EM Algorithm

Shu-Ching Chang Hyung Jin Kim

December 9, 2007

1 Introduction

It's very important for us to understand the data structure before doing the data analysis. However, most of the time, there may exist of a lot of missing values or incomplete information in the data subject to the analysis. For example, survival time data always have some missing values because of death or job transfer. These kinds of data are called censored data. Since these data might obtain some incomplete but useful information, if we ignore them in the analysis, it's risky for us to get some biased results. The EM algorithm has the ability to deal with missing data and unidentified variables, so it is becoming useful in a variety of incomplete-data problem.

In this project, we investigate the EM algorithm for estimating parameters in application to missing data and mixture density problems. [4] shows the combination of EM algorithm and Bootstrap improves Satellite Image Fusion. Since Bootstrap approach uses the resampling concepts that can be applied to check the adequacy of standard measures of uncertainty and give quick approximate solutions, we are also interested in comparing the parameter estimations obtained form the EM algorithm to the combination of EM and Bootstrap. Therefore, after using the EM algorithm to find the unknown parameters from the data, we use the parameters estimated from EM to do the parametric Bootstrap to find the same unknown parameters from the data and compare them to results we obtain from the EM algorithm. Then, we can observe whether or not these two methods can get compatible results.

The paper is organized as follows. In Section 2, we introduce the EM algorithm including Expectation (E) Step and Maximization (M) Step. In Section 3, we apply the EM method to real data analysis and then we use the estimates obtained form the EM algorithm to the parametric Bootstrap. Finally, section 4 contains the summary and discussion of the results we obtained.

2 EM Algorithm

The EM algorithm has two main applications. The first case occurs when the data has missing values due to limitations or problems with the observation process. The second case occurs when the likelihood function can be obtained and simplified by assuming that there is an additional but missing parameters.

With missing values or parameters in the data which is generated by some distribution under assumption, we call the data, X, the incomplete data. And, we assume that the complete data, $Z = (X, Y)$ exists with Y being missing data and that a joint density function also exists as follows:

$$p(z|\theta) = p(x, y|\theta) = p(y|x, \theta) * p(x|\theta)$$

where θ is a set of unknown parameters from a distribution including a missing parameter.

With the density function, we now define the complete-data likelihood as:

$$L(\theta|Z) = L(\theta|X, Y) = p(X, Y|\theta)$$

And, the original likelihood $L(\theta|X)$ is called the incomplete-data likelihood function. Since the missing data Y is unknown under a certain distribution by assumption, we can think of $L(\theta|X, Y)$ as a function of a random variable, Y, with constant values, X and θ.

$$L(\theta|X, Y) = f_{(X,\theta)}(Y)$$

Using the complete-data log-likelihood function with respect to the missing data Y given the observed data X, the EM algorithm finds its expected value as well as the current parameter estimates at the E Step and maximizes the expectation at the M step. By repeating the E and M step, the algorithm is guaranteed to converge to a local maximum of the likelihood function with each iteration increasing the log-likelihood.

2.1 Expectation (E) Step

First, we define the expectation of the complete-data log-likelihood function as:

$$Q(\theta, \theta^{(i-1)}) = E[log\ p(X, Y|\theta)|X, \theta^{(i-1)}] \tag{1}$$

where $\theta^{(i-1)}$ is a set of current parameters estimates that we use to evaluate the expectation and to increase Q with the new θ for optimization. Here, X and $\theta^{(i-1)}$ are known constants and θ is a variable to be adjusted. Since Y is a missing random variable under an assumed distribution, $f(y|X, \theta^{(i-1)})$. Then, the expectation in the equation (1) can be written as:

$$E[log\ p(X, Y|\theta)|X, \theta^{(i-1)}] = \int_{y\in\Omega} log\ p(X, y|\theta) * f(y|X, \theta^{(i-1)})dy \tag{2}$$

where Ω is the space of values where y can take values on and $f(y|X, \theta^{(i-1)})$ is the marginal distribution of the missing data Y depending on observed data and current parameters.

2.2 Maximization (M) Step

At the M step, we maximize the expectation we obtain in the E step. That is to find:

$$\theta^{(i)} = \underset{\theta}{argmax}\ Q(\theta, \theta^{(i-1)})$$

Maximizing the equation (1) becomes either easy or hard depending on the form of $p(X, y|\theta)$. For instance, if $p(X, y|\theta)$ is a simple normal distribution where $\theta = (\mu, \sigma^2)$, we set the derivative of $log(L(\theta|X, Y))$ equal to zero and solve directly for μ and σ^2. This is an easy example, but we need to resort to more elaborate techniques for more complicated ones.

3 Example

The data used for the example is called `faithful` implemented in R. It contains waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. From this data, we use the waiting time part. The histogram of the waiting time resembles the mixture Gaussian distribution; short and long waiting times. We set indicators which mixture component each observation belongs to as a missing data and the EM algorithm will find the proportion of observations belonging to each normal distribution along with other unknown parameters for means and variances.

The density for the mixture of two Gaussian populations is

$$f_W(w|\theta) = p * \frac{1}{\sigma_1} * \varphi\left(\frac{w - \mu_1}{\sigma_1}\right) + (1 - p) * \frac{1}{\sigma_2} * \varphi\left(\frac{w - \mu_2}{\sigma_2}\right)$$

And, our unknown parameters are $p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ where μ_1 and σ_1^2 indicate the mean and the variance from the normal distribution with the shorter waiting time and μ_2 and σ_2^2 represent the mean and the variance from the longer waiting time. Our p represents the proportion an observation comes from the normal distribution with the shorter waiting time. We let $\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

The indicator variable as a missing data is

$$Y_i = \begin{cases} 1 & W_i \text{ belongs to distribution of shorter waiting times} \\ 0 & W_i \text{ belongs to distribution of longer waiting times} \end{cases}$$

where Y_i is Bernoulli distribued with parameter p.

Therefore, the likelihood expression for the complete-data is given by:

$$L_n(\theta|W,Y) = \prod_{i=1}^{n} p^{Y_i} * (1-p)^{1-Y_i} * \frac{1}{\sigma_1^{Y_i}} \varphi\left(\frac{W_i - \mu_1}{\sigma_1}\right)^{Y_i} * \frac{1}{\sigma_2^{1-Y_i}} \varphi\left(\frac{W_i - \mu_2}{\sigma_2}\right)^{1-Y_i}$$

And, the corresponding log-likelihood function for the density from the data `faithful` becomes:

$$\begin{aligned} l_n(\theta|W,Y) = & \sum_{i=n}^{n} Y_i * log(p) + \sum_{i=n}^{n} (1-Y_i) * log(1-p) \\ & - \frac{1}{2} \sum_{i=1}^{n} Y_i * log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} Y_i * (W_i - \mu_1)^2 \\ & - \frac{1}{2} \sum_{i=1}^{n} (1-Y_i) * log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2} \sum_{i=1}^{n} (1-Y_i) * (W_i - \mu_2)^2 \end{aligned}$$

From now, we apply the EM algorithm and find the expectation of Y_i. Since the conditional distribution of Y_i given W is

$$Y_i|W_i, \theta^{(k)} \sim Bin(1, p_i^{(k)})$$

with

$$p_i^{(k)} = \frac{p^{(k)} \frac{1}{\sigma_1^{(k)}} \varphi\left(\frac{w - \mu_1^{(k)}}{\sigma_1^{(k)}}\right)}{p^{(k)} \frac{1}{\sigma_1^{(k)}} \varphi\left(\frac{W_i - \mu_1^{(k)}}{\sigma_1^{(k)}}\right) + (1-p^{(k)}) \frac{1}{\sigma_2^{(k)}} \varphi\left(\frac{W_i - \mu_2^{(k)}}{\sigma_2^{(k)}}\right)}.$$

where $p^{(k)}$ is a set of known or estimated parameters at kth step. $p^{(0)}$ is an initial values. Thus, by the property of the Binomial distribution, the conditional mean is

$$E(Y_i|W_i, \theta^{(k)}) = p_i^{(k)}.$$

By substituting $p^{(k)}$ for Y_i, we obtain the expectation function as

$$\begin{aligned} Q(\theta|\theta^{(k)}) = & \sum_{i=1}^{n} p_i^{(k)} * log(p) + \sum_{i=n}^{n} (1-p_i^{(k)}) * log(1-p) \\ & - \frac{1}{2} \sum_{i=1}^{n} p_i^{(k)} * log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} p_i^{(k)} * (W_i - \mu_1)^2 \\ & - \frac{1}{2} \sum_{i=1}^{n} (1-p_i^{(k)}) * log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2} \sum_{i=1}^{n} (1-p_i^{(k)}) * (W_i - \mu_2)^2 \end{aligned}$$

In the maximization step, setting the first derivatives of $Q(\theta|\theta^{(k)})$ with respect to each parameter equal to zero results in following equations for each parameter.

$$
\begin{aligned}
p^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n} p_i^{(k)} \\
\mu_1^{(k+1)} &= \frac{\sum_{i=1}^{n} p_i^{(k)} W_i}{\sum_{i=1}^{n} p_i^{(k)}} \\
\mu_2^{(k+1)} &= \frac{\sum_{i=1}^{n} (1-p_i^{(k)}) W_i}{\sum_{i=1}^{n} (1-p_i^{(k)})} \\
(\sigma_1^{(k+1)})^2 &= \frac{\sum_{i=1}^{n} p_i^{(k)} (W_i - \mu_1^{(k+1)})^2}{\sum_{i=1}^{n} p_i^{(k)}} \\
(\sigma_2^{(k+1)})^2 &= \frac{\sum_{i=1}^{n} (1-p_i^{(k)}) (W_i - \mu_2^{(k+1)})^2}{\sum_{i=1}^{n} (1-p_i^{(k)})}
\end{aligned}
$$

For the initial value of θ, we decide to have the following values from the Figure 1.

$$p^{(0)} = 0.4, \mu_1^{(0)} = 40, \mu_2^{(0)} = 90, \sigma_1^{(0)} = 4, \sigma_2^{(0)} = 4$$

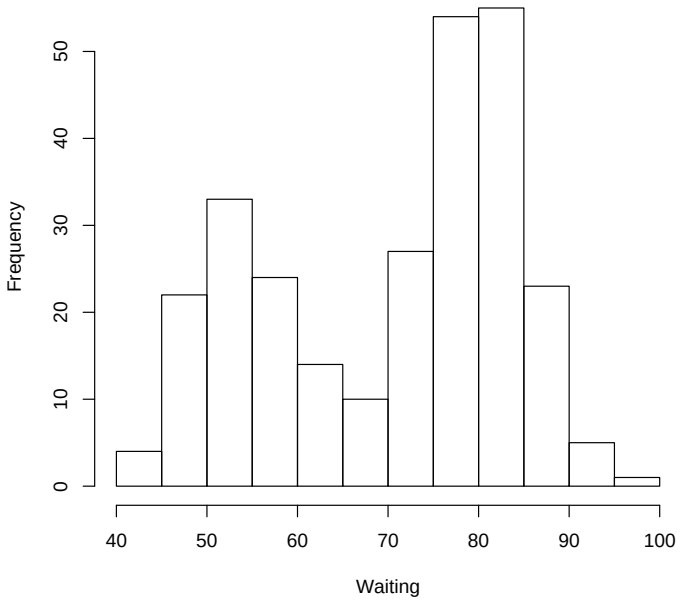

Figure 1: Histogram of Waiting

After the EM algorithm, our estimates for unknown parameters are described in the Table 1. In addition, Table2 shows the results form the combination of EM algorithm and Bootstrap.

p	μ_1	μ_2	σ_1^2	σ_2^2
0.35	54.22	79.91	29.86	35.98

Table 1: Estimates Using EM Algorithm

μ_1	μ_2	σ_1^2	σ_2^2
54.26	79.88	29.74	36.03

Table 2: Estimates Using the Combination of EM Algorithm and Bootstrap

4 Discussion

In our study, we review some literature about EM algorithm, which has shown great performance in practice to deal with missing data and mixture density problems. In section 3, we utilize EM algorithm to experiment on one data called "faithful" which has the characteristics of missing values and mixture densities . Since Bootstrap is a re-sampling method that improves estimator properties notably in small sample, we also want to know if the bootstrap method combined with EM algorithm can improve the accuracy of the EM algorithm alone. Then we use the estimates form the EM algorithm to do parametric Bootstrap and compare the results from these two approaches.

From Figure 1, we can clearly find that there exists of two mixture normal distributions for the waiting time of faithful dataset. We apply the EM algorithm to this dataset in order to find unknown parameters for means and variances. In Table 1, we find about 35% an observation comes from the normal distribution with the shorter waiting time. The mean of normal distribution with the shorter and longer waiting time are separately about 54.22 and 79.91 that are much closed to the result of the histogram plot of waiting time we obtain in Figure 1. In addition, the variances we get form the EM algorithm seem appropriate. We may reasonably conclude that the parameter estimations of mixture density dataset form the EM algorithm almost approximate to the parameters of real dataset. Therefore, EM algorithm seems a good procedure to help us find the characteristics of one data, especially for mixture density dataset. Table 2 shows the results form the combination of EM algorithm with Bootstrap that also obtains the parameter estimations approximated to the ones of original dataset.

Many interesting problems may be worth future study, such as creating a criterion to decide which procedure could get more precise parameter estimators to the original dataset. In addition, since it's important to choose appropriate initial values in the EM algorithm, finding a procedure of choosing initial values is also needed. Finally, we might try to find a more suitable formula of the combination of EM algorithm and bootstrap.

5 Appendix

```
> data(faithful)
```

```
> attach(faithful)
>
> ## EM Algorithm
>
> W = waiting
>
> s = c(0.5, 40, 90, 16, 16)
>
> em = function(W,s) {
+   Ep = s[1]*dnorm(W, s[2], sqrt(s[4]))/(s[1]*dnorm(W, s[2], sqrt(s[4])) +
+        (1-s[1])*dnorm(W, s[3], sqrt(s[5])))
+   s[1] = mean(Ep)
+   s[2] = sum(Ep*W) / sum(Ep)
+   s[3] = sum((1-Ep)*W) / sum(1-Ep)
+   s[4] = sum(Ep*(W-s[2])^2) / sum(Ep)
+   s[5] = sum((1-Ep)*(W-s[3])^2) / sum(1-Ep)
+   s
+ }
>
> iter = function(W, s) {
+       s1 = em(W,s)
+       for (i in 1:5) {
+           if (abs(s[i]-s1[i]) > 0.0001) {
+               s=s1
+               iter(W,s)
+           }
+           else s1
+       }
+       s1
+ }
>
> iter(W,s)
[1]  0.3507784 54.2179838 79.9088649 29.8611799 35.9824271
>
> p = iter(W, s)
>
> p1<-p[1]
> p2<-p[2]
> p3<-p[3]
> p4<-p[4]
> p5<-p[5]
>
> Boot<-function(B){
+       r<-0
+       k<-0
+       bootmean1 <-rep(0, B)
```

```
+       bootvar1<-rep(0, B)
+       bootmean2<-rep(0, B)
+       bootvar2<-rep(0, B)
+       for(i in 1:B){
+             p<-runif(1, 0, 1)
+             if(p<p1){
+                    boot1<-rnorm(p1*272, p2, sqrt(p4))
+                    bootmean1[i]<-mean(boot1)
+                    bootvar1[i]<-var(boot1)
+                    r<-r+1
+              }
+              else{
+                      boot2<-rnorm((1-p1)*272, p3, sqrt(p5))
+                      bootmean2[i]<-mean(boot2)
+                      bootvar2[i]<-var(boot2)
+                      k<-k+1
+              }
+         }
+        meanbootm1<-sum(bootmean1)/r
+        meanbootvar1<-sum(bootvar1)/r
+        meanbootm2<-sum(bootmean2)/k
+        meanbootvar2<-sum(bootvar2)/k
+        list(meanbootm1= meanbootm1, meanbootvar1= meanbootvar1,
+             meanbootm2= meanbootm2, meanbootvar2= meanbootvar2 )
+ }
>
> Boot(1000)
$meanbootm1
[1] 54.26174

$meanbootvar1
[1] 29.74411

$meanbootm2
[1] 79.88433

$meanbootvar2
[1] 36.02818
```

References

[1] Jeff A. Blimes, International Computer Science Institute.
Computer Science Division, Department of Electrical Engineering and Computer Science. (April 1998).

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977).
Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1-38.

[3] Richard A. Redner and Homer F. Walker (April 1984)
Mixture Densities, Maximum Likelihood and the Em Algorithm. SIAM Review, 26(2), 195-239.

[4] Tijani Delleji, Mourad Zribi, and Ahmed Ben Hamida (2007)
On the EM Algorithm and Bootstrap Approach Combination for Improving Satellite Image Fusion. International Journal of Signal Processing, volume 4, No.1, ISSN 1304-4478.

[5] Geoffrey J. McLachlan and Thriyambakam Krishnan (1997)
The EM Algorithm and Extensions.John Wiley & Sons, Inc.

[6] Michiko Watanabe and Kazunori Yamaguchi (1991)
The EM Algorithm and Related Statistical Models. STATISTICS: A DEKKER series of TEXTBOOKS and MONOGRAPHS

[7] Kate Cowles, The University of Iowa.
Lecture Note 11. (September 24, 2006).