Bayesian gaussian finite mixture model

J Mirra and S Abdullah
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
Corresponding author’s email: sarini@sci.ui.ac.id

Abstract. In data analysis it is common to assume that data are generated from one population. However, due to some underlying factors, data might come from several sources that should be considered as several sub-populations. Partitioning methods such as k-means clustering, as well as hierarchical clustering, are two commonly used methods for identifying data grouping. The grouping is done based on distance between observations. We discuss alternative means of data grouping, based on data distribution, known as Finite Mixture Model. Parameter estimation will be done using the Bayesian approach. Incorporation of prior distribution on the parameter of interest will complete data information from the sample, resulting in updated information in the form of posterior distribution. Gaussian distribution is assumed for the sampling model. Markov Chain Monte Carlo with Gibbs Sampling algorithm is implemented for sampling from the posterior distribution. Data on wave sensitivity on monkeys’ eyes were used to illustrate the method.

Keywords: Monte Carlo, prior distribution, posterior distribution, sampling model, simulation

1. Introduction
Statistical methods that are often used in data grouping include the k-means and hierarchical clustering method. These methods assume that data comes from one population, that is, data follows a certain distribution. Thus, the groupings are based solely on the distance between observations. We argue that this approach is not suitable for heterogeneous data. Partitioning data based on the distribution seemed to be more appropriate, as distribution preserve patterns in data, which is often ignored in methods that rely on distance measurement.

For example, as illustrated in our simulated data displayed in figure 1, data in the region of the blue box will be considered as one group, since their values are close, if we are using the distance-based method for grouping. However, it is clearly seen that part of this data is more likely under the density line of the first distribution (the one with the higher density for mode), and the rest is under the density of the second distribution (the one with the lower density for mode). Thus, a density-based clustering is sought for grouping this type of data.

Finite mixture models (FMM) [1] is flexible to model large heterogeneous data [2]. The heterogeneity might be due to some underlying, or latent, factors that assuming data comes from just one population may lead to unreliable analysis results. Although the number of groups, hereafter will be called sub-populations, are usually unknown, in FMM it is given. Therefore, some options for sub-population number will be given, starting from 1, and will stop until the optimal number is reached. The optimal number is determined using the Goodness of Model (GoF) fit information criterion.
While frequentist approach is common in conventional statistical parameter estimation, it relies heavily on data quality. For example, in maximum likelihood method, the estimated parameters are obtained by maximizing the likelihood function, which is formed by the joint distribution of the sample data. On the other hand, Bayesian approach offers a measure of flexibility in parameter estimation [3]. By flexibility we meant that the estimated parameters are obtained not based solely on data, but also taking into account other sources of information, such as experts’ opinions, or the results from similar previous studies to complete the information from a data sample. This additional information is known as prior, accommodated in the model through specification of prior distribution for the parameters of interest. Combining the prior information and sampling model (from sample data) results in a posterior distribution, that can be viewed as an updated information in which our inference will be based on.

Since in addition of sample data it also considers the prior information, we consider the Bayesian approach to be more comprehensive than the frequentist approach, and shall thus implement this method in this study. While others may argue that there is a possibility of unreliable prior information that will eventually lead to less accurate results compared to just relying on data, under the Bayesian procedure, the resulting posterior is optimal, considering the condition set by the prior [4]. Moreover, if the quality of prior information is questionable, we can always set the prior distribution to be vague, known as a flat or uninformative prior [5]; thus, the posterior will be dominated by the information from the sample data, as in frequentist approach. This property leads to the flexibility of the Bayesian approach.

An analytic solution to the posterior distribution is often difficult to obtain. Thus, we implemented numerical simulation through Markov Chain Monte Carlo (MCMC) with Gibbs Sampling to sample parameters from their posterior distribution [6-8]. As an example of application of Gaussian Bayesian FMM, we use data on the sensitivity level wave of monkeys’ eyes [9].

2. Statistical methods

2.1. Finite mixture model

Let data consists of $N$ observations on one numerical measurement, say $y$, that is, $y = (y_1, ..., y_n, ..., y_N)$. Assume that data might come from $K$ sub-populations ($K$ will be determined...
later) with each population is represented by the density \( p_k(\theta_k) \) with \( \theta_k \) is the parameter(s) of sub-population \( k, k = 1, ..., K \). Then, the finite mixture model is as follows:

\[
 f(y|\eta, \theta) = \sum_{k=1}^{K} \prod_{i=1}^{n} \eta_k p_k(y_i|\theta_k) \tag{1}
\]

where \( f(y|\eta, \theta) \) is the density function, \( \eta = (\eta_1, \eta_2, ..., \eta_K) \) is the weight of each sub-population, such that \( \sum_{k=1}^{K} \eta_k = 1, 0 \leq \eta_k \leq 1 \), and \( \theta = (\theta_1, \theta_2, ..., \theta_K) \) is the mean parameter in each subpopulation.

This equation describes the combining of probability density function of each formed sub-population. Each sub-population is assigned the proportion (weight) that reflects the chance of each group representing the overall data. Assuming Gaussian distribution for the sampling model, \( p_k(\theta_k) \) takes the form of a Gaussian probability density function, where the parameter \( \theta_k \) may differ between sub-populations.

According to the model, in order to assign each observation to the suitable sub-population, the parameter \( \eta_k \) needs to be estimated. Furthermore, once all the sub-populations have been formed, the characteristics of those sub-populations must be determined, thus requiring the estimation of parameter \( \theta_k \). Parameter estimation through Bayesian approach will be discussed in the subsequent section.

2.2. Bayesian gaussian finite mixture model

According to the sampling model in equation 1, the parameters to be estimated are the sub-population weights \( \eta_k \) and the sub-population characteristics \( \theta_k \). Thus, prior distribution as an initial information on those parameters should be specified, namely the prior distribution. Since \( \eta \) is the weight, it is reasonable to choose Dirichlet distribution [10] -which can be viewed as generalization of a two-class Bernoulli distribution for \( \eta \), that is

\[
 \eta \sim \text{Dir}(\alpha)
\]

where \( \alpha = (\alpha_1, ..., \alpha_K) \) is the hyper-parameter representing the initial belief on how to distribute all \( N \) observations into \( K \) sub-populations. The functional form of Dirichlet distribution is

\[
 p(\eta) \propto \prod_{k=1}^{K} \eta_k^{\alpha_k-1}, \text{where } \propto \text{ means proportional to. We only consider the functional form, and ignore the constant terms, due to the difficulty on estimating the combined constant terms later in the posterior distributions (known as the normalizing constant). However, as the distributional shape is determined by the functional form, and the solution is based on numerical simulation, this approach is acceptable.}

Meanwhile, the prior for \( \theta_k \), the mean of sub-population \( k \) is assumed to follow a Gaussian distribution with mean \( \theta_0 \) and variance \( \sigma^2_k \), as the following:

\[
 p(\theta_k) = \left( \frac{1}{\sqrt{2\pi} \sigma_k^2} \right)^k \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{k=1}^{m} (\theta_k - \theta_0)^2 \right\} \tag{3}
\]

Rewriting the sampling model in equation 1, we obtain the likelihood as the following:

\[
 p(y|\theta_k, \sigma^2_k) = \prod_{i=1}^{N} p(y_i|\theta_k, \sigma^2_k) = \sum_{k=1}^{K} \prod_{i=1}^{N} p(y_i|\theta_k, \sigma^2_k) \tag{4}
\]
Applying the Bayes rule,

$$p(\eta, \theta, y) \propto p(\theta)p(\eta) \times f(y|\theta, \eta)$$

we obtain the posterior distribution of \(\eta\) and \(\theta\) as the following:

$$p(\eta, \theta|y) \propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{k=1}^{K}(\theta_k - \theta_0)^2\right\} \cdot \prod_{k=1}^{K} \eta_k^{x_k-1} \cdot \exp\left\{-\frac{\sum_{j}(y_j - \theta_0)^2}{2\sigma^2}\right\} \cdot \prod_{k=1}^{K} \eta_k^{x_k}$$

(6)

Posterior distribution for proportion \(\eta\) follows a Dirichlet distribution, \(Dir(x + \alpha)\), and posterior distribution for mean \(\theta\) is a normal distribution. Samples of parameters will be drawn from the posterior distribution using numerical approach, through MCMC with Gibbs Sampling.

2.3. Markov Chain Monte Carlo

MCMC simulation is often used in Bayesian inference, especially if the posterior distributions of the parameters observed have a complicated form. Using MCMC will accrue to a sequence of random samples which correlates, that is, the value of the \(j\)-th sample of \(\theta\), for each of its element, \(\theta_k\), is updated by sample from the previous iteration. The sampling procedure is using Gibbs sampling, which the algorithm is presented below (algorithm 1).

**Algorithm 1. MCMC with Gibbs Sampling.**

**Input:** Initial values for \(\theta\) and \(\eta\): \(\theta_1^{(0)}, ..., \theta_K^{(0)},\) and \(\eta_1^{(0)}, \eta_2^{(0)}, ..., \eta_K^{(0)},\) and the number of iterations \(M\).

**Sampling Steps:**

Sample \(\theta_1^{(1)}\) from a complete posterior distribution of \((\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_K^{(0)}, \eta_1^{(0)}, ..., \eta_K^{(0)}, y)\).

Sample \(\theta_2^{(1)}\) from a complete posterior distribution of \((\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_K^{(0)}, \eta_1^{(0)}, ..., \eta_K^{(0)}, y)\).

...

Sample \(\theta_K^{(1)}\) from a complete posterior distribution of \((\theta_K|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{K-1}^{(1)}, \eta_1^{(0)}, ..., \eta_K^{(0)}, y)\).

Sample \(\eta_1^{(1)}\) from a complete posterior distribution of \((\eta_1|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{K-1}^{(1)}, \eta_2^{(1)}, ..., \eta_K^{(0)}, y)\).

Sample \(\eta_2^{(1)}\) from a complete posterior distribution of \((\eta_2|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{K-1}^{(1)}, \eta_1^{(1)}, \eta_3^{(0)}, ..., \eta_K^{(0)}, y)\).

...

Sample \(\eta_K^{(1)}\) from a complete posterior distribution of \((\eta_K|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{K-1}^{(1)}, \eta_1^{(1)}, ..., \eta_{K-1}^{(1)}, \eta_K^{(0)}, y)\).

**Repeat** Sampling steps until \(M\) iterations are completed.

**Output:** A sequence of \(\{\theta_m\}_{m=1}^{M}\) and \(\{\eta_m\}_{m=1}^{M}\).
Not all samples are used for inference. The first \( M_0 < M \) samples were considered as burn-in; usually the values are still fluctuating (not convergent). The remaining \( M - M_0 \) samples that show convergence will be used for further analysis.

Since the number of sub population is specified when building the model, an evaluation on the fitness of the model needs to be done. Model assessment is conducted using Akaike’s Information Criterion (AIC). The AIC for model with \( K \) sub-population is,

\[
AIC_K = -2 \log(p(y|\eta^*, \theta^*)) + 2v_K
\]  

(7)

where \(-2 \log(p(y|\eta^*, \theta^*))\) is the fitted model’s deviance measuring the fitness of the model to data \((\eta^*, \theta^*)\) are the estimated values of \((\eta, \theta)\), and \(2v_K\) is the number of parameters to be estimated, measuring the complexity of the model. A small value of AIC indicates fitness of the model; thus, an optimal model is the result of a trade-off between maintaining model fitness and simplicity.

3. Results and discussion

We implemented the method on wavelength sensitivity in a monkey’s eyes [9], consists of 48 measurements. A sample of 150,000 iterations were drawn from the posterior distribution, including 50,000 first iterations as burn-in. The number of sub-populations was initially set to 2 and was increased one at a time, then stopped when the optimum number of sub-populations is reached. The AIC for each model was recorded. Fitting model with the assumption that data come from 2 sub population, we obtained the following result, as summarized in table 1.

Around 60% of the data was more likely to be assigned to sub-population 1, the group with mean sensitivity wavelength of 536.8, with its 95% range of values (credibility interval) from 535.1 to 538.6. The remaining 40% of the data was in sub-population 2, a group with a higher sensitivity wavelength at 548.9 (mean), with the 95% credibility interval from 546.2 to 551.2.

The non-overlap of the coverage values of both sub populations, as also shown in figure 2, implies the goodness of fit of the model. The AIC of 271.9 cannot be solely used to assess the model’s fit, as it needs to be compared with the other model for the relative comparison. However, checking on the convergence of the iteration, FMM with 2 sub population seems fit the data well, as the implied by the unimodality of density of each parameter shown in figure 3.

Table 1. Estimated model’s parameters for FMM with 2 sub population. P[.] is the proportion of data that assigned to sub population [.], lambda[.] is the mean of sensitivity wavelength for sub population [.], Std. dev. is the standard deviation of the estimates.

| Node   | Mean | Std. dev. | 2.5%  | Median | 97.5%  |
|--------|------|-----------|-------|--------|--------|
| P[1]   | 0.5983 | 0.087    | 0.4199 | 0.601  | 0.7599 |
| P[2]   | 0.4017 | 0.087    | 0.2407 | 0.399  | 0.5803 |
| Deviance | 266.9 | 10.86    | 255.8 | 263.9  | 299.0  |
| lambda[1] | 536.8 | 0.9065   | 535.1 | 536.7  | 538.6  |
| lambda[2] | 548.9 | 1.295    | 546.2 | 548.9  | 551.2  |
| Sigma  | 3.777 | 0.6311   | 2.935 | 3.664  | 5.441  |
Figure 2. Boxplot of wavelengths sensitivity values (lambda) in two sub populations based on FMM.

Figure 3. Density plot of model’s parameters for FMM with $K = 2$. x axis represents sample of size 100,000 of the corresponding parameter from its posterior distribution, the y axis represents the density of the parameter. $P[1]$ and $P[2]$ are the weights for sub-populations 1 and 2, respectively; deviance is the deviance associated with the corresponding formed sub-populations, lambda [1] and lambda [2] are the wavelength sensitivity values for sub-populations 1 and 2, respectively; sigma is the standard deviation of wavelength sensitivity in each sub-population.

Proceeding further with $K = 3$, assuming that data might be better modelled with 3 sub-populations, the distribution of wavelength sensitivity in those three sub-populations is shown in figure 4. According to figure 4, it is seen that there are overlapping between sub-population 1 and sub-population 2, and also between sub-population 2 and sub-population 3. The AIC is 263, smaller than the AIC for model with 2 sub-populations, yet the difference seems insignificant.

Further checking on the characteristics of each sub-population is summarized in table 2. The 95% credible interval for wavelength sensitivity in sub-population 1 is between 532 and 537.9, overlapping with the credible interval for sub-population 2 (536.5 to 548.2). A similar condition occurs for data in sub-populations 2 and 3. Therefore, we stopped fitting the model to 3 sub-populations, and conclude that a 2 sub-population model is a better fit to the data.
Figure 4. Distribution of wavelength sensitivity with assumption that data was from three sub-populations. Lambda is the wavelength sensitivity measurement in the sub-populations.

Table 2. Estimated model’s parameters for FMM with 3 sub-populations. P[.] is the proportion of data that assigned to sub population [,]; lambda [,] is the mean sensitivity wavelength for sub-population [,]. Std. dev. is the standard deviation of the estimates.

| Node   | Mean  | Std. dev. | 2.5%    | Median | 97.5% |
|--------|-------|-----------|---------|--------|-------|
| P[1]   | 0.4227| 0.1538    | 0.07998 | 0.4408 | 0.675 |
| P[2]   | 0.2377| 0.143     | 0.01424 | 0.2233 | 0.5574|
| P[3]   | 0.3395| 0.09233   | 0.1588  | 0.3393 | 0.5211|
| Deviance| 255.7 | 16.96     | 224.9   | 257.0  | 291.6 |
| lambda[1]| 535.5 | 1.525     | 532.0   | 535.6  | 537.9 |
| lambda[2]| 541.2 | 3.073     | 536.5   | 541.0  | 548.2 |
| lambda[3]| 549.6 | 1.29      | 547.1   | 549.7  | 551.9 |
| Sigma  | 3.3   | 0.7202    | 2.206   | 3.23   | 4.992 |

4. Conclusion
In this study we demonstrated data groupings based on density through the finite mixture model, with Bayesian approach. As the number of optimal sub-populations is not known, then several options should fit, starting with 2 sub-populations and adding one at a time. The model is then further assessed by the trade-off between the information criterion and interpretability of the resulting sub-population, that is, the optimal sub-groupings should not overlap, as well as the convergence of the iterations. Implementation on wavelength sensitivity on a monkey’s eyes resulted in 2 sub-populations, implying that data was more likely to be generated by two different processes.
References
[1] McLachlan G J and Peel D 2004 Finite Mixture Models (London: John Wiley & Sons)
[2] Kappe E, DeSarbo W S and Medeiros M C 2019 J. Bus. Econ. Stat. 38 1-24
[3] Wagenmakers E J, Lee M, Lodewyckx T and Iverson G J 2008 Bayesian Evaluation of Informative Hypotheses (New York: Springer) pp 181-207
[4] Hoff P D 2009 A First Course in Bayesian Statistical Methods Volume 580 (New York: Springer)
[5] Lambert P C, Sutton A J, Burton P R, Abrams K R and Jones D R 2005 Stat. Med. 24 2401-28
[6] Smith A F and Roberts G O 1993 J. R. Stat. Soc. Ser. B Methodol. 55 3-23
[7] Gelfand A E 2000 J. Am. Stat. Assoc. 95 1300-4
[8] Carlo C M 2004 Markov Chain Monte Carlo and Gibbs Sampling (New York: Springer Verlag) p. 581
[9] Bowmaker J K, Jacobs G H and Mollon J D 1987 Proc. Royal Soc. B 231 383-90
[10] Huang J 2005 Maximum Likelihood Estimation of Dirichlet Distribution Parameters available at https://www.semanticscholar.org/paper/Maximum-Likelihood-Estimation-of-Dirichlet-Huang/f2d1ae94b3ba484d89b95365b5500a1ef0c87150