Journal of Data Analysis and Information Processing, 2019, 7, 141-152
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Bayesian Non-Parametric Mixture Model with Application to Modeling Biological Markers

Mercy K. Peter[1], Levi Mbugua[2], Anthony Wanjoya[3]

[1]Department of Mathematics, Pan African University Institute for Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial science, Technical University of Kenya, Nairobi, Kenya
[3]Department of Statistics and Actuarial science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: mercypeterkalo@gmail.com, lmbugua@tukenya.ac.ke, awanjoya@gmail.com

How to cite this paper: Peter, M.K., Mbugua, L. and Wanjoya, A. (2019) Bayesian Non-Parametric Mixture Model with Application to Modeling Biological Markers. *Journal of Data Analysis and Information Processing*, **7**, 141-152.
https://doi.org/10.4236/jdaip.2019.74009

Received: August 22, 2019
Accepted: September 13, 2019
Published: September 16, 2019

Abstract

The effect of treatment on patient's outcome can easily be determined through the impact of the treatment on biological events. Observing the treatment for patients for a certain period of time can help in determining whether there is any change in the biomarker of the patient. It is important to study how the biomarker changes due to treatment and whether for different individuals located in separate centers can be clustered together since they might have different distributions. The study is motivated by a Bayesian non-parametric mixture model, which is more flexible when compared to the Bayesian Parametric models and is capable of borrowing information across different centers allowing them to be grouped together. To this end, this research modeled Biological markers taking into consideration the Surrogate markers. The study employed the nested Dirichlet process prior, which is easily peaceable on different distributions for several centers, with centers from the same Dirichlet process component clustered automatically together. The study sampled from the posterior by use of Markov chain Monte carol algorithm. The model is illustrated using a simulation study to see how it performs on simulated data. Clearly, from the simulation study it was clear that, the model was capable of clustering data into different clusters.

Keywords

Bayesian Non-Parametric, Nested Dirichlet Process, Biomarker, Clustering, Surrogate Markers, Dirichlet Process, Markov Chain Monte Carlo

1. Introduction

To model hierarchical data when the distribution is not known is a big problem and has affected many researchers dealing with big data [1]. This is because of the disparity within the data, to account for the heterogeneity a Bayesian non-parametric model is necessary as it leads to flexible density estimates which are capable of identifying clusters of individuals with similar biomarker characteristics. Bayesian non-parametric mixture model is a good fit to model biological markers because it exhibits flexibility when modeling data which has a skewed and multi-modal distribution. The reason behind this is because data sets become bigger every day and require flexible models which can expand with the data. Mixture methods approach allows for probabilistic approach of clustering data points to different clusters [2]. The model also gives support to out of sample cluster assignments through computing the posterior probabilities for new data points.

In clinical trials, the importance of a treatment is either to decrease the burden of the disease for the patient or to eliminate the disease. To identify a biomarker which is changed by a treatment is not easy due to difficulties associated with the disease mechanisms. If a biomarker which is affected by the treatment has been identified, coming up with the association of the biomarker and the outcome is not easy because of the changes in the variability of the biomarker, patient response, and evaluation methods used. Thus, it is important to identify the changes each individual exhibit and whether there are changes or no changes as a result of the treatment [3]. The responses of individuals to treatment may be related, and identifying of groups of individuals sharing similar characteristics is of important.

Many authors have applied the Bayesian non-parametric procedures to study various categories of biomarkers ranging from prognostic, predictive, phamacodynamic, and surrogate endpoints. For example, [4] studied the prognostic biomarkers and showed how they related to the clinical outcome using the Bayesian non-parametric procedures. Additionally, [3] studied the prognostic biomarkers using Bayesian parametric procedures, and finally [5] studied the surrogate endpoints using the Bayesian methods. These studies identified the need to study biomarkers and determine how they are related with the clinical outcome.

Bayesian non-parametrics have a wide application in many areas especially big data analytics. Bayesian non-parametric methods are widely used to solve problems where the size of the data changes leading to growth of the dimension of interest, for instance, in problems where the number of features varies with increase in the observed data. Also, they are commonly used in clustering and the number of clusters depends on the data being used. In general, in Bayesian non-parametrics models the number of parameters increases as the size of the data grows.

The study of [4] applied the Bayesian non-parametrics in modeling biological markers. In the study the model assumed measurements of the biomarkers were taken continuously before the subjects under study are introduced to treatment and after the patient has been given some treatment. In the study the measurements were not depended on covariates and the survival result was due to measurement of the change, though, different distributions could give the same outcome.

Accordingly, [1] developed an integrative Bayesian predictive modeling framework to identify individual pathological brain states depending on the choice of fluoro-deoxyglucose positron emission tomography (PET) imaging Biomarkers and evaluated the relation of the states with a clinical outcome. The study would identify patient subgroup characterized by different biomarkers to produce the clinical outcome. The strategy also identified imaging Biomarkers with pathological states of the individuals and assumed that the latent individual state gets its values from one of the pathological states, and one of the states was a reference point. The latent random variables were independent and identically distributed taking a multinomial distribution. On the mixture weights a Dirichlet prior was used, considering a where the Gaussian distribution was considered, the mean was taken as one of the parameters to model the latent state specific random effect and to characterize the mean metabolic profile for individuals within the latent state. The Variance-covariance matrix captured the association between regions for individuals with latent state. A likelihood function was also established.

Additionally, [6] developed a Bayesian model to sample inference with availability of inverse-probability weights. The study used a hierarchical method where the distribution of the weights from the non-sampled units was modeled and included predictors in a non-parametric Gaussian process. Simulation study was used to check how the procedure performed and compared to the classical design-based estimator. The study concluded that Bayesian non-parametric finite population estimator is more appropriate compared with the classical estimator. Also, [7] compared the hierarchical Bayes model for biomarker subset effects in clinical trials to the profile likelihood method, to make references to the threshold parameter using bootstrap. The method provided improved sample properties for probability coverage at 95% confidence interval.

Therefore, the importance of modeling surrogate markers in this study is to be able to determine the relationship between the baseline biomarker and the samples taken after an individual has been given some treatment. Bayesian non-parametric methods are flexible methods and will accurately indicate the relationship to show whether there is any change and be able to identify groups of individuals which have similar characteristics through clustering [8]. Also, the method is capable of showing whether after treatment the distribution of the biomarker changed through increase, decrease or it did not change at all.

The other parts of the paper are arranged as follows; in Section 2, discussion of the general modeling framework is done. Section 3, discusses the proposed model by detailing the nested Dirichlet process model for characterizing patient profiles. In Section 4, the hierarchical model is formulated. Section 5, describes the posterior computation. Section 6, is a simulation study to assess the performance of the model. Finally, the conclusion is in Section 7.

2. General Modeling Framework

Let T denote the treatment effect, X represent the baseline biomarker, Y denote the post treatment values, and E the clinical outcome, and Z are the covariates which are present. If $p(.)$ is a distribution, for instance, $P(E \mid X,Y,Z,T)$, is a conditional distribution. If the treatment impact T is put into consideration, then the biomarker distribution will be affected. To address this then the inpatient change from X to Y is necessary. To assess the inpatient change, then putting into consideration of the relationship between X and Y because of the inpatient effects is necessary. Due to the effect the treatment has on Y and the effect of the covariate to X or Y thus it leads to, $P(Y \mid X,Z,T)$ and $P(X \mid Z)$, though the distribution can either be highly disperse and complex. The model in this study will involve representation of a biomarker profile as $\Delta = \Delta(p(X,Y \mid Z,T))$, to symbolize the change made on the biomarker because of treatment, incorporating them to the model to include the impact of the change on the outcome E. The model is also able to classify groups of individuals with various changes in Biomarker profiles depending on how the impacts of T and the change Δ have on E. Thus, employing the probabilistic factorization then;

$$p(E,X,Y \mid Z,T) = p(E \mid X,Y,Z,T)\, p(X,Y \mid Z,T) \tag{1}$$

From Equation (1), the following assumptions are made;

1) $p(E \mid X,Y,Z,T) = p(E \mid \Delta,Z,T)$, which implies, with the effect of the covariates and the treatment, the impact of the (X, Y) on E is indicated by the change.

2) Also, the distribution of X and Y may be depended on the covariate, then the study assumes that both do not depend on the covariates.

A hierarchical Bayesian non-parametric model is employed for $p(X,Y \mid T)$ and for the $p(E \mid \Delta,Z,T)$; a non-parametric regression model in the Bayesian case is employed, to give adaptable cluster estimates for individual's specific distributions of Δ and their clusters. A hierarchical structure is obtained through making assumption of the individual's specific Dirichlet processes being samples that are conditionally independent and obtained from a hyperprior which is also a Dirichlet process.

3. Proposed Model

Here the structure of the data is developed and the general model introduced. The subjects are indexed by $i = 1,\cdots,N$. Assuming E_i is time-to-event outcome, let E_i^0 be the observed time of the event with $\varepsilon_i = 1$, if $E_i^0 = E_i$, and 0 if $E_i^0 < E_i$. For $E^0 = (E_1^0,\cdots,E_N^0)$, $\varepsilon = (\varepsilon_1,\cdots,\varepsilon_N)$, and $Z_i = (Z_{1i},Z_{2i},\cdots,Z_{ki})$ be the baseline covariates with $Z = (Z_1,Z_2,\cdots,Z_N)$. For the i^{th} individual let n_i and m_i be the measurement frequencies of the levels of the biomarker obtained before treatment and after. Let $X_i = (X_{i1},\cdots,X_{ini})$ and $Y_i = (Y_{i1},\cdots,Y_{imi})$ be the individuals pre and post-treatment biomarker values, where, $X = (X_1,X_2,\cdots,X_N)$ and $Y = (Y_1,Y_2,\cdots,Y_N)$.

The functional $\Delta = \Delta(p(X_i,Y_i \mid T_i))$, is a representation of the individual change for the levels of the biomarker before and after treatment. Where T_i is the treatment given to the i^{th} individual and Δ is some measure of distributional distance. The distributional distance is defined on a sample space cumulative density function (Cdf) of one-dimensional random variables, which is the distributional distance between the two cdf's F_X and F_Y in the space of cumulative density function. The vertical quantile function is;

$$Q_{X,Y}\left(p\right) = F_Y\left(F_X^{-1}\left(p\right)\right) \text{ for } p \in \left(0,1\right) \tag{2}$$

where, Equation (2) is a quantile function of order p which is a representation of the functional for the density plot. The quantile function allows for comparison of various functions for all the distributions. For instance, $Q_{X,Y}\left(p\right) = 0.5$ is used in median tests. Also, the vertical quantile function is associated with the Receiver Operating Characteristic (ROC) curve represented as

$$\text{ROC}\left(p\right) = 1 - F_Y\left(F_X^{-1}\right)\left(1-p\right),$$

where F_X and F_Y are the cdf's of the diagnostic variables in the populations. Here, the interest is not to assess the diagnostic performance for a biomarker; however, to evaluate the targeted treatment, the vertical quantile function is estimated by taking into consideration the distribution functions F_{Xi} and F_{Yi} for the subject levels of biomarker for different individuals. Therefore, the distributional change is;

$$\Delta = \int_0^1 Q_{X,Y}\left(p\right)\mathrm{d}p = EF_Y\left(F_X\left(Y\right)\right) = p\left(X < Y\right) \tag{3}$$

Equation (3) corresponds to the area under the curve which is majorly applied in diagnostic studies. Thus, Δ represents the change of the distribution of the biomarkers for the i^{th} individual, because of the treatment administered to the subject. A posterior estimate with $p\left(X_{ij} < Y_{ik} \mid \text{data}\right) > 0.5$, means that the individual's distribution has moved to the right, that is, there is a biomarker increase. Also, $p\left(X_{ij} < Y_{ik} \mid \text{data}\right) < 0.5$, shows a change to the left side, hence a decrease in the biomarker levels, and $p\left(X_{ij} < Y_{ik} \mid \text{data}\right) \approx 0.5$, indicates no remarkable change. Thus, from Equation (1) the patient level data likelihood is;

$$p\left(T_i^0, \varepsilon_i, X_i, Y_i \mid Z_i, T_i, \beta, \theta\right) = p\left(T_i^0, \varepsilon_i \mid \Delta_i, T_i, \beta\right) p\left(X_i, Y_i \mid T_i, \theta\right) \tag{4}$$

For β a vector of parameters for regression modeling and θ parameterizes the hierarchical model.

$$p\left(T^0, \varepsilon \mid X, Y, Z, T, \beta\right) = p\left(T^0, \varepsilon \mid \Delta, Z, T, \beta\right) \tag{5}$$

Thus, T^0, follows one of the common distribution like the log-normal, where the linear component is a function of the change (Δ), the treatment (T) and the covariates (Z). The model $P\left(X, Y \mid T, \theta\right)$ should be adaptable so as to cover many biomarker distributions which are possible and can either be skewed or multimodal and take account of the variability between subjects. To explain the variability then a hierarchical Bayesian non-parametric framework is used. The model allows flexible density which can identify groups of subjects characterized by individual's biomarker profile. Therefore, the study assumes that measurements of the biomarker are samples are obtained from unknown individuals distributions with $X_{i1},\cdots,X_{ini}\overset{ind}{\sim}F_{Xi}$ and $Y_{i1},\cdots,Y_{imi}\overset{ind}{\sim}F_{Yi}$, Where X_i and Y_i are vectors of subject specific measurements. F_{Xi} and F_{Yi} are modeled separately using mixtures of Gaussian distribution denoted by mean μ and standard deviation δ, that is $N(\mu,\delta)$. The pdf and cdf are denoted by $\phi(.,\mu,\delta)$ and $\Phi(t,\mu,\delta)$. The mixture components are defined as w_i for each component with the constraint such that, $\sum_{i=1}^{k}w_i=1$, implying that the total probability distribution will normalize to 1. Thus, the Gaussian mixture model is represented as;

$$p(x)=\sum_{i=1}^{k}w_iN(x\mid\mu_i,\delta_i) \tag{6}$$

$$N(x\mid\mu_i,\delta_i)=\frac{1}{\delta_i\sqrt{2\pi}}\exp\left(\frac{-(x-\mu_i)^2}{2\delta_i^2}\right)$$

Assuming a DP with a concentration parameter α and a base distribution G_0. Then, for each individual $i=1,\cdots,N$ it follows that;

$$\begin{gathered}Y_{ik}\mid\mu_{Yik},\delta_{Yik}\overset{ind}{\sim}N(\mu_{Yik},\delta_{Yik}),k=1,\cdots,m_i\\ X_{ij}\mid\mu_{Xij},\delta_{Xij}\overset{ind}{\sim}\prod_{j=1}^{p}N(\mu_{Xij},\delta_{Xij}),j=1,\cdots,n_i\\ \mu_{Yik},\delta_{Yik},\mu_{Xij},\delta_{Xij}\mid G\overset{iid}{\sim}G,G\sim\mathrm{DP}(\alpha,G_0)\end{gathered} \tag{7}$$

where, $\alpha=1$, and $G_0=N(\mu,\delta)$.

Let $\theta_{Xij}=(\mu_{Xij},\delta_{Xij})$ and $\theta_{Yij}=(\mu_{Yij},\delta_{Yij})$. Under the mixture model θ_{Xij} and θ_{Yij} are sampled from some mixing distributions G_{Xi} and G_{Yi} as follows;

$$\begin{gathered}\theta_{Xi1},\cdots,\theta_{Xini}\mid G_{Xi}\overset{ind}{\sim}G_{Xi}\\ \theta_{Yi1},\cdots,\theta_{Yimi}\mid G_{Yi}\overset{ind}{\sim}G_{Yi}\end{gathered} \tag{8}$$

This means that the conditionals on the realizations of G_{Xi} and G_{Yi}, thus, the distributions for the X_i and Y_i are the following mixtures;

$$\begin{gathered}f_X(x_i\mid G_{Xi})=\int\prod_{j=1}^{ni}\phi(x_{ij};\theta_{Xij})G_{Xi}(\mathrm{d}\theta_{Xij})\\ f_Y(y_i\mid G_{Yi})=\int\prod_{j=1}^{mi}\phi(y_{ik};\theta_{Yik})G_{Yi}(\mathrm{d}\theta_{YiK})\end{gathered} \tag{9}$$

To assess the change in the distribution of Y_i verses X_i in terms of Δi so as to be able to investigate the association of the change with the outcome and classify groups of subjects which have the same biological responses. Additionally, a prior model is defined on G_{Xi} and G_{Yi} and it involves the Dirichlet process (DP), which is commonly preferred prior probability model due to its clustering capability. [9] expressed this as $G\sim\mathrm{DP}(\alpha,G_0)$, which is a random distribution G following a DP that has a base distribution $E(G)=G_0$ and a concentration parameter α. α Shows significant properties, that is, how G varies about the mean (base distribution), where a smaller value of α shows high uncertainty and vice versa.

Since G are discrete samples, they have some positive probability ties, as some θ_{Yij}'s and θ_{Xij}'s in Equation (8) may be equal. A DP can be easily used to estimate G_{Xi} and G_{Yi} for each subject, though it lacks the clustering properties of the distributions for all individuals or among the pre and post treatment values obtained. Clustering is necessary so as to identify the change after the treatment for all the individuals. Thus, the change is obtained by assuming that G_{Xi} and G_{Yi} are realizations of a common Dirichlet process mixture model (DPMM). In a DPMM the individual's realizations of G_{Xi} and G_{Yi} are shared across and for each subject's pre and post treatment values. Therefore, G_{Xi} and G_{Yi} are independent conditional samples from the same the Dirichlet process, then;

$$G_{Xi} \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_{ok}} \text{ and } G_{Yi} \sim \sum_{k=1}^{\infty} \pi_k \delta_{G_{ok}} \tag{10}$$

where G_{0k} is a realization from a common DP prior that is $\mathrm{DP}\left(\gamma, G_0^*\right)$ which has a base distribution G_0^* and a concentration parameter γ, then;

$$G_{ok}(.) = \sum_{p=1}^{\infty} w_{pok}^* \delta_{\theta_{pok}}^* (.) \tag{11}$$

Here $\theta_{pok}^* \sim G_0^*$. Therefore each G_{Xi} and G_{Yi} is automatically obtained from a collection of different distributions that is the G_{0k}'s.

4. Formulation of the Hierarchical Model

The hierarchical model is formulated using the nested Dirichlet Process (nDP) which is as follows;

$$\left(G_{Xi}, G_{Yi}\right) \sim \mathrm{DP}\left(\alpha, \gamma, G_0^*\right) \tag{12}$$

In the earlier discussions, it is clearly expressed that Δi is a functional of $p\left(X_i, Y_i \mid Z_i, T_i\right)$, which in the nDP is determined by the realizations G_{Xi} and G_{Yi} in Equation (10). Since the Dirichlet process given by Equation (10) has a discrete support and π_k's in the equation cannot be neglected, then it shows a non-trivial probability where $G_{Xi} = G_{Yi}$, which means that the treatment has no biological impact on patient *i*, this is clearly shown through the posterior estimate of the Δi. Additionally, there is also non-trivial probability that $\left(G_{Xi}, G_{Yi}\right) = \left(G_{Xi}, G_{Yi}\right)$ for $i \neq i'$ which implies $\Delta i = \Delta i'$, implying the biomarkers profiles for individuals *i* and *i'* are in one cluster.

To complete the model the base distribution G_0^* is specified and it is defined as a Normal-Inverse Gamma (N-IG) distribution for the mean and precision parameters in the Normal model and α and γ are assigned independent Gamma priors, thus, the hierarchical probabilistic model.

4.1. Biomarker Profiles Likelihood

The Biomarker Profiles Likelihood is as below;

$$\begin{gathered} Y_{ik} \mid \mu_{Yik}, \delta_{Yik} \overset{ind}{\sim} N\left(\mu_{Yik}, \delta_{Yik}\right), k = 1, \cdots, m_i \\ X_{ij} \mid \mu_{Xij}, \delta_{Xij} \overset{ind}{\sim} \prod_{j=1}^{p} N\left(\mu_{Xij}, \delta_{Xij}\right), j = 1, \cdots, n_i \\ \Delta_i = E\left(G_{Yi}\left(Y_{ik}\right)\right) = p\left(X_{ij} < Y_{ik}\right) \end{gathered} \tag{13}$$

4.2. The Model and the Priors

$$\theta_{Xij} = \left(\mu_{Xij}, \delta_{Xij}\right)^E \text{ and } \theta_{Yik} = \left(\mu_{Yik}, \delta_{Yik}\right)^E$$
$$\theta_{Xij} \mid G_{Xi} \sim G_{Xi} \text{ and } \theta_{YiK} \mid G_{Yi} \sim G_{Yi}$$
$$\left(G_{Xi}, G_{Xi}\right) \sim \text{nDP}\left(\alpha, \gamma, G_0^*\right) \tag{14}$$
$$\alpha \sim \text{Gam}\left(a_\alpha, b_\alpha\right), \gamma \sim \text{Gam}\left(a_\gamma, b_\gamma\right)$$
$$G_0^* \sim \text{N-IG}\left(\mu_0, k_0, a_0, d_0\right)$$
$$\left(\beta, \delta_E\right) \sim \text{N-IG}\left(\mu_1, k_1, a_1, d_1\right)$$

where, the fixed hyper parameters are; μ_0, k_0, a_0, d_0, μ_1, k_1, a_1, d_1.

5. Posterior Computation

To compute the joint posterior distribution for model parameters, this is done computationally. Thus Markov Chain Monte Carlo (MCMC) algorithm for posterior inference is used. The full conditional to update the nDP are gotten using the method described by [10] depending on a truncated Dirichlet process. At each iteration, for the baseline distribution G_0^*, parameters are continuously updated based on all the samples represented by the biomarker values. The algorithm is developed using a truncation of a Dirichlet process to give approximate truncation to the stick breaking process of a Dirichlet process leading to method of computation in finite mixture models.

This assumes that, individuals are clustered into K groups and for every individual the observations on the biomarker level can be clustered into L groups. To provide support for the estimation of both clinical and biological effects together, the proposed model accounts completely for the uncertainty of the random quantities, together with variability of the Δ_i's to express the variation of the population. In every iteration the Gibbs sampling algorithm gives samples of the distribution of the biomarker (G_{Xi}, G_{Yi}) for every individual, used to get the biomarker profile Δ_i. This can be easily illustrated by Considering the model as in Equation (13) and Equation (14), to obtain samples from the posterior after the burn in, every value that is sampled (Δ_i^*) is obtained by getting the average for the estimates of the posterior $\left(G_{Xi}^*, G_{Yi}^*\right)$ of the subjects distributions of biomarkers. From Equation (9) and Equation (10), it is clear that every mixing distribution G_{0k}, $F\left(t \mid G_{0k}\right) = \Phi\left(t;\theta\right) G_{0k}\left(d\theta\right) = \sum_{i=1}^{\infty} w_{lr}^* \Phi\left(t; \theta_{lr}^*\right)$ for $F = G_{Xi}$ or G_{Yi}, to obtain the biomarker profile for the posterior then, Equation (15) is applicable;

$$\Delta_i^* = EG_{Yi}^*\left(G_{Xi}^*\left(Y_{ik}\right)\right) = \int G_{Xi}^*\left(y\right) dG_{Yi}^*\left(y\right) dy \tag{15}$$

where $G_{Xi}^* = \sum_{l=1}^{\infty} w_l^* \delta_{\theta_l}^*$ and $G_{Yi}^* = \sum_{l=1}^{\infty} w_{l'}^* \delta_{\theta_{l'}}^*$ thus the estimate of the posterior biomarker profile is obtained by dividing the mean with the posterior values which is;

$$\Delta_l^* = \sum_l \sum_{l'} w_l^* w_{l'}^* \left(1 - \Phi\left(\frac{\mu_l^* - \mu_{l'}^*}{\sqrt{\delta_l^{2*} + \delta_{l'}^{2*}}}\right)\right) \tag{16}$$

6. Simulation Study

In this section, a simulation study is presented to show the capability as well the ability of the nested Dirichlet Process when modeling biological markers to give accurate density estimates by obtaining strength from different centers. In the simulation, N samples are obtained from a mixture of four Gaussian distribution.

$$f(x) = \sum_i^k w_i f_k(x) \tag{17}$$

Equation (17), is a representation of a mixture of Gaussian distribution with w_i mixing weight for every component, and $f_k(x)$ is the component which can be represented by any distribution. Here, the components are represented by a normal distribution such that the mixture distribution becomes;

$$f(x) = \sum_i^k w_i N\left(\mu_i, \delta_i^2\right) \tag{18}$$

The study generates J = 40 samples each of size 100, for $j = 1, \cdots, 40$. Every sample is obtained from a mixture of k = 4 Gaussian mixtures summarized in Table 1, and plotted in Figure 1.

The true distributions are plotted in Figure 1.

Distribution S1 and S2 are asymmetric with a mixture of two Gaussian components with different weights. For distribution S3 and S4, they share three

Table 1. The components of various Gaussian distributions.

Distribution	Component 1			Component 2			Component 3			Component 4		
	w	μ	δ	w	μ	δ	w	μ	δ	w	μ	δ
S1	0.85	0	1	0.15	4	3	-	-	-	-	-	-
S2	0.65	0	1	0.35	4	3	-	-	-	-	-	-
S3	0.4	0	1	0.3	−2	3	0.3	3	3	-	-	-
S4	0.39	0	1	0.29	−2	3	0.29	3	3	0.03	11	3

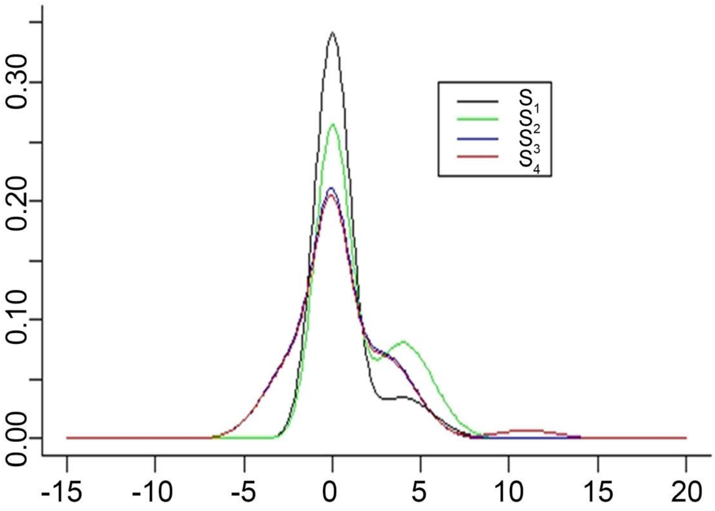

Figure 1. Plot for the true distributions used in the simulation study.

mixture components which are located at the origin, with difference only on the fourth component of the distribution S4.

The true cluster memberships are plotted in Figure 2. It represents the true cluster membership s_i for J = 40 samples through plotting I ($s_i = s_j$) for all the pairs. The samples are ordered by their true clusters. Therefore, these are simulation conditions with well separated true distributional clusters.

The same number of samples has been simulated for the each of the four true distributions. To obtain the posterior simulation, then; the precision parameters α and γ are both fixed to 1 and the Normal-Inverse Gamma distribution which is the baseline measure (base distribution) are $\mu_0 = 0$, $\lambda = 0.01$, $a = 3$, and $b = 1$, such that; NIG(0, 0.01, 3, 1). Therefore, a priori $E\left(\mu \mid \delta^2\right) = 0$, $V\left(\mu \mid \delta^2\right) = 100\delta^2$, $E\left(\delta^2\right) = 1$, and $V\left(\delta^2\right) = 3$.

The algorithm described in Section 5 is used to obtain the samples of the posterior distribution using the nested Dirichlet Process. The study runs MCMC chain with 12,000 iterations, discarding the first 2000 iterations and thinning out to save one in every 10 iterations.

The estimated distributions $E\left(F_k \mid y\right)$ for each distributional cluster are represented in Figure 3. Figure 3 is an image of Figure 1. This is a clear indication that the prior and the posterior samples obtained after the MCMC draws are the same and reflect the distribution where each of the observation has been obtained from. The posterior draws are drawn from all the distributions with all the components. Hence, the posterior and the prior distribution are the same. Thus, in this case when using the Bayesian non-parametric mixture model it reflects the individual biomarker distributions before treatment taking the same form as the after treatment measurements drawn from different centers.

Also, the posterior cluster memberships takes the same form as the true cluster memberships as clearly shown in Figure 4.

The posterior co-clustering probabilities take the same form as the true cluster membership. The model developed is able to classify groups of individuals from different centers (distributions) to one group. The individuals are placed into the groups as per the prior information which is available. Hence, the diagram displays four clusters similar to the estimated distribution as shown in Figure 4.

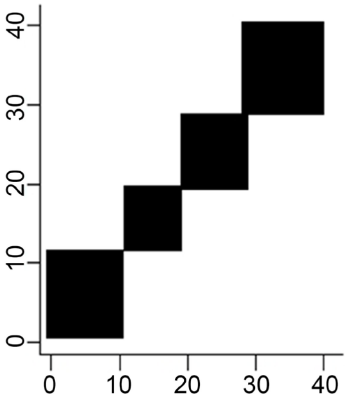

Figure 2. Representation of the true cluster membership.

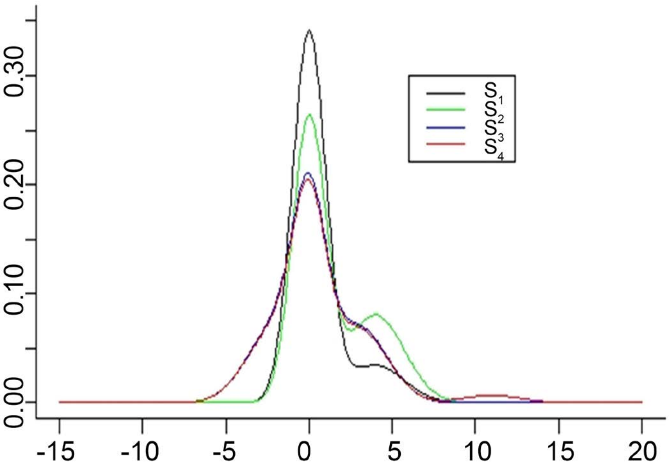

Figure 3. Representation of the estimated distribution.

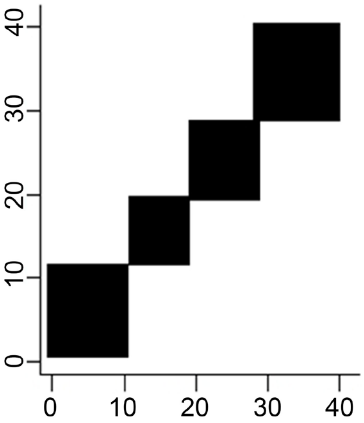

Figure 4. Representation of the posterior co-clustering probabilities for all the distributions.

7. Conclusions

We introduced a model using the truncated nested Dirichlet process to identify groups of individuals who respond similarly to the same treatment for a specified biological marker. An MCMC algorithm has been used to estimate the posterior inference. Since the nDP is a non-parametric model, it has the capability of grouping all the observations from the mixture depending on the entire distribution, rather than selecting particular features of the distribution. In the simulation study the proposed method for biological markers showed a good performance in differentiating the unimodal distributions from the multimodal distributions.

The proposed procedure in this paper reveals that Bayesian non-parametric mixture model can be used to obtain flexible estimates of the individual biomarkers and characterize the heterogeneity of how the subjects are responding to treatment. The proposed procedure only considers measurements taken before introduction of treatment and after the treatment. Biomarkers sometimes can change with time, thus a more structured model can be developed by use of longitudinal biomarker values to account for individuals biomarker processes. This work can be extended to model the relationship between two or more groups of data after the individuals have been clustered. Also, the procedure did not take into consideration of the covariates which might affect the biomarkers. This can also be incorporated so as to see whether they have any effect.

Acknowledgements

We thank the Editor and the referee for their comments. Our research is funded by the African Union, this support is greatly appreciated.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Chiang, S., Guindani, M., Yeh, H.J., Dewar, S., Haneef, Z., Stern, J.M. and Vannucci, M. (2017) A Hierarchical Bayesian Model for the Identification of Pet Markers Associated to the Prediction of Surgical Outcome after Anterior Temporal Lobe Resection. *Frontiers in Neuroscience*, **11**, 669. https://doi.org/10.3389/fnins.2017.00669

[2] Orbanz, P. and Teh, Y.W. (2011) Bayesian Nonparametric Models. Springer, Berlin, 81-89. https://doi.org/10.1007/978-0-387-30164-8_66

[3] Morita, S., Thall, P.F., Bekele, B.N. and Mathew, P. (2010) A Bayesian Hierarchical Mixture Model for Platelet-Derived Growth Factor Receptor Phosphorylation to Improve Estimation of Progression-Free Survival in Prostate Cancer. *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), **59**, 19-34. https://doi.org/10.1111/j.1467-9876.2009.00680.x

[4] Graziani, R., Guindani, M. and Thall, P.F. (2015) Bayesian Nonparametric estimation of Targeted Agent Effects on Biomarker Change to Predict Clinical Outcome. *Biometrics*, **71**, 188-197. https://doi.org/10.1111/biom.12250

[5] Cowles, M. (2004) Evaluating Surrogate Endpoints for Clinical Trials: A Bayesian Approach. Technique Report, University of Iowa, Iowa City, IA.

[6] Si, Y., Pillai, N.S. and Gelman, A. (2015) Bayesian Nonparametric Weighted Sampling Inference. *Bayesian Analysis*, **10**, 605-625. https://doi.org/10.1214/14-BA924

[7] Chen, B.E., Jiang, W. and Tu, D. (2014) A Hierarchical Bayes Model for Biomarker Subset Effects in Clinical Trials. *Computational Statistics and Data Analysis*, **71**, 324-334. https://doi.org/10.1016/j.csda.2013.05.015

[8] Kai, C. and Wenshan, C. (2012) Spike-and-Slab Dirichlet Process Mixture Models. *Open Journal of Statistics*, **2**, 512-518. https://doi.org/10.4236/ojs.2012.25066

[9] Ferguson, T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209-230. https://doi.org/10.1214/aos/1176342360

[10] Rodriguez, A., Dunson, D.B. and Gelfand, A.E. (2008) The Nested Dirichlet Process. *Journal of the American Statistical Association*, **103**, 1131-1154. https://doi.org/10.1198/016214508000000553