Humboldt Universität zu Berlin

Numerical Introductory Course Final Project

Expectation Maximization Algorithm

Ching-Hsi Lee

leeching@hu-berlin.de

Abstract

The main aim of this report is to summarize and explore the Expectation Maximization (EM) algorithm, one of the most widely applied numerical method that building model under presence of missing value. Beside fundamental introduction on EM algorithm (Section 3), we also cover the widely extension of EM algorithm (Section 5), including clustering method Gaussian Mixture Model. In the last section, we illustrate EM method by some empirical analysis on actual dataset (Section 7).

Contents

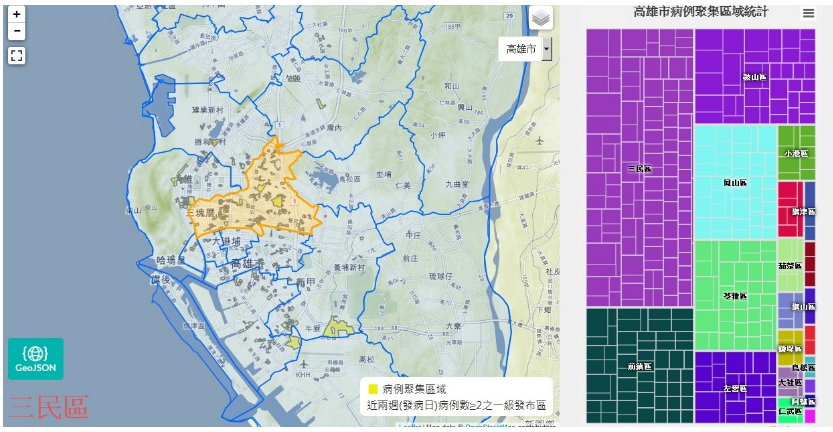

Figure 1: Dengue fever incidence in Kaohsiung City

1 Motivation

In real world situation, there were always modelling problem from missing or truncated dataset. For example in an epidemiology study: Dengue fever outbreak in a sub-tropical city [1].

We collect data for incidence and population numbers from different area i in the city, and make assumption on possibly model [2] behind.

$$\begin{cases} x_i \sim Poisson(\lambda_i), \text{ Population of area } i \\ y_i \sim Poisson(\beta\lambda_i), \text{ Incidence of Dengue fever in area } i \end{cases}$$

The goal for the study is to use Maximum Likelihood (ML) method estimate parameter λ_i, β from collecting data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, and make inference based on this model estimator. The parameter ML estimation based on complete-data is $\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$ and $\hat{\lambda_j} = \frac{x_j + y_j}{\hat{\beta}+1}, j = 1, \ldots, n$

However, in case of incomplete-data, we don't have population number for first area x_1. The data collected become $\{(x_1 = \blacksquare, y_1)(x_2, y_2), \ldots, (x_n, y_n)\}$. Based on the situation, the maximum likelihood of model parameter will be recursive and unsolvable. We will do a parameter comparison in Appendix 9.1.1

The complicated situation is the place to apply Expectation Maximization Algorithm.

2 Structure

The structure for this report is based on arrangement of presentation, with implemented figure and explanation. The first part for this report is example of simulated Poisson distribution on Dengue fever. Second part illustrate the literature on Expectation Maximization (EM) Method and formalized its method by example on Dengue fever. The third part is application on EM algorithm, and we emphasize on one of most applied clustering method on Gaussian Mixture Model (GMM) [3]. We will continue illustrate EM method on other type of data, other extension based on idea of Expectation Maximization, and comparison of strength and weakness. We then include empirical analysis on 3 classic dataset in R, and its handy package "mclust". The final part is conclusion on EM method.

3 Literature & Formalize EM Method

EM Algorithm is a Maximum Likelihood method under incomplete-data. It first rooted back in work done in 1950's without apparently naming [4]. Later on, it was improved and illustrated its applications in a classical seminar work [5], but the prove of convergence was incorrect. In the beginning of 1980, a valid prove of convergence was raised [6, 7] and various overview of EM algorithm had been proposed [8, 9].

3.1 Data Structure

The dataset applied on EM algorithm is usually dataset with missing, truncated or censored variable, which is also called augmented data $\mathcal{Z}$. Those latent variable with observable variable $\mathcal{W}$ in combination is called Data Mixture $\mathcal{X} = (\mathcal{W}, \mathcal{Z})$.

Continuing on Dengue fever example, the observed but incomplete-data $\mathcal{W}$ is number we collected. The augmented data $\mathcal{Z}$ will be the number of population in first area $x_1 = \blacksquare$.

$$\mathcal{X} = \begin{cases} \mathcal{W} = \{y_1, (x_2, y_2), \ldots, (x_n, y_n)\} \text{ observed data} \\ \mathcal{Z} = x_1 = \blacksquare \text{ augmented data} \end{cases}$$

3.2 What is EM algorithm? How does it work?

Our general goal in constructing model is aimed to find the most probable maximum likelihood estimators $\hat{\theta}$ under missing value presented in Data Mixture $\mathcal{X}$. However, based on incomplete-data $\mathcal{W}$, the objective function $lnL(\theta|\mathcal{W})$ is sometimes lead to a recursive parameter or hard to find a close form solution.

EM algorithm is based on ML method, bypassing the difficulty dealing with intensive computation. We divide EM algorithm into Expectation step and Maximization step.

On Expectation step, we first take "conditional Expectation" on log-likelihood function $lnL(\theta|\mathcal{W})$ given first initial parameter $\theta^{(1)}$ and observable dataset $\mathcal{W}$, and acquire the maximization objective function $Q(\theta|\theta^{(1)})$ from former part.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta}\ lnL(\theta|\mathcal{W}) = \arg\max_{\theta} E[lnL(\theta|\mathcal{W})|\theta, \mathcal{W}] \\ &= \arg\max_{\theta} E[lnL(\theta|\mathcal{W}, \mathcal{Z})|\theta^{(1)}, \mathcal{W}] - E[ln(\frac{f(\mathcal{W}, \mathcal{Z}|\theta)}{f(\mathcal{W}|\theta)})|\theta^{(1)}, \mathcal{W}] \\ &\triangleq \arg\max_{\theta}\ Q(\theta|\theta^{(1)}) - \text{irrelevant term}\end{aligned}$$

On Maximization step, we derive the $\hat{\theta}$ estimation that "Maximize" the conditional Expectation objective function $Q(\theta|\theta^{(1)})$, given an initial parameter $\theta^{(1)}$

The mechanism for EM algorithm is taking conditional expectation of log-likelihood function while average over all possible missing value $\mathcal{Z}$. And thus achieve the most probable estimation of parameter. While we optimize the log-likelihood function given incomplete-data, we are at the same time optimize the conditional expectation over log-likelihood function given Data Mixture, and omit the irrelevant term.

When we find the maximized realization of parameter $\hat{\theta}$ from $Q(\theta|\theta^{(1)})$ is equivalent to find MLE from incomplete-data log-likelihood function $lnL(\theta|\mathcal{W})$.

3.3 EM monotonicity

To achieve such mechanism, we need the monotonic property of EM sequence [5, 6, 7, 10]. That is, when we find maximized realization sequence from

$Q(\theta|\theta^{(1)})$, we are equivalently find monotone increasing sequence of estimator from log-likelihood function $lnL(\theta|\mathcal{W})$.

Theorem 3.1 Monotonic EM sequence
Let $Q(\theta|\theta^{(1)}) = E[lnL(\theta|\mathcal{W},\mathcal{Z})|\theta^{(1)},\mathcal{W}]$ Q-function be objective of maximization. If $Q(\theta|\theta^{(m)}) \geq Q(\theta^{(m)}|\theta^{(m)})$, then $lnL(\theta|\mathcal{W}) \geq lnL(\theta^{(m)}|\mathcal{W})$.

sketch of proof in Appendix 9.2

3.4 Formalize a EM algorithm

For easier illustration and comparison in later sections, we formalize the EM algorithm in two steps.

E-step: Given estimate from m-th iteration$\theta^{(m)}$, calculate the conditional expectation Q-function,

$$Q(\theta|\theta^{(m)}) = E[lnL(\theta|\mathcal{W},\mathcal{Z})|\theta^{(m)},\mathcal{W}]$$

M-step: Find the (m+1)-th θ,

$$\theta^{(m+1)} = \arg\max_\theta Q(\theta|\theta^{(m)})$$

3.5 Illustrate on Dengue fever

We proceed on Dengue fever study, our objective parameter set is $\theta = (\beta,\lambda_1,\ldots,\lambda_n)$, where i area population x_i is derived from $Poisson(\lambda_i)$. y_i incidence of Dengue fever in area i is $Poisson(\beta\lambda_i)$ distributed. β is incidence discount factor from population.

Our data mixture is,

$$\mathcal{X} = \begin{cases} \mathcal{W} = \{y_1,(x_2,y_2),\ldots,(x_n,y_n)\} \text{ observed data} \\ \mathcal{Z} = x_1 = \blacksquare \text{ augmented missing data} \end{cases}$$

First starting on **E-step**, we calculate the conditional expectation given initial parameter $\theta^{(1)}$ and observed data $\mathcal{W}$, then acquire Q-function

$$Q(\beta,\lambda_1,\ldots,\lambda_n|\theta^{(1)}) = \sum_{x_1=0}^{\infty} ln(\prod_{i=1}^{n} \frac{e^{-\beta\lambda_i}(\beta\lambda_i)^{y_i}}{y_i!}\frac{e^{-\lambda_i}(\lambda_i)^{x_i}}{x_i!})\frac{e^{-\lambda_1^{(1)}}(\lambda_1^{(1)})^{x_1}}{x_1!}$$

Then continued on **M-step**, we maximize Q-function and denote it as second step parameter $\theta^{(2)}$

$$\theta^{(2)} = \arg \max_{\beta,\lambda_1,\ldots,\lambda_n} \sum_{x_1=0}^{\infty} ln(\prod_{i=1}^{n} \frac{e^{-\beta\lambda_i}(\beta\lambda_i)^{y_i}}{y_i!} \frac{e^{-\lambda_i}(\lambda_i)^{x_i}}{x_i!}) \frac{e^{-\lambda_1^{(1)}}(\lambda_1^{(1)})^{x_1}}{x_1!}$$

We could find the closed form solution for second step parameter $\hat{\theta}^{(2)} = (\hat{\beta}^{(2)}, \hat{\lambda_1}^{(2)}, \hat{\lambda_j}^{(2)})$, for $j > 1$, derived from **M-step** as below,

$$\hat{\theta}^{(2)} = \begin{cases} \hat{\beta}^{(2)} = \dfrac{\sum_{i=1}^{n} y_i}{\lambda_1^{(1)} + \sum_{i=2}^{n} x_i} \\ \hat{\lambda}_1^{(2)} = \dfrac{\lambda_1^{(1)} + y_1}{\hat{\beta}^{(2)} + 1} \\ \hat{\lambda}_j^{(2)} = \dfrac{x_j + y_j}{\hat{\beta}^{(2)} + 1}, j = 2, \ldots, n \end{cases}$$

The EM solution is easily reached once provide initial value $\theta^{(1)}$, in compared to MLE derived from incomplete-data $(\hat{\beta}^{\text{iMLE}}, \hat{\lambda}_1^{\text{iMLE}}, \hat{\lambda}_j^{\text{iMLE}})$ as shown in Appendix 9.1.1. We could find the optimized solution for EM estimator by continue iteration over EM algorithm till some pre-setting ϵ, such that $||\hat{\theta}^{(m+1)} - \hat{\theta}^{(m)}||^2 < \epsilon$. This is also promised by monotonic increasing property of EM algorithm.

3.6 Simulation from Dengue fever

The simulation result from Dengue fever is in Figure 2. We had model parameter as $\beta = 0.6, \lambda_1 = 450, \lambda_2 = 120$ labeled as red line in $n = 10$ areas (omit $\lambda_j, j > 2$ here). We could observe the simulating parameter reach its optimum at 15 steps of iteration in all three parameters, which means in simple Poisson model mixture, the EM algorithm could give a convenient parameter estimation in short period of time.

4 Application on EM

In this section, we will shortly mention few applications on EM algorithm. For its applicability and extensive on various field [11], we could better construct model upon missing, censored data or apply on clustering problem.

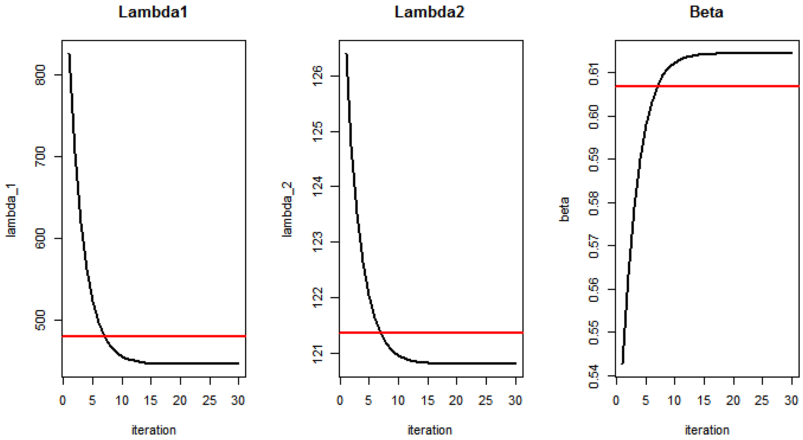

Figure 2: Dengue fever Poisson Mixture Model EM simulation with parameter $\beta = 0.6, \lambda_1 = 450, \lambda_2 = 120, n = 10$

In Financial field, EM algorithm can apply on estimating risk of portfolio, incomplete-data house price model. Another widely application is EM clustering on Gaussian Mixture Model (GMM), which recognized group label as missing value and evaluate the most proper model. Computer Scientist use EM algorithm to implement computer vision, nature language process, web document identification, handwriting recognition and Bayesian network model. Psychiatrist apply it on item response theory. EM algorithm also help research projects like micro-array, medical image analysis and fuzz image segmentation. With profound development in EM algorithm, there are also modification of EM algorithm such as α-EM, Expectation Conditional Maximization (ECM) and Generalized Expectation Maximization (GEM).

Upon all the methods and application above, we will further emphasize on extensive application of EM algorithm in next section. We first focus on Gaussian Mixture Model accompany with its example on Geyser dataset. In order to better understand the setting based on EM and functional applicability of EM, we further delve into other model setting extending from EM algorithm such as Hidden Markov Model (HMM), Maximization-maximization model and Maximum a Posterior (MAP) model.

5 EM Extensions

We first introduce Gaussian Mixture Model, the most widely applied extension from EM algorithm. Then further compare HMM, M-M and MAP model with original EM algorithm.

5.1 GMM: Gaussian Mixture Model problem

The goal for GMM is to solve clustering problem on multivariate normal model, which is also called EM clustering. Assuming data from 2 normal distribution, we could first deem clustering label $z_i \in (1,2)$ of each data point as missing value. Then use EM algorithm to estimate parameters $\theta = \{\mu_1, \Sigma_1, \mu_2, \Sigma_2, \tau_1, \tau_2\}$ for each cluster, τ_1, τ_2 is cluster probability of observed data.

$$\begin{cases} x_i|z_i = 1 \sim \mathcal{N}(\mu_1, \Sigma_1), \text{ with probability } p(z_i = 1) = \tau_1 \\ x_i|z_i = 2 \sim \mathcal{N}(\mu_2, \Sigma_2), \text{ with probability } p(z_i = 2) = \tau_2 = 1 - \tau_1 \end{cases}$$

Data mixture from observed incomplete-data and missing data can be expressed as below, we alter the missing value from EM algorithm to missing "labels"

$$\mathcal{X} = \begin{cases} \mathcal{W} = \{x_1, \ldots, x_n\} \text{ observed data} \\ \mathcal{Z} = \blacksquare = \{z_1, \ldots, z_n\} \text{ augmented data} \end{cases}$$

We need to define an additional Gaussian component, the posterior probability of sample x_i in $z_i = 1$ clusters is for example as below

$$P(z_i = 1|\mathcal{W}, \mu_1, \Sigma_1, \mu_2, \Sigma_2, \tau_1, \tau_2) = \frac{\tau_1 \phi(x_i|\mu_1, \Sigma_1)}{\tau_1 \phi(x_i|\mu_1, \Sigma_1) + \tau_2 \phi(x_i|\mu_2, \Sigma_2)}$$

And the objective log-likelihood function is 2 weighted log-normal density function between two clusters. We could then proceed on **E-step**, calculate conditional expectation given $(\theta^{(1)}, \mathcal{W})$ and obtain Q-function.

$$\begin{aligned} & Q(\mu_1, \Sigma_1, \mu_2, \Sigma_2, \tau_1, \tau_2|\theta^{(1)}) \\ & = \sum_{i=1}^{n} \sum_{j=1}^{2} (ln\tau_j - \frac{1}{2} ln|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)) \\ & \quad \times \frac{\tau_j^{(1)} \phi(x_i|\mu_1^{(1)}, \Sigma_1^{(1)})}{\tau_1^{(1)} \phi(x_i|\mu_1^{(1)}, \Sigma_1^{(1)}) + \tau_2^{(1)} \phi(x_i|\mu_2^{(1)}, \Sigma_2^{(1)})} + constant \end{aligned}$$

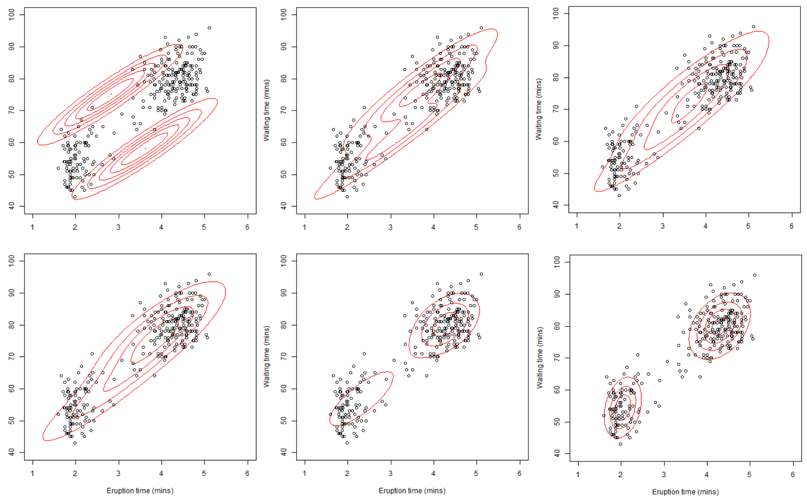

Figure 3: Old Faithful Geyser cluster spline movement

In **M-step**, we will maximize Q-function and find second step parameter $\theta^{(2)} = \arg\max_{\mu_1,\Sigma_1,\mu_2,\Sigma_2,\tau_1,\tau_2} Q(\mu_1, \Sigma_1, \mu_2, \Sigma_2, \tau_1, \tau_2|\theta^{(1)})$, the close form solution could be written as below, denote $n_j^{(1)} = \sum_{i=1}^{n} P(z_i = j|\theta^{(1)}, \mathcal{W})$, the EM clustering parameter for estimators are

$$\theta^{(2)} = \begin{cases} \hat{\tau_j}^{(2)} = \dfrac{n_j^{(1)}}{n}, \ j = 1, 2 \\ \hat{\mu}_j^{(2)} = \dfrac{1}{n_j^{(1)}} \displaystyle\sum_{i=1}^{n} P(z_i = j|\theta^{(1)}, \mathcal{W}) x_i, \ j = 1, 2 \\ \hat{\Sigma}_j^{(2)} = \dfrac{1}{n_j^{(1)}} \displaystyle\sum_{i=1}^{n} P(z_i = j|\theta^{(1)}, \mathcal{W})(x_i - \hat{\mu}_j^{(2)})(x_i - \hat{\mu}_j^{(2)})^T, \ j = 1, 2 \end{cases}$$

5.2 Simulation from Old Faithful Geyser

We illustrate the GMM method by famous Old Faithful Geyser eruption data in this section. We could observe the cluster spline movement of iteration in Figure 3 [12], starting from first random point then converge gradually to ML center. The iteration will reach its theoretic mean around 25 steps for each cluster in Figure 4, showing with respect to its x-y-axis (blue, red line is presuming threshold for each cluster).

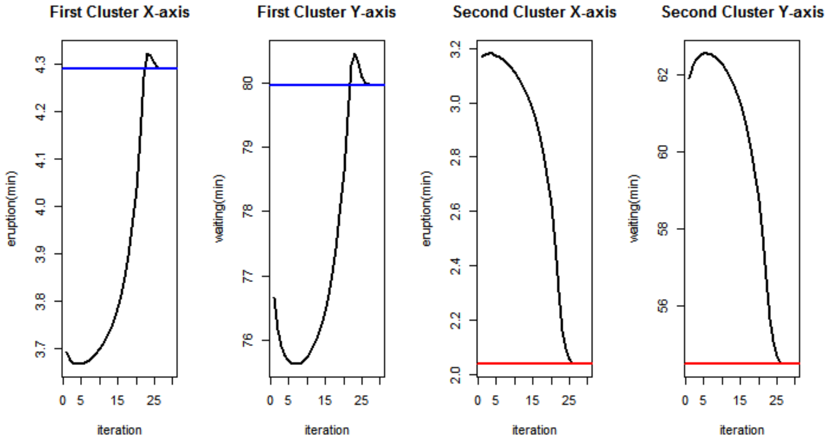

Figure 4: Old Faithful Geyser EM simulation

5.3 HMM: Hidden Markov Model

Hidden Markov Model (HMM) is aimed to model parameter θ from a random sequence $\mathcal{W}$ and its hidden state $\mathcal{Z}$ [13], while the Data mixture from observed incomplete-sequence and missing sequence state can express as below

$$\mathcal{X} = \begin{cases} \mathcal{W} = \{x_1, \ldots, x_T\} \text{ observed squence of length } T \\ \mathcal{Z} = \blacksquare = \{z_1, \ldots, z_T\} \text{ hidden state} \end{cases}$$

Under additional assumption of Markov property, z_t and $x_t|z_t$ is conditional independent of other x_s, for all $s \neq t$. The state probability $\pi = P(z_1)$ and transition probability $P_{ij} = P(z_t|z_{t-1})$ could formulate a joint-likelihood function $f(\mathcal{W}, \mathcal{Z}|\theta) = \prod_{t=1}^{T} p(x_t|z_t, \theta)P(z_1)\prod_{t=2}^{T} P(z_t|z_{t-1})$, which become the objective of HMM, different from log-likelihood function of standard EM.

5.4 M-M: Maximization-Maximization

This method replace the first estimation step by another maximization step from the standard EM, and forms the M-M algorithm. It aims to maximize a better lower bound $F(\tilde{P}, \theta) = L(\theta) - D_{KL}(\tilde{P}||P_\theta)$ to true likelihood $L(\theta)$ [14], where $\tilde{P} \sim \tilde{p}(w, z)$ denote current guessed distribution of $(\mathcal{W}, \mathcal{Z})$ and $P_\theta \sim p(w, z|w, \theta)$ denote likelihood of dataset given θ. $D_{KL}(\tilde{P}||P_\theta)$ stands for Kullback-Leibler divergence.

The M-M method works by maximizing objective function $F(\tilde{P}, \theta)$ and iterating between $\tilde{P}$ and θ. Formalize in the Appendix 9.1.2.

5.5 MAP: Maximum a Posterior

MAP method maximize a posterior $\hat{\theta}_{\text{MAP}}$ by modifying maximization step of standard EM, such that $\hat{\theta}_{\text{MAP}} = \arg\max_\theta lnL(\theta|\mathcal{W}) + lnp(\theta)$, where $p(\theta)$ is prior distribution of θ. The method formalized in the Appendix 9.1.3 [15]

We could observe the difference from standard EM algorithm, MAP adding an extra $lnp(\theta)$ term.

6 Strength and Weakness of EM

EM algorithm is a widely applied method in various field, because of its simplicity on calculation. EM is also an easy tool to estimate parameter upon incomplete-data problem.

While EM algorithm still had its limitation, it need model assumption like MLE, as it is the background method for parameter estimation. EM have a rather slow convergence speed compare to other estimation method [3], and we could not know its exact converge rate. EM method could only find local optimum, it is possibly no result in higher dimension due to its form of conditional Expectation function. We need to check concavity of log-likelihood function or try various initial value to find its convergence.

7 Empirical Analysis

In order to thoroughly understand EM algorithm, we have conducted few empirical analysis on various dataset in above section, for example on Dengue simulated Poisson data and Geyser bivariate normal data. In this section, we are going to shortly introduce a package "mclust" in R programming language, which is the simplest way at-hand to try EM algorithm. "mclust" package is aimed for model clustering on Gaussian Mixture dataset, it apply EM algorithm for finite clustering, classification and density estimation.

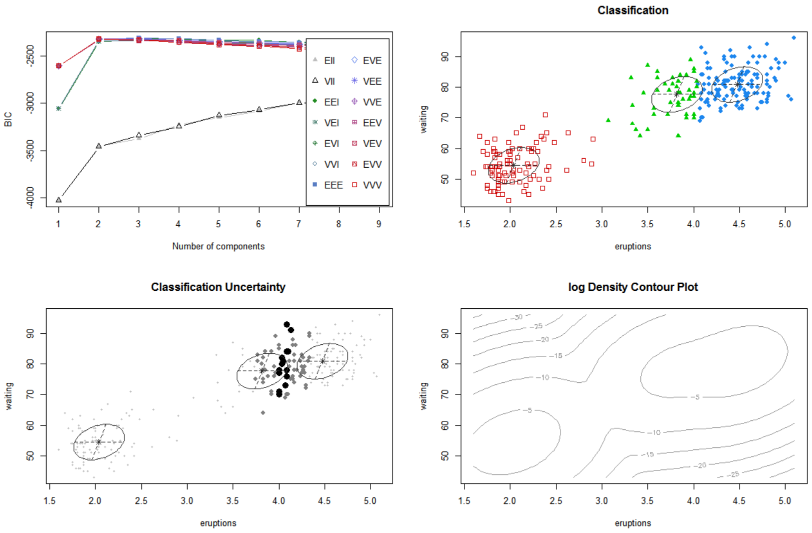

Figure 5: mclust on Geyser dataset

7.1 Example on mclust package: Faithful

We once again apply Old Faithful Geyser dataset on mclust package, the result first provide possible number of cluster (3 here, Figure 5), then plot the according clustering plot, possibility for each data point and density spline for it.

There will be two more extra examples on mclust package. It is analysis on Wreath dataset and Iris dataset in the Appendix 9.3.

From all result above generated by "mclust" package, we could find the readiness of mclust package on EM algorithm. It is the best way to explore EM algorithm than to program the iteration from scratch. Once the dataset is ready, it could help on clustering identifying, and density estimation.

8 Conclusion

EM algorithm is a widely using tool to estimate parameter upon missing data. The algorithm is based on MLE, but avoid the burden of recursive solution in parameter finding. EM method is well-developed and have versatile application upon different model assumption. One can find applicable method to furnish parameter estimation upon incomplete-data. However, despite all the benefit from EM algorithm, it still has its weakness. One could implement it but under consideration of its limitation at the same time.

9 Appendix

9.1 Equations & Methods

9.1.1 MLE solution compared with EM solution

We could first find the solution for ML estimation with missing value from Section 1.

$$\begin{cases} \hat{\beta}^{\text{iMLE}} &= \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\lambda}_i^{\text{iMLE}}} \\ \hat{\lambda}_1^{\text{iMLE}} &= \dfrac{y_1}{\hat{\beta}^{\text{iMLE}}} \\ \hat{\lambda}_j^{\text{iMLE}} &= \dfrac{x_j + y_j}{\hat{\beta}^{\text{iMLE}} + 1}, j = 2, \ldots, n \end{cases}$$

And the EM solution with missing value in the same model assumption.

$$\begin{cases} \hat{\beta}^{(2)} &= \dfrac{\sum_{i=1}^{n} y_i}{\lambda_1^{(1)} + \sum_{i=2}^{n} x_i} \\ \hat{\lambda}_1^{(2)} &= \dfrac{\lambda_1^{(1)} + y_1}{\hat{\beta}^{(2)} + 1} \\ \hat{\lambda}_j^{(2)} &= \dfrac{x_j + y_j}{\hat{\beta}^{(2)} + 1}, j = 2, \ldots, n \end{cases}$$

The EM solution is easily reached once provide initial value $\theta^{(1)}$, and similar to ML estimator without missing value in Section 1, but replace $x_1 = \blacksquare$ by first step parameter $\lambda_1^{(1)}$. On the other hand, ML estimator $\hat{\theta}^{\text{iMLE}}$ from incomplete-data is more complicated, we need to solve the recursion additionally.

9.1.2 M-M: Maximization-Maximization

The M-M method works by maximizing objective function $F(\tilde{P}, \theta)$ and iterating between $\tilde{P}$ and θ. Formalize as below,

M-step: Given estimate from m-th iteration $\theta^{(m)}$, maximize objective function $F(\tilde{P}, \theta^{(m)})$ to find

$$\tilde{P}^{(m+1)} = \arg\max_{\tilde{P}} F(\tilde{P}, \theta^{(m)})$$

M-step: Maximize $F(\tilde{P}^{(m+1)}, \theta)$ over θ,

$$\theta^{(m+1)} = \arg\max_{\theta} F(\tilde{P}^{(m+1)}, \theta)$$

9.1.3 MAP: Maximum a Posterior

The MAP method formalize as below, we could see the obvious difference to original EM algorithm.

E-step: Given estimate from m-th iteration $\theta^{(m)}$, calculate the conditional expectation Q-function,

$$Q(\theta|\theta^{(m)}) = E[lnL(\theta|\mathcal{W}, \mathcal{Z})|\theta^{(m)}, \mathcal{W}]$$

M-step: Find the (m+1)-th MAP θ,

$$\theta^{(m+1)} = \arg\max_{\theta}(Q(\theta|\theta^{(m)}) + lnp(\theta))$$

9.2 Proof

Theorem 3.1 Monotonic EM sequence
Let $Q(\theta|\theta^{(1)}) = E[lnL(\theta|\mathcal{W}, \mathcal{Z})|\theta^{(1)}, \mathcal{W}]$ Q-function be objective of maximization. If $Q(\theta|\theta^{(m)}) \geq Q(\theta^{(m)}|\theta^{(m)})$, then $lnL(\theta|\mathcal{W}) \geq lnL(\theta^{(m)}|\mathcal{W})$.

sketch of proof

$\theta = \arg\max_{\theta} Q(\theta|\theta^{(m)})$ implies that $Q(\theta|\theta^{(m)}) \geq Q(\theta^{(m)}|\theta^{(m)})\ \forall m$

we denote $K(\mathcal{Z}|\theta, \mathcal{W}) = \dfrac{f(\mathcal{W}, \mathcal{Z}|\theta)}{f(\mathcal{W}|\theta)}$, and rewrite the log-likelihood function

$$\begin{aligned} lnL(\theta|\mathcal{W}) =& Q(\theta|\theta^{(m)}) - E[ln(K(\mathcal{Z}|\theta, \mathcal{W}))|\theta^{(m)}, \mathcal{W}] \\ \geq& Q(\theta|\theta^{(m)}) - E[ln(K(\mathcal{Z}|\theta^{(m)}, \mathcal{W}))|\theta^{(m)}, \mathcal{W}] \\ \geq& Q(\theta^{(m)}|\theta^{(m)}) - E[ln(K(\mathcal{Z}|\theta^{(m)}, \mathcal{W}))|\theta^{(m)}, \mathcal{W}] \\ =& lnL(\theta^{(m)}|\mathcal{W}) \end{aligned}$$

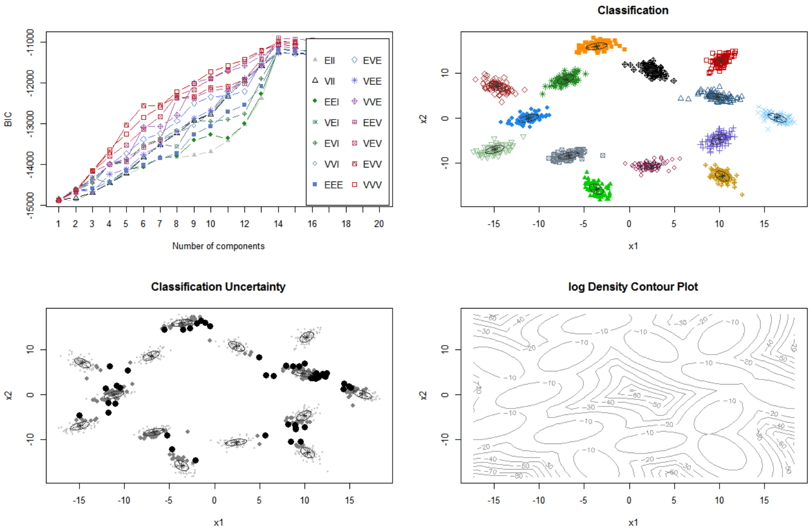

Figure 6: mclust on Wreath dataset

9.3 Empirical analysis

9.3.1 Extra example on mclust package: Wreath

Wreath dataset is bivariate normal with 14 clusters. From analysis result of mclust package, it can easily distinguish by EM algorithm as in Figure 6. This example shows that mclust package could easily handle cluster number more the 3.

9.3.2 Extra example on mclust package: Iris

In the famous Iris dataset, we could also easily separate its species by mclust package. We could observe the multivariate clustering from EM algorithm for more than one attribute ($p = 4$).

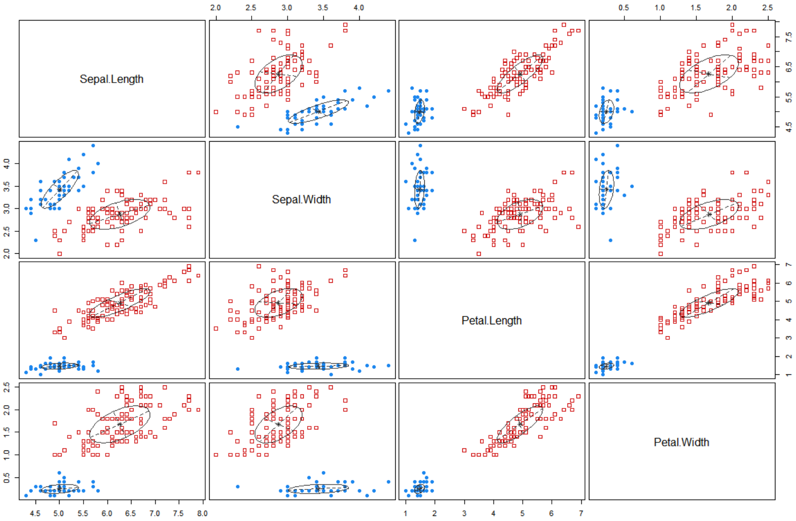

Figure 7: mclust on Iris dataset

References

[1] Taiwan Centers for Disease Control:
available on http://www.cdc.gov.tw, (2015)

[2] Casella, G. and Berger, R. L. (2001): Statistical Inference -2nd ed. Duxbury. ISBN 0-534-24312-6

[3] Xu, L. and Jordan, M. I.: On Convergence Properties of the EM Algorithm for Gaussian Mixtures. Neural Computation. Vol 8, 129-151 (1996)

[4] Hartley, H. O.: Maximum Likelihood Estimation from Incomplete Data. Biometrics, Vol. 14, No. 2, 174-194 (1958)

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B, Vol. 39, No. 1, 1-38 (1977)

[6] Boyles, R. A.: On the Convergence of the EM Algorithm. Journal of the Royal Statistical Society. Series B, Vol 45, 47-50 (1983)

[7] Wu, C. F. J.: On the Convergence Properties of the EM Algorithm. The Annuals of Statistics. Vol 11, 95-103 (1983)

[8] Little, R. J. A. and Rubin, D. B. (1987): Statistical Analysis with Missing Data. New York: Wiley. ISBN: 978-0-471-18386-0

[9] Tanner, M. A. (1996): Tools for Statistical Inference: Observed Data and Data Augmentation Method. New York: Springer-Verlag. ISBN 812034930x

[10] Chen, Y. and Gupta, M. R.: EM Demystified: An Expectation-Maximization Tutorial. University of Washington, Department of Electrical Engineering. (2010)

[11] Expectation Maximization algorithm, Wikipedia:
available on http://en.wikipedia.org

[12] 3mta3, Wikimedia Commons:
available on http://commons.wikimedia.org, (2009)

[13] Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, Vol 77, 257-286 (1989)

[14] Neal, R. M. and Hinton, G. E.: A View of the EM Algorithm That Justifies Incremental, Sparse and Other Variants. in M. I. Jordan (editor) Learning in Graphical Models. Dordrecht: Kluwer Academic Publishers, 355-368 (1998)

[15] Fraley, C. and Raftery, A. E.: Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Journal of Classification, Vol 24, 155-181 (2007)