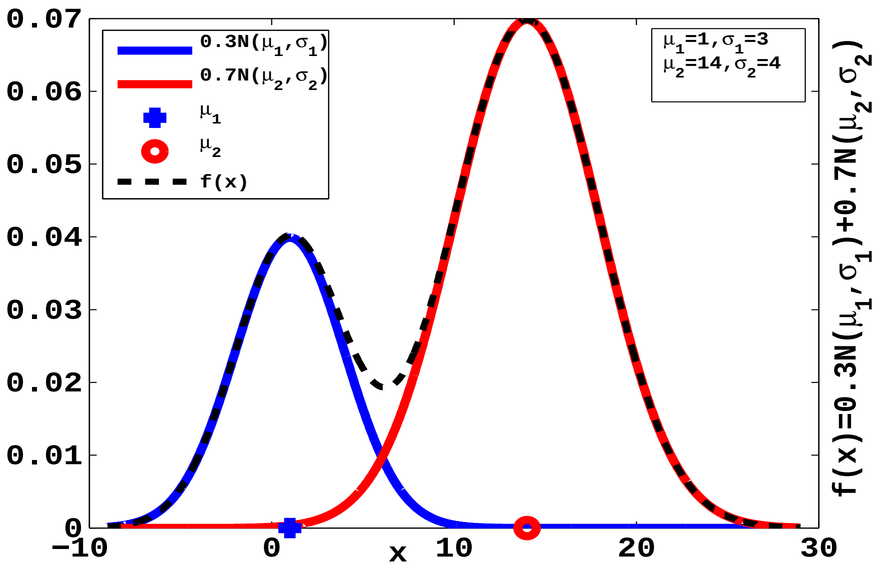
- Consider a density model

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x), \quad \lambda_k \geq 0, \text{ and } \sum_{k=1}^{K} \lambda_k = 1$$

  where each f_k is a density function.
- Since each f_k is a density, given the conditions on λ_k, f is a convex combination of densities and hence is itself a density.
- Mixture densities are useful when data distribution is multimodal.

- Most standard densities are unimodal.

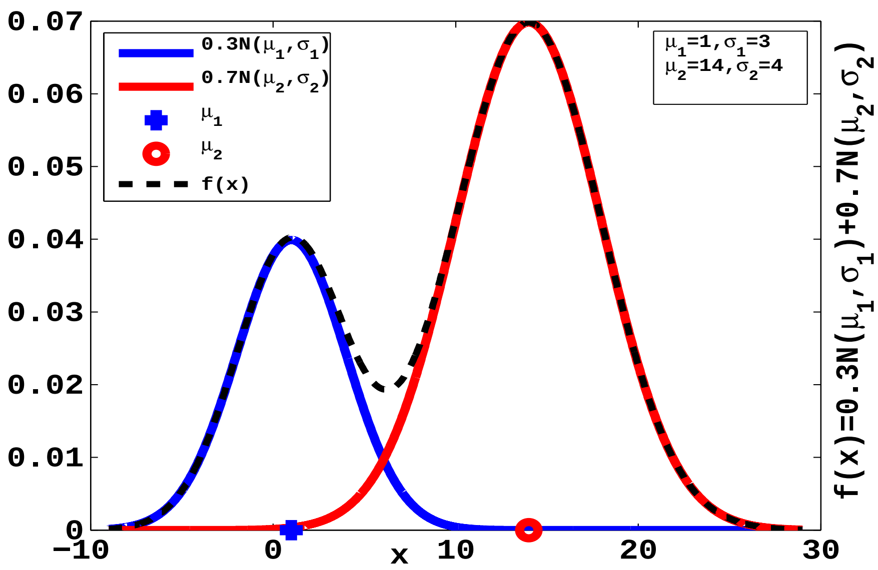
- Most standard densities are unimodal.
- For example, consider the normal density.



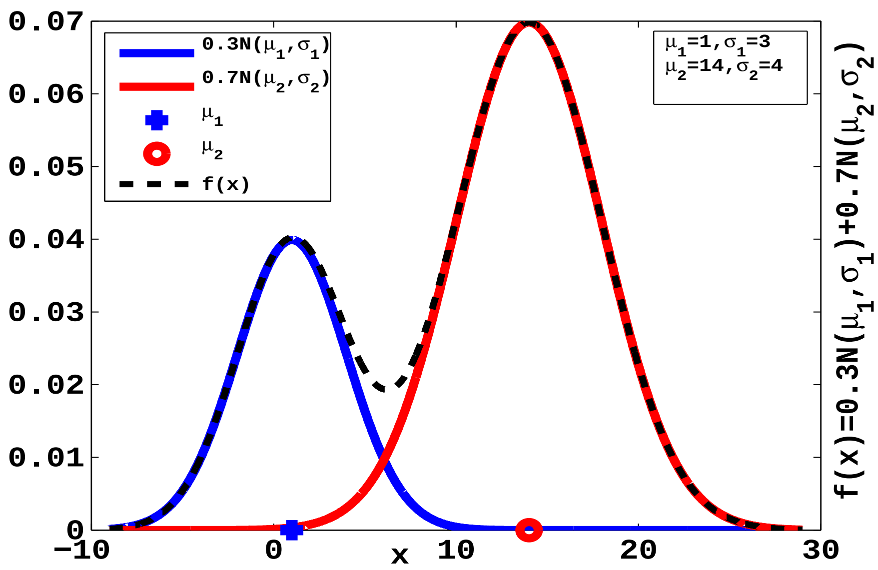
This is unimodal.

- Now let us consider a mixture of two normal densities

- Now let us consider a mixture of two normal densities

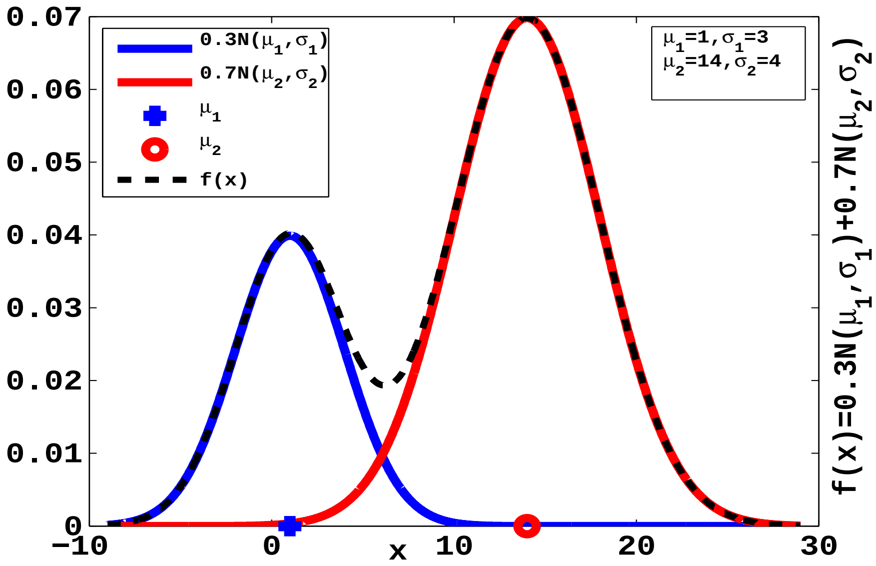


- Now let us consider a mixture of two normal densities



- This is a multimodal density

- Now let us consider a mixture of two normal densities



- This is a multimodal density
- When data density is multi-modal, we can often approximate it with mixture of gaussians.

ML estimation of mixture models

- Consider a mixture of normal densities

$$f(x \mid \theta) = \sum_{k=1}^{K} \lambda_k \, f_k(x)$$

where each f_k is $\mathcal{N}(\mu_k, \, \Sigma_k)$.

ML estimation of mixture models

- Consider a mixture of normal densities

$$f(x \mid \theta) = \sum_{k=1}^{K} \lambda_k \, f_k(x)$$

  where each f_k is $\mathcal{N}(\mu_k,\ \Sigma_k)$.

- The parameter vector, θ, consists of all λ_k, which are called mixing coefficients, and all the parameters of the constituent densities, namely, $\mu_k,\ \Sigma_k$, $k = 1, \cdots, K$.

- Let $\mathcal{D} = \{x_1, \cdots, x_n\}$ be a sample of n *iid* data from this density.

- Let $\mathcal{D} = \{x_1, \cdots, x_n\}$ be a sample of n *iid* data from this density.
- Then the likelihood function is

$$L(\theta \mid \mathcal{D}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \lambda_k \, f_k(x_i) \right]$$

- The log likelihood is given by

$$l(\theta \mid \mathcal{D}) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \lambda_k \, f_k(x_i) \right]$$

- The log likelihood is given by

$$l(\theta \mid \mathcal{D}) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} \lambda_k \, f_k(x_i)\right]$$

- Since there is a sum inside the log function, the densities f_k being from exponential family, does not simplify log likelihood.

- The log likelihood is given by

$$l(\theta \mid \mathcal{D}) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \lambda_k \, f_k(x_i) \right]$$

- Since there is a sum inside the log function, the densities f_k being from exponential family, does not simplify log likelihood.
- Maximizing log likelihood could become a difficult optimization problem.

Mixture of two one dimensional densities

- Consider one dimensional case with $K = 2$. Let, for $j = 1, 2$,

$$\phi(x \,|\, \theta_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(- \frac{(x - \mu_j)^2}{2\sigma_j^2} \right), \quad \theta_j = (\mu_j, \sigma_j)$$

Mixture of two one dimensional densities

- Consider one dimensional case with $K = 2$. Let, for $j = 1, 2$,

$$\phi(x \,|\, \theta_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(- \frac{(x - \mu_j)^2}{2\sigma_j^2} \right), \quad \theta_j = (\mu_j, \sigma_j)$$

- The density model is

$$f(x \mid \theta) = \lambda_1 \, \phi(x \mid \theta_1) + \lambda_2 \, \phi(x \mid \theta_2)$$

where $\theta = (\theta_1, \ \theta_2, \ \lambda_1, \ \lambda_2)$

- The log likelihood is

$$l(\mathcal{D} \mid \theta) = \sum_{i=1}^{n} \ln(\lambda_1\, \phi(x_i \mid \theta_1) + \lambda_2\, \phi(x_i \mid \theta_2))$$

- The log likelihood is

$$l(\mathcal{D} \mid \theta) = \sum_{i=1}^{n} \ln(\lambda_1 \, \phi(x_i \mid \theta_1) + \lambda_2 \, \phi(x_i \mid \theta_2))$$

- We need to maximize this with respect to θ.

- The log likelihood is

$$l(\mathcal{D} \mid \theta) = \sum_{i=1}^{n} \ln(\lambda_1 \, \phi(x_i \mid \theta_1) + \lambda_2 \, \phi(x_i \mid \theta_2))$$

- We need to maximize this with respect to θ.
- Let us calculate the partial derivatives of l.

- The log likelihood is

$$l(\mathcal{D} \mid \theta) = \sum_{i=1}^{n} \ln(\lambda_1\, \phi(x_i \mid \theta_1) + \lambda_2\, \phi(x_i \mid \theta_2))$$

- We need to maximize this with respect to θ.
- Let us calculate the partial derivatives of l.
- First note that

$$\frac{\partial\, \phi(x \mid \theta_j)}{\partial \mu_s} = \frac{\partial\, \phi(x \mid \theta_j)}{\partial \sigma_s} = 0, \quad \text{if} \;\; j \neq s.$$

By differentiation we get, for $j = 1, 2$,

$$\frac{\partial\, \phi(x \mid \theta_j)}{\partial \mu_j} = \phi(x \mid \theta_j)\, \frac{(x - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial\, \phi(x \mid \theta_j)}{\partial \sigma_j} = \phi(x \mid \theta_j) \left[\frac{(x - \mu_j)^2}{\sigma_j^3} - \frac{1}{\sigma_j} \right]$$

By differentiation we get, for $j = 1, 2$,

$$
\begin{aligned}
\frac{\partial\, \phi(x \mid \theta_j)}{\partial \mu_j} &= \phi(x \mid \theta_j)\, \frac{(x - \mu_j)}{\sigma_j^2} \\
\frac{\partial\, \phi(x \mid \theta_j)}{\partial \sigma_j} &= \phi(x \mid \theta_j)\, \left[\frac{(x - \mu_j)^2}{\sigma_j^3} - \frac{1}{\sigma_j} \right]
\end{aligned}
$$

Now we have

$$
\frac{\partial\, l(\mathcal{D} \mid \theta)}{\partial \mu_j} = \sum_{i=1}^{n} \frac{\lambda_j\, \phi(x_i \mid \theta_j)\, \frac{(x_i - \mu_j)}{\sigma_j^2}}{\lambda_1\, \phi(x_i \mid \theta_1) \; + \; \lambda_2 \phi(x_i \mid \theta_2)}
$$

- Define γ_{ij}, $i = 1, \cdots, n$, $j = 1, 2$,

$$\gamma_{ij} = \frac{\lambda_j \, \phi(x_i \mid \theta_j)}{\lambda_1 \, \phi(x_i \mid \theta_1) \; + \; \lambda_2 \phi(x_i \mid \theta_2)}$$

- Define γ_{ij}, $i = 1, \cdots, n$, $j = 1, 2$,

$$\gamma_{ij} = \frac{\lambda_j \, \phi(x_i \mid \theta_j)}{\lambda_1 \, \phi(x_i \mid \theta_1) \; + \; \lambda_2 \phi(x_i \mid \theta_2)}$$

- Then we get

$$\frac{\partial \, l(\mathcal{D} \mid \theta)}{\partial \mu_j} \; = \; \sum_{i=1}^{n} \gamma_{ij} \, \frac{(x_i - \mu_j)}{\sigma_j^2}$$

- Define γ_{ij}, $i = 1, \cdots, n$, $j = 1, 2$,

$$\gamma_{ij} = \frac{\lambda_j \, \phi(x_i \mid \theta_j)}{\lambda_1 \, \phi(x_i \mid \theta_1) \; + \; \lambda_2 \phi(x_i \mid \theta_2)}$$

- Then we get

$$\frac{\partial \, l(\mathcal{D} \mid \theta)}{\partial \mu_j} \; = \; \sum_{i=1}^{n} \gamma_{ij} \, \frac{(x_i - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial \, l(\mathcal{D} \mid \theta)}{\partial \sigma_j} \; = \; \sum_{i=1}^{n} \gamma_{ij} \, \left[\frac{(x_i - \mu_j)^2}{\sigma_j^3} \; - \; \frac{1}{\sigma_j} \right]$$

- Hence the ML estimates satisfy, for $j = 1, 2$,

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij} \, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- Hence the ML estimates satisfy, for $j = 1, 2$,

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- First, we like to note that these are not really estimates. The RHS in the above equations depends on the unknown parameter values.

- Hence the ML estimates satisfy, for $j = 1, 2$,

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij} \, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- First, we like to note that these are not really estimates. The RHS in the above equations depends on the unknown parameter values.
- However, there is an interesting structure here.

$$
\begin{aligned}
\hat{\mu}_j &= \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}} \\
\hat{\sigma}_j^2 &= \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}
\end{aligned}
$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- These are similar to the ‘sample mean estimates’.

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- These are similar to the ‘sample mean estimates’.
- It is a sample mean with ‘weight’ γ_{ij} for x_i.
  γ_{ij} are sometimes called responsibility coefficients.

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- These are similar to the ‘sample mean estimates’.
- It is a sample mean with ‘weight’ γ_{ij} for x_i.
  γ_{ij} are sometimes called responsibility coefficients.
- If there is only one component in the mixture, these become the usual ML estimates.

- Let us also find maximizers of log likelihood with respect to λ_j.

- Let us also find maximizers of log likelihood with respect to λ_j.
- Since we have a constraint $\lambda_1 + \lambda_2 = 1$, this is a constrained optimization.

- Let us also find maximizers of log likelihood with respect to λ_j.
- Since we have a constraint $\lambda_1 + \lambda_2 = 1$, this is a constrained optimization.
- So, we need to equate to zero, the partial derivatives of

$$l(\mathcal{D} \mid \theta) + \eta(\lambda_1 + \lambda_2 - 1)$$

where η is the Lagrange multiplier.

- Let us also find maximizers of log likelihood with respect to λ_j.
- Since we have a constraint $\lambda_1 + \lambda_2 = 1$, this is a constrained optimization.
- So, we need to equate to zero, the partial derivatives of
$$l(\mathcal{D} \mid \theta) + \eta(\lambda_1 + \lambda_2 - 1)$$
where η is the Lagrange multiplier.
- By equating to zero the partial derivative of the above with respect to λ_1, we get

$$\sum_{i=1}^{n} \frac{\phi(x_i \mid \theta_1)}{\lambda_1 \, \phi(x_i \mid \theta_1) + \lambda_2 \phi(x_i \mid \theta_2)} + \eta = 0$$

or

$$\sum_{i=1}^{n} \frac{\gamma_{i1}}{\lambda_1} + \eta = 0$$

$$\sum_{i=1}^{n} \frac{\phi(x_i \mid \theta_1)}{\lambda_1\, \phi(x_i \mid \theta_1) + \lambda_2 \phi(x_i \mid \theta_2)} + \eta = 0$$

or

$$\sum_{i=1}^{n} \frac{\gamma_{i1}}{\lambda_1} + \eta = 0$$

- we get a similar equation for derivetive w.r.t. λ_2.

$$\sum_{i=1}^{n} \frac{\phi(x_i \mid \theta_1)}{\lambda_1 \, \phi(x_i \mid \theta_1) + \lambda_2 \phi(x_i \mid \theta_2)} + \eta = 0$$

or

$$\sum_{i=1}^{n} \frac{\gamma_{i1}}{\lambda_1} + \eta = 0$$

- we get a similar equation for derivetive w.r.t. λ_2.
- Now, using $\lambda_1 + \lambda_2 = 1$, we get

$$\eta = \eta(\lambda_1 + \lambda_2) = -\sum_{i=1}^{n} (\gamma_{i1} + \gamma_{i2}) = -n$$

- Hence, the ML estimates for λ_j satisfy

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

- Hence, the ML estimates for λ_j satisfy

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

- Putting all these together we get

- The ML estimates for $\mu_j, \sigma_j, \lambda_j$, $j = 1, 2$, satisfy

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}, \quad \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- The ML estimates for $\mu_j, \sigma_j, \lambda_j$, $j = 1, 2$, satisfy

$$\hat{\mu}_j = \frac{\sum_{i=1}^n \gamma_{ij} \, x_i}{\sum_{i=1}^n \gamma_{ij}}, \quad \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^n \gamma_{ij}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \gamma_{ij} \, (x_i - \mu_j)^2}{\sum_{i=1}^n \gamma_{ij}}$$

- The structure of equations is interesting.

- The ML estimates for $\mu_j, \sigma_j, \lambda_j$, $j = 1, 2$, satisfy

$$\hat{\mu}_j \;=\; \frac{\sum_{i=1}^{n} \gamma_{ij}\, x_i}{\sum_{i=1}^{n} \gamma_{ij}}, \quad \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

$$\hat{\sigma}_j^2 \;=\; \frac{\sum_{i=1}^{n} \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

- The structure of equations is interesting.
- These are not expressions for estimates.

- The ML estimates for $\mu_j, \sigma_j, \lambda_j$, $j = 1, 2$, satisfy

$$
\begin{aligned}
\hat{\mu}_j &= \frac{\sum_{i=1}^n \gamma_{ij}\, x_i}{\sum_{i=1}^n \gamma_{ij}}, \quad \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^n \gamma_{ij} \\
\hat{\sigma}_j^2 &= \frac{\sum_{i=1}^n \gamma_{ij}\, (x_i - \mu_j)^2}{\sum_{i=1}^n \gamma_{ij}}
\end{aligned}
$$

- The structure of equations is interesting.
- These are not expressions for estimates.
- However, we can solve for estimates using, e.g., Gauss-Siedel iteration.

$$
\begin{aligned}
\mu_j^{(k+1)} &= \frac{\sum_{i=1}^n \gamma_{ij}^{(k)} \; x_i}{\sum_{i=1}^n \gamma_{ij}^{(k)}}, \quad \lambda_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}^{(k)} \\
(\sigma_j^2)^{(k+1)} &= \frac{\sum_{i=1}^n \gamma_{ij}^{(k)} \, (x_i - \mu_j^{(k)})^2}{\sum_{i=1}^n \gamma_{ij}^{(k)}} \\
\gamma_{ij}^{(k+1)} &= \frac{\lambda_j^{(k+1)} \, \phi(x_i \mid \theta_j^{(k+1)})}{\sum_{j=1}^2 \lambda_j^{(k+1)} \, \phi(x_i \mid \theta_j^{(k+1)})}
\end{aligned}
$$

- It is easy to generalize this to mixture of K Gaussians.

- What we have done so far is a special case of general procedure.

- What we have done so far is a special case of general procedure.
- In many cases the ML estimation of mixture of densities gives rise to such an iterative optimization procedure.

- What we have done so far is a special case of general procedure.
- In many cases the ML estimation of mixture of densities gives rise to such an iterative optimization procedure.
- We now look at this general procedure.

- Our density model was

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

(while we stick to 2-component mixture, it is easily generalized to K components).

- Our density model was

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

(while we stick to 2-component mixture, it is easily generalized to K components).

- In our sample each x_i is drawn *iid* according to this distribution.

density model: $f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$

$$\text{density model:} \quad f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

- To generate x_i, we first choose a component density, with probabilities λ_j, and then generate it from the corresponding $\phi(x \mid \theta_j)$.

density model: $f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$

- To generate x_i, we first choose a component density, with probabilities λ_j, and then generate it from the corresponding $\phi(x \mid \theta_j)$.
- If we knew which x_i are generated from which component density, then the estimation of all parameters is very easy.

density model: $f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$

- To generate x_i, we first choose a component density, with probabilities λ_j, and then generate it from the corresponding $\phi(x \mid \theta_j)$.
- If we knew which x_i are generated from which component density, then the estimation of all parameters is very easy.
- Let us first formalize this notion.

Missing Information

- Let random variables Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, denote the information of which component density each sample comes from.

Missing Information

- Let random variables Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, denote the information of which component density each sample comes from.
- For each i, $Z_{ij} = 1$ if x_i came from j^{th} component density.

Missing Information

- Let random variables Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, denote the information of which component density each sample comes from.
- For each i, $Z_{ij} = 1$ if x_i came from j^{th} component density.
- We would have $\sum_j Z_{ij} = 1, \ \forall i$.

Missing Information

- Let random variables Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, denote the information of which component density each sample comes from.
- For each i, $Z_{ij} = 1$ if x_i came from j^{th} component density.
- We would have $\sum_j Z_{ij} = 1, \; \forall i$.
- Also, we have

$$P[Z_{ij} = 1] = \lambda_j, \; \forall i; \quad \text{and} \quad f(x_i \,|\, Z_{ij} = 1) = \phi(x_i \,|\, \theta_j)$$

Missing Information

- Let random variables Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, denote the information of which component density each sample comes from.
- For each i, $Z_{ij} = 1$ if x_i came from j^{th} component density.
- We would have $\sum_j Z_{ij} = 1, \ \forall i$.
- Also, we have

$$P[Z_{ij} = 1] = \lambda_j, \ \forall i; \quad \text{and} \quad f(x_i \,|\, Z_{ij} = 1) = \phi(x_i \,|\, \theta_j)$$

We can think of Z_{ij} as the 'missing information'.

- Let Z_i denote the vector with components Z_{ij}.

- Let Z_i denote the vector with components Z_{ij}.
- Denote $\mathcal{D}^c = \{(x_1, Z_1), \cdots, (x_n, Z_n)\}$.

- Let Z_i denote the vector with components Z_{ij}.
- Denote $\mathcal{D}^c = \{(x_1, Z_1), \cdots, (x_n, Z_n)\}$.
- Our data consists of only x_i. But suppose the sample data was $\mathcal{D}^c$.

- Let Z_i denote the vector with components Z_{ij}.
- Denote $\mathcal{D}^c = \{(x_1, Z_1), \cdots, (x_n, Z_n)\}$.
- Our data consists of only x_i. But suppose the sample data was $\mathcal{D}^c$.
- Then estimation is easy. For example,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} Z_{i1}\, x_i}{\sum_{i=1}^{n} Z_{i1}}, \quad \hat{\mu}_2 = \frac{\sum_{i=1}^{n} Z_{i2}\, x_i}{\sum_{i=1}^{n} Z_{i2}}$$

- Let Z_i denote the vector with components Z_{ij}.
- Denote $\mathcal{D}^c = \{(x_1, Z_1), \cdots, (x_n, Z_n)\}$.
- Our data consists of only x_i. But suppose the sample data was $\mathcal{D}^c$.
- Then estimation is easy. For example,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n Z_{i1}\, x_i}{\sum_{i=1}^n Z_{i1}}, \quad \hat{\mu}_2 = \frac{\sum_{i=1}^n Z_{i2}\, x_i}{\sum_{i=1}^n Z_{i2}}$$

- These are very similar to earlier equations.

Complete and incomplete data

- The general situation is as follows.

Complete and incomplete data

- The general situation is as follows.
- The data that we have is ‘incomplete’

Complete and incomplete data

- The general situation is as follows.
- The data that we have is ‘incomplete’
- This is because of some ‘hidden’ or ‘missing’ data.

Complete and incomplete data

- The general situation is as follows.
- The data that we have is ‘incomplete’
- This is because of some ‘hidden’ or ‘missing’ data.
- If we are given the complete data then ML estimation is easy.

Complete and incomplete data

- The general situation is as follows.
- The data that we have is ‘incomplete’
- This is because of some ‘hidden’ or ‘missing’ data.
- If we are given the complete data then ML estimation is easy.
- In our example, x_i is the incomplete data.

Complete and incomplete data

- The general situation is as follows.
- The data that we have is ‘incomplete’
- This is because of some ‘hidden’ or ‘missing’ data.
- If we are given the complete data then ML estimation is easy.
- In our example, x_i is the incomplete data.
- (x_i, Z_i) constitutes the complete data and Z_i constitute the missing or hidden data/variables.

The EM Algorithm

- The EM algorithm is an efficient iterative procedure for ML estimation in such situations.

The EM Algorithm

- The EM algorithm is an efficient iterative procedure for ML estimation in such situations.
- The algorithm basically has two steps: ‘Expectation’ and ‘Maximization’

The EM Algorithm

- The EM algorithm is an efficient iterative procedure for ML estimation in such situations.
- The algorithm basically has two steps: ‘Expectation’ and ‘Maximization’
- Hence the name of the algorithm.

The EM Algorithm

- The EM algorithm is an efficient iterative procedure for ML estimation in such situations.
- The algorithm basically has two steps: ‘Expectation’ and ‘Maximization’
- Hence the name of the algorithm.
- As per our notation, x_i, $i = 1, \cdots, n$ is the incomplete data and (x_i, Z_i), $i = 1, \cdots, n$ is the complete data.

- Let $f(x, Z \mid \theta)$ be the density for the complete data.

- Let $f(x, Z \mid \theta)$ be the density for the complete data. That is, the complete data is n iid samples from this density model.

- Let $f(x, Z \mid \theta)$ be the density for the complete data. That is, the complete data is n iid samples from this density model.
- Thus, the complete data log likelihood is

$$l(\theta \mid \mathcal{D}^c) = \ln \left(\prod_{i=1}^{n} f(x_i, Z_i \mid \theta) \right)$$

- Let $f(x, Z \mid \theta)$ be the density for the complete data. That is, the complete data is n iid samples from this density model.
- Thus, the complete data log likelihood is

$$l(\theta \mid \mathcal{D}^c) = \ln \left(\prod_{i=1}^{n} f(x_i, Z_i \mid \theta) \right)$$

- As earlier, we would also denote $\mathcal{D}^c$ by $(\mathbf{x}, \mathbf{Z})$.

- Let $f(x, Z \mid \theta)$ be the density for the complete data. That is, the complete data is n iid samples from this density model.
- Thus, the complete data log likelihood is

$$l(\theta \mid \mathcal{D}^c) = \ln\left(\prod_{i=1}^{n} f(x_i, Z_i \mid \theta)\right)$$

- As earlier, we would also denote $\mathcal{D}^c$ by $(\mathbf{x}, \mathbf{Z})$.
- Hence the complete data loglikelihood is also denoted by $\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta))$.

- The two steps of EM algorithm are as follows:

- The two steps of EM algorithm are as follows:

E-step : Compute $Q(\theta,\ \theta^{(k)})$ which is expectation of the complete data loglikelihood w.r.t. the conditional distribution of hidden variables conditioned on incomplete data and current value of θ as $\theta^{(k)}$.

- The two steps of EM algorithm are as follows:

E-step : Compute $Q(\theta,\ \theta^{(k)})$ which is expectation of the complete data loglikelihood w.r.t. the conditional distribution of hidden variables conditioned on incomplete data and current value of θ as $\theta^{(k)}$.

$$Q(\theta,\ \theta^{(k)}) = E_{\mathbf{Z}|\mathbf{x},\theta^{(k)}}\ \ln(f(\mathbf{x}, \mathbf{Z} \mid \theta))$$

- The two steps of EM algorithm are as follows:

E-step : Compute $Q(\theta,\ \theta^{(k)})$ which is expectation of the complete data loglikelihood w.r.t. the conditional distribution of hidden variables conditioned on incomplete data and current value of θ as $\theta^{(k)}$.

$$Q(\theta,\ \theta^{(k)}) = E_{\mathbf{Z}|\mathbf{x},\theta^{(k)}}\ \ln(f(\mathbf{x}, \mathbf{Z} \mid \theta))$$

M-step : Compute next value of θ as $\theta^{(k+1)}$ by maximizing $Q(\theta,\ \theta^{(k)})$ over θ.

- The two steps of EM algorithm are as follows:

E-step : Compute $Q(\theta,\ \theta^{(k)})$ which is expectation of the complete data loglikelihood w.r.t. the conditional distribution of hidden variables conditioned on incomplete data and current value of θ as $\theta^{(k)}$.

$$Q(\theta,\ \theta^{(k)}) = E_{\mathbf{Z}|\mathbf{x},\theta^{(k)}}\ \ln(f(\mathbf{x},\mathbf{Z} \mid \theta))$$

M-step : Compute next value of θ as $\theta^{(k+1)}$ by maximizing $Q(\theta,\ \theta^{(k)})$ over θ.

$$\theta^{(k+1)} = \arg\max_{\theta}\ Q(\theta,\ \theta^{(k)})$$

Example of EM

- Let us consider the example of estimating a two component Gaussian density.

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

Example of EM

- Let us consider the example of estimating a two component Gaussian density.

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

- The x_i, $i = 1, \cdots, n$, is the given data which is the incomplete data here.

Example of EM

- Let us consider the example of estimating a two component Gaussian density.

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

- The x_i, $i = 1, \cdots, n$, is the given data which is the incomplete data here.
- The Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, that we defined earlier are the hidden variables or the missing data.

Example of EM

- Let us consider the example of estimating a two component Gaussian density.

$$f(x \mid \theta) = \sum_{j=1}^{2} \lambda_j \, \phi(x \mid \theta_j)$$

- The x_i, $i = 1, \cdots, n$, is the given data which is the incomplete data here.
- The Z_{ij}, $i = 1, \cdots, n$, $j = 1, 2$, that we defined earlier are the hidden variables or the missing data.
- Recall that Z_{ij} is the indicator whether or not x_i came from the j^{th} component of the mixture.

- By definition of Z_{ij}, we have

$$P[Z_{ij} = 1] = \lambda_j, \ \forall i; \quad \text{and} \quad f(x_i \mid Z_{ij} = 1) = \phi(x_i \mid \theta_j)$$

- By definition of Z_{ij}, we have

$$P[Z_{ij} = 1] = \lambda_j, \ \forall i; \quad \text{and} \quad f(x_i \mid Z_{ij} = 1) = \phi(x_i \mid \theta_j)$$

- Recall $Z_i = (Z_{i1}, \ Z_{i2})$.

- By definition of Z_{ij}, we have

$$P[Z_{ij} = 1] = \lambda_j, \ \forall i; \quad \text{and} \quad f(x_i \mid Z_{ij} = 1) = \phi(x_i \mid \theta_j)$$

- Recall $Z_i = (Z_{i1}, \ Z_{i2})$. Hence

$$f(Z_i|\theta) = \prod_{j=1}^{2} (\lambda_j)^{Z_{ij}}, \quad \text{and} \quad f(x_i|Z_i, \theta) = \prod_{j=1}^{2} (\phi(x_i|\theta_j))^{Z_{ij}}$$

- Hence density of complete data is

$$f(x_i,\ Z_i \mid \theta) = \prod_{j=1}^{2} (\lambda_j \ \phi(x_i \mid \theta_j))^{Z_{ij}}$$

- Hence density of complete data is

$$f(x_i,\ Z_i \mid \theta) = \prod_{j=1}^{2} (\lambda_j\ \phi(x_i \mid \theta_j))^{Z_{ij}}$$

- Thus complete data likelihood is

$$f(\mathbf{x},\ \mathbf{Z} \mid \theta) = \prod_{i=1}^{n} \left[\prod_{j=1}^{2} (\lambda_j\ \phi(x_i \mid \theta_j))^{Z_{ij}} \right]$$

- The complete data log likelihood is

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

- The complete data log likelihood is

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

- Note that we now have ‘sum of log’ rather than ‘log of sum’

- The complete data log likelihood is

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

- Note that we now have ‘sum of log’ rather than ‘log of sum’
- It is easy to see how knowledge of the ‘hidden’ variables makes the ML estimation easy.

Example: E-step

- For the E-step, we have to take expectation of $\mathbf{Z}$ w.r.t. distribution conditioned on $\mathbf{x}$ at a given value of θ.

Example: E-step

- For the E-step, we have to take expectation of $\mathbf{Z}$ w.r.t. distribution conditioned on $\mathbf{x}$ at a given value of θ.
- We have, for any θ',

$$E[Z_{ij} \mid \mathbf{x}, \theta'] \;=\; P[Z_{ij} = 1 \mid \mathbf{x},\ \theta'] = P[Z_{ij} = 1 \mid x_i,\ \theta']$$

Example: E-step

- For the E-step, we have to take expectation of $\mathbf{Z}$ w.r.t. distribution conditioned on $\mathbf{x}$ at a given value of θ.
- We have, for any θ',

$$
\begin{aligned}
E[Z_{ij} \mid \mathbf{x}, \theta'] &= P[Z_{ij} = 1 \mid \mathbf{x},\ \theta'] = P[Z_{ij} = 1 \mid x_i,\ \theta'] \\
&= \frac{f(x_i \mid Z_{ij} = 1,\ \theta')\ P[Z_{ij} = 1]}{\sum_{j=1}^{2} f(x_i \mid Z_{ij} = 1,\ \theta')\ P[Z_{ij} = 1]}
\end{aligned}
$$

Example: E-step

- For the E-step, we have to take expectation of $\mathbf{Z}$ w.r.t. distribution conditioned on $\mathbf{x}$ at a given value of θ.
- We have, for any θ',

$$\begin{aligned} E[Z_{ij} \mid \mathbf{x}, \theta'] &= P[Z_{ij} = 1 \mid \mathbf{x},\ \theta'] = P[Z_{ij} = 1 \mid x_i,\ \theta'] \\ &= \frac{f(x_i \mid Z_{ij} = 1,\ \theta')\ P[Z_{ij} = 1]}{\sum_{j=1}^{2}\ f(x_i \mid Z_{ij} = 1,\ \theta')\ P[Z_{ij} = 1]} \\ &= \frac{\lambda_j\ \phi(x_i \mid \theta'_j)}{\sum_{j=1}^{2}\ \lambda_j\ \phi(x_i \mid \theta'_j)} \end{aligned}$$

- Thus, $E[Z_{ij} \mid \mathbf{x}, \theta'] = \gamma_{ij}(\theta')$ where

$$\gamma_{ij}(\theta') = \frac{\lambda_j \, \phi(x_i \mid \theta'_j)}{\sum_{j=1}^{2} \lambda_j \, \phi(x_i \mid \theta'_j)}$$

- Thus, $E[Z_{ij} \mid \mathbf{x}, \theta'] = \gamma_{ij}(\theta')$ where

$$\gamma_{ij}(\theta') = \frac{\lambda_j \, \phi(x_i \mid \theta'_j)}{\sum_{j=1}^{2} \lambda_j \, \phi(x_i \mid \theta'_j)}$$

- This is the same γ_{ij} that we defined earlier.

- Thus, $E[Z_{ij} \mid \mathbf{x}, \theta'] = \gamma_{ij}(\theta')$ where

$$\gamma_{ij}(\theta') = \frac{\lambda_j \ \phi(x_i \mid \theta'_j)}{\sum_{j=1}^{2} \ \lambda_j \ \phi(x_i \mid \theta'_j)}$$

- This is the same γ_{ij} that we defined earlier.
- This notation emphasizes the fact that the value of γ_{ij} depends on the parameter vector.

- Thus, $E[Z_{ij} \mid \mathbf{x}, \theta'] = \gamma_{ij}(\theta')$ where

$$\gamma_{ij}(\theta') = \frac{\lambda_j \, \phi(x_i \mid \theta'_j)}{\sum_{j=1}^{2} \lambda_j \, \phi(x_i \mid \theta'_j)}$$

- This is the same γ_{ij} that we defined earlier.
- This notation emphasizes the fact that the value of γ_{ij} depends on the parameter vector.
- Now we need to do this expectation on the complete data log likelihood which is

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

- Thus, under the E-step, we get

$$Q(\theta,\ \theta^{(k)}) \ = \ \sum_{i=1}^{n} \left[\sum_{j=1}^{2} E[Z_{ij} \mid \mathbf{x},\ \theta^{(k)}] \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

$$\ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) = \sum_{i=1}^{n} \left[\sum_{j=1}^{2} Z_{ij} \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

- Thus, under the E-step, we get

$$\begin{aligned} Q(\theta,\ \theta^{(k)}) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} E[Z_{ij} \mid \mathbf{x},\ \theta^{(k)}] \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right] \\ &= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)}) \ \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right] \end{aligned}$$

Example: the M-step

- In the M-step, we find $\theta^{(k+1)}$ that maximizes (over θ),

Example: the M-step

- In the M-step, we find $\theta^{(k+1)}$ that maximizes (over θ),

$$Q(\theta,\ \theta^{(k)}) \ = \ \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)}) \ \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

Example: the M-step

- In the M-step, we find $\theta^{(k+1)}$ that maximizes (over θ),

$$
\begin{aligned}
Q(\theta,\ \theta^{(k)}) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)})\ \ln(\lambda_j\ \phi(x_i \mid \theta_j)) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)}) \,[\ln(\lambda_j)\ -\ \ln(\sigma_j \sqrt{2\pi}) \\
&\qquad\qquad -\ \frac{(x_i - \mu_j)^2}{2\sigma_j^2}]
\end{aligned}
$$

Example: the M-step

- In the M-step, we find $\theta^{(k+1)}$ that maximizes (over θ),

$$Q(\theta,\ \theta^{(k)}) \ = \ \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)}) \ \ln(\lambda_j \, \phi(x_i \mid \theta_j)) \right]$$

$$= \ \sum_{i=1}^{n} \sum_{j=1}^{2} \gamma_{ij}(\theta^{(k)}) \, [\ln(\lambda_j) \ - \ \ln(\sigma_j \sqrt{2\pi}) \ - \ \frac{(x_i - \mu_j)^2}{2\sigma_j^2}]$$

- This is now a simple optimization problem.

- For example, $\frac{\partial Q}{\partial \mu_1} = 0$ gives us

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^k) \, \frac{(x_i - \mu_1)}{\sigma_1^2} = 0$$

- For example, $\frac{\partial Q}{\partial \mu_1} = 0$ gives us

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^k) \, \frac{(x_i - \mu_1)}{\sigma_1^2} = 0$$

- Hence we get

$$\mu_1^{k+1} = \frac{\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \, x_i}{\sum_{i=1}^{n} \gamma_{i1}(\theta^k)}$$

- For example, $\frac{\partial Q}{\partial \mu_1} = 0$ gives us

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^k) \, \frac{(x_i - \mu_1)}{\sigma_1^2} = 0$$

- Hence we get

$$\mu_1^{k+1} = \frac{\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \, x_i}{\sum_{i=1}^{n} \gamma_{i1}(\theta^k)}$$

- This is same as the iterative algorithm we derived earlier.

- Similarly, $\frac{\partial Q}{\partial \sigma_1} = 0$ gives

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \left[- \frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3}\right] = 0$$

- Similarly, $\frac{\partial Q}{\partial \sigma_1} = 0$ gives

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \; [- \frac{1}{\sigma_1} \; + \; \frac{(x_i - \mu_1)^2}{\sigma_1^3}] = 0$$

- Hence we get

$$(\sigma_1^2)^{(k+1)} = \frac{\sum_{i=1}^{n} \; \gamma_{i1}(\theta^{(k)}) \; (x_i - \mu_1^{(k)})^2}{\sum_{i=1}^{n} \; \gamma_{i1}(\theta^{k})}$$

- Similarly, $\frac{\partial Q}{\partial \sigma_1} = 0$ gives

$$\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \, [- \frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3}] = 0$$

- Hence we get

$$(\sigma_1^2)^{(k+1)} = \frac{\sum_{i=1}^{n} \gamma_{i1}(\theta^{(k)}) \, (x_i - \mu_1^{(k)})^2}{\sum_{i=1}^{n} \gamma_{i1}(\theta^{k})}$$

- Once again same as earlier algorithm.

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(k)} x_i}{\sum_{i=1}^n \gamma_{ij}^{(k)}}, \quad \lambda_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \gamma_{ij}^{(k)}$$

$$(\sigma_j^2)^{(k+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(k)} (x_i - \mu_j^{(k)})^2}{\sum_{i=1}^n \gamma_{ij}^{(k)}}$$

$$\gamma_{ij}^{(k+1)} = \frac{\lambda_j^{(k+1)} \phi(x_i \mid \theta_j^{(k+1)})}{\sum_{j=1}^2 \lambda_j^{(k+1)} \phi(x_i \mid \theta_j^{(k+1)})} = \gamma_{ij}(\theta^{(k+1)})$$

- So, this is actually the EM algorithm.