Chapter

Assessing Heterogeneity of Two-Part Model via Bayesian Model-Based Clustering with Its Application to Cocaine Use Data

Ye-Mao Xia, Qi-Hang Zhu and Jian-Wei Gou

Abstract

The purpose of this chapter is to provide an introduction to the model-based clustering within the Bayesian framework and apply it to asses the heterogeneity of fractional data via finite mixture two-part regression model. The problems related to the number of clusters and the configuration of observations are addressed via Markov Chains Monte Carlo (MCMC) sampling method. Gibbs sampler is implemented to draw observations from the related full conditionals. As a concrete example, the cocaine use data are analyzed to illustrate the merits of the proposed methodology.

Keywords: model-based clustering, finite mixture model, two-part model, Markov Chain Monte Carlo sampling, cocaine use data

1. Introduction

A recurring theme in the statistical analysis is to separate the unstructured data into groups to detect the similarity or discrepancy within or between groups. This is especially true in the fields, e.g., discriminant analysis [1–3], pattern recognition [4, 5], gene expression [6–8], machine learning [9], and artificial intelligence [10]. In the literature, the clustering problem is often formulated within the *cluster analysis* framework, which is generally categorized into two classes: the non-probabilistic framework and the probabilistic framework. The non-probabilistic clustering method, including the K-means method [9, 11, 12] and the hierarchical/agglomerative clustering algorithms [13–15], is based on the *distance* between any two observations or groups. It clusters data by merging or removing observations according to the "closeness" specified by the distance. This method is more general since it does not impose any distributional assumptions on data, hence having greater flexibility in the real applications. Instead, the non-probabilistic clustering algorithm, also termed the *model-based clustering*, groups data by positing a probability model on data and then clustering data via configuration function related to the model. Compared with the non-probabilistic framework, the model-based methods enable us to assess the statistical properties of the solutions, e.g., how many clusters are there, how well the configuration function works, and how robust the method is against the model deviation and so on. There is rich literature on this issue. Among them, finite mixture model (FMM, [16–18]) perhaps is the most popular choice and has often been proposed and studied in the context of clustering (see a short review in Fraley and Raftery [2]). FMM assumes that each cluster is identified with a probability distribution indexed by the cluster-specific parameter(s), and each observation is related to clusters via configuration or membership function. The statistical task is the inference about the number of clusters, the estimation of the unknown parameters, and the allocation of observations.

In this chapter, we pursue a Bayesian model-based method to address the heterogeneity of fraction data. Fractional data are very common in the social and economical surveys. A distinguished feature of fractional responses is that its measurements are responded on a scale in the unity interval [0,1] but suffer from excessive zeros and unities on the boundaries. In understanding such type of data, the commonly used method is to separate the whole data into three parts: two corresponding to the zeros and unities respectively, and one corresponding to the continuously positive values. Two separative logistic models are suggested to model two discrete value parts respectively while single normal linear regression model is formulated for the continuous value part. This method, though more appealing, ignores the instinct association across different parts and readily leads to inconsistence of the occurrence probabilities on each part. Instead, we propose a three-category multinomial model for the occurrence variable, in which the usual separated models can be considered as the marginal models of our proposal. Such modeling always ensures the probabilities on each part to be proper, thus avoiding parameter constraints, see for example, [19]. To assess the heterogeneity underlying data, we formulate the problem into a finite mixture analysis of which each component is specified by two-part regression model. In view of the model complexity, we implement Markov Chains Monte Carlo sampling method to implement posterior analysis. Block Gibbs sampler is implemented to draw observations from the target distributions. The posterior inference including parameters estimates, model selection, and the configuration determination of observations are obtained based on the simulated observations.

The chapter is organized as follows. Section 2 introduces a general model-based clustering method to address the heterogeneity of regression model within the Bayesian framework. In Section 3, we apply the proposed method to the fractional data. Section 4 presents a cocaine use study. And Section 5 concludes the chapter.

2. Method description

2.1 General framework

Suppose that for $i = 1, 2, \cdots, n$, y_i is an observed response, each associated with an m dimensional fixed covariates $\mathbf{x}_i = (x_{i1}, \cdots, x_{im})$. In the context of regression analysis, the interest mainly focuses on exploring the pattern of the influence of $\mathbf{x}_i$ on y_i and predicting the mean of a future response y in terms of a new $\mathbf{x}$. This is usually achieved by formulating $\{\mathbf{x}_i, y_i\}$ as $\mathbb{E}(y_i|\mathbf{x}_i) = m(\mathbf{x}_i)$ for some mean function $m(\cdot)$. In the parametric fitting framework, the function $m(\mathbf{x})$ is assumed to be related to $\mathbf{x}$ via linking function as the form of

$$m(\mathbf{x}) = h(\mathbf{x}^T\boldsymbol{\beta}) \tag{1}$$

which induces the so-called generalized linear model [20] for $\{\mathbf{x}_i, y_i\}$, where $\boldsymbol{\beta}$ is the regression coefficients used to quantify the uncertainty about m, and $h(\cdot)$ is the known linking function used to link the mean and the predictors.

More often, the single relationship such as Eq. (1) may not be sufficient when the patterns among the subjects take on the heterogeneity such as clustering. The heterogeneous data occur when the observations are generated from the different populations of which the number of populations and the membership of each observation to the population are unknown. The main objective is to separate data into different clusters to detect the possible similarity within clusters or the discrepancy between clusters. This is generally accomplished by defining a cluster's membership/configuration function $\mathcal{K} : \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \mapsto \{1, \cdots, K\}$ such that $K_i = \mathcal{K}((\mathbf{x}_i, y_i)) = k$ if $(\mathbf{x}_i, y_i)$ belongs to the cluster k, where K is assumed to be less than n. The discrepancy between any two clusters is characterized by the cluster-specific parameters such as intercepters, regression coefficients, and/or disperse parameters.

The model-based clustering assumes that given the clusters membership K_i, $(\mathbf{x}_i, y_i)$ within the cluster k has the following sampling density

$$(y_i | K_i = k, \mathbf{x}_i) \overset{ind.}{\sim} f_k(y_i | \mathbf{x}_i^T \boldsymbol{\beta}_k, \quad \tau_k) \tag{2}$$

while K_i is specified by

$$\mathbb{P}(K_i = k) = \pi_k \tag{3}$$

where f_k, maybe independent of k, is the probability density function, $\boldsymbol{\beta}_k$ and $\boldsymbol{\tau}_k$ are the cluster-specific regression coefficients and the disperse parameters, respectively, and π_k is the mixing proportion identifying the proportion of the component k over the entire population. It is assumed that $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1.0$.

Two important issues arise when formulating data clustering problem as Eqs. (2) and (3). One is related to the number of clusters, and the other is pertained to the determination of configurations. Within the Bayesian framework, several methods have been proposed for the first issue. One can, for example, follow [21] and treat K to be random and assign a prior to it. The reversible jump MCMC method (RJMCMC, [21, 22]) can be implemented to conduct the joint analysis of K with other random quantities. Another method is along the lines with the hypothesis test procedure and routinely to estimate K via model comparison/selection procedure. This perhaps is the most popular choice in the model-based clustering context, in which various measures such as the Akaike information criterion (AIC) [23], the corrected AIC (AICc) [24, 25], the Bayesian information criterion (BIC) [26], the integrated completed likelihood (ICL) [27], and Bayes factor (BF, [28, 29]) can be adopted to select a suitable model. It is worth pointing out that the deviance information criterion (DIC) [30] may not be appropriate for the mixture model comparison. The well-known software WinBUGS® [31] for Bayesian analysis does not provide DIC results for mixture analysis. In addition, many authors suggested modeling heterogeneous data into the mixture of Dirichlet process (MDP, [32, 33]). However, as discussed in Ishwaran and James [34], DP fitting often overestimates the number of clusters and readily leads to model over fitting.

For the second issue, the complexity of problem depends on the methods adopted in the analysis. In the frequency framework, for example, the configuration of observation i is often achieved by maximizing $\mathbb{P}(K_i = k | \mathbf{Y}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Xi}})$ over $k = 1, \cdots, K$, where $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\Xi}}$ are the maximum likelihood estimates (MLE) obtained via, e.g., the expectation-maximization algorithm (EM, [35]). In the next section, we will present a Bayesian procedure for determining $\mathcal{K}$. Compared with the frequency approach, the nice feature of the Bayesian approach is its flexibility to utilize prior information for achieving better results. Also, the sampling-based Bayesian methods depend less on the asymptotic theory and hence have the potential to produce reliable results even with small sample size.

Let $\mathbf{Y}$ be the set of all observed responses and $\mathbf{X}$ be the set of fixed covariates; Write Ξ as the collection of $\boldsymbol{\beta}_k$ and $\boldsymbol{\tau}_k$. Integrating over K_i produces a K-component mixture model for y_i, which is given by

$$p(y_i|\boldsymbol{\pi}, \Xi, \mathbf{x}_i) = \sum_{k=1}^{K} \pi_k f_k(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_k, \boldsymbol{\tau}_k). \tag{4}$$

The log-likelihood of the observed data conditional on K is given by

$$\mathcal{L}(\pi, \Xi|\mathbf{Y}, \mathbf{X}) = \sum_{k=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_k, \boldsymbol{\tau}_k)\right). \tag{5}$$

As an illustration, **Figure 1** presents a three-component normal linear mixture regression model with one covariate. It can be seen clearly that the density function illustrates strong heterogeneity. The regression line is obviously different from those of components, which indicates that single model is unappreciate in fitting such data. In what follows, we suppress $\mathbf{X}$ for notational simplicity.

2.2 Bayesian model-based clustering via MCMC

Bayesian analysis for analyzing Eqs. (2) and (3) especially $\mathcal{K}$ requires the specification of a prior distribution $p(\boldsymbol{\pi}, \Xi)$ for the parameters of the mixture model. By model convention, it is naturally to assume that $\boldsymbol{\pi}$ and Ξ are independent, and the components among Ξ are also independent. In particular,

$$\boldsymbol{\beta}_k \overset{iid.}{\sim} N_m(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\tau}_k^{-1} \overset{iid.}{\sim} W(\rho_0, \boldsymbol{R}_0) \tag{6}$$

in which $W(\rho_0, \boldsymbol{R}_0^{-1})$ is the Wishart distribution with the degrees of freedom ρ_0 and the scale matrix $\boldsymbol{R}_0$, and reduces to the scaled Chi-square distribution when $\boldsymbol{\tau}_k$ is a univariate; $\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_0$, ρ_0 and $\boldsymbol{R}_0$ are the hyper-parameters, which are treated to fixed and known. In the real applications, if no extra information can be available, the values of these hyper-parameters are often taken to ensure $\boldsymbol{\beta}_k$ and $\boldsymbol{\tau}_k$ to be dispersed enough. For example, one can set $\boldsymbol{\Sigma}_0 = \lambda_0\mathbf{I}$ with large λ_0 (Throughout, we use $\mathbf{I}$ to signify an identify matrix). In this case, the values of $\boldsymbol{\beta}_0$ are not really important and can be set to any values, e.g., zeros. Note that for the mixture models, Diebolt and Robert [36]

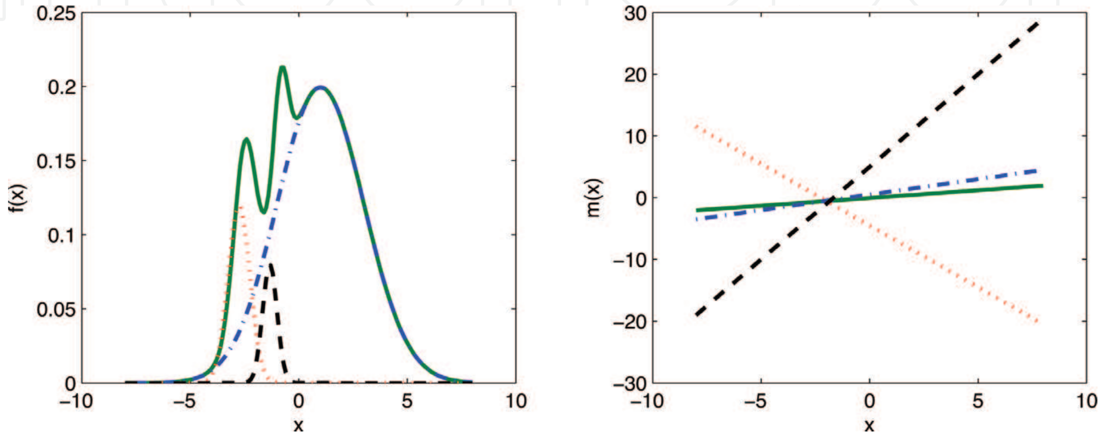

Figure 1.
Plot of the three-component normal mixture model $0.3N(-4 - 2x, 1) + 0.5N(0.5 + 0.5x, 1) + 0.2N(4.5 + 3x, 1)$*. Left panel: Plot of the density functions of the mixture as well as their three weighted components ; right panel: plots of regression lines. Mixture model: solid line "–" component one: dotted lines "…" component two: dashed lines "− −" and component three: dotted-dashed lines "−·"*

(see also, for example, [37]) showed that using fully non-informative prior distributions may lead to improper posterior distributions and hence is strictly prohibitive.

We assign a symmetric Dirichlet distribution to $\boldsymbol{\pi}$ as follows

$$\boldsymbol{\pi}|\alpha \sim D_K(\alpha, \cdots, \alpha) \tag{7}$$

in which $\alpha(>0)$ is the hyper-parameter, which is treated to fixed and unknown. In the applications, we can take sensitive analysis by setting smaller and larger values for α. See section 4 for more details.

Let $\mathbf{K} = \{K_1, \cdots, K_n\}$ be the collection of all configurations. A Bayesian procedure for model-based clustering mainly focuses on exploring the behavior of the posterior of $\mathbf{K}$ given data, which is given by

$$p(\mathbf{K}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{K})p(\mathbf{K}) \tag{8}$$

where $p(\mathbf{Y}|\mathbf{K})$ is the marginal distribution of $p(\mathbf{Y},\ \pi, \Xi|\mathbf{K})$ with $\boldsymbol{\pi}$ and Ξ being integrated out. Generally, no closed form can be available for this target distribution. Markov Chain Monte Carlo [38, 39] sampling method can be used to conduct posterior analysis. In particular, one can follow the routine in Tanner and Wong [40] and treat the latent quantities $\{\boldsymbol{\pi}, \mathbf{K},\ \Xi\}$ as the missing data and augment them with the observed data. Posterior analysis is carried out based on the joint distribution $p(\boldsymbol{\pi}, \mathbf{K}, \Xi|\mathbf{Y})$. In this case, block Gibbs sampler [41, 42] can be implemented to draw observations from such target distribution. The Gibbs sampler is iteratively implemented by drawing: (i) Ξ from $p(\Xi\ |\boldsymbol{\pi}, \mathbf{K}, \mathbf{Y})$; (ii) $\boldsymbol{\pi}$ from $p(\boldsymbol{\pi}\ |\mathbf{K}, \Xi, \mathbf{Y})$ and $\mathbf{K}$ from $p(\mathbf{K}|\ \boldsymbol{\pi}, \Xi, \mathbf{Y})$ till convergence. The convergence can be monitored by the "estimated potential scale reduction" (EPSR) values [43] or by plotting the traces of estimates against iterations under different starting values. Note that except for (i), all full conditionals involved in the Gibbs sampler are standard. However, drawing Ξ in (i) depends on the specific form of the density function f_k and sometimes requires implementing Metropolis-Hastings algorithm (MH, [44, 45]) or rejection sampling [46].

2.3 Label switching

Formulating the model-based clustering problem into mixture model Eq. (2) faces the model identification. A statistical model is said to be identified if the observed likelihood is uniquely determined by unknown parameters. A less identified model may be problematic and will distort the estimates of unknown parameters. It is easily showed that the observed likelihood of data is only determined up to the permutation of the component labels. As a matter of fact, suppose that there are the pair $\{\boldsymbol{\pi}^{(1)},\ \ \Xi^{(1)}\}$ and $\{\boldsymbol{\pi}^{(2)},\ \ \Xi^{(2)}\}$ such that

$$p\left(y|\boldsymbol{\pi}^{(1)},\ \ \Xi^{(1)}\right) = p\left(y|\boldsymbol{\pi}^{(2)}, \Xi^{(2)}\right) \tag{9}$$

then there exists a permutation $\nu : \{1, 2, \cdots, K\} \mapsto \{1, 2, \cdots, K\}$ such that $\pi_k^{(1)} = \pi_{\nu(k)}^{(2)}$, $\boldsymbol{\beta}_k^{(1)} = \boldsymbol{\beta}_{\nu(k)}^{(2)}$ and $\boldsymbol{\tau}_k^{(1)} = \boldsymbol{\tau}_{\nu(k)}^{(2)}$. In this setting, we can not distinguish $\mathcal{K}$ and $\nu \circ \mathcal{K}$ in terms of data ("$\circ$" denotes the operator of function composition). With this in mind, any exchangeable priors on $\boldsymbol{\pi}$ and Ξ like Eqs. (6) and (7) produces symmetric and multi-modal posterior distributions with up to $K!$ copies of each "genuine" mode, which induces the so-called label switching problem on Bayesian estimate. Traditional approaches to eliminating such exchangeability is to impose identifiability constraints on the parameter space. However, as pointed out by Frühwirth-Schnatter [18], an unappropriate identifiability constraint may not be able to eliminate label switching. Many efforts have been devoted to coping with this issue, see Chapter 11 in Lee [47] for a review. Among them, the relabeling algorithm [48] is more appealing due to its simplicity and flexibility. The relabeling sampling procedure takes a decision-theoretical approach and requires specifying an appropriate loss function to measure the loss in terms of the classification probability. The model identification problem is addressed via postprocessing the MCMC output to minimize the posterior expected loss. Specifically, let $\boldsymbol{\theta}$ be the collection of $\boldsymbol{\Xi}$ and $\boldsymbol{\pi}$, and write $\mathbf{Q} = \{q_{ik}(\boldsymbol{\theta})$ as the matrix of allocation probabilities of order $n \times K$ with $q_{ik}(\theta) = \mathbb{P}(K_i = k|\mathbf{Y},\ \ \boldsymbol{\theta})$. In the context of clustering, the loss function can be defined on the cluster label $\mathcal{K}$ as follows

$$\mathcal{L}_0(\mathcal{K};\ \ \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log q_{iK_i}(\boldsymbol{\theta}). \tag{10}$$

Given that $\boldsymbol{\theta}^{(1)}, \cdots,\ \ \boldsymbol{\theta}^{(M)}$ are the sampled parameters and let $\nu_1, \cdots, \nu_M$ be the permutation applied to them. The relabeling algorithm proceeds by selecting initial values for the ν_ms, which are generally taken to be the identity permutations, then iterating the following steps until a fixed point is reached.

a. Choose $\hat{\mathcal{K}}$ to minimize $\sum_{m=1}^{M}\mathcal{L}_0\left(\mathcal{K}, \nu_m\left(\boldsymbol{\theta}^{(m)}\right)\right)$;

b. For $m = 1, 2, \cdots, M$, choose ν_m to minimize $\mathcal{L}_0\left(\hat{\mathcal{K}}, v_m\left(\boldsymbol{\theta}^{(m)}\right)\right)$.

2.4 Posterior inference

Once the label switching is taken care of, the MCMC samples can be used to draw posterior inference. For example, the joint Bayesian estimate of $\boldsymbol{\theta}$ can be obtained easily via the corresponding sample means of the generated observations via ergodic average as follows:

$$\hat{\boldsymbol{\beta}}_k = M^{-1}\sum_{m=1}^{M}\boldsymbol{\beta}_k^{(m)}, \hat{\boldsymbol{\tau}}_k = M^{-1}\sum_{m=1}^{M}\ \ \boldsymbol{\tau}_k^{(m)}, \quad \text{and}\, \hat{\pi}_k = M^{-1}\sum_{m=1}^{M}\pi_k^{(m)} \tag{11}$$

The consistent estimates of the covariance matrix of estimates can be obtained via sample covariance matrix.

Given the observations $\{\mathbf{K}^{(m)} : m = 1, 2, \cdots, M\}$ drawn from the posterior $p(\mathbf{K}|\mathbf{Y})$ via MCMC sampling, serval methods can be available for arriving at a point estimate of the clustering using draws from the posterior clustering distribution. The simplest method, known as the maximum a posteriori (MAP) clustering, is to select the observed clustering that maximizes the density of the posterior clustering distribution, i.e.,

$$\hat{\mathcal{K}} : \hat{K}_i = \text{argmax}_{k=1,\cdots,K}\mathbb{P}(K_i = k|\mathbf{Y}) \tag{12}$$

in which $\mathbb{P}(K_i = k|\mathbf{Y})$ can be approximated by

$$\mathbb{P}(K_i = k|\mathbf{Y}) \approx M^{-1}\sum_{m=1}^{M} I\left\{K_i^{(m)} = k\right\}. \tag{13}$$

A more appreciate alternative to MAP is based on the pairwise probability matrix, an $n \times n$ association matrix $\delta(\mathcal{K})$ with the (i,j)th element formed by the indicator of whether the subject i is clustered with subject j. Element-wise averaging of these association matrices yields the pairwise probability matrix of clustering, denoted $\hat{\psi}$. Medvedovic and Sivaganesan [49] and Medvedovic et al. [50] suggested a clustering estimate of $\mathcal{K}$ by using the pairwise probability matrix $\hat{\psi}$ as a distance matrix in hierarchical/agglomerative clustering. However, as augured by Dahl [51], such routine seems counterintuitive to apply an ad hoc clustering method on top of a model which itself produces clusterings. In the context of Dirichlet process mixture-based clustering, Dahl [51] proposed a least-squares model-based clustering method by using draws from a posterior clustering distribution. Specifically, the least-squares clustering $\mathcal{K}_{LS}$ is the observed clustering $\mathcal{K}_{LS}$, which minimizes the sum of squared deviations of its association matrix $(\mathcal{K})$ from the pairwise probability matrix:

$$\hat{\mathcal{K}}_{LS} = \text{argmin}_{\mathcal{K}\in\left\{\mathbf{K}^{(1)},\cdots,\mathbf{K}^{(m)}\right\}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\delta(i,j)(\mathcal{K}) - \hat{\psi}(i,j)\right)^2. \tag{14}$$

Dahl [51] showed that the least-squares clustering has the advantage over those in Medvedovic and Sivaganesan [49] since it utilizes the information from all the clusterings and is intuitively appealing for the “average” clustering instead of forming a clustering via an external, ad hoc clustering algorithm.

3. Assessing heterogeneity of two-part model

In this section, we first proposed a two-part regression model for the fractional data especially for the U shaped fractional data and then extend the method discussed above to the current situation to address the possible heterogeneity of the population underlying data.

3.1 Two-part model for U shaped fractional data

Suppose that for subject/individual $i(=1,\cdots,n)$, y_i is an univariate fractional response taking values in $[0,1]$; $\mathbf{x}_i$ is an $m\times 1$ fixed covariate vector denoting various explanatory factors under consideration. Usually, y_i suffers from excess zeros and ones on the boundaries, and the whole data set takes on the U shape. In modeling such data, we introduce a three-category indicator variable d_i and a continuous intensity variable z_i such that

$$d_i = \begin{cases} 1 & \text{if} \quad y_i = 0 \\ 2 & \text{if} \quad y_i = 1 \\ 3 & \text{if} \quad 0 < y_i < 1 \end{cases} \quad \text{and} \quad z_i = \begin{cases} h(y_i) & \text{if} \quad 0 < y_i < 1 \\ \text{irrelevant} & \text{if} \quad y_i = 0,1 \end{cases} \tag{15}$$

where $h(\cdot)$ is any monotone increasing function such that $z_i \in (-\infty, +\infty)$. That is, we break the data set into three parts: two parts corresponding to zeros and ones respectively and one part corresponding to the continuous values between 0 and 1. We formulate a two-part model for y_i by first specifying a baseline-category logits model [52] for d_i and then a conditional continuous model for z_i. The baseline-category logits model is assumed that conditional upon $\mathbf{x}_i$, d_is are independent satisfying the following logits models simultaneously: for $j = 1, 2$,

$$\log \frac{\mathbb{P}(d_i = j|\mathbf{x}_i)}{\mathbb{P}(d_i = 3|\mathbf{x}_i)} = \mathbf{x}_i^T \boldsymbol{\alpha}_j \tag{16}$$

where $\boldsymbol{\alpha}_j$ is an $m \times 1$ regression coefficients vector. We use category $d_i = 3$ as the reference for the ease of parameters interpretation. For example, the magnitude of $\alpha_{j\ell}$ in $\boldsymbol{\alpha}_j$ indicates that the increase of one unit in $x_{i\ell}$ will increase $e^{\alpha_{j\ell}}$ times chance of $d_i = j$ over that of $d_i = 3$.

The conditional continuous model for z_i is given by

$$p(z_i|d_i = 3, \mathbf{x}_i) = p^z\left(z_i|\mathbf{x}_i^T\boldsymbol{\gamma}, \tau\right) \tag{17}$$

or equivalently

$$p\left(y_i|0 < y_i < 1, \mathbf{x}_i\right) = p^z\left(h\left(y_i\right)|\mathbf{x}_i^T\boldsymbol{\gamma}, \psi\right)|\dot{h}\left(y_i\right)| \tag{18}$$

where $\dot{h}(s) = dh/ds$, $p^z(u|a, \tau)$ is the normal density with mean a and variance $\tau > 0$, and $\boldsymbol{\gamma}$ like that in Eq. (16), is the regression coefficient vector. Although the identical covariates are taken in Eqs. (16) and (17), this is not necessary in practice. Each equation can own their covariates. This can be achieved by imposing particular structure on the regression coefficients. For example, we can exclude x_{i1} from Eq. (17) by restricting γ_1 in $\boldsymbol{\gamma}$ to be zero.

It follows from Eqs. (16) and (17) that marginal distribution of y_i is given by

$$p\left(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau\right) = q_{i1}\delta_0 + q_{i2}\delta_1 + \left(1 - q_{i1} - q_{i2}\right)p\left(y_i|0 < y_i < 1, \mathbf{x}_i, \ \ \boldsymbol{\gamma}, \tau\right) \tag{19}$$

where $q_{ij} = \mathbb{P}\left(d_i = j|\mathbf{x}_i, \ \ \boldsymbol{\alpha}_j\right)(\ j = 1, 2)$ is the response probability specified by Eq. (16) and $\boldsymbol{\beta}$ is the regression parameters constituted by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ and $\boldsymbol{\gamma}$.

3.2 Assessing heterogeneity of two-part model

To detect the possible heterogeneity among y_i, we extend the model Eq. (18) to the mixture case by assuming that conditional upon $K_i = k$, d_i and z_i satisfy Eqs. (16) and (17) with $\boldsymbol{\alpha}_j$ replaced by $\boldsymbol{\alpha}_{jk}$ and $(\boldsymbol{\gamma}, \tau)$ by $(\boldsymbol{\gamma}_k, \tau_k)$ respectively. This indicates that the mixture component f_k in Eq. (1) in Section 2 is given by Eq. (19) with $\boldsymbol{\beta} = \boldsymbol{\beta}_k$ and $\tau = \tau_k$.

For the Bayesian analysis, the general forms of full conditionals involved in the model-based clustering have been given in Section 2. We here only focus on the technical details of the conditional distribution of $\boldsymbol{\Xi}$ in (i) in the Gibbs sampler.

We assume that the prior of τ_k is the same as that in Eq. (6), while the priors of $\boldsymbol{\beta}_k$ are taken as $p(\boldsymbol{\beta}_k) = p(\boldsymbol{\alpha}_{k1})p(\boldsymbol{\alpha}_{k2})p(\boldsymbol{\gamma}_k)$, in which

$$p(\boldsymbol{\alpha}_{k\ell}) \overset{D}{=} N_m(\boldsymbol{\alpha}_{\ell 0}, \boldsymbol{\Sigma}_{\alpha\ell 0})(\ell = 1, 2), \quad p(\boldsymbol{\gamma}_k) \overset{D}{=} N_m\left(\boldsymbol{\gamma}_0, \ \boldsymbol{\Sigma}_{\gamma 0}\right). \tag{20}$$

where $\boldsymbol{\alpha}_{\ell 0}$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\Sigma}_{\alpha\ell 0}$ and $\boldsymbol{\Sigma}_{\gamma 0}$ are the hyper-parameters treated to be known.

Gibbs sampling $\boldsymbol{\Xi}$ now becomes drawing $\boldsymbol{\alpha}_k$, $\boldsymbol{\gamma}_k$ and τ_k alternatively from the full conditional distributions $p(\boldsymbol{\alpha}_k|\mathbf{K}, \mathbf{Y})$, $p(\boldsymbol{\gamma}_k|\tau_k, \mathbf{K}, \mathbf{Y})$ and $p(\tau_k|\ \boldsymbol{\gamma}_k, \mathbf{K}, \mathbf{Y})$ respectively. By some algebras, it can be shown that

$$\begin{aligned} p(\boldsymbol{\alpha}_k|\mathbf{K}, \mathbf{Y}) &\propto p(\boldsymbol{\alpha}_k)\prod_{K_i = k} p(d_i|\mathbf{x}_i, \ \ \boldsymbol{\alpha}_k), \\ p(\boldsymbol{\gamma}_k|\tau_k, \mathbf{K}, \mathbf{Y}) &\propto p(\boldsymbol{\gamma}_k)\prod_{K_i = k} p\left(z_i|d_i = 3, \mathbf{x}_i^T\boldsymbol{\gamma}_k, \tau_k\right), \\ p(\tau_k|\ \boldsymbol{\gamma}_k, \mathbf{K}, \mathbf{Y}) &\propto p(\tau_k)\prod_{K_i = k} p\left(z_i|d_i = 3, \mathbf{x}_i^T\boldsymbol{\gamma}_k, \tau_k\right) \end{aligned} \tag{21}$$

in which the full conditionals of γ_k and τ_k are easily obtained and given by

$$p(\gamma_k|\tau_k, \mathbf{K}, \mathbf{Y}) \overset{D}{=} N\left(\hat{\gamma}_k, \hat{\boldsymbol{\Sigma}}_{\gamma k}\right) \tag{22}$$

$$p\left(\tau_k^{-1}|\gamma_k, \mathbf{K}, \mathbf{Y}\right) \overset{D}{=} Gamma\left(\hat{\alpha}_k, \hat{\beta}_k\right) \tag{23}$$

in which

$$\begin{aligned}
&\hat{\boldsymbol{\Sigma}}_{\gamma k} = \left(\sum_{K_i=k:d_i=2} \mathbf{x}_i\mathbf{x}_i^T/\tau_k + \boldsymbol{\Sigma}_{\gamma 0}^{-1}\right)^{-1}, \\
&\hat{\gamma}_k = \hat{\boldsymbol{\Sigma}}_k\left(\boldsymbol{\Sigma}_{\gamma 0}^{-1}\ \gamma_0 + \sum_{K_i=k,\, d_i=3} \mathbf{x}_i z_i/\tau_k\right), \\
&\hat{\alpha}_k = \alpha_0 + n_k/2, \\
&\hat{\beta}_k = \beta_0 + \sum_{K_i=k,\, d_i=3}\left(z_i - \mathbf{x}_i^T\boldsymbol{\gamma}_k\right)^2/2
\end{aligned} \tag{24}$$

and $n_k = \#\{K_i = k, d_i = 3\}$.

However, drawing $\boldsymbol{\alpha}_{k\ell}$ is more tedious since its distribution loses the standard form. We first note that

$$p(\boldsymbol{\alpha}_{k\ell}|\,\boldsymbol{\alpha}_{k,-\ell}, \mathbf{K}, \mathbf{Y}) \propto p(\boldsymbol{\alpha}_{k\ell}) \prod_{K_i=k}^{n} \frac{\exp\left(\tilde{d}_{i\ell}\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)} \tag{25}$$

where $\tilde{d}_{i\ell} = I\{d_i = \ell\}$ and $C_{ik\ell} = \log\left(1.0 + \exp\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k,-\ell}\right)\right)$; $\boldsymbol{\alpha}_{k,-\ell}$ denotes the set $\boldsymbol{\alpha}_k$ with $\boldsymbol{\alpha}_{k\ell}$ removed. Following the similar routine in Polson, Scott, and Windle [53], we recast the logistic function Eq. (25) as follows

$$\begin{aligned}
\frac{\exp\left(\tilde{d}_{i\ell}\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k1} - C_{ik\ell}\right)} &= 2^{-1}\exp\left\{\kappa_{i1}\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)\right\} \\
&\times\int_0^\infty \exp\left\{-\frac{1}{2}\omega_{i\ell}\left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)^2\right\} p_{\mathrm{PG}}(\omega_{i\ell})\,\mathrm{d}\omega_{i\ell}
\end{aligned} \tag{26}$$

in which $\kappa_{i\ell} = \tilde{d}_{i\ell} - 1/2$ and p_{PG} is the well-known $\mathrm{PG}(1, 0)$ density function [53]. If one introduces n independent Pólya-Gamma variables $\omega_{i\ell}$ into the current analysis, then,

$$p(\omega_{i\ell}|\,\boldsymbol{\alpha}_{k\ell}, \boldsymbol{\alpha}_{k,-\ell}, \mathbf{K}, \mathbf{Y}) \overset{D}{=} \mathrm{PG}\left(1, \left(\mathbf{x}_i^T\boldsymbol{\alpha}_{k\ell} - C_{ik\ell}\right)\right) \tag{27}$$

$$p(\boldsymbol{\alpha}_{k\ell}|\,\boldsymbol{\alpha}_{k,-\ell},\ \ \Omega, \mathbf{K}, \mathbf{Y}) \overset{D}{=} N\left(\hat{\boldsymbol{\alpha}}_{k\ell}, \hat{\boldsymbol{\Sigma}}_{\alpha k\ell}\right) \tag{28}$$

where

$$\hat{\boldsymbol{\Sigma}}_{\alpha k\ell} = \left(\sum_{K_i=k}\mathbf{x}_i\mathbf{x}_i^T\omega_{i\ell} + \boldsymbol{\Sigma}_{\alpha\ell 0}^{-1}\right)^{-1}, \quad \hat{\boldsymbol{\alpha}}_{k\ell} = \hat{\boldsymbol{\Sigma}}_{\alpha k\ell}\left(\boldsymbol{\Sigma}_{\alpha\ell 0}^{-1}\boldsymbol{\alpha}_{\ell 0} + \sum_{K_i=k}\mathbf{x}_i\eta_{ik\ell}\right) \tag{29}$$

with $\eta_{ik\ell} = \kappa_{i\ell} + C_{ik\ell}\omega_{i\ell}$. Consequently, drawing $\boldsymbol{\alpha}_{k\ell}$ is accomplished by first drawing $\omega_{i\ell}$ from the Pólya gamma distribution and then drawing $\boldsymbol{\alpha}_{k\ell}$ from the normal distribution. The draw of $\omega_{i\ell}$ is a little intractable since its density function involves the infinite sum. By taking advantage of series sampling method [54], Polson et al. [53] devised a rejection algorithm for generating observations from such type of distribution. Their method can be adapted to draw $\omega_{i\ell}$, see also [55].

4. A real example

In this section, a small portion of cocaine use data is analyzed to illustrate the practical value of the proposed methodology. The data are obtained from the 322 cocaine use patients who were admitted in 1988–89 to the West Los Angeles Veterans Affairs Medical Center. The original data set is made up of 68 measurements in which 17 items were assessed at four unequally spanned time points. In this study, we mainly focus on the measurements 1 year after treatment and ignore the initial effects at the baseline. The measurements cover the information on the cocaine use, treatment received, psychological problems, social status, employments, and so on. Among them, the measurement "cocaine use per month" (denoted by CC) plays a critical role since it identifies the severity of cocaine use of patients and therefore is treated as the dependent response. The CC is originally measured by 0–30 points but suffered from small portion of fractions. We identify CC/30 as the fraction response in [0,1]. In view of that the missing data are presented, we delete the subjects with missing values. The total sample size is 228. A primary analysis shows that CC/30 has excessive zeros and ones. **Figure 2** gives the histograms of CC/30 and their fractional values in (0,1) via logistic transformation. It can be seen clearly that there is a large number of zeros and unities accumulated on the boundaries. The proportions of zeros and unities are about 15 and 4%, respectively. Moreover, panel (b) in **Figure 2** indicates that single parametric model may be unappreciate for fitting the continuous valued variable.

To explore the effects of exogenous factors on the cocaine use, the following measurements are selected as the explanatory variables: the occupational status of a patient (x_1). This is a binary indicator: 1 for employment and 0 for non-employment; the level of technical proficiency of patients engaged in work (x_2): scaled on 0–4 points and the patient's lifestyle (x_3) with five-point scale. To unify the scales, all covariates are standardized. However, a preliminary analysis shows that there exists strong multiple collinearity among these covariates. The minimum eigenvalue of sample covariance matrix equals to 0.06284, which approaches zero. We remove such collinearity by implementing principle component analysis (PCA)

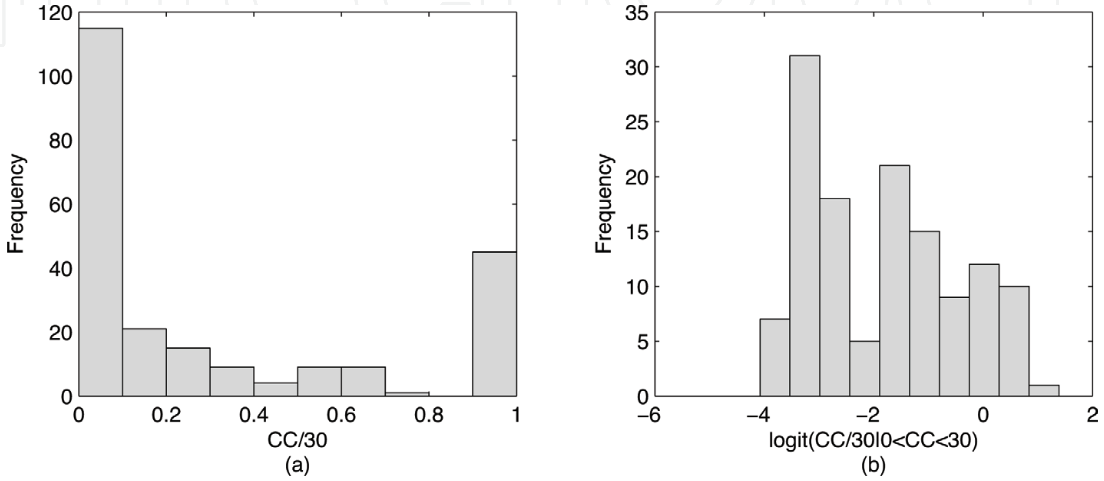

Figure 2.
Plots of CC in cocaine use data: (a) Histograms of CC/30; and (b) histograms of CC/30 on logistic transformation conditional on CC/30 in (0,1).

and treat the scores of the first two components (still denoted by x_1 and x_2) as our explanatory variables. These two principle components can be interpreted as the levels related to the patients' occupation and their live life.

To formulate a two-part model for the observed responses, we identity $\mathrm{CC}_i/30$ with d_i and z_i, where d_i is the three-category indicator indicating the state of cocaine use after one year treatments: quitting cocaine successfully (state 1), insisting on cocaine use every day in a month (state 2) and taking the cocaine occasionally (state 3); z_i is the intensity variable representing the numer of days of cocaine use in a month. We assess the effects of exogenous factors x_1 and x_2 on the cocaine use via Eqs. (16) and (17), respectively.

We proceed data analysis by first fitting data to the K-component mixture two-part models with $K = 1, 2, \cdots, 6$. The model fits are assessed via AIC, AICc, and BIC, which are defined as $-2\log p\left(\mathbf{Y}|\hat{\boldsymbol{\theta}}_K\right)$ penalized by $2d_K$, $2n(d_K+1)/(n-d_K-2)$, and $d_K \log n$ respectively, where $\hat{\theta}_K$ is the MLE of $\boldsymbol{\theta}_K$ and d_K is the dimension of unknown parameters under the model K. In view of that the Bayesian estimates and the ML estimates are close to each other, we replace the ML estimates by their Bayesian counterparts in evaluating AIC, AICc, and BIC. For computation, we take $\alpha = n^{-1}$, n^0, n^1, and n^2 in Eq. (7), which represents our knowledge about $\boldsymbol{\pi}$ *a prior*. Note that for large value of α, the Dirichlet distribution places most of the mass on its center and the prior Eq. (7) tends to be informative. However, for small α, the Dirichlet distribution concentrates the mass on the boundaries of sampling space and the distribution tends to be degenerated and sparse. As a result, some components in $\boldsymbol{\pi}$ reduces to zeros. When $\alpha = 1$, $D_K(\alpha, \cdots, \alpha)$ becomes an uniform distribution on the simplex $\mathbb{S}_K$. For the inputs of the hyper-parameters involved in the priors Eq. (20), we take $\boldsymbol{\alpha}_{0\ell} = \boldsymbol{\gamma}_0 = \mathbf{0}_3$, $\boldsymbol{\Sigma}_{\alpha\ell 0} = \boldsymbol{\Sigma}_{\gamma 0} = 100\mathbf{I}_3$, $\alpha_{\gamma 0} = 2.0$ and $\beta_{\gamma 0} = 2.0$. These values ensure the priors Eq. (20) to be inflated enough and represent the weak information on the parameters.

The relabeling MCMC algorithm described in Section 2 is implemented to draw observations from the posterior. The convergence of algorithm is monitored by plotting the traces of estimates against iterations under three starting values. **Figure 3** presents the values of EPSR of unknown parameters against the number of

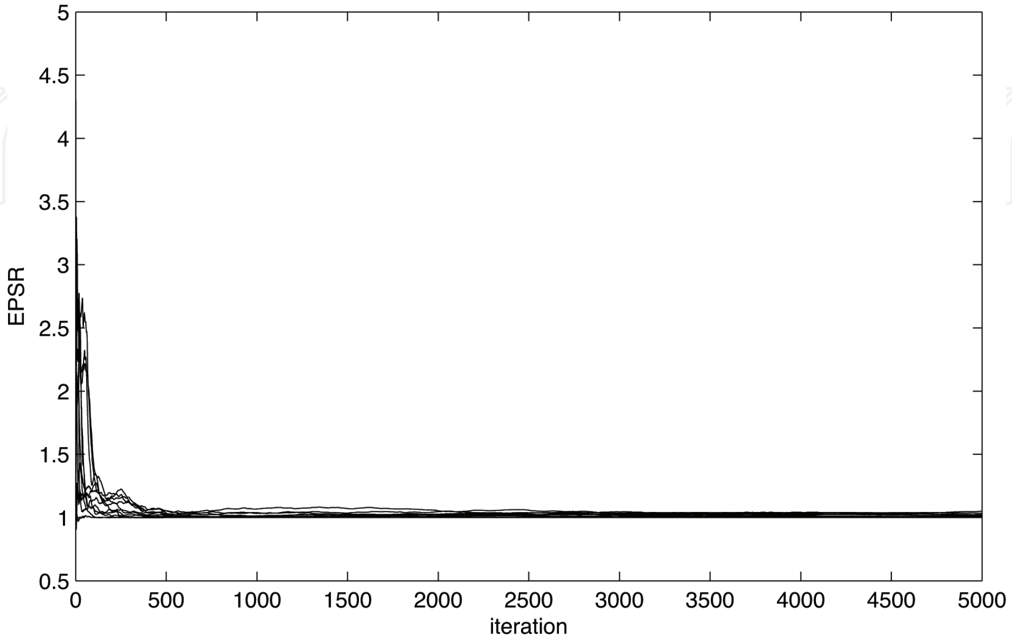

Figure 3.
Plots of values of EPSR of estimates of unknown parameters against the number of iterations under three different staring values in the cocaine use example: K = 2.

	Model	$\alpha = 1/n$	$\alpha = n^0$	$\alpha = n$	$\alpha = n^2$
AIC	$K = 1$	921.3887	–	–	–
	$K = 2$	923.0580	907.4485	901.9474	901.4380
	$K = 3$	929.0698	926.5423	956.4945	994.5039
	$K = 4$	990.7506	949.5966	1014.5477	1006.4228
	$K = 5$	989.2483	971.4688	1069.1561	1037.5005
	$K = 6$	1097.6853	1007.4899	1091.8049	1086.8491
AICc	$K = 1$	882.4025	–	–	–
	$K = 2$	885.5434	869.9339	864.4329	863.9234
	$K = 3$	875.9005	873.3730	903.3253	941.3347
	$K = 4$	925.3159	884.1618	949.1130	940.9880
	$K = 5$	915.5836	897.8041	995.4914	963.8357
	$K = 6$	1020.6483	930.4529	1014.7679	1009.8120
BIC	$K = 1$	995.6821	–	–	–
	$K = 2$	995.0742	979.4647	973.9637	973.4542
	$K = 3$	1038.8088	1036.2814	1066.2335	1104.2429
	$K = 4$	1138.2125	1097.0585	1162.0096	1153.8846
	$K = 5$	1174.4330	1156.6534	1254.3408	1222.6851
	$K = 6$	1320.5928	1230.3973	1314.7124	1309.7565

Table 1.
Summary statistics of AIC, AICcc, and BIC for model selection in cocaine use data analysis.

iterations under three different starting values with $K = 2$. It shows that the convergence of the proposed algorithm is fast and the values of EPSR are less than 1.2 in less than 1000 iterations. Hence, 3000 observations, after removing the initial 2000 iterations, are collected for calculating AIC, AICc, and BIC. The resulting summary is given in **Table 1**.

Examination of **Table 1** shows that all measures favor the model with $K = 2$. This indicates that the proposed model with two groups seems to give a better fit to the data. It also indicates that large α favors the model fit. Furthermore, we calculate the posterior predictive density estimate of z_i under the elected model. Results (not represented here for saving spaces) show that our method can be successful in capturing the skewness and modes of data. We also follow [56] to plot the estimated residuals $\hat{\delta}_i = z_i - \hat{\boldsymbol{\gamma}} \mathbf{x}_i^T$ and find that these plots lie within two parallel horizontal lines that are centered at zero, with nonlinear or quadratic trends detected. This roughly indicates that the proposed linear model Eq. (18) is adequate.

Table 2 presents the estimates of unknown parameters associated with corresponding standard deviation (SD) estimates under $K = 2$. Based on **Table 2**, we can find the following facts: (i) for Part one, we observe that except for $\hat{\alpha}_{23}$, the Bayesian estimates of unknown parameters within two clusters have the same signs but their magnitudes are more different. For example, the estimate of α_{11} within Cluster one is -1.540 with SD 0.587 while equals to -0.732 with SD 0.481 within Cluster two. This indicates that the baselines of logits Eq. (16) exist obvious difference. For α_{23}, the estimates between two clusters have the opposite signs. Recall that α_{23} quantifies the magnitude of effects of live life on the probability $\mathbb{P}(d_i = 2)$ over $\mathbb{P}(d_i = 3)$ on log scale. This shows that increasing the level of live life will lead to an

Para.	Component I		Component II	
	Est.	SD.	Est.	SD
α_{11}	−1.540	0.587	−0.732	0.481
α_{12}	0.150	0.317	0.604	0.322
α_{13}	0.261	0.703	0.188	0.601
α_{21}	−1.337	0.480	−1.059	0.545
α_{22}	−0.166	0.355	−0.229	0.418
α_{23}	0.232	0.378	−0.184	0.411
γ_1	−2.779	0.144	−0.490	0.215
γ_2	−0.029	0.080	−0.011	0.154
γ_3	0.087	0.144	0.179	0.240
τ	0.674	0.150	0.924	0.234

Table 2.
Summary statistics for the Bayesian estimates of unknown parameters in the cocaine use data.

opposite effect among two clusters; (ii) for Part two, although all the estimates within two clusters have the same signs but the levels of effects among them are obviously different. The estimates of γ_1 is -2.779 with SD 0.144 in the cluster one and attains -0.490 associated with SD 0.215 in the Cluster two. This indicates that the baseline of cocaine use in Cluster one is 50 times as much as that in Cluster two; and (iii) investigation of the estimate of τ also indicates that there exists the different amount of the fluctuation among two clusters.

5. Discussion

This chapter introduces a general Bayesian model-based clustering procedure for the regression model and proposed a Bayesian method for assessing the heterogeneity of fractional data within the mixture of two-part regression model framework. The heterogeneous fractional data arise mainly from two resources: one is that the excessive zeros and ones are accumulated upon the boundaries, and the other is that the underlying population may consist more than one components. For the first issue, we propose a novel two-part model, in which a three-category multinomial regression is suggested to model the occurrence probabilities of each part, and a conditional normal linear regression is used to fit the continuous positive values on logit scale. Such formulation is more appealing since it can ensure the probabilities on each part to be consistent and and at the same time maintains the coherent association across parts. For the second problem, we resort to the finite mixture model in which the cluster-specific components are specified via two-part model. MCMC sampling method is adopted to carry out the posterior analysis. The number of clusters and the configuration of observations are addressed based on the simulated observations from the posterior. We illustrate the proposed methodology in the analysis of cocaine use data.

When interest is concentrated upon the estimates, model identification is surely an important issue since it involves whether or not the estimates of component-specific quantities are meaningful. For a finite mixture model, model identification mainly stems from the label switching, in which the likelihood and the posterior are invariant under label permutation. Many efforts have devoted to alleviating such indeterminacy. Among them, parameters' constraints may be the most popular choice. However, an unappreciated constraint fails to deal with the label switching. In this case, one can follow the routine in Frühwirth-Schnatter [18] and implement random permutation sampling to find the suitable identifiability constraints. The random permutation sampler is similar to the unconstrained MCMC sampling but only at each sweep, the labels $\{1, \cdots, K\}$ are randomly permutated. The permutation aims to deliver a sample that explores the whole unconstrained parameter space and jumps between the various labeling subspaces in a balanced fashion. The output of such balanced sample can help us to find a suitable identifiability constraint. A more detailed discussion on model identification in the mixture context can be referred to, for example, [18, 57]. Instead, we resort to the relabeling algorithm for simplicity. Compared with the random permutation sampling, the relabeling method requires implementing MCMC samplng only once, thus saving the computation cost.

The methodology developed in this chapter can be extended to the case where latent factors are included to identify the unobserved heterogeneity due to some fixed convariates absent. Another possible extension is to establish a dynamic LVM, wherein model parameters vary across times. These issues may raise theoretical and computational challenges and therefore require further investigation.

Acknowledgements

The work presented here was fully supported by grant from the National Natural Science Foundation of China (NNSF 11471161). The authors are thankful to Professor Xin-Yuan Song, the Chinese University of Hong Kong, for using her cocaine use data in the real example.

Conflict of interest

The authors have no conflicts of interest to disclose.

Author details

Ye-Mao Xia[1], Qi-Hang Zhu[2] and Jian-Wei Gou[1]*

1 Department of Applied Mathematics, Nanjing Forestry University, Nanjing, China

2 College of Economics and Management, Nanjing Forestry University, Nanjing, China

*Address all correspondence to: gjw1983@139.com

References

[1] McLachlan GJ. Discriminant Analysis and Statistical Pattern Recognition. New York: John Wiley; 1992. DOI: 10.1002/0471725293.ch3

[2] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002; **97**(458):611-631. DOI: 10.2307/3085676

[3] Andrews JL, McNicholas PD. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. Statistics and Computing. 2012;**22**(5):1021-1029. DOI: 10.1007/s11222-011-9272-x

[4] Ripley BD. Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge Univeristy Press; 1996. DOI: 10.1080/00401706.1997.10485099

[5] Paalanen P, Kamarainen JK, Ilonen J, Kälviäinen H. Feature representation and discrimination based on Gaussian mixture model probability densities Practices and algorithms. Pattern Recognition. 2006;**39**(7):1346-1358. DOI: 10.1016/j.patcog.2006.01.005

[6] Qin LX, Self SG. The clustering of regression models method with applications in gene expression data. Biometrics. 2006;**62**:526-533

[7] McNicholas PD, Murphy TB. Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics. 2010; **21**:2705-2712. DOI: 10.1093/bioinformatics/btq498

[8] Yuan M, Kendziorski C. A unified approach for simultaneous gene clustering and differential expression identification. Biometrics. 2006;**62**:1089-1098

[9] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2002;**24**(7): 881-892. DOI: 10.1109/TPAMI.2002.1017616

[10] Mahmoudi MR, Akbarzadeh H, Parvin H, Nejatian S, Alinejad-Rokny H. Consensus function based on cluster-wise two level clustering. Artificial Intelligence Review. 2021;**54**:639-665. DOI: 10.1007/s10462-020-09862-1

[11] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Cam LML, Neyman J, editors. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics & Probability. Vol. 1. Berkeley, CA: University of California Press; 1967. pp. 281-297

[12] Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. Journal of the Royal Statistical Society, Series C. 1979;**28**(1):100-108. DOI: 10.2307/2346830

[13] Anderberg MR. Cluster Analysis for Applications. New York: Academic Press; 1973

[14] Everitt BS, Landau S, Leese M. Cluster Analysis. 4th ed. London: Hodder Arnold; 2001

[15] Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 2th ed. New Jersey: Prentice Hall; 1988

[16] Titterington DM, Smith AFM, Makov UE. Statistical Analysis of Finite Mixture Distributions. Chichester: John Wiley and Sons; 1985. DOI: 10.2307/2531224

[17] McLachlan GJ, Peel D. Finite Mixture Models. New York: John Wiley; 2000. DOI: 10.1002/0471721182

[18] Frühwirth-Schnatter S. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association. 2001, 2001; **96**(453):194-209. DOI: 10.1198/016214501750333063

[19] Fang KN, Ma SG. Three-part model for fractional response variables with application to Chinese household health insurance coverage. Journal of Applied Statistics. 2013;**40**(5):925-940. DOI: 10.1080/02664763.2012.758246

[20] McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman and Hall; 1989. DOI: 10.1007/978-1-4899-3242-6

[21] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;**82**(71):17-32. DOI: 10.1093/biomet/82.4.711

[22] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society, Series B. 1997; **59**:731C792. DOI: 10.1111/1467-9868.00095

[23] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F, editors. Second International Symposium on Information Theory. Budapest, Hungary: Akad¨¦mia Kiad¨®; 1973. pp. 267-281. DOI: DOI.10.1007/978-1-4612-1694-0-15

[24] Sugiura N. Further analysis of the data by Akaikes information criterion and the finite corrections. Communications in Statistics-Theory and Methods. 1978;**A7**:13-26

[25] Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. Biometrika. 1989;**76**:297-307. DOI: 10.1093/biomet/76.2.297

[26] Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978;**6**:461-464. DOI: 10.1214/aos/1176344136

[27] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;**22**(7):719-725. DOI: 10.1109/34.865189

[28] Berger JO. Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag; 1985. DOI: 10.1007/978-1-4757-4286-2

[29] Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995;**90**:773-795. DOI: 10.1080/01621459.1995.10476572

[30] Spiegelhalter DJ, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B. 2002; **64**:583-640. DOI: 10.1111/1467-9868.00353

[31] Spiegelhalter DJ, Thomas A, Best NG, Lunn D. WinBUGS User Manual. Version 1.4. Cambridge, England: MRC Biostatistics Unit; 2003. DOI: 10.1001/jama.284.24.3187

[32] Ferguson TS. A Bayesian analysis of some nonparametric problems. The Annals of Statistics. 1973;**1**(2):209-230. DOI: 10.1214/aos/1176342360

[33] Antoniak CE. Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. The Annals of Statistics. 1974;**2**:1152-1174. DOI: 10.1214/aos/1176342871

[34] Ishwaran H, James LF. Gibbs sampling methods for stickbreaking priors. Journal of the American Statistical Association. 2001;**96**: 161-173. DOI: 10.1198/016214501750332758

[35] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B. 1977;**39**:1-38

[36] Diebolt J, Robert CP. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, Series B. 1994;**56**: 363-375. DOI: 10.1111/j.2517-6161.1994.tb01985.x

[37] Roeder K, Wasserman L. Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association. 1997; **92**:894-902. DOI: 10.1080/01621459.1997.10474044

[38] Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984;**6**:721-741. DOI: 10.1109/TPAMI.1984.4767596

[39] Geyer CJ. Practical Markov chain Monte Carlo. Statistical Science. 1992;**7**: 473-511. DOI: 10.1214/ss/1177011137

[40] Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation(with discussion). Journal of the American statistical Association. 1987;**82**:528-550. DOI: 10.2307/2289463

[41] Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association. 1990; **85**:398-409. DOI: 10.1080/01621459.1990.10476213

[42] Ishwaran H, Zarepour M. Markov chain Monte Carlo in approximation Dirichlet and beta-parameter process hierarchical models. Biometrika. 2000; **87**:371-390

[43] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;**7**: 457-472. DOI: 10.2307/2246093

[44] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics. 1953;**21**:1087-1092. DOI: 10.1063/1.1699114

[45] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; **57**(1):97-109. DOI: 10.1093/biomet/57.1.97

[46] Gilks WR, Wild P. Adaptive rejection sampling for gibbs sampling. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1992;**41**(2): 337-348. DOI: 10.2307/2347565

[47] Lee SY. Structural Equation Modeling: A Bayesian Approach. New York: John Wiley & Sons; 2007

[48] Stephens M. Dealing with label-switching in mixture models. Journal of the Royal Statistical Society, Series B. 2000;**62**:795-809. DOI: 10.1111/1467-9868.00265

[49] Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics. 2002;**18**(9): 1194-1206. DOI: 10.1093/bioinformatics/18.9.1194

[50] Medvedovic M, Yeung KY, Bumgarner RE. Bayesian mixture model based clustering of replicated microarray data. Bioinformatics. 2004; **20**(8):1222-1232. DOI: 10.1093/bioinformatics/bth068

[51] Dahl DB. Model-based clustering for expression data via a Dirichlet process mixture model. In: Do KA, Müller P, Vannucci M, editors. Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press; 2006. DOI: 10.1017/CBO9780511584589.011

[52] Agresti A. Categorical Data Analysis. 2nd ed. New York: John Wiley & Sons; 2003

[53] Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using pólya Gamma latent variables. Journal of the American Statistical Association. 2013, 2013;**108**(504): 1339-1349. DOI: 10.1080/01621459.2013.829001

[54] Devroye L. The series method in random variate generation and its application to the Kolmogorov-Smirnov distribution. American Journal of Mathematical and Management Sciences. 1981;**1**:359-379. DOI: 10.1080/01966324.1981.10737080

[55] Gou JW, Xia YM, Jiang DP. Bayesian analysis of two-part nonlinear latent variable model: Semiparametric method. Statistical Modeling. Published on line. 2021. DOI: 10.1177/1471082X211059233

[56] Xia YM, Tang NS, Gou JW. Generalized linear latent models for multivariate longitudinal measurements mixed with hidden Markov models. Journal of Multivariate Analysis. 2016; **152**:259-275. DOI: 10.1016/j.jmva.2016.09.001

[57] Jasra A, Holmes CC, Stephens DA. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science. 2005;**20**(1):50-67. DOI: 10.1214/088342305000000016