MTH210

Likelihood Based Estimation

Likelihood function

Suppose $X_1, ..., X_n$ is a random sample from a given sitribution with density $f(x|\theta)$, for $\theta \in \Theta$. After obtained the real data, from F, we want to estimate θ and assess the quality of this estimator. One useful method is the *maximum likelihood estimation(MLE)*. Let $\mathbb{X} = (X_1, ..., X_n)$

$$L(\theta|\tilde{X} = \tilde{x}) = f(\tilde{x}|\theta) = f(x_1, ..., x_n|\theta)$$

> 💡 Note that $L(\theta|\tilde{x})$ is not a distribution over θ, it is just a function, that quantifies how likely a value of θ is.

Maximum Likelihood Estimation

$$\hat{\theta}_{MLE} = arg\ max_{\theta \in \Theta} L(\theta|\tilde{x})$$

It is the"most likely" value of θ habing observed the data. $\hat{\theta}_{MLE}$ is the maximum likelihood estimator of θ

Definitions:

Concave Function(1D) : a function $h(x)$ is concave if $h''(x) \leq 0$ for all x.

Concave Function : a function $h(\tilde{x})$ is concave if the hessian matrix $\nabla^2 h(\tilde{x})$, is **negative semi definite** for all $\tilde{x}$. That is, if all eigenvalues of the Hessian are non- positive or $\tilde{a}^T(\nabla^2 h(\tilde{x}))\tilde{a} \leq 0,\ \forall \tilde{a}$.

$$\nabla_{\tilde{x}} = \begin{bmatrix} d/dx_1 \\ d/dx_2 \\ . \\ . \\ . \\ d/dx_n \end{bmatrix} \ \& \ \nabla^2 = \begin{bmatrix} d^2/dx_1^2 & . & . & . & d^2/(dx_1)(dx_n) \\ . & d^2/dx_2^2 & . & . & d^2/(dx_2)(dx_n) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & \frac{d^2}{dx_n^2} \end{bmatrix}$$

> 💡 Fo n iid observations from bernoulli(p) $\tilde{X}$: the estimator $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Regression

Let $Y_1, ..., Y_n$ be observations known as response. Let $x_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$ be the ith corresponding vector of <u>covariates</u> for the ith observation. Let $\beta \in \mathbb{R}^p$ be the *regression coefficient* so that for $\sigma^2 > 0$, $Y_i = x_i^T\beta + \epsilon_i$, where ϵ_i ~ $\mathcal{N}(0, \sigma^2)$. Define $\tilde{X} := (x_1^T\ x_2^T\ , \ldots, x_n^T)^T$. Now,

$$\tilde{Y} = \begin{bmatrix} y_1 \\ \cdot \\ y_i \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix} = \tilde{X}\tilde{\beta} + \epsilon \sim \mathcal{N}_n(X\beta, \sigma^2\mathbb{I}_n)$$

This model is built to estimate β, which measures the linear effect of X on Y

> 💡 Review of some matrix-vector differentiation. For $\tilde{x}, \tilde{a} \in \mathbb{R}^p, \text{and } p \times p \text{ matrix } A:$
>
> -
> $\nabla\tilde{x}^T\tilde{a} = \nabla\tilde{a}^T\tilde{x} = \tilde{a}$
>
> -
> $\nabla x^TAx = (A + A^T)x$ ⇒ If A is symmetric ⇒ $\nabla x^TAx = 2Ax$

MLE for Linear Regression

Let us understand the linear relationship b/w X and β. Then the equation is as follows

$L(\beta, \sigma^2|y) = \prod_{i=1}^{n} f(y_i|\tilde{X}, \beta, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}\}$

Thereby obtain the log likelihood function, and note that $(y - X\beta)^T(y - X\beta) = (y^T - \beta^TX^T)(y - X\beta) = y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta$., upon setting $dl/d\beta$ & $dl/d\sigma^2 = 0$, we obtain

$$\hat{\beta}_{MLE} = (X^TX)^{-1}X^Ty$$

$$\hat{\sigma^2}_{MLE} = \frac{(y - X\hat{\beta}_{MLE})^T(y - X\hat{\beta}_{MLE})}{n}$$

$$\begin{aligned} L(\beta, \sigma^2|y) &= \prod_{i=1}^{n} f(y_i|X, \beta, \sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2}\frac{(y - X\beta)^T(y - X\beta)}{\sigma^2}\right\} \\ \Rightarrow l(\beta, \sigma^2) := \log L(\beta, \sigma^2|y) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(y - X\beta)^T(y - X\beta)}{\sigma^2} \end{aligned}$$

Note that

$$\begin{aligned} (y - X\beta)^T(y - X\beta) &= (y^T - \beta^TX^T)(y - X\beta) \\ &= y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta \\ &= y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta\,. \end{aligned}$$

Using this we have (recall your multivariable calculus courses)

$$\begin{aligned} \frac{dl}{d\beta} &= -\frac{1}{2\sigma^2}\left[-2X^Ty + 2X^TX\beta\right] = \frac{X^Ty - X^TX\beta}{2\sigma^2} \overset{set}{=} 0 \\ \frac{dl}{d\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{(y - X\beta)^T(y - X\beta)}{2\sigma^4} \overset{set}{=} 0\,. \end{aligned}$$

The first equation leads to $\hat{\beta}_{\text{MLE}}$ satisfying

$$X^Ty - X^TX\hat{\beta}_{\text{MLE}} = 0 \Rightarrow \hat{\beta}_{\text{MLE}} = (X^TX)^{-1}X^Ty\,,$$

if $(X^TX)^{-1}$ exists. And $\hat{\sigma}^2_{MLE}$ is

$$\hat{\sigma}^2_{MLE} = \frac{(y - X\hat{\beta}_{\text{MLE}})^T(y - X\hat{\beta}_{\text{MLE}})}{n}\,.$$

<u>**Verify:** that the Hessian matrix is negative definite, and thus the objective function is concave.</u>

> 💡 When does $(X^TX)^{-1}$ does not exist ?
> If
> $p > n$ i.e. **The number of observations is less than the number of parameters**, since X is $n \times p$, so X^TX is $p \times p$ of rank $n < p$. So X^TX is not full rank and thus cannot be inverted. In this case MLE does not exist and other estimator required. *PENALIZED Regression*

!!!Verification of negative definite hessian

Penalized Regression

Note that for the linear regression setup : $\hat{\beta}_{MLE} = arg\ min_{\beta}(y - X\beta)^T(y - X\beta)$. Suppose that (X^TX) is not-invertible(p > n), then we dont know how to estimate β. In such cases, we may use penalized likelihood, that penalizes the coefficients β, so that some of the βs are "pushed towards zero", or essentially making them unimportant, thereby removing singularity from X^TX. Thus instead of looking at the likelihood, we consider the penalized likelihood. Thus the final penalized(negative) log-likelihood function is :

$$Q(\beta) = -logL(\beta|y)\ +\ P(\beta), \text{where } P(\beta) \text{ is penalization function}$$

> 💡 Note that :
> 1. We want to minimize the negative log-likelihood i.e.
> $Q(\beta)$
> 2.
> **The penalization function assigns large values for large** β, thus the opimization problem favors small values of β.

Example : Ridge Regression

Ridge penalization term is $P(\beta) = \lambda\beta^T\beta/2$, for some $\lambda > 0$. (λ will be user-chosen, will study in cross validation)

$$\hat{\beta}_{\text{ridge}} = arg\ min_{\beta}\{\frac{(y - X\beta)^T(y - X\beta)}{2} + \frac{\lambda}{2}\beta^T\beta\} = (X^TX + \lambda\mathbb{I}_p)^{-1}X^Ty.$$

!!! Verify that the hessian matrix is positive definite, the final ridge solution always exists even if X^TX is not invertible.

> 💡 Note that $(X^TX + \lambda I_p)$ is always positive definite for $\lambda > 0$, since for $a \in \mathbb{R}^p \neq 0$, $a^T(X^TX + \lambda I_p)a = a^TX^TXa + \lambda a^Ta > 0$

No Closed-Form MLEs

Example - Gamma Distribution

$l(\alpha) := logL(\alpha|x) = -n\ log(\Gamma(\alpha)) + (\alpha - 1)\sum log\ x_i - \sum x_i \Rightarrow \frac{dl(\alpha)}{\alpha} = -n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum log\ x_i = set = 0$
Solving the above analytically is not possible : look at the double derivative
$= -n\frac{d^2}{d\alpha^2}log(\Gamma(\alpha))$ i.e. concave function.

> 💡 $\frac{d^2}{d\alpha^2}log(\Gamma(\alpha))$ is the *polygamma function of order 1*, which is always **positive.**
>
> $\frac{d^{m-1}}{d\alpha^{m-1}}log(\Gamma(\alpha))$ is the polygamma function of order 0.

Thus the need of OPTIMIZATION METHODS :

Numerical Optimization Methods

$f(\theta)$ be an *objective function*. We want to solve the maximization, $\theta^* = arg\ max_\theta f(\theta) = arg\ max_\theta e^{f(\theta)} = arg\ max_\theta log(f(\theta)) = arg\ max_\theta [f(\theta) + 100]$.

We will generate a sequence of $\{\theta_{(k)}\}$ such that the goal is for $\{\theta_{(k)}\} \to \theta^*$ in a deterministic manner (non - random convergence)

If the objective function is concave, then all methods will guarentee a global maxima !

Taylor - Series Approximation

for a univariate function $f(\theta)$, its Taylor series representation around a point θ_0 is

$$f(\theta) = f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + f''(\theta_0)\frac{(\theta - \theta_0)^2}{2!} + ...$$

Linear approximation only requires the first two terms in the taylor-series approximation. Thus $f(\theta) = [f'(\theta_0)]\theta + f(\theta_0) - f'(\theta_0)\theta_0$, which describes a line with intercept b and slope m. Note that $(\theta_0, f(\theta_0))$ lies on the line.

Quadratic Approximation requires only the first 3 terms, and upon solving $f(\theta)$ can be expressed as a quadratic curve with parameter θ, and this curve also passes through $(\theta_0, f(\theta_0))$ ⇒ concavity and convexity can be checked by checking the coefficient of θ^2.

Taylor's Approximation in higher dimensions

- Linear Approximation :

$$f(\theta) = f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0)$$

- Quadratic Approximation

$$f(\theta) \approx f(\theta_0) + (\theta - \theta_0)^T \nabla f(\theta_0) + (\theta - \theta_0)^T \nabla^2 f(\theta_0)(\theta - \theta_0) =: \tilde{f}_Q(\theta)$$

Newton - Raphson's Method

$\theta^* = arg\ max_\theta f(\theta)$, Consider the quadratic taylor series approximation around $\theta_{(k)}$ and let this quadratic curve has notation $\tilde{f}_Q(\theta)$. Now the Newton-Raphson algorithm finds the optima of the curve. Take a derivative w.r.t θ, and set it to zero. The optima occurs at $\theta_{(k)} - \frac{f'(\theta_{(k)})}{f''(\theta_{(k)})}$. Thus using iterations

$$\theta_{(k+1)} = \theta_{(k)} - \frac{f'(\theta_{(k)})}{f''(\theta_{(k)})}$$

Algorithm :
1.
Choose starting value of $\theta_{(0)}$ and tolerance ϵ
2. For any k find
$\theta_{(k+1)} = \theta_{(k)} - \frac{f'(\theta_{(k)})}{f''(\theta_{(k)})}$
3. If
$|f'(\theta_{(k+1)})| < \epsilon$, then Return $\theta_{(k+1)}$ and stop
4. Else continue to step 2

if the objective function is concave N-R method converges to global maxima, otherwise it converges to a local optima or diverges

Example (Gamma Distribution continued)

$$l(\alpha) := -nlog(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} logx_i \; - \sum x_i$$

Set a reasonable α_0. Then iterate $\alpha_{(k+1)} = \alpha_{(k)} - \frac{f'(\alpha_{(k)})}{f''(\alpha_{(k)})}$, where f'' and f' are calculated in the previous section + concave function. For a good starting α_0, we know that the mean of $\text{Gamma}(\alpha, 1)$ is α, thus given a realisation $\tilde{X}$, a good starting value of $\alpha_0 = n^{-1}\sum_{i=1}^{n} X_i$.

WHATS THE DIFFERENCE BETWEEN SOLVING IT ANALYTICALLY AND USING OPTIMIZATION METHODS :
- Note that We are unable to solve for
α^* by setting $f'(x)$ to zero, as it includes functions such as first derivative of Γ function, which is not possible to solve
- However in Optimization methods, we only want to insert values of parameter in the function and then apply the sequence of equations to get the optimal value of
α^*.

Example (Cauchy) - Notes

Find log-likelihood. Show that the $l''(\mu)$ can be positive or negative. So it is not a concave function ⇒ the N-R method does not guarentee to converge to the global maxima. Also $\mu_0 = \text{Median}(X_i)$

▼ Solution -Location Cauchy distribution with mode at $\mu \in \mathbb{R}$

$f(x|\mu) = \frac{1}{\pi}\frac{1}{(1+(x-\mu)^2)}$,

$l(\mu) := log(L(\mu|X) = -nlog\pi - \sum_{i=1}^{n} log(1 + (X_i - \mu)^2)$

$l'(\mu) = 2\sum_{i=1}^{n} \frac{X_i - \mu}{1+(X_i-\mu)^2}$, $l''(\mu) = 2\sum_{i=1}^{n}[2\frac{(X_i-\mu)^2}{[1+(X_i-\mu)^2]^2} - \frac{1}{1+(X_i-\mu)^2}]$

⇒ not a concave function, thus NR method does not guarentee a global maxima, and may als

NR - Higher Dimensions

Always check that $\nabla^2 f$ is negative definite, *to know if there is a unique maximum.*

$$\theta_{(k+1)} = \theta_{(k)} - [\nabla^2 f(\theta_{(k)})]^{-1}\nabla f(\theta_{(k)})$$

Algorithm :

1. Choose starting value $\theta_{(0)}$ and tolerance ϵ
2. For any k find
$\theta_{(k+1)} = \theta_{(k)} - [\nabla^2 f(\theta_{(k)})]^{-1} \nabla f(\theta_{(k)})$.
3. if
$||f'(\theta_{(k+1)}|| < \epsilon$ then return $\theta_{(k+1)}$ and stop
4. else return to step 2

Example - Ridge Regression

Note that $Q(\beta) = \frac{(y-X\beta)^T(y-X\beta)}{2} + \frac{\lambda}{2}\beta^T\beta$ is quadratic, and the NR method approximation of the objective function to a quadratic approximation(which will be the same quadratic in this case), and upon following the iteration formulaes of the NR method, we reach the same conclusion as earlier which is $\beta_{(k+1)} = (X^TX + \lambda I_p)^{-1}X^Ty$

7.3 Gradient Ascent (Descent)

Consider the objective function $f(\theta)$ that you want to maximize and suppose $\theta*$ is the true maximum.

Taylor's series approximation at a fixed θ_0, is $f(\theta) \approx f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + f''(\theta_0)(\theta - \theta_0)^2/2$

If $f''(\theta)$ is unavailable use, fouble derivative = $-1/t$, for some $t > 0$, i.e. assume concavity and quadratic.

Maximize the taylor series approximation w.r.t θ, differentiate and set to zero to obtain $\theta_{(k+1)} = \theta_{(k)} + tf'(\theta_{(k)})$

The iteration can be stopped when difference between consecutive thetas $< \epsilon$

Note

For concave functions, there exists a t such that gradient ascent converges to the global maxima. In general (when the function is not concave), there exists a t such that gradient ascent converges to a local maxima, as long as you don't start from a local minima.

Example- Location Cauchy Distribution

$l(\mu) := logL(\mu|X) = -nlog\pi - \sum_{i=1}^{n} log(1 + (X_i - \mu)^2)$

Algorithm :

1.

Choose t(say 0.3), and Set $\mu_0 = Median(X_i)$, since the mean of cauchy does not exist

2. Determine

$\mu_{(k+1)} = \mu_{(k)} + (0.3)(2\sum_{i=1}^{n} \frac{X_i - \mu}{1+(X_i-\mu)^2})$

3. Stop when

$|l'(\mu_{(k+1)}| < \epsilon$

Higher Dimensions

$\theta_{k+1} = \theta_k + t\nabla f(\theta_k)$, example of Logistic Regression in Notes

7.4 MM (Minorize/Maximize)Algorithm

Consider Obtaining a solution to $\theta* = argmax_\theta f(\theta)$

Consdier a minorizing function $\tilde{f}(\theta|\theta_k)$ such that

- $f(\theta_k) = g(\theta_k|\theta_k)$

- $f(\theta) \geq g(\theta|\theta_k)$ for all other θ

$$\theta_{(k+1)} = arg\ max_{\theta} g(\theta|\theta_{(k)})$$

This algorithm has the ascent property in that every update increases the objective value, as $f(\theta_{k+1}) \geq g(\theta_{k+1}|\theta_k) \geq g(\theta_k|\theta_k) = f(\theta_k)$

How to obtain such minorizing functions- One common way

Use the remainder form of taylor' series expansion

$$f(\theta) = f(\theta_{(k)}) + f'(\theta_{(k)})(\theta - \theta_k) + \frac{1}{2} f''(z)(\theta - \theta_k)^2$$

where z is some constant between θ_k and θ(MVT)

If we can lower bound $f''(z) > L$, then

$$g(\theta|\theta_k) = f(\theta_k) + f'(\theta_k)(\theta - \theta_k) + \frac{1}{2} L(\theta - \theta_k)^2$$

Clearly $f(\theta) > g(\theta|\theta_k), \forall \theta$

Iterates are $\theta_{(k+1)} = \theta_{(k)} - \frac{f'(\theta_{(k)})}{L}$

Example - Location Cauchy

Notes

8 EM Algorithm

8.1 The Expectation-Maximization Algorithm

Suppose we have a vector of parameters θ, and we have only observed the data $(X_1, \ldots, X_n)$ and some part of data was not observed say $Z_1, \ldots Z_m$. For $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{z} = (z_1, .., z_m)$

- f(x|θ) denotes the marginal distribution of the observed incomplete data
- f(x, z|θ) denotes the joint distribution of the unobserved complete data

Objective is to maximize $l(\theta|\mathbf{x}) := log f(\mathbf{x}|\theta) = log \int f(\mathbf{x}, \mathbf{z}|\theta) d\nu_z$, where $\int .d\nu_z$ denotes integral based on whether Z is continuous or discrete

Consider a starting value θ_0. Then for any (k + 1) iteration

1. E-Step : Compute the Expectation of the complete expected likelihood:

$$q(\theta|\theta_{(k)}) = E_{Z|X}[log f(\mathbf{x}, \mathbf{z}|\theta)|\mathbf{X} = x, \theta_{(k)}] = \int log f(x, z|\theta) f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz$$

the expectation is computed w.r.t to the conditional distribution of Z given X = x for the current iterate $\theta_{(k)}$.

2. M-Step : Computer Maximization iteration $\theta_{(k+1)} = arg\ max_{\theta \in \Theta} q(\theta|\theta_{(k)})$
3. Stop when absolute difference < ϵ

8.2 EM Algorithm for Censored data

Suppose n bulbds, Failure times of each light bulb is $X_1, \ldots X_n \sim Exp(\lambda)$. m of the lightbulbs were not recorded. Define $E_j = I(X_j < T)$, where T is the time after which the failure time recorded. So the observed data is

$E_1 = 1, \ldots E_m = 1, X_{m+1}, \ldots X_n$. Note that $E_i \sim Bern(p)$, where $p = 1 - e^{-\lambda T}$. We have to find MLE for λ.

$L(\lambda, E_1, \ldots, E_m, X_{m+1}, \ldots X_n) = f(E_1, \ldots E_m, X_{m+1}, \ldots X_n|\lambda) = \prod_{i=1}^{m} \mathbb{P}(E_i = 1). \prod_{j=m+1}^{n} f(x_j|\lambda) = (1 - e^{-\lambda T})^m \lambda^{n-m} \text{exp} - \lambda \sum_{j=m+1}^{n} x_j$

Complete Likelihood : $l_{\text{comp}}(\lambda|x_1, \ldots, x_n) = logf(\mathbf{x}|\lambda) = nlog\lambda - \lambda \sum_{i=1}^{n} x_i$.

In order to implement the EM algorithm, we need the conditional distribution of the unobserved data, given the observed data. Unobserved data is $X_1, \ldots X_m : f(X_1, \ldots X_m|E_1, .., E_m) = \prod_{i=1}^{m} f(X_i|E_i) = \prod_{i=1}^{m} \frac{\lambda e^{-\lambda x_i}}{1 - e^{-\lambda T}}$

1. E-Step - find the expectation of the complete likelihood given the observed data.

   $q(\lambda|\lambda_k) = E(logf(X_1, \ldots X_n|\lambda)|E_1, .., E_m, X_{m+1}, \ldots, X_n)$
2. M-Step : $\lambda_{k+1} = argmax_{\lambda}[.]$, which is easy update

8.3 EM Theory

Theorem : The EM Algorithm is an MM algorithm and thus has an ascent property.

Proof : First find a minorizing function. The objective function is $logf(\mathbf{x}|\theta)$. So find a $g(\theta|\theta_k)$, such that $g(\theta_k|\theta_k) = logf(\mathbf{x}|\theta_{(k)})$, and in general $logf(\mathbf{x}|\theta) \geq g(\theta|\theta_k)$. We will show that $g(\theta|\theta_k) = q(\theta|\theta_k) + \text{constants}$, then maximizing g is same as maximizing q(the M step).

$$\text{let } g(\theta|\theta_{(k)}) = \int_z log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + logf(\mathbf{x}|\theta_{(k)}) - \int_z logf(\mathbf{x}, \mathbf{z}|\theta_{(k)}) f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz.$$

Clearly at $\theta = \theta_{(k)} \Rightarrow g(\theta_k|\theta_k) = logf(\mathbf{x}|\theta_k)$

To show the minorizing property : We can establish the inequatlity using Jensen's Inequality.

8.4 Gaussian mixture model

Suppose $X_1, .., X_n \sim F$, where F is the mixture of normal distributions so that the density $f(x|\theta) = \sum_{j=1}^{C} \pi_j f_j(x|\mu_j, \sigma_j^2)$, where $f_i(x|\mu_i, \sigma_i^2)$ is the density of $N(\mu_i, \sigma_i^2)$ distribution for $i = 1, 2, \ldots, C.$ This is a mixture of normals with C classes or clusters or components.

Thus suppose the complete data was of the form $(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n)$, where each $Z_i = k$ means that X_i is from population k. Let $\theta = (\mu_1, \ldots, \mu_C, \sigma_1^2, \ldots, \sigma_C^2, \pi_1, \ldots, \pi_C)$

$[X_i|Z_i = c] \sim N(\mu_c, \sigma_c^2)$, and $\mathbb{P}(Z_i = c) = \pi_c$.

- Observed data $X_1, \ldots, X_n$: $X_i \sim f(x_i|\theta) = \sum_{j=1}^{C} \pi_j f_j(x_i|\mu_j, \sigma_j^2)$
- Unobserved data are iid $Z_1, \ldots, Z_n$ s.t. $Pr(Z_i = k) = \pi_k$
- Thus using EM algorithm to estimate the MLE of θ : we need to estimate $q(\theta|\theta_k)$, thus wwe require conditional distribution of Z given X, thus $Pr(Z_i = c|X_i = x_i) = \frac{f_c(x|\mu_c, \sigma_c^2)\pi_c}{f(x_i) = \sum_{j=1}^{C} \pi_j f_j(x_i|\mu_j, \sigma_j^2)} := \gamma_{i,c}$ - By bayes theorem

Thus for any kth iterate step $\theta_{(k)} = (\mu_{1,k}, ..., \mu_{c,k}, \sigma_{1,k}^2, ...\sigma_{c,k}^2, \pi_{1,k}, ..., \pi_{C,k})$

Now $q(\theta|\theta_k) = E_{Z|X}[log(f(\mathbf{x}, \mathbf{z}|\theta)|\mathbf{X} = \mathbf{x}, \theta_{(k)}]$