# Gaussians Mixture Model in EM Algorithm

*Yuance He & Yiyi Xu*

*10/25/2018*

## Abstract

Use the Newton's theorem to find the mel of the $\mu$ and $\sigma$ of the smoker and non-smoker and then find the Normal Mixture of the data. Then use the same way to find the Normal Mixture of the smoker, non-smoker, male and female in order to find the percentage of each type to fend the influence of each factor.

# 1 Introduction

We has a set of data contains gender, smoker/non-smoker, region and individual medical costs billed by health insurance. And we will analysis how gender and smoker/non-smoker influence the cost. So our goal here is to find the Normal Mixture Model $y = \alpha N(\mu_1, \sigma_1) + (1 - \alpha)N(\mu_2, \sigma_2)$, where $\alpha$ belongs to $(0, 1)$, and Smoker ~ $N(\mu_1, \sigma_1^2)$, Non-Smoke ~ $N(\mu_2, \sigma_2^2)$ by MLE based on the data sample we had. Therefore, the result can help insurance company to determine how to rearrange the price based on whether customer is smoker or non-smoker,male or female.

# 2 Math Equations

Step 1: Use Newton's Theorem to find the $\mu$ and $\sigma$ Find the $\mu$ and $\sigma$ of the smoker and non-smoker. Smoker and non-smoker each satisfty different normial distribution.

$$f(x = x_i) = f(x_i | \mu_k, \sigma_k)$$

$$L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i | \mu_k, \sigma_k^2)$$

$$l(\mu, \sigma) = \ln(L) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \mu}{\sigma})^2$$

The two Partial derivatives simplify to

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = -\sum_{i=1}^{n}\frac{x_i - \mu}{\sigma} - \frac{1}{\sigma}$$

$$\frac{\partial l(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

so

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Step 2:Use EM algorithm to find the percent of Normal Mixture Assume $\alpha$ is the percent of smoker and non-smoker. Smoker and non-smoker each satisfty different normial distribution.

$$P(x = x_i) = \sum_{\alpha_k} P(x_i | \alpha_k, \mu_k, \sigma_k^2)$$

Then the log likelihood function:

$$L(\alpha) = \prod_{i=1}^{n} \sum_{\alpha_k} P(x_i | \alpha_k, \mu_k, \sigma_k^2)$$

$$l(\alpha) = \ln(L) = \sum_{i=1}^{n} \log(\sum_{\alpha_k} P(x_i | \alpha_k, \mu_k, \sigma_k^2))$$

assume $\alpha$ follow the distrubtion Q, then we have

$$l(\alpha) = \sum_{i=1}^{n} \log(\sum_{\alpha_k} Q(\alpha_k) \frac{P(x_i | \alpha_k, \mu_k, \sigma_k^2)}{Q(\alpha_k)})$$

$$l(\alpha) = \sum_{i=1}^{n} \log[(1 - \alpha) N(\mu_1, \sigma_1^2) + \alpha N(\mu_2, \sigma_2^2)$$

By Jensen’s Inequality,

$$l(\alpha) \geq \sum_{i=1}^{n} (\sum_{\alpha_k} Q(\alpha_k) \log \frac{P(x_i | \alpha_k, \mu_k, \sigma_k^2)}{Q(\alpha_k)})$$

where the probability of sample $x_i$ belongs to typle k is

$$Q(\alpha_k) = \frac{P(x_i, \alpha_k)}{\sum_{\alpha_k} P(x_i, \alpha_k)} = P(\alpha_k | x_i)$$

M-Step

$$\begin{aligned}
f &= \sum_{i=1}^{m} \sum_{j=1}^{k} Q_i(\alpha = j) \log \frac{P(x_i | \alpha = j)}{Q(\alpha = j)} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} Q_i(\alpha = j) \log \frac{P(x_i) P(\alpha = j)}{Q(\alpha = j)} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^i \log \frac{N(\mu_j, \sigma_j^2)}{w_j^i} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^i \log \frac{e^{-0.5(x_i - \mu_j)^T} (x_i - \mu_j)) \psi_j}{w_j^i \sqrt{s\pi} (\sigma^2)^{0.5} \sigma_j^2}
\end{aligned}$$

$$
\begin{aligned}
0 &= \frac{\partial f}{\partial \psi} \\
&= \sum_{i=1} m \frac{w_j^i}{\psi_j} + \beta \\
1 &= \sum_{i=1}^{m} \psi_i
\end{aligned}
$$

$$\psi_j = \frac{\sum_{i=1}^{m} w_j^i}{m}$$

# 3 Analysis

```
library(readxl)
Projectdata <- data.frame(
  read_excel("C:/Users/Yuance He/Documents/final-project-yuance-yiyi-group-project/5361
 Table.xlsx"))
smoker <- Projectdata[which(Projectdata$smoker=='yes'),]
nonsmoker <- Projectdata[which(Projectdata$smoker=='no'),]
x <- smoker$charges
y <- nonsmoker$charges
###Newton method
loglike <- function(data,mu,sigma){
  sum(dnorm(data,mu,sigma,log = TRUE))
}
Mfirstderiv <- function(data,mu,sigma){
  L1 <- (1/(sigma^2))*sum(data-mu)
  return(L1)
}
Msecondderiv <- function(data,mu,sigma){
  L2 <- (-length(data))/(sigma^2)
  return(L2)
}
Sfirstderiv <- function(data,mu,sigma){
  l1 <- (-length(data)/sigma) + (sum((data-mu)^2)/sigma^3)
  return(l1)
}
Ssecondderiv <- function(data,mu,sigma){
  l2 <- (length(data)/sigma^2) - (3*sum((data-mu)^2)/sigma^4)
  return(l2)
}
Newton <- function(data,mu.init,sigma.init,max,tol){
  mu.current <- mu.init
  sigma.current <- sigma.init
  for (i in 1:max){
    mu.update <- mu.current - Mfirstderiv(data,mu.current,
                sigma.current)/Msecondderiv(data,mu.current,sigma.current)
    mu.dif <- abs(mu.update-mu.current)
    if(mu.dif < tol) break
    mu.current <- mu.update
  }
  for (j in 1: max){
    sigma.update <- sigma.current- Sfirstderiv(data,mu.current,
                sigma.current)/Ssecondderiv(data,mu.current,sigma.current)
    sigma.dif <- abs(sigma.update-sigma.current)
    if(sigma.dif < tol) break
    sigma.current <- sigma.update
  }
  return(c(mu.current,sigma.current,i+j))
}

smokerresult <- matrix(0,1,3)
nsmokerresult <- matrix(0,1,3)
smokerresult[1,] <- Newton(x,30000,8000,max = 200,tol = 1e-5)
nsmokerresult[1,] <- Newton(y,7200,5000,max = 200,tol = 1e-5)
colnames(smokerresult) <- c("Mu","Sigma","# of iteration")
colnames(nsmokerresult) <- c("Mu","Sigma","# of iteration")
## Table of MLE of smokers
knitr::kable(smokerresult)
```

| Mu | Sigma | # of iteration |
|---:|---:|---:|
| 32495.45 | 6841.929 | 8 |

```
## Table of MLE of nonsmokers
knitr::kable(nsmokerresult)
```

| Mu | Sigma | # of iteration |
|---:|---:|---:|
| 12275.54 | 3967.321 | 9 |

```
normalmix <- function(data,mu1,sigma1,mu2,sigma2,delta,max,tol){
  p1 <- p2 <- rep(0,length(data))
  for (i in 1: max) {
    for (j in 1: length(data)) {
      p1[j] <- delta * dnorm(data[j],mu1,
      sigma1)/(delta * dnorm(data[j],mu1,sigma1)+(1-delta)* dnorm(data[j],mu2,sigma2))
      p2[j] <- 1-p1[j]
    }
    delta.new <- mean(p1)
    if(abs(delta.new-delta)<tol){return(c(delta,i))}
    delta <- delta.new
  }
}

normalmix(Projectdata$charges,mu1 = 32495.45,sigma1 = 6841.929,
          mu2 = 12275.54,sigma2 = 3967.321,delta = 0.5,max = 500,tol = 1e-5)
```

```
## [1] 0.5676571 5.0000000
```

```
malesmoker <- smoker[which(smoker$sex=='male'),]
femalesmoker <- smoker[which(smoker$sex=='female'),]
ms <- malesmoker$charges
fs <- femalesmoker$charges

Newton(ms,30000,7000,max = 200,tol = 1e-5)
```

```
## [1] 32504.159  7002.222     5.000
```

```
Newton(fs,30000,7000,max = 200,tol = 1e-5)
```

```
## [1] 32486.690  6676.759     7.000
```

```
normalmix(smoker$charges,mu1 = 32504.159,sigma1 = 7002.222, mu2 = 32486.69,sigma2 = 667
6.759, delta = 0.5,max=1000,tol = 1e-5)
```

```
## [1]   0.4672901 856.0000000
```

```
malenonsmoker <- nonsmoker[which(nonsmoker$sex=='male'),]
femalenonsmoker <- nonsmoker[which(nonsmoker$sex=='female'),]
mns <- malenonsmoker$charges
fns <- femalenonsmoker$charges

Newton(mns,12000,4000,max = 200,tol = 1e-5)
```

```
## [1] 12286.253  4313.735     7.000
```

```
Newton(fns,12000,4000,max = 200,tol = 1e-5)
```

```
## [1] 12265.246  3603.412     7.000
```

```
normalmix(nonsmoker$charges,mu1 = 12286.253,sigma1 = 4313.735, mu2 = 12265.246,sigma2 =
3603.412, delta = 0.5,max=1000,tol = 1e-5)
```

```
## [1]   0.264637 121.000000
```

# 4 Summary and Discussion

By the result, we can see that the smoker has a higher $\mu$ and $\sigma$ than the non-smoker. Which means that the smoker always has worse health and higher cost. The graph below shows the relationship of the sample and the estimated distribution: Most people cost between 10000~ 15000 and 30000~ 35000. So the factor (smoker or non-smoker ) has a large influence of the health.

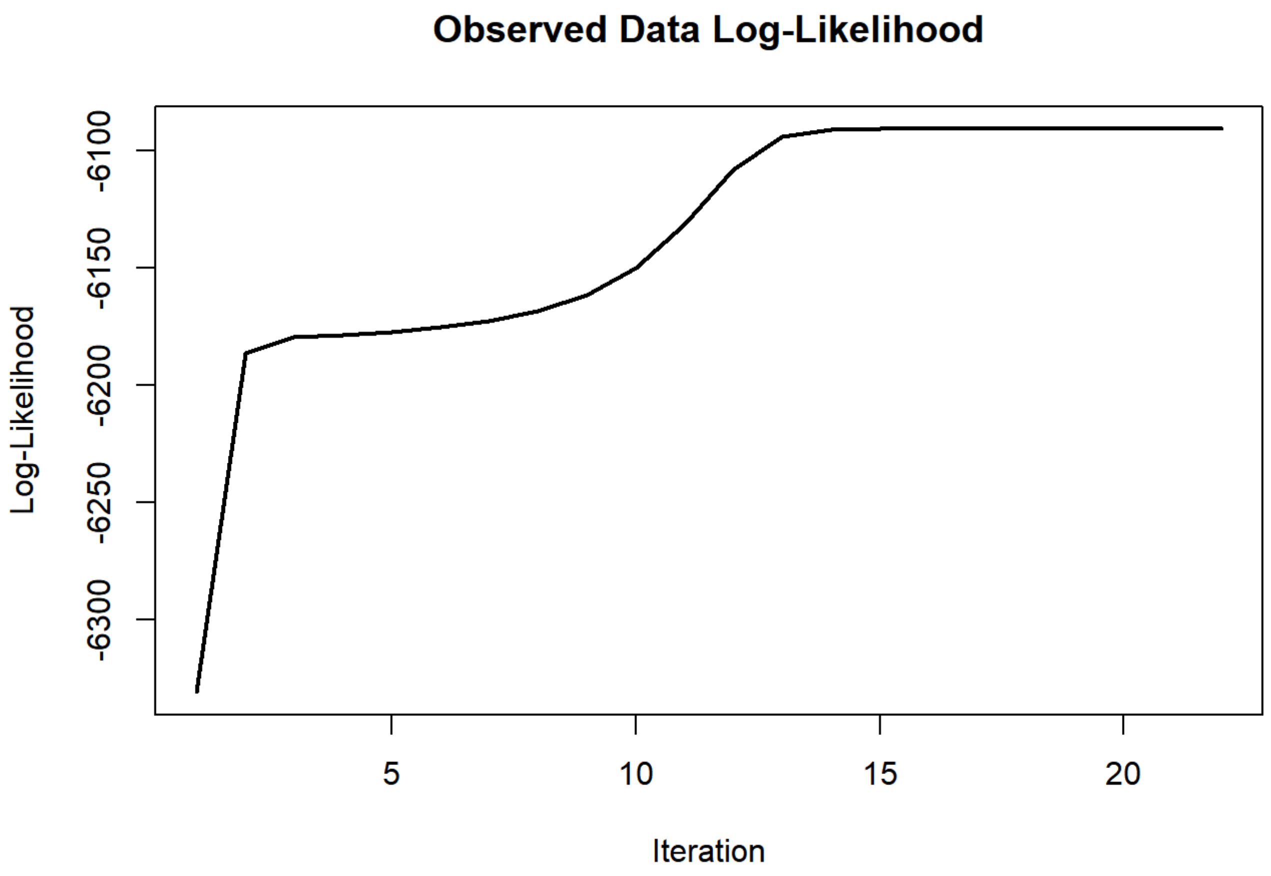
```
library(mixtools)
```

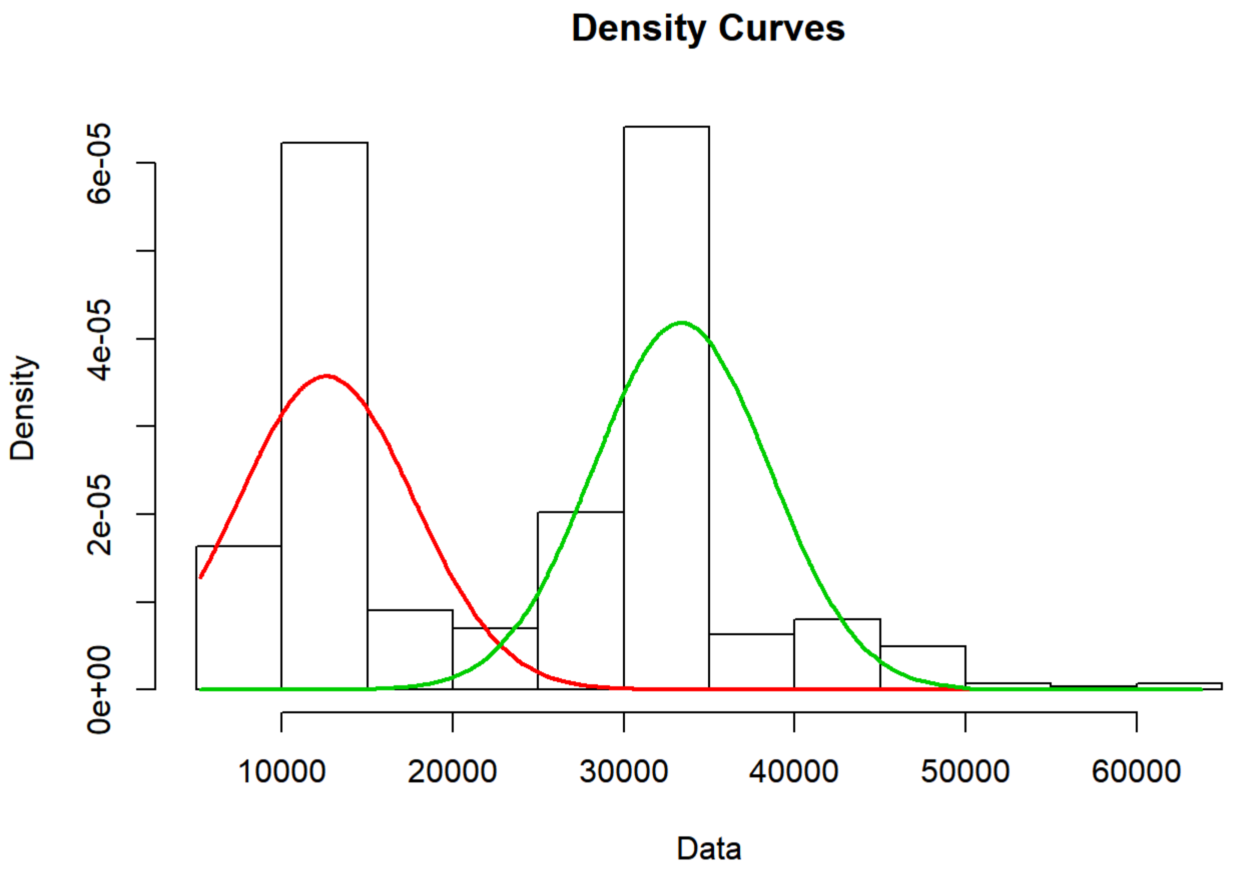
```
## mixtools package, version 1.1.0, Released 2017-03-10
## This package is based upon work supported by the National Science Foundation under Gr
ant No. SES-0518772.
```

```
charges <- normalmixEM(Projectdata$charges,epsilon = 1e-08,arbvar = FALSE, fast = TRUE)
```

```
## number of iterations= 21
```

```
plot(charges, density=TRUE)
```


Observed Data Log-Likelihood
Log-Likelihood
-6100
-6150
-6200
-6250
-6300
5
10
15
20
Iteration


Density Curves
Density
0e+00
2e-05
4e-05
6e-05
10000
20000
30000
40000
50000
60000
Data

After analysing the result of the influence of the gender, we will see that male has a higher $\mu$ than female but the difference is tiny. From research we get that the female always has a longer lifetime than male, so we guess one of the reason that may cause such result is female has a better health sitution than the male, that's also explains why the male has a higher $\mu$ that female. Due to our model, insurance company can estimate reserve fee by simply seperate people into four types.

# Reference

Choi, Miri. "Medical Cost Personal Datasets." RSNA Pneumonia Detection Challenge | Kaggle, 21 Feb. 2018, www.kaggle.com/mirichoi0218/insurance.

Hogg, Robert V., et al. Introduction to Mathematical Statistics. Pearson, 2016.