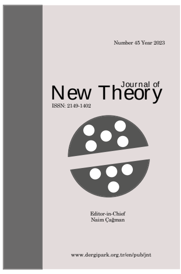
Journal of
New Theory
ISSN: 2149-1402

45 (2023) 57-72
Journal of New Theory
https://dergipark.org.tr/en/pub/jnt
Open Access



On Finite and Non-Finite Bayesian Mixture Models

Rasaki Olawale Olanrewaju[1], Sodiq Adejare Olanrewaju[2], Adedeji Adigun Oyinloye[3], Wasiu Adesoji Adepoju[4]

Article Info
Received: 21 Sep 2023
Accepted: 23 Nov 2023
Published: 31 Dec 2023
doi:10.53570/jnt.1358754
Research Article

Abstract − In this paper, a Bayesian paradigm of a mixture model with finite and non-finite components is expounded for a generic prior and likelihood that can be of any distributional random noise. The mixture model consists of stylized properties-proportional allocation, sample size allocation, and latent (unobserved) variable for similar probabilistic generalization. The Expectation-Maximization (EM) algorithm technique of parameter estimation was adopted to estimate the stated stylized parameters. The Markov Chain Monte Carlo (MCMC) and Metropolis–Hastings sampler algorithms were adopted as an alternative to the EM algorithm when it is not analytically feasible, that is, when the unobserved variable cannot be replaced by imposed expectations (means) and when there is need for correction of exploration of posterior distribution by means of acceptance ratio quantity, respectively. Label switching for exchangeability of posterior distribution via truncated or alternating prior distributional form was imposed on the posterior distribution for robust tailoring inference through Maximum a Posterior (MAP) index. In conclusion, it was deduced via simulation study that the number of components grows large for all permutations to be considered for subsample permutations.

Keywords *Bayesian paradigm, expectation-maximization, MCMC, proportional allocation, metropolis–hastings*

Mathematics Subject Classification (2020) 62F15, 62H30

1. Introduction

The essence of Bayesian methods, inference, and statistics in time-series analysis, mixture models, econometric, machine learning, and data science had received great attention since the propounded of Bayes' theorem by Thomas Bayes in the early 1980s [1]. Its great advantage is via its simplicity and bedrock of simplifying Bayes' theorem in a discretized or continuous form. However, it provides subjective pre-judgment and pre-knowledge (prior information) about the data that is to be incorporated into likelihoods, in order to aid sequential learning, decision-making, and prediction [2]. In other words, an inescapable requirement of Bayesian methods is to rightfully specify prior distributions for all involved parameters in the model that are usually regarded as unknown quantities. However, there have been debates and controversies over the choice of priors that truly advocate for likelihood(s) of interest [3]. Although, different types of priors have been proposed, types like, conjugate and non-conjugate priors; horseshoe prior, improper priors (otherwise refer to

[1]olanrewaju_rasaq@yahoo.com (Corresponding Author); [2]sodiqadejare19@gmail; [3]adedeji.oyinloye@bouesti.edu.ng; [4]adepojuwasiu@gmail.com
[1]Department of Business Analytics and Value Network, Africa Business School, Mohammed VI Polytechnic University, Rabat, Morocco
[2]Department of Statistics, Faculty of Science, University of Ibadan, Ibadan, Nigeria
[3]Department of Mathematical Sciences, College of Sciences, School of Pure and Applied Sciences, Bamidele Olumilua University of Education, Science and Technology, Ikere-Ekiti, Nigeria
[4]Department of Mathematics Education, Faculty of Education, University of Ibadan, Ibadan, Nigeria

as σ-finite measure), Zellner's *G*-prior, *G*-Priors, non-informative *G*-Priors, Jeffreys' prior, but there has not been a clear distinction as regards the ideal one [4]. However, prior distributions to be considered depend on how they inclusively or exclusively contained in the mixture model, the model to be considered, their involvements in the likelihood coefficients; and their bearing resulting inference via sensitivity analysis. Consequently, it is not all the time that conjugate priors defined for certain likelihoods do give posterior forms of the likelihoods. In addition, the deduction that some non-informative priors usually accompany undefined posteriors irrespective of the sample size is a clear indicator of the complexity of Bayesian inference for some models [5].

According to [6], prior distributions are being specified based on principles, relying on asymptotes, approximations, algorithms' flexibilities, and ignorance about the parameters. This makes it feasible for emergence of any inferential probabilistic prior with its corresponding likelihood to yield closed form solution and limiting distribution for the embedded parameters. Contrary to the adoption of priors' principles, [7] proposed Approximate Bayesian Computation–Population Monte Carlo (ABC–PMC) algorithm as an alternative technique for finite mixture model inferential. [7] adopted a kernel function as a substitute for prior distribution and explicitly highlighted how the problem of label switching can be solved with the use of the adopted kernel. In extension, [8] adopted Bayes factor to find required number of K-components that will be associated to a finite mixture model. The adopted Bayes factor ratio was incorporated in parametric family of finite mixtures and that of nonparametric via `strongly identifiable' Dirichlet Process Mixture (DPM) model and inferred that scalable evidence estimation technique for non-conjugate Dirichlet Process mixtures will be needed to derive the parametric and nonparametric processes. Prediction is one of the key factors that distinguished Bayesian paradigm, because of its ability to take into account all involved parameters and integrate them into posterior distribution in order to iteratively estimate their reliable and inferential solutions [9].

The continual usage of Bayesian methods in mixture models and econometrics in general, is because of the repercussion in flexibility and efficient algorithms used in conducting inferential inference through estimating unknown quantities. However, various powerful Bayesian computational techniques have been designed to estimate posterior solutions analytically and intractably for dimensional models. Among the techniques designed are Markov Chain Monte Carlo (MCMC) methods, Gibbs sampler and the Metropolis-Hastings sampler, Arnason–Schwarz Gibbs Sampler; and Stan implemented Hamiltonian Monte Carlo called "Stan" algorithms. The MCMC method is a veritable revolution and procedure for implementing a class of computational algorithms that can be easily applied to almost every model. The idea behind MCMC method is to generate analytically intractable posterior solutions via Markov Chain that converge to a chain of selections from the posterior distribution [10,11]. Once one of the chain drawings is available or successful, predictive inference will also be achieved. There are various ways of designing Markov Chain depending on the structure of the problem, once the chain exists, Gibbs sampler and the Metropolis-Hastings sampler; Arnason–Schwarz Gibbs Sampler; or Stan implemented Hamiltonian Monte Carlo can then be employed. Adding of additional auxiliary variable(s) (data augmentation) by the sampler (MCMC sampler) usually facilitate the implementation and analysis to be conducted on the augmented domain on not only the unknown quantities (model parameters), but also on unobserved variables (latent variables) and missing observational structure [12, 13]. This accessible reference and ability to incorporate additional data augmentation by MCMC sampler makes Bayesian paradigm to be the feasible method for mixture models (mixture models that involve incorporation of latent variable for regime switching, sample size allocation, and proportional allocation (mixing weights)).

This article covers a wide range of mixture models for simple probability distribution to be made more complex and less informative by a mechanism that combines several known or unknown same distribution. This composition is what is called mixture model or mixture of distributions. Inference made about the known and unknown quantities (parameters) of the ingredients of the mixture model and proportional allocations (mixing weights) is what is usually referred to as mixture estimation. In relation to machine learning, the repossession of the source distribution of each observation from the mixture of distributions is usually termed as classification (that is, distinguishing unsupervised classification from supervised classification). This technique requires advanced and sophisticated computational tools since the composition of the posterior distribution might not be easily computed. However, this article covers theoretical cases, as well as simulation studies for generic finite and non-finite Bayesian mixture models for common likelihoods with their prior distributions for known and unknown number of components, proportional allocation, and allotted sample size for specific approximation of Expectation-Maximization (EM) parameter estimation. The EM parameter estimation technique will be alternatively updated via MCMC, Gibbs sampler and the Metropolis-Hastings algorithms. The problem of identifying exchangeable posterior distribution will be treated via label switching, with its associated total allocation sample size being carve-out via Monte Carlo approximation with the use of Maximum a Posteriori (MAP) estimator.

2. Method

[14-16] propounded that a finite mingle of mixture model of similar probabilistic distribution to be a generalization of,

$$h_i(y) = \lambda_1 g_1(y) + \lambda_2 g_2(y) + \cdots + \lambda_k g_k(y) \tag{1}$$

$$h_i(y) = \sum_{i=1}^{k} \lambda_i \, g_i(y) \tag{2}$$

That is, a mixture model of same distribution is nothing but a convex combination, such that, $h_i(y)$ is the complete function of the mixture generalization model, with $g_i(y)(i \in \{1,2,3,\cdots,k\})$ being the known and inferential probabilistic distribution for each allocation (λ_k) with their corresponding unknown proportions on sample points $(y_1,\cdots,y_n)$ on g_i components, such that, $\lambda_i \geq 0 \ni \sum_{i=1}^{k} \lambda_i \approx 1$, for $i \in \{1,2,3,\cdots,k\}$, for a drawn or selected sample size of size (n).

However, in a parametric setting, where $g_i(y)(i \in \{1,2,3,\cdots,k\})$ can take any distributional form, forms like Gaussian, exponential, Beta, or student-*t* distribution with unknown coefficients (parameters, say ϑ_i), Equation 2 can then be rewritten as:

$$h(y) = \sum_{i=1}^{k} \lambda_i \, g_i(y \,|\, \vartheta_i) \tag{3}$$

With λ_i as the mixing weights (or proportional allocation) and ϑ_i the component coefficients for $(i \in \{1,2,3,\cdots,k\})$. The idea of mixing allocations from a parametric point of view makes it possible to associate component coefficients with missing data structure (unobserved or latent variable), while in a subjective prejudgment manner (Bayesian paradigm), they are known to be related observations. This noticeable assertion might not be germane in a computational setting that involves likelihood function or construction of prior distribution, pertinent in the interpretation of posterior results.

One of the reasons (motivations) for constructing mixtures of same distributions is to usefully extend "standard" distributions statistically in an approach that envisions observations as several unobserved (latent) sub-populations (strata). Conditioning it on the setting, the inferential goal is to associate selected samples (n_i) or drawings from mixtures of finite or non-finite, but with components of the same distribution to reassemble selected groups (usually refer to as cluster) by estimating the unobserved component, say "s", to provide estimators for unknown coefficients for several groups, or to estimate the number of k -groups.

2.1. Procedure for Generic Mixture Likelihoods and Posteriors Drawings

Assuming an Independent and Identically Distributed (IID) sample drawings of $(y_1, \cdots, y_n)$ was drawn from the mixture of distributions with, proportional allocations, and component parameters of Equation 3. Then, the likelihood is such that,

$$\ell(\vartheta, \gamma|y) = \prod_{j=1}^{n} \sum_{i=1}^{k} \lambda_i g(y_j|\vartheta_i) \tag{4}$$

Equation 4 is the likelihood that contains k^n-terms of $\lambda_i g(y_j|\vartheta_i)$. The computational acumen of Equation 4 depends on the feasibility of the order $O(nk)$ of Equation 3 that can cater for the analytical solution of either the Bayes estimators or Maximum Likelihood (ML) estimators.

Let $T(\vartheta, \lambda)$ be the prior distribution of any form for the likelihood distribution, then the posterior distribution of $(\vartheta, \lambda|y)$ can be added-up for a multiplicative constant, say,

$$T(\vartheta, \lambda|y) \infty \left(\prod_{j=1}^{n} \sum_{i=1}^{k} \lambda_i g\big(y_j\big|\vartheta_i\big) \right) T(\vartheta, \lambda) \tag{5}$$

For $T(\vartheta, \lambda|y)$ that can be computed for guess values of the parameters in $T(\vartheta, \lambda)$ at a contrivance order of $O(nk)$. The derivation of $T(\vartheta, \lambda|y)$ posterior outputs and expectations (means) of the mixture distribution coefficients of interest can only be achieved in an exponential time order of $O(nk)$. Incorporating the latent (unobserved or missing variable) intuition into the mixture posterior distribution of Equation 5 for each y_i in association to the latent variable "s" indicated a Markov chain component of distribution that can be generated. This makes it to be seen as hierarchical structure associated with the mixture model to be:

$$s_i|\lambda \sim M_k(\lambda_1, \cdots, \lambda_k)$$

where M_k denotes the Multinomial distribution for $y_i|\, s_i, \vartheta \sim g\big(y|\vartheta_{s_i}\big)$.

The complete data, that is, (y_i, s_i) is the complete likelihood corresponding to the unobserved variable, s_i is

$$\ell(\vartheta, \lambda|y, s) = \prod_{j=1}^{n} \lambda_{s_j} g(y_j|\vartheta_{s_j}) \tag{6}$$

Then, the posterior distribution that added-up for a multiplicative constant is given as

$$T(\vartheta, \lambda|y, s) \infty \left(\prod_{j=1}^{n} \lambda_{s_j} g(y_j|\vartheta_{s_j}) \right) T(\vartheta, \lambda) \tag{7}$$

where $s = \{s_1, s_2, s_3, \cdots, s_n\}$ for k^n-terms of $\Xi = \{1,2,3, \cdots, k\}^n$ possible values of the specified vector of "s". Having ascertained proportional allocation for each mixture distribution to be (λ_k). In a similar manner, sample size allocation can also be ascertained via partitioning (decomposition) to R via $\mathrm{R} = \bigcup_{i=1}^{r} \Xi_i$ for a given allocation size vector of $(n_1, n_2, n_3, \cdots, n_k)$, where, $\sum_{i=1}^{k} n_k$, the number of observations allotted to each component, then partition sets can be worked-out as:

$$\Xi_j = \left\{ s\colon \sum_{i=1}^{n} I_{s_i=1}, \cdots, \sum_{i=1}^{n} I_{s_k=k} \right\} \tag{8}$$

Ξ_j consist of all proportional allocations with their corresponding or given allocation sizes $(n_1, n_2, n_3, \cdots, n_k)$, such that, the partitioning sets with $j = (n_1, n_2, n_3, \cdots, n_k)j$ can be conceptualized as a lexicographical ordering of $(n_1, n_2, n_3, \cdots, n_k)$'s.

$$j = 1, \quad (n_1, n_2, n_3, \cdots, n_k) = (n, 0, \cdots, 0)$$

$$j = 2, \quad (n_1, n_2, n_3, \cdots, n_k) = (n - 1, 1, \cdots, 0)$$

$$j = 3, \quad (n_1, n_2, n_3, \cdots, n_k) = (n - 1, 0, 1, \cdots, 0)$$

and

$$j = 4, \quad (n_1, n_2, n_3, \cdots, n_k) = (n - 1, 1, 0, 1, \cdots, 0)$$

So that the posterior distribution of (θ, λ) can be rewritten in a close form as:

$$T(\theta, \lambda|y) = \sum_{j=1}^{r} \sum_{s \in \Xi_r} \eta(s) T(\theta, \lambda|y, s) = \sum_{s \in \Xi} T(\theta, \lambda|y, s) \tag{9}$$

Such that $\eta(s)$ is the marginal posterior likelihood of the allotted "s" conditioned on sample points' domain of "y". This can be derived after integrating out "ϑ" and "λ", such that, the close form of the Bayes estimator of (ϑ, λ) is

$$E^T[\vartheta, \lambda|y] = \sum_{j=1}^{r} \sum_{s \in \Xi_r} \eta(s)\, E^T[\vartheta, \lambda|y] \tag{10}$$

Decomposing Equation 9, for an inferential point of view. It connotes that the posterior distribution takes into account each possible partition of "s" in the dataset and allocate a posterior likelihood of $\eta(s)$ to these partitions, as well construct a posterior distribution for the embedded coefficients conditioned on the allocations.

Employing the Expectation-Maximization (EM) algorithm procedure for completion of parameter estimation mechanism that involves latent (unobserved) variable. The "E" stands for expectation and "M" connotes maximization steps that involve convergence of local maximum of likelihood.

Iteratively in time variant "t", the E-step computational function corresponds to

$$Q\{(\vartheta^{(t)}, \lambda^{(t)}), (\vartheta, \lambda)\} = E_{(\vartheta^{(t)}, \lambda^{(t)})}[\log \ell\,(\vartheta, \lambda|y, s)|y] \tag{11}$$

$\log \ell\,(\vartheta, \lambda|y, s)$is regarded as the likelihood of the joint distribution of "y" and "s", such that, imposed means of the coefficients to be calculated under the conditional distribution of "s" given "y" for the value of $(\vartheta^{(t)}, \lambda^{(t)})$. The second case, which is the M-step, is the maximization of $Q\{(\vartheta^{(t)}, \lambda^{(t)}), (\vartheta, \lambda)\}$ in (ϑ, λ) with convergence solution of $(\vartheta^{(t+1)}, \lambda^{(t+1)})$ (see [17,18]).

2.2. MCMC Solutions as an Alternate for Expectation-Maximization (EM) Algorithm

In situation where the first step (that is, the E-step) of the EM-algorithm is not analytically feasible, that is, when the unobserved variable "s" cannot be replaced by imposed expectations (means) for the joint distribution of Equation 5, then the full conditional distribution of "s" given "y" will be evaluated as,

$$T(s|y, \vartheta, \lambda) \infty \prod_{j=1}^{n} \lambda_{s_j} g\,(y_i|\vartheta_{s_i}) \tag{12}$$

Equation 12 can be computed for guess value of $T(\vartheta,\lambda)$ at a contrivance order of $O(n)$ for standard distributions of $g(\cdot\,|\vartheta)$.

It is to be noted that if $T(\vartheta,\lambda)$ is a conjugate prior, then its full conditional posterior can also be worked-out via Gibbs sampler. Assuming "ϑ" and "λ" are independent a priori, then conditioning "s", the vectors "y" and "λ" are independent; that is,

$$T(\lambda|s,y)\infty T(\lambda)g(s\,|\,\lambda)g(y\,|\,s)\infty T(\lambda)g(s\,|\,\lambda)\infty)T(\lambda|s) \tag{13}$$

It is to be noted that "ϑ" is independent posterior of the form "λ" given "s" and "y", with density $T(\vartheta|s,y)$ for successive simulation of "s" and (λ,ϑ) conditional on one another, as well on the data points (y).

This implies that Gibbs sampler will be ideal under the umbrella of latent variable (unobserved variable) by simulation under data augmentation. The simulation of $\vartheta j's$ depends solely on sampling density of $g(\cdot\,|\vartheta)$ coupled with the prior, $T(\vartheta,\lambda)$.

The marginal distribution of s_i's is nothing but a multinomial distribution of $M_k(\lambda_1,\cdots,\lambda_k)$, which allow a conjugate prior on "λ", such that, $\lambda=(\lambda_1,\cdots,\lambda_k)$, where "$\lambda$" follows a Dirichet distribution, that is, $\lambda\sim\wp(\delta_1,\cdots,\delta_k)$ with density

$$\frac{\Gamma(\delta_1+\cdots+\delta_k)}{\Gamma(\delta_1)\cdots\Gamma(\delta_k)}\lambda_1^{\delta_1}\cdots\lambda_k^{\delta_k}$$

on k-real number line $\Re^k$,

$$\wp=\left\{(\lambda_1,\cdots,\lambda_k)\in[0,1]^k;\sum_{i=1}^{k}\lambda_i=1\right\} \tag{14}$$

Its own sample size allocation can be denoted as

$$n_i=\sum_{\iota=1}^{n}I_{s_\iota=i}\,(1\le i\le k)$$

for posterior distribution of "λ" given "s" that is,

$$\lambda|s\sim\wp(n_1+\delta_1,\cdots,n_k+\delta_k) \tag{15}$$

2.2.1. Algorithm for the Gibbs Sampler

Guess an initialization (that is starting values) for "λ" and "ϑ": That is, choose $\lambda^{(0)}$ and $\vartheta^{(0)}$ arbitrarily. Note that $0\le\lambda^{(0)}\le 1$.

Iteration $t(t\ge 1)$:

i. For $i\in\{1,2,3,\cdots,n\}$, generate $s_i^{(t)}\ni P(s_i=j|\lambda,\vartheta)\infty\lambda_j^{(t-1)}g\left(y_i|\vartheta_j^{(t-1)}\right)$

ii. Generate $\lambda^{(t)}$ according to $\left(\lambda|s^{(t)}\right)$

iii. Generate $\vartheta^{(t)}$ according to $\left(\vartheta|s^{(t)},y\right)$

The simulation of ϑ_j's exists only for conjugate prior. The intricacy in the simulation of the ϑ_j's depends on the sampling density $g(\cdot\,|\vartheta)$ as well as the prior distribution of $T(\vartheta,\lambda)$.

2.3. Metropolis–Hastings Algorithm as an Alternate to Expectation-Maximization (EM) and MCMC Algorithms

Knowing that the likelihood of mixture model is usually in a closed-form manner when computation is in $O(nk)$ order and time variant "t" and the posterior distribution is thus up to a multiplicative constant. One can alternatively switch to Metropolis–Hastings algorithm, as long as a new quantity "Υ" provides a correct exploration for the posterior distribution with acceptance ratio,

$$\gamma(y \mid \vartheta', \lambda') = \frac{T(\vartheta', \lambda' | y)}{T(\vartheta, \lambda | y)} \frac{\Upsilon(\vartheta, \lambda | \vartheta', \lambda')}{\Upsilon(\vartheta', \lambda' | \vartheta, \lambda)} \wedge 1 \tag{16}$$

computed in $O(nk)$ time.

2.3.1. Algorithm for the Metropolis–Hastings

Guess an initialization (that is starting values) for $y^{(0)}$: That is, choose $y^{(0)}$ arbitrarily.

Iteration $t (t \geq 1)$:

i. Given $y^{(t-1)}$, generate $\vartheta' \sim \lambda'\left(y^{(t-1)}, y\right)$

ii. Given $y^{(t-1)}$, generate $\lambda' \sim \vartheta'\left(y^{(t-1)}, y\right)$

iii. Compute $\gamma\left(y^{(t-1)} \mid \vartheta', \lambda'\right) = \min\left(\frac{T(\vartheta', \lambda' | y)}{T(\vartheta, \lambda | y)} \frac{\Upsilon(\vartheta, \lambda | \vartheta', \lambda')}{\Upsilon(\vartheta', \lambda' | \vartheta, \lambda)} \wedge 1\right)$

iv. With probability $\gamma\left(y^{(t-1)} \mid \vartheta', \lambda'\right)$, accept $\vartheta'$and set $y^{(t)} = \vartheta'$; or accept $\lambda'$and set $y^{(t)} = \lambda'$ otherwise reject ϑ' and set $y^{(t)} = y^{(t-1)}$ or reject λ' and set $y^{(t)} = y^{(t-1)}$

The distribution of ϑ', λ' is called the instrumental distribution for acceptance–rejection method. ϑ', λ' and T are proportionality constants in the calculation of "γ". The merit of this approach in comparison to Gibbs sampler is that it does not necessarily need the usage of conditional distributions of "T". The Metropolis-Hastings algorithm proposed that the distribution of ϑ' to provide correct exploration of the posterior surface, since the acceptance ratio

$$\frac{T(\vartheta', \lambda' | y)}{T(\vartheta, \lambda | y)} \frac{\Upsilon(\vartheta, \lambda | \vartheta', \lambda')}{\Upsilon(\vartheta', \lambda' | \vartheta, \lambda)} \wedge 1$$

2.4. Label Switching for Exchangeability of Posterior Distribution

In scenarios, where either robust alternative prior distribution is needed for more tailoring inference, label switching is what is termed to be required. The ability to identify exchangeable posterior distribution answers the problem of imposing identifiability restriction to estimate the unknown quantities (parameters). A typical example is by defining components via ordering means, allocation sample size, or proportional allocation in a mixture model. From Bayesian perspective, this is nothing but truncating the source or first used prior distribution from $T(\vartheta, \lambda)$ to

$$T(\vartheta, \lambda) I_{\vartheta_1 \leq \cdots \leq \vartheta_k} \tag{17}$$

The resolution to label switching problem is to shun imposition of the restriction mingling that consist arbitrarily drawing of the $k!$ (k factorial).

2.5. Monte Carlo Approximation for Estimating Maximum a Posteriori (MAP)

Given a MCMC of total allocation sample of size of say "N", We might be interested in finding the Monte Carlo approximation of the MAP estimator by taking $\vartheta^{(i*)}, \lambda^{(i*)}$, such that,

$$i^* = \underset{i=1,\cdots,N}{\arg\max} \, T\{(\lambda, \vartheta)^{(i)} \mid y\} \tag{18}$$

The approximate MAP estimate would act as pivot that yields good approximation for the mode and when reordering iterations with respect to the mode. It is to be noted that Equation 18 is for simulation value that produces maximal posterior density.

In case where the reordering is based on Euclidean distance in the parameter space domain of ϑ, one can employ the distance in the domain of the allotted proportions. Assuming Ψ_k is a k −permutation set and $r \in \Psi_k$, minimizing "r" in an entropy Euclidean distance by adding the relative entropies between $P\left(s_j = t|\vartheta^{(i*)}, \lambda^{(i*)}\right)'s$ and $P\left(s_j = t|\, r\{\vartheta^{(i)}, \lambda^{(i)}\}\right)$ such that,

$$f(i,r) = \sum_{j=1}^{n} \sum_{t=1}^{k} P\left(s_j = t|\vartheta^{(i*)}, \lambda^{(i*)}\right) \log \left[\frac{P\left(s_j = t|\vartheta^{(i*)}, \lambda^{(i*)}\right)}{P\left(s_j = t|\, r[\vartheta^{(i)}, \lambda^{(i)}]\right)}\right] \tag{19}$$

See the permutations of selection of reordering the MCMC output algorithm below.

```
Algorithm for Pivotal Reordering
For iteration of $i \in \{1,2,3,\cdots,N\}$
i. Compute $r_i = \underset{v' \in \Psi_k}{\arg\min} \, f(i,r)$
ii. Set $\left(\vartheta^{(i)}, \lambda^{(i)}\right) = r_i\left(\vartheta^{(i)}, \lambda^{(i)}\right)$
```

Therefore, after the reordering steps from Equation 16 to 18, the Monte Carlo estimate of the posterior expectation can be written as $E^T\left[\vartheta_j|y\right] = \sum_{i=1}^{N} \frac{\vartheta_j^{(i)}}{N}$, where $E^T\left[\vartheta_j|y\right]$ (or its approximation) can be compared with Ψ_k in order to check for convergence.

2.6. Mixtures with an Unknown Number of Components

It is to be noted that the number of homogeneous components (k-components) connotes the degree of approximation, and cannot be fixed in advance, except one ascertained the number of components (proportional allocation) via visualization or other detection techniques. Even from the classical approach perspective, the number of homogeneous clustering (usually via the mean) within the population of interest is usually not ascertained and first-hand inference is usually employed to determine the number of components. For example, in financial stock where the number of different patterns of studied stock evolution that may be unknown to analyst (unknown homogeneous components) (for more details, see [19,20]).

In this type of computational resolution, the number of models is infinite and requires special type of MCMC exploration with variability inference. Mixture models with unknown number of proportional allocations are usually referred to as variable dimensional models that require special simulation technique called reversible jump and collection of 2^k-sub-models. It usually requires high degree of formalization, sensitive calibration, and approximated marginal likelihoods in this kind of special case of mixture model. The enumeration of mixture model with unknown components depends on sampling approximation via the marginal likelihood of whole range of potentials models as

$$f_J(y|\varsigma_J) = \prod_{i=1}^{n}\sum_{j=1}^{J} \lambda_j \, f(y_i|\vartheta_j) \tag{20}$$

such that $\varsigma_J = (\vartheta, \lambda) = (\vartheta_1, \cdots, \vartheta_J, \lambda_1, \cdots, \lambda_J)$. $f_J(y|\varsigma_J)$ evolve round the marginal likelihood integral via the sampling approximation,

$$c_J(y) = \int f_J(y|\,\varsigma_J)T_J(\varsigma_J)\partial\varsigma_J \tag{21}$$

"J" connotes the model index, that is, the infinitesimal case of the components in a different representation that starts from another arbitrary density say d_J, then

$$\Gamma = \int d_J(\varsigma_J)\partial\varsigma_J = \int \frac{d_J(\varsigma_J)}{f_J(y|\varsigma_J)T_J(\varsigma_J)} f_J(y|\varsigma_J)T_J(\varsigma_J)\partial\varsigma_J \tag{22}$$

$$\Gamma = c_J(y) \int \frac{d_J(\varsigma_J)}{f_J(y|\varsigma_J)T_J(\varsigma_J)} T_J(\varsigma_J)\partial\varsigma_J \tag{23}$$

This connotes that the estimate of $c_J(y)$ is

$$c_J(\hat{y}) = \frac{1}{\frac{1}{T}\sum_{t=1}^{T}\left[\frac{d_J(\varsigma_J^{(t)})}{f_J(y|\varsigma_J^{(t)})T_J\left(\varsigma_J^{(t)}\right)}\right]} \tag{24}$$

where $\varsigma_J^{(t)}$ are products of the MCMC sampler pointed at $T_J\left(\varsigma_J^{(t)}\right)$.

3. Simulation Studies

A simulated dataset of 1000 observations from independent Gaussian randomly selected observations with true mean values $(\mu_1, \mu_2, \lambda_1) = (2.3,0,0.7)$ were considered, such that, Dirichet variates were used as prior, we consider a two-component Gaussian mixture model of

$$\lambda_1 N(\mu_1, 1) + \lambda_2 N(\mu_2, 1) \ni \lambda_2 = (1 - \lambda_1) \tag{25}$$

In this scenario, the Gaussian coefficients (parameters) are identifiable. This connotes that μ_1 and μ_2 cannot be bewildered for each other, such that, λ_1 is not equal to 0.5. If λ_1 is equal to 0.5, it implies that $\lambda[N(\mu_1, 1) + N(\mu_2, 1)]$. The log-likelihood surfaces of Figure 1 below give the image representation of Equation 25. Two modes were exhibited and expounded, such that the upper chamber with larger mode is noted to be closer to the neighborhood of the true values of the average coefficients simulated. The mode with the lower chamber possessed an inverse separation of the dataset of the two clusters. For better understanding of the lower chamber mode, if a limit of $\lambda_1 = \lambda_2 = 0.5$ is set, it means that there is high likelihood that the two equivalent modes will be approximately equal, that is $(\mu_1, \mu_2) = (\mu_2, \mu_1)$. If λ_1 will be different from 0.5, the lower chamber mode becomes smaller and smaller in comparison with the larger chamber. It is to be noted that the starting guessing points in both cases of μ_1 and μ_2 are saddled points between the two modes.

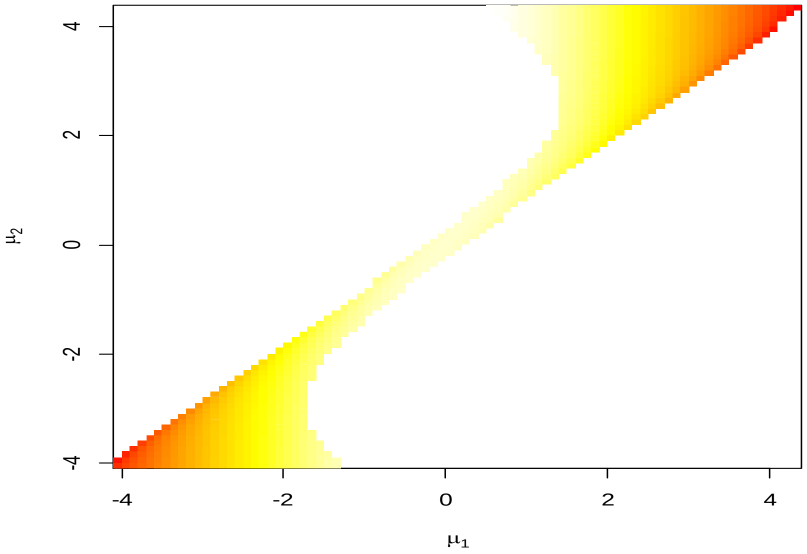


Figure 1. Log-likelihood of the mixture of distribution in Equation 25

A special case of Equation 25 is when two different independent Gaussian priors of $\mu_1 \sim \mathbb{N}(0,4)$ and $\mu_2 \sim \mathbb{N}(2,4)$ are considered from the simulated dataset, then the posterior allocation weight vector of "s" is given as

$$\overline{y}(s) = \frac{1}{1000}\sum_{i=1}^{1000} I_{s_i=1} y_i \text{ and } \overline{y}(s) = \frac{1}{1000-n_1}\sum_{i=1}^{1000} I_{s_i=2} y_i$$

Its variance is equal to

$$\overset{\wedge}{\sigma_1^2}(s) = \sum_{i=1}^{1000} I_{s_i=1}\left(y_i - \overline{y}_1(s)\right)^2 \text{ and } \overset{\wedge}{\sigma_2^2}(s) = \sum_{i=1}^{1000} I_{s_i=2}\left(y_i - \overline{y}_2(s)\right)^2$$

The log-likelihood of the posterior distribution was carved-out in Figure 2 below (contour plot function that exhibits an additional mode on the likelihood surface). It simply connotes that each of the two partitions of "s" of the simulated dataset, such that 0.0002 and 0.0006 allocates a posterior probability to each of the partition, and afterwards construct a posterior distribution for the conditional coefficients on μ_1 and μ_2. The conditional posterior distribution of the s_i's given (μ_1, μ_2), for $i \in \{1,2,3,\cdots,n\}$

$$P(s_i = 1|\mu_1, y_i) \infty \lambda \exp(-0.5(y_i - \mu_1)^2)$$

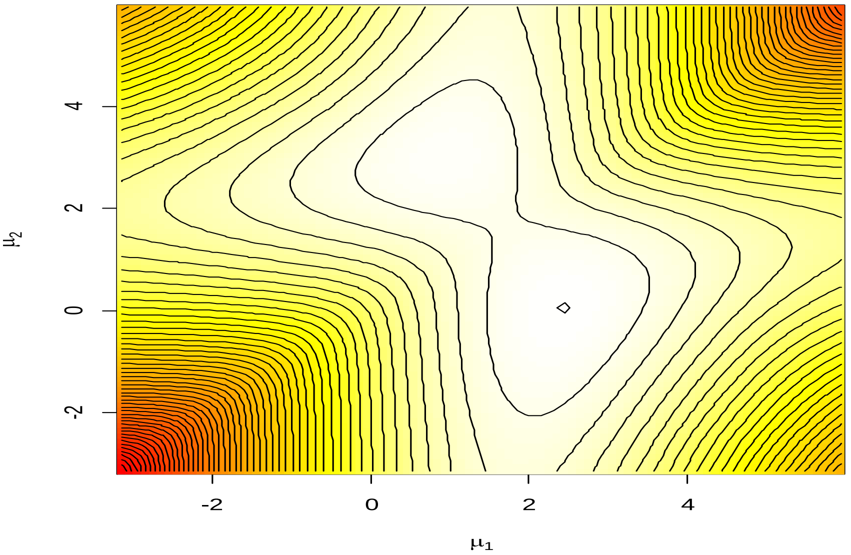


Figure 2. Contour plot of two different independent Gaussian priors of $\mu_1 \sim \mathbb{N}(0,4)$;$\mu_2 \sim \mathbb{N}(2,4)$

Considering a more general case of a mixture of two Gaussian distributions with their parameters unknown, such that, $\lambda_1 N(\mu_1, \sigma_1^2) + (1 - \lambda_1) N(\mu_2, \sigma_2^2)$ and for the conjugate prior distribution ($j \in \{1,2\}$). The same starting guessing points in both cases of μ_1 and μ_2 were the saddled points between the modes. Gaussian random walk of scaled unity was adopted because of its smaller magnitude require to paddle more iterations for proper modal region to be reached. For the posterior associated with Equation 25, the Gaussian random walk proposal is $\overset{\wedge}{\mu_1} \sim N\left(\mu_1^{(t-1)}, \gamma^2\right)$ and $\overset{\wedge}{\mu_2} \sim N\left(\mu_2^{(t-1)}, \gamma^2\right)$ which leads to the acceptance probability of

$$r = \min\left\{1, \frac{T\left(\overset{\wedge}{\mu_1}, \overset{\wedge}{\mu_2} | y\right)}{\left(\mu_1^{(t-1)}, \mu_2^{(t-1)} | y\right)}\right\}$$

"γ" was chosen to achieve a reasonable acceptance rate. However, Metropolis–Hastings algorithm checkmated the drawback of Gibbs sampler of insignificantly smaller index that can trap in phenomenon as the scale. This corresponds to $0.25N(2.35,1) + 0.75N(0.02,1)$ of the likelihood surface of Figure 3. The Gibbs sampler was based on 10,000 iterations in agreement with the likelihood surface. It was deduced that the Gibbs sampler ended-up in trapping the lower mode.

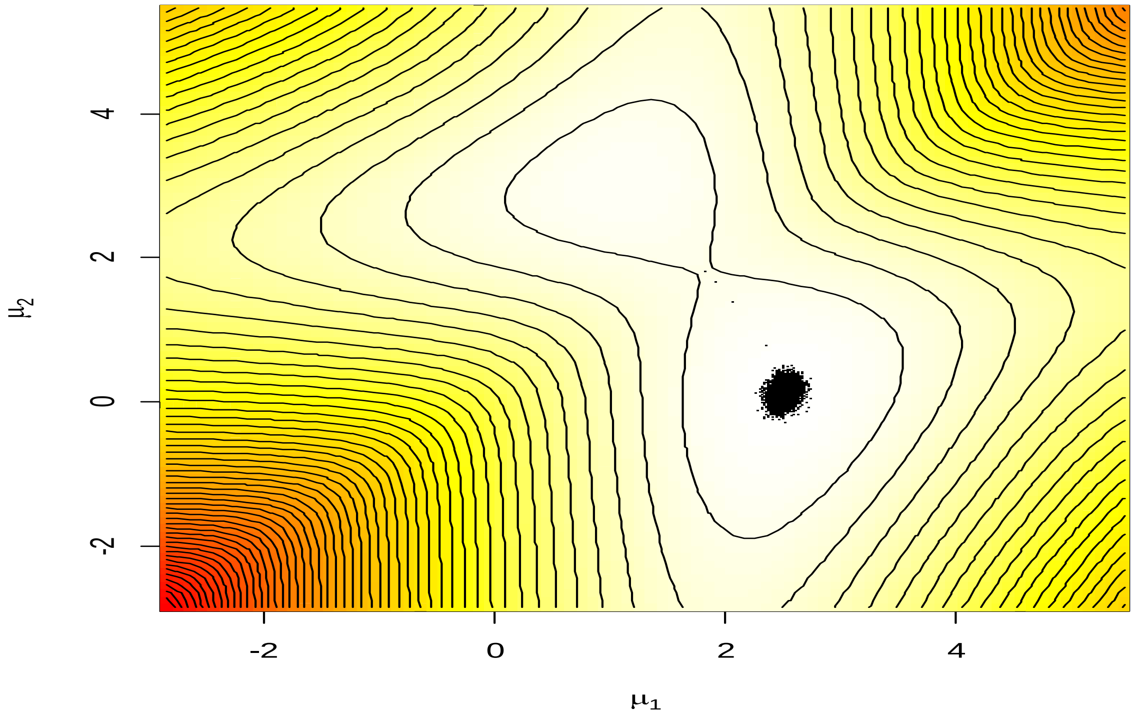


Figure 3. Log-likelihood surface and the corresponding Gibbs sampler for the model, based on 10,000 iterations

It is to be noted that the starting point of the Gibbs sampler in Figure 3 is ($\mu_1 = 0.005$ and μ_2 =0.005). It clearly indicates that the unconstrained random walk of Metropolis-Hastings remains justifiable for constrained parameters, but not efficient when the Markov Chain moves closer to the boundary of the parameter domain of Figure 4. It also needs to be noted that the parameter domain moves slowly by conditioning the proposed values to be incompatible with the constraints, thus leading to the rejection of the Metropolis-Hastings acceptance ratio. For label switching under invariant permutation indices of components, the Gaussian mixture of $0.25N(2.35,1) + 0.75N(0.002,1)$ and $0.75N(0.002,1) + 0.25N(2.35,1)$ are similar. This does not tantamount to $0.75N(0.002,1)$ distribution that can be called the first component of the mixture model. However, the component parameters ϑ_i are not identifiable marginal, such that, $\vartheta_1 = 0.002$ maybe 2.35 as well. In this case, the quantities (ϑ_1, λ_1) and (ϑ_2, λ_2) are exchangeable.

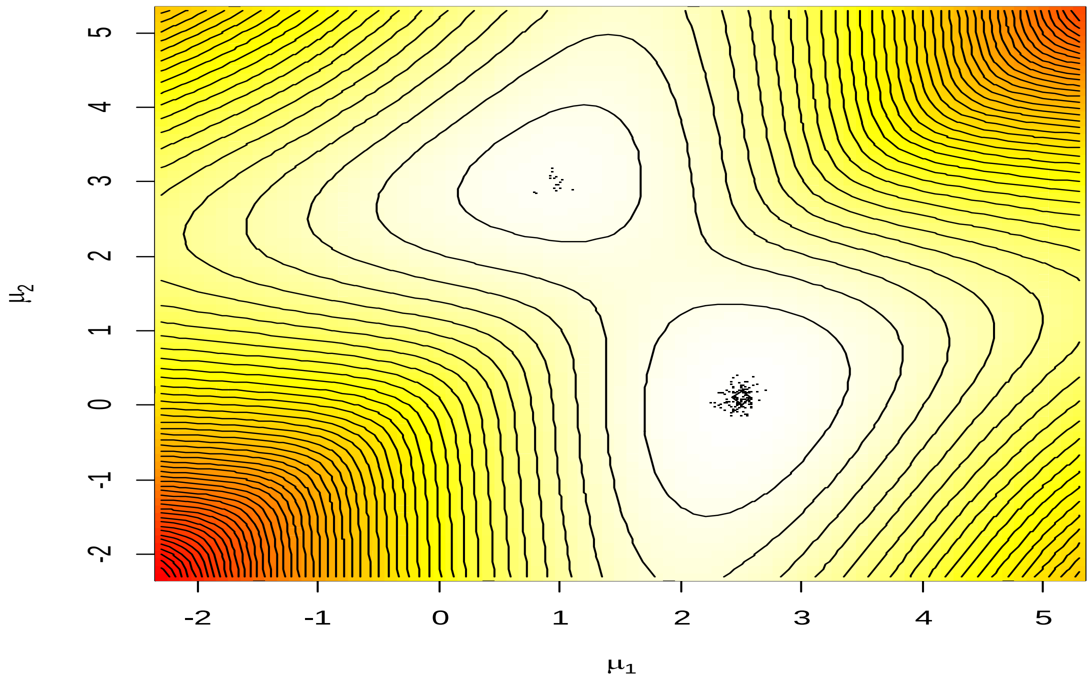


Figure 4. 10,000-iteration outcome of the random walk metropolis-hastings sample on the log-likelihood surface with guessing starting point of (0.5,0.4)

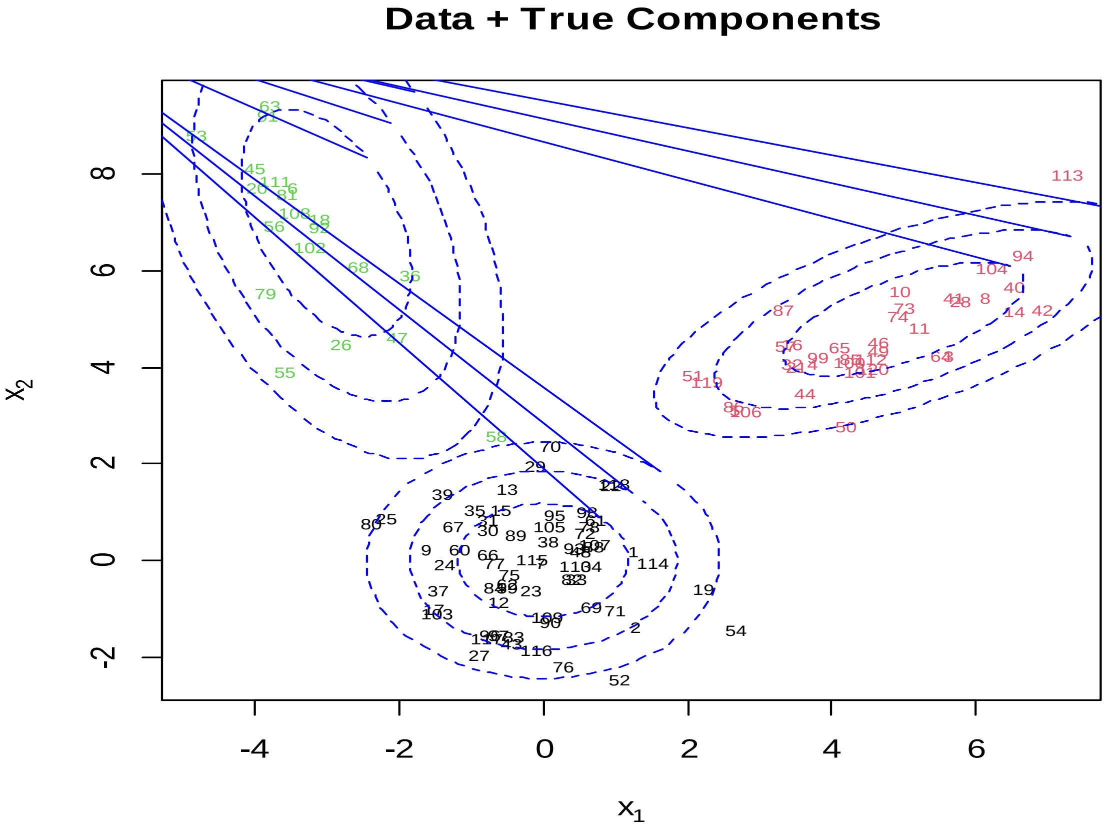


Figure 5. Contour plot of the three-component multivariate mixture model of data + true components

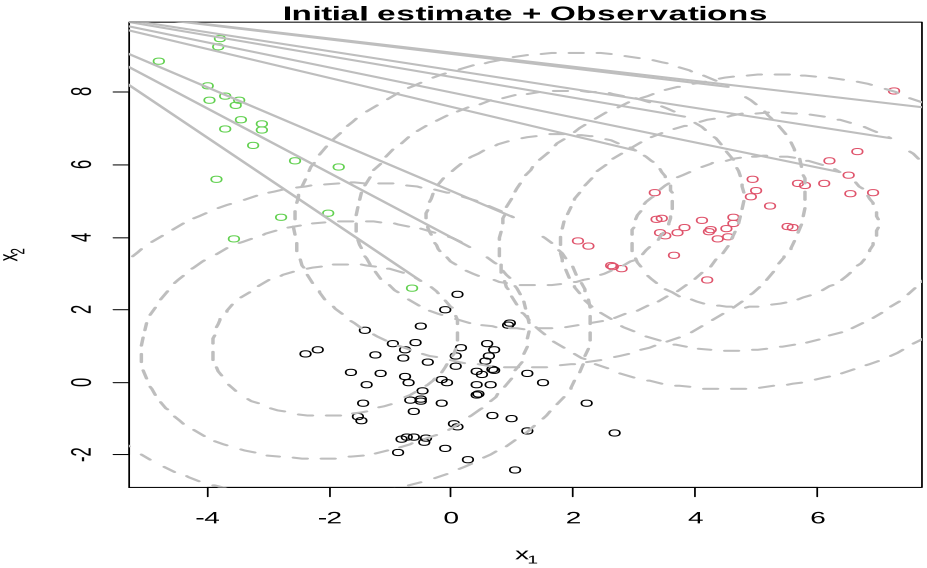


Figure 6. Contour plot of the three-component multivariate mixture model of Initial Estimate + Observations

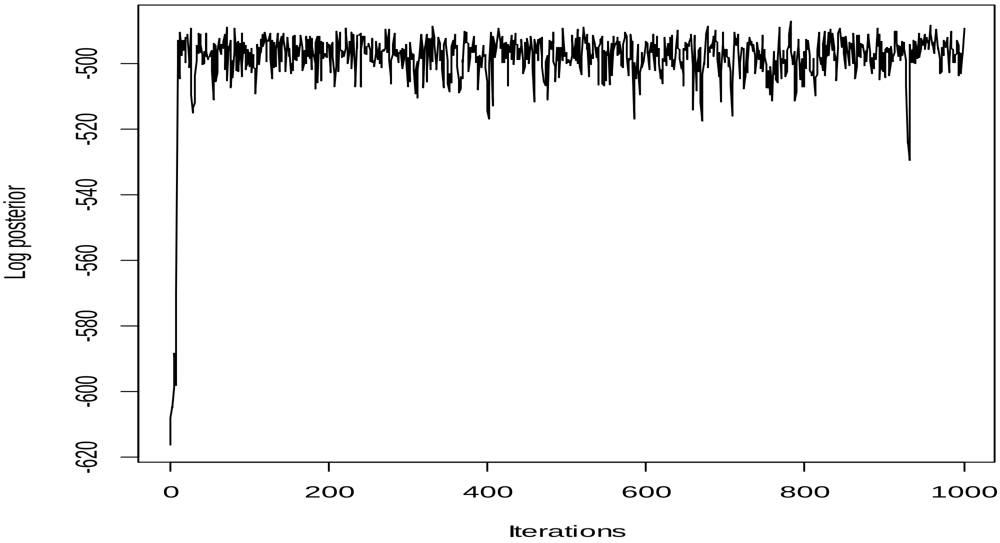


Figure 7. Plot the log-posterior distribution for various samples

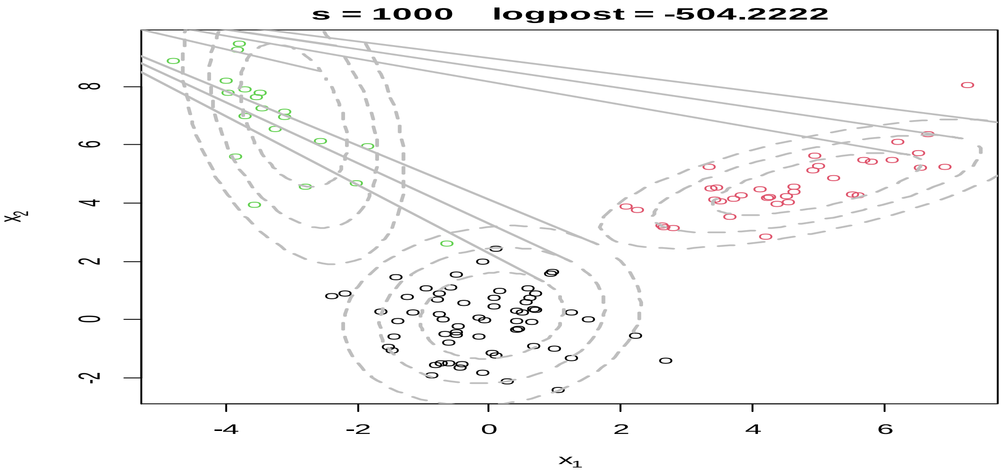


Figure 8. Plot the density estimate for the last iteration of the MCMC

3.1. Discussion of Simulation Results

MCMC algorithm of mixture models of three-component of 2-variates Gaussian components was tested to produce Figures 5 to 8, respectively. The true weights for the three associated components are 0.5, 0.3, and 0.2 respectively. Their true means are (0,0), (5,5), and (-3,7) for the first, second and third components respectively for the simulated data of sample of size 120, such that their true sigma (variance) are (1,0), (2,0.9), and (1,-0.9) respectively. The starting guess value of the weights for each component was assigned to by equal weight of repetition (1,3)/3 via iteration of the sampler. From Figure 5, this is the cluster of three contour plots of the simulated data. The first contour of cluster is denser with clustered of numbers in black color, where the second contour with red character number is lesser than the first one. The third contour at the left corner possessed a scanty cluster of numbers in green color. However, there are three numbers that can be regarded as point outliers: 54, 58, and 133 for the first, second and third contour respectively. The three-compartmental multivariate contour plot of Figure 6 is similar to that of Figure 5 and Figure 8 but it was notably contaminated with outliers, positively abnormal values to indicate that there is possibility of absolving a robust noisy prior-posterior distribution for proper capture. In comparison, the logarithm of the posterior was estimated to be -504.222, such that, the outliers possessed by the posterior density of Figure 7 carved-out six (6) point outliers in contrast to by the true data.

4. Conclusion

This article studies the generic procedure of mixture models with similar inferential probabilistic distribution in a Bayesian setting was proposed. Mixture of similar distributions via Bayesian paradigm was expounded in a finite and non-finite setting, such that the proportional allocation, sample size allocation, and mixing weights for the posterior distribution were carved-out for *k*-components. The parameter estimation of the generic posterior distribution of proportional allocation, sample size allocation, and mixing components coupled with the embedded latent (unobserved) variable was carried-out via the EM algorithm. Metropolis–Hastings and MCMC algorithms were alternately employed in place of the EM algorithm under some conditions. Monte Carlo approximation technique was employed to estimate MAP, such that label switching for exchangeability of posterior distribution was carried-out under different prior for known and unknown components with finite and non-finite mixture. In conclusion, it was deduced that the number of components grows large for all permutations to be considered for subsample of permutations simulated. Further study can be extended to generic procedure of mixture models with different inferential probabilistic distributions from both a parametric, non-parametric and Bayesian point of view. In addition, in scenarios where different prior-likelihood distributions are merged or convoluted, label switching for exchangeability of posterior distribution for finite and non-finite mixture needs to be further studied. In extension, the finite and non-finite mixture models via Bayesian paradigm can be extended to time-varying processes like networking autoregressive and mixture autoregressive processes.

Author Contributions

All the authors equally contributed to this work. They all read and approved the final version of the paper.

Conflict of Interest

All the authors declare no conflict of interest.

References

[1] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, Bayesian Data Analysis, 3rd Edition, Chapman and Hall, New York, 2013.

[2] G. Wioletta, *The Advantages of Bayesian Methods over Classical Methods in the Context of Credible Intervals*, Information Systems in Management 4 (1) (2015) 53–63.

[3] C. Charlton, J. Rasbash, W. J. Browne, M. Healy, B. Cameron, *MLwiN. In: Centre for Multilevel Modeling* (2020), https://www.bristol.ac.uk/cmm/, Accessed 20 Sep 2023.

[4] J. E. Johndrow, A. Smith, N. Pillai, N. Dunson, *MCMC for Imbalanced Categorical Data*, Journal of the American Statistical Association 114 (527) (2019) 1394–1403.

[5] R. O. Olanrewaju, S. A. Olanrewaju, L. A. Nafiu, *Multinomial Naive Bayes Classifier: Bayesian versus Non-parametric Classifier Approach*, European Journal of Statistics 2 (8) (2022) 1–14.

[6] R. O. Olanrewaju, *Bayesian Approach: An Alternative to Periodogram and Time Axes Estimation for Known and Unknown White Noise*, International Journal of Mathematical Sciences and Computing 2 (5) (2018) 22–33.

[7] U. Simola, J. Cisewski-Kehe, L. R. Wolpert, *Approximate Bayesian Computation for Finite Mixture Models*, Journal of Statistical Computation and Simulation 91 (6) (2021) 1155–1174.

[8] A. Hairault, C. P. Robert, J. Rousseau, *Evidence Estimation in Finite and Infinite Mixture Models and Applications* (2022) 43 pages, https://arxiv.org/abs/2205.05416.

[9] A. R. Hassan, R. O. Olanrewaju, Q. C. Chukwudum, S. A. Olanrewaju, S. E. Fadugba, *Comparison Study of Generative and Discriminative Models for Classification of Classifiers*, International Journal of Mathematics and Computer Simulation 16 (12) (2022) 76–87.

[10] M. Betancourt, *A Conceptual Introduction to Hamiltonian Monte Carlo* (2017) 60 pages, https://arxiv.org/abs/1701.02434.

[11] J. F. Ojo, R. O. Olanrewaju, S. A. Folorunsho, *Bayesian Logistic Regression Using Gaussian Naıve Bayes (GNB)*, Journal of Medical and Applied Biosciences 9 (2) (2017) 1–18.

[12] R. O. Olanrewwaju, L. O. Adekola, E. Oseni, S. A. Phillips, A. A. Oyinloye, *Disintegration of Price Ordered Probit Model: An Application to Prices of Cereal Crops in Nigeria*, African Journal of Applied Statistics 7 (1) (2020) 781–804.

[13] S. Virolainen, *A Mixture Autoregressive Model Based on Gaussian and Student-t-Soft Distributions*, Studies in Nonlinear Dynamics & Econometrics 26 (4) (2022) 559–580.

[14] R. O. Olanrewaju, A. G. Waititu, L. A. Nafiu, *Bull and Bear Dynamics of the Nigeria Stock Returns Transitory via Mingled Autoregressive Random Processes*, Open Journal of Statistics 11 (2021) 870–885.

[15] R. O. Olanrewaju, A. G. Waititu, L. A. Nafiu, *On the Estimation of k-Regimes Switching of Mixture Autoregressive Model via Weibull Distributional Random Noise*, International Journal of Probability and Statistics 10 (1) (2021) 1–8.

[16] J. F. Ojo, R. O. Olanrewaju, *On Mixture Auto-Regressive (MAR) Using Naira-Dollar Exchange Rates*, Journal of Nigeria Association Mathematical Physics 38 (12) (2016) 155-165.

[17] R. O. Olanrewaju, S. A. Olanrewaju, *An Alternative Mean Variance Portfolio Theoretical Framework: Nigeria Banks' Market Shares Analysis*, Global Journal of Business, Economics, and Management 11 (3) (2021) 220–234.

[18] R. O. Olanrewaju, *On the Application of Generalized Beta-G Family of Distributions to Prices of Cereals*, Journal of Mathematical Finance 11 (4) (2021) 670–685.

[19] R. O. Olanrewaju, M. A. Jallow, S. A. Olanrewaju, *An Analysis of the Atlantic Ocean Random Cosine and Sine Alternate Wavy ARIMA Functions*, International Journal of Intelligent Systems and Applications 14 (5) (2022) 22–34.

[20] J. F. Olanrewaju, R. O. Olanrewaju, S. A. Folorunso, *Performance of all Nigeria Banks' Shares using Student-t Mixture Autoregressive Model*, Journal of Engineering and Applied 9 (1) (2017) 69–82.