Communications for Statistical Applications and Methods
2015, Vol. 22, No. 6, 625–637

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.6.625
Print ISSN 2287-7843 / Online ISSN 2383-4757

Dirichlet Process Mixtures of Linear Mixed Regressions

Minjung Kyung[1,a]

[a]Department of Statistics, Duksung Women's University, Korea

Abstract

We develop a Bayesian clustering procedure based on a Dirichlet process prior with cluster specific random effects. Gibbs sampling of a normal mixture of linear mixed regressions with a Dirichlet process was implemented to calculate posterior probabilities when the number of clusters was unknown. Our approach (unlike its counterparts) provides simultaneous partitioning and parameter estimation with the computation of the classification probabilities. A Monte Carlo study of curve estimation results showed that the model was useful for function estimation. We find that the proposed Dirichlet process mixture model with cluster specific random effects detects clusters sensitively by combining vague edges into different clusters. Examples are given to show how these models perform on real data.

Keywords: normal mixture, cluster specific random effect, model-based cluster, linear mixed regression, Dirichlet process

1. Introduction

Clustering algorithms attempt to find a partition of a finite set of objects in to a not necessarily predetermined number of nonempty subsets. Many methods have been proposed in the literature. An alternative form assumes a mixture model with an unknown number of components. Each mixture component depends on a parameter vector which could have common and peculiar components. The Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) has been widely used for the parameter estimation. Clusters are formed based on the posterior probability of cluster for each subject after a number of clusters have been chosen (Banfield and Raftery, 1993; Dasgupta and Raftery, 1998; Fraley and Raftery, 2002; McLachlan and Basford, 1988; McLachlan and Peel, 2000). This method has been termed as model-based clustering (MBC) by Dasgupta and Raftery (1998).

As a special case of clustering methods, mixtures of linear regression have been discussed in the literature. Mixtures of regression were introduced by Quandt (1958) as the switching regressions problem. Quandt and Ramsey (1978) introduced the moment generating function estimator defined as the estimator which minimizes the sum of squares for differences between the theoretical and sample moment generating functions. The consistency and asymptotic normality of the estimator are proved. Kiefer (1978) showed that for the mixture of regressions problem the likelihood equations have consistent root despite the unbounded likelihood function.

A (generalized) linear *mixed* model can be specified to accommodate outcome variables conditional on mixtures of possibly correlated random and fixed effects (Breslow and Clayton, 1993;

This research was supported by the Duksung Women's University research grant (No. 3000002179).

[1] Department of Statistics, Duksung Women's University, 33 Samyang-ro 144-gil, Dobong-gu, Seoul 01369, Korea.
Email: mkyung@duksung.ac.kr

Published 30 November 2015 / journal homepage: http://csam.or.kr

© 2015 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

Buonaccorsi, 1996; Wang, *et al.*, 1998; Wolfinger and O'Connell, 1993). In many cases, the observations are correlated and there may be other underlying phenomena that contribute to the resulting variability. For example, observations are repeatedly measured for each subject in longitudinal designs, which induces a correlation structure. Subjects in the same cluster may show similar behavior in clustered designs, while behavior will be different between clusters. The standard assumption of independence no longer holds in these cases.

Mixtures of linear mixed models have been recently and widely used in various fields for clustering and function estimation; consequently, random effects detect hidden structure and mixtures capture multimodality and skewness of distributions. Likelihood maximization through the EM algorithm has been used for estimation and the optimal number of components was determined by comparing different mixture models using information criteria such as AIC and BIC. Markov chain Monte Carlo (MCMC) algorithms are also developed based on a stochastic search algorithm for finding partitions of the data with a high posterior probability.

From a Bayesian perspective, the most common choice prior for the clustering structure with unknown number of components is the Dirichlet process (DP). DP mixture models were introduced by Ferguson (1973), who defined the process and investigated basic properties. Blackwell and MacQueen (1973) showed that the marginal distribution of the DP is equal to the distribution of the n^{th} step of a Pólya urn process. It means that for the DP, if a new observation is obtained, it either has the same value of a previously drawn observations, or it has a new value drawn from a distribution G_0, the base measure. The frequency of new components from G_0 is controlled by α, the precision parameter. In particular, they proved that for $\psi_1, \dots, \psi_n$ iid from $G \sim \mathcal{DP}$, the joint distribution of $\boldsymbol{\psi}$ is a product of successive conditional distributions of the form:

$$\psi_i|\psi_1, \dots, \psi_{i-1}, \quad m \sim \frac{m}{i-1+m}\,\phi_0(\psi_i) + \frac{1}{i-1+m}\sum_{l=1}^{i-1}\delta(\psi_l = \psi_i), \tag{1.1}$$

where δ denotes the Dirac delta function.

The DP, a nonparametric prior and the models with DP priors are treated as hierarchical models in a Bayesian framework. Realizations of the DP are discrete (with probability one), even given support over the full real line, and are treated like countably infinite mixtures. The implementation of the DP mixture models has been made feasible by the modern method of Bayesian computation and efficient algorithms. Escobar and West (1995) provided a Gibbs sampling algorithm for the estimation of posterior distribution for all model parameters and the direct evaluation of predictive distributions. They also discussed inferences about the precision parameter using a gamma prior. MacEachern and Müller (1998) presented a framework of Gibbs sampling with non-conjugate priors using auxiliary parameters; in addition, Neal (2000) provided an extended and more efficient Gibbs sampler to handle general DP mixture models with non-conjugate priors using a set of auxiliary parameters.

We consider mixtures of linear mixed models with an unknown number of components, where the response distribution is a normal mixture with cluster-specific random effects so that observations from the same cluster are correlated. We develop a Bayesian clustering procedure based on a DP prior. The proposed approach (unlike its counterparts) provides simultaneous partitioning and parameter estimation with the computation of the classification probabilities. We note that Bayesian DP linear mixture models with cluster specific random effects consider the hidden structure in the simultaneous detection of multimodality. Cluster-specific random effects are assumed to be independent between clusters and can be seen to control all unobserved group characteristics that are shared by group members. We compare the proposed model to mixtures of linear mixed models with subject-specific random effects that account for individual heterogeneity.

In this paper, we extend a method for clustering based on mixtures of linear mixed models with cluster-specific random effects using the DP prior. We also offer to provide accurate estimation of the number of clusters. Section 2 describes the proposed DP mixture model of linear mixed regression with cluster-specific random effects. Section 3 derives a Gibbs sampler for the model parameters and the clusters of the DP. Section 4 illustrates empirical results by simulation and the comparison of the Bayesian DP mixture of cluster-specific random effects model to the Bayesian DP mixture of subject-specific random effects. Section 5 provides an application of our methodology to real data. Finally, Section 6 provides the concluding remarks.

2. Dirichlet Process Mixture Model of Linear Mixed Regression

A mixture of linear mixed model with cluster-specific random effects for i^{th} response Y_i can be written

$$\begin{aligned} Y_i|i \in C_k &= f\left(\boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_k\right) + \epsilon_{ik}, \\ \epsilon_{ik}|i \in C_k &\sim F\left(0, \sigma_k^2\right), \quad i = 1, \ldots, n, \\ \eta_k &\sim G, \end{aligned}$$

where C_k is a set of index for cluster k, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ is vector of independent variables for i^{th} subject, $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{ik}, \ldots, \boldsymbol{\beta}_{pk})'$ is the regression parameter for cluster k, η_k is the cluster-specific random effects of cluster k, and σ_k^2 the variance of errors in cluster k. Here, f, F and G are taken to be normal.

2.1. Dirichlet process mixture model

The DP mixture model is known as a Bayesian nonparametric mixture model. It can be expressed in the form of

$$f(\,\cdot\,|G, \theta) = \int k(\,\cdot\,|z, \theta)\,\mathrm{d}G(z), \quad G \sim \mathrm{DP}\left(\alpha, G_0\right), \tag{2.1}$$

where z is the latent cluster index, θ is the parameter of the density function $k(\,\cdot\,|z, \theta)$, α is a positive concentration parameter, and G_0 is a specified probability measure on $(\mathcal{X}, \mathcal{B})$. The hierarchical formulation with latent mixing parameter z_i associated with response y_i is

$$\begin{aligned} y_i|z_i, \theta &\sim k\left(y_i; z_i, \theta\right), \quad i = 1, \ldots, n, \\ z_i|G &\sim \text{i.i.d. } G, \\ G|\alpha, \phi &\sim \mathrm{DP}\left(\alpha, G_0\left(\,\cdot\,|\phi\right)\right), \\ \theta, \alpha, \phi &\sim \pi(\theta)\pi(\alpha)\pi(\phi), \end{aligned}$$

where $\pi(\theta)$ is a prior on a parameter with hyperpriors on α and hyper-parameters ϕ, $\pi(\alpha)$ and $\pi(\phi)$.

In the DP mixture model, α controls the prior distribution of the number of distinct latent cluster indices. The expected number of prior clusters, κ, can be expressed as

$$\kappa = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}. \tag{2.2}$$

When we integrate over the DP, as done algorithmically according to Blackwell and MacQueen (1973), the right-hand-side of (2.2) is the expected number of clusters, given the prior distribution on m. Neal (2000, p.252) shows this as the probability in the limit of a unique table seating, conditional on the previous table seatings, which makes intuitive sense since this expectation depends on individuals sitting at unique tables to start a new (sub)cluster in the algorithm.

Rather than estimating α, a better strategy is to include α directly in the Gibbs sampler, as the maximum likelihood estimate from the likelihood function of α can be very unstable (Kyung *et al.*, 2010). A prior distribution results in a unique value of the posterior mode that needs to be considered. One of the candidates was a gamma distribution with the shape parameter a and scale parameter b. We choose the gamma candidate by using an approximate mean and variance of the prior distribution to set the parameters of the candidate. To get the approximate mean and variance, we will use the Laplace approximation of Tierney and Kadane (1986). We use the approximation as the first and second moments of the candidate gamma distribution. Details are in Kyung *et al.* (2010).

Using the constructive definition of the DP, the probability density (2.1) can be represented as a countable mixture of parametric densities,

$$f(\cdot|G,\theta) = \sum_{l=1}^{\infty} w_l k(\cdot|z_l,\theta),$$

where w_l is the weight on cluster l. The given formulation provides a link between the limits of finite mixtures, with priors for the weights given by a Dirichlet distribution, and DP mixture models. Thus, the K finite mixture model will have the form of

$$\sum_{k=1}^{K} w_k k(y|z_k), \tag{2.3}$$

with $(w_1,\ldots,w_K) \sim \text{Dirichlet}(\alpha/K,\ldots,\alpha/K)$ and $z_k \sim$ i.i.d. G_0 for $k = 1,\ldots,K$.

2.2. Dirichlet process mixture of linear mixed regression

Consider a response variable Y_i in cluster k. With design vector $\boldsymbol{x}_i$ and cluster-specific random effect η_k, it can be expressed as

$$\begin{aligned} Y_i|z_i = k, \mathbf{w}, \boldsymbol{\theta} &\sim N\left(\boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_k, \sigma_k^2\right), \\ \eta_k|z_i = k, \tau^2 &\sim N\left(0, \tau^2\right), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$ is a vector of model parameters with each $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2)$. Here, we let $\mathbf{z} = (z_1,\ldots,z_n)$ be a vector of a latent allocation variable with probability $P(z_i = k) = w_k$ with $\mathbf{w} = (w_1,\ldots,w_K)$ and $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_K)$ be a length K vector of cluster-specific random effects.

We consider conjugate priors on the coefficients and on the variance of the response $\mathbf{Y}$. The hierarchical formulation of DP mixture of linear mixed regression with cluster-specific random effect is

$$\begin{aligned} Y_i|z_i = k, \mathbf{w}, \boldsymbol{\theta} &\sim N\left(\boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_k, \sigma_k^2\right) \\ \eta_k|z_i = k &\sim N\left(0, \tau^2\right) \\ z_i|G &\sim \text{i.i.d. } G \\ G|\alpha, \mathbf{b}, d, a, b &\sim \text{DP}\left(\alpha, G_0\left(\boldsymbol{\beta}_k, \sigma_k^2|\mathbf{b}, d, a, b\right)\right) \\ G_0\left(\boldsymbol{\beta}_k, \sigma_k^2|\mathbf{b}, d, a, b\right) &= \text{MVN}_p\left(\boldsymbol{\beta}_k|\mathbf{b}, d\sigma_k^2\boldsymbol{I}\right)\text{IG}\left(\sigma_k^2|a, b\right) \end{aligned}$$

$$\begin{aligned} \tau^2|a_1, b_1 &\sim \mathrm{IG}\left(a_1, b_1\right) \\ \alpha|a_2, b_2 &\sim \mathrm{Gamma}\left(a_2, b_2\right), \end{aligned} \tag{2.5}$$

where MVN_p is a p-dimensional multivariate normal distribution and IG is a inverse gamma distribution. For the parameters of the prior distribution of τ^2, a_1, b_1 can be fixed as a numerical value which make the prior distribution as in the flat form, because the prior parameters are not sensitive to get the proper posterior distribution. The parameters of the prior distribution of α has been discussed in Section 2.1.

We are more concerned about making inferences in applications in regards to partition-specific parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ where $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2)$, cluster-specific random effects $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ and α as well as the partition labels $\mathbf{z}$. Our approach (unlike its counterparts) provides simultaneous partitioning and parameter estimation with classification probabilities.

Heavy computation is required and a search algorithm is needed to determine the optimal clusters without knowing the number of clusters. Richardson and Green (1997) developed a new methodology for a fully Bayesian mixture analysis making use of reversible jump MCMC methods. Booth *et al.* (2008) proposed a stochastic search algorithm to cluster multivariate data using an objective function.

3. Sampling Scheme of Clustering with Bayesian Mixture Model

We describe a general Gibbs sampling scheme that iteratively generates the partition-specific parameters $\boldsymbol{\theta}$, cluster-specific random effects $\boldsymbol{\eta}$ and α as well as the partition labels $\mathbf{z}$. We describe our sampling scheme by simulating from posterior full conditional distributions, which arise by combining the likelihoods with the corresponding prior full conditionals.

Given the data, the full posterior of the DP mixture model is

$$\pi\left(G, \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{z}, \tau^2, \alpha|\mathbf{y}\right) = \pi\left(G|\mathbf{z}, \alpha\right)\pi\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \tau^2, \mathbf{z}, \alpha|\mathbf{y}\right).$$

Here, $\pi(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{z}, \alpha|\mathbf{y})$ is the marginal posterior of the full parameter vector $(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{z}, \alpha)$ and $G|\mathbf{z}, \alpha \sim \mathrm{DP}(\alpha^*, G_0^*)$, where $\alpha^* = \alpha + n$,

$$G_0^*\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2\right) = \frac{\alpha}{\alpha + n} G_0\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2|\mathbf{b}, d, a, b\right) + \frac{1}{\alpha + n}\sum_{i=1}^{n} \delta_{\boldsymbol{\beta}_i, \sigma_i^2}\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2\right),$$

and $\delta(\cdot)$ is a Dirac measure. We update the number of clusters and the cluster index from the probabilities such that for the i^{th} observation, the probability that the i^{th} observation is in the pre-existing cluster is the number of observations in that specific cluster over $\alpha + n$, and the probability that the i^{th} observation will form a new cluster is $\alpha/(\alpha + n)$.

The posterior update steps are given as follow. We iterate between these three steps until convergence:

1. For each $i = 1, \ldots, n$, given α, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ and the partition labels except for the i^{th} observation, generate $\boldsymbol{\beta}_i, z_i$ from

$$\begin{aligned} &\boldsymbol{\beta}_i, \sigma_i^2, z_i|\boldsymbol{\beta}_{-i}, \boldsymbol{\sigma}_{-i}^2, \boldsymbol{\eta}, \alpha, \boldsymbol{X}, \mathbf{y} \\ &\sim \frac{\alpha q_0}{\alpha q_0 + \sum_{k=1}^{K^*} n_k^* p_k} h\left(\boldsymbol{\beta}_i, \sigma_i^2|\boldsymbol{\eta}, \mathbf{b}, d, a, b, y_i, \boldsymbol{x}_i\right) + \sum_{k=1}^{K^*} \frac{n_k^* q_k}{\alpha q_0 + \sum_{k=1}^{K^*} n_k^* q_k} \delta_{\boldsymbol{\beta}_k{}^*, \sigma_k^{*2}}\left(\boldsymbol{\beta}_i, \sigma_i^2\right), \end{aligned}$$

where K^* is the number of clusters in $\{z_j : j \neq i\}$, n_k^* is the number of elements in cluster k with $\{z_j : j \neq i\}$,

$$
\begin{aligned}
q_0 &= \int f\left(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}\right) g_0\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2|\mathbf{b}, d, a, b\right) \mathrm{d}\boldsymbol{\beta}\mathrm{d}\sigma^2 \\
&\quad \text{where } g_0\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2|\mathbf{b}, d, a, b\right) = \mathrm{MVN}_p\left(\boldsymbol{\beta}|\mathbf{b}, d\sigma^2\boldsymbol{I}\right)\ \mathrm{IG}\left(\sigma^2|a, b\right) \\
&= \left|\boldsymbol{I} + d\boldsymbol{x}'\boldsymbol{x}\right|^{-\frac{1}{2}} \frac{\Gamma\left(a + \frac{1}{2}\right)}{\Gamma(a)\Gamma\left(\frac{1}{2}\right)} \frac{b^a\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\left[b + \frac{1}{2}\frac{(y-\eta-\boldsymbol{x}\mathbf{b})^2}{\left(1+d\sum_{j=1}^p x_{ij}^2\right)}\right]^{a+\frac{1}{2}}}, \\
q_k &= (2\pi\sigma_k)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_k^2}\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k - \eta_k\right)^2\right\}
\end{aligned}
$$

and

$$
h\left(\boldsymbol{\beta}_i, \sigma_i^2|\boldsymbol{\eta}, \mathbf{b}, d, a, b, y_i, \boldsymbol{x}_i\right) = \mathrm{MVN}_p\left(\tilde{\boldsymbol{\beta}},\ \sigma_i^2\tilde{\boldsymbol{\Sigma}}\right)\ \mathrm{IG}\left(a^*, b^*\right),
$$

where

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \left(\boldsymbol{x}_i'\boldsymbol{x}_i + \frac{1}{d}\boldsymbol{I}\right)^{-1}\left\{\boldsymbol{x}_i'\left(y_i - \eta_i\right) + \frac{1}{d}\mathbf{b}\right\}, \qquad \tilde{\boldsymbol{\Sigma}} = \left(\boldsymbol{x}_i'\boldsymbol{x}_i + \frac{1}{d}\boldsymbol{I}\right)^{-1}, \\
a^* &= a + \frac{1}{2}, \quad \text{and} \quad b^* = b + \frac{1}{2}\frac{\left(y_i - \eta_i - \boldsymbol{x}_i\mathbf{b}\right)^2}{\left(1 + d\sum_{j=1}^p x_{ij}^2\right)}.
\end{aligned}
$$

2. Given $\boldsymbol{\theta}$, K, $\mathbf{z}$ and τ^2, generate the cluster-specific random effects η_k for $k = 1, \ldots, K$ from

$$
\eta_k|\boldsymbol{\theta}, \tau^2, \alpha, \boldsymbol{X}, \mathbf{y} \sim N\left(\frac{n_k\tau^2}{n_k\tau^2 + \sigma_k^2}\sum_{z_i=k}\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta}_k\right),\ \left(\frac{n_k}{\sigma_k^2} + \frac{1}{\tau^2}\right)^{-1}\right),
$$

where K is the number of components and n_k is the number of elements in cluster k, which have been updated in Step 1.

3. Given $\boldsymbol{\theta}$, K, $\mathbf{z}$ and $\boldsymbol{\eta}$, generate τ^2 from

$$
\tau^2|\boldsymbol{\theta}, \boldsymbol{\eta}, \alpha, \boldsymbol{X}, \mathbf{y} \sim \mathrm{IG}\left(a_1 + \frac{K}{2},\ b_1 + \frac{1}{2}\sum_{k=1}^K \eta_k^2\right),
$$

where η_k's are cluster-specific random effects which are updated in Step 2.

4. Given $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, τ^2, K and $\mathbf{z}$, generate the concentration parameter α of the DP from

$$
\begin{aligned}
\pi\left(\alpha|\boldsymbol{\theta}, K, \mathbf{z}, \boldsymbol{X}, \mathbf{y}\right) &\propto \pi(\alpha)\alpha^K\frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \\
&\propto \pi(\alpha)\alpha^{K-1}\left(\alpha + n\right)\int_0^1 \xi^\alpha(1 - \xi)^{n-1}\mathrm{d}\xi,
\end{aligned}
$$

where $\pi(\alpha) = \text{Gamma}\,(\alpha|a_2, b_2)$. This implies that $\pi(\alpha|K)$ is the marginal distribution from a joint distribution for α and a continuous auxiliary variable ξ such that

$$\pi\,(\alpha, \xi|K) \propto \pi(\alpha)\alpha^{K-1}\,(\alpha + n)\,\xi^{\alpha}(1 - \xi)^{n-1}$$

for $\alpha > 0$ and $0 < \xi < 1$. Hence, the extended Gibbs sampler is

$$\begin{aligned}&\xi|\alpha, K \sim \text{Beta}\,(\alpha + 1, n)\,,\\&\alpha|\xi, K \sim \pi_\alpha \text{Gamma}\,(a_2 + K, b_2 - \log\,(\xi)) + (1 - \pi_\alpha)\,\text{Gamma}\,(a_2 + K - 1, b_2 - \log\,(\xi))\,,\end{aligned}$$

with weights π_α defined by $\pi_\alpha/\,(1 - \phi_\alpha) = (a_2 + K - 1)/\,\{n\,(b_2 - \log\,(\xi))\}$.

Escobar and West (1995) proved the convergence theorems of the normal mixture DP model. Our model is an extension of the normal mixture DP model with a cosine orthonormal basis system. The proofs would be straightforward extensions of Escobar and West (1995) with bounds of the expected posterior distributions as constraints with respect to α.

4. Curve Estimation with Simulated Data

We first evaluated the performance of our method with simulated data (where the classes are known) since real data sets are generally noisy and their clusters may not be fully reflective of the class information. We consider four different data sets of different allocation probability and different linear coefficients with subject specific or cluster specific random effects. We generate the length n X_1 from normal distribution with mean -3 and variance 0.001, X_2 from normal with mean 2 and variance 0.1. We combine 1, X_1, and X_2, then consider the design matrix $\boldsymbol{X} = (1, X_1, X_2)$ as fixed. We generate $n = 100$ samples of $K = 3$ clusters. We simulated data according to the following regression model:

$$Y_i|z_i = k = \boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_i + \epsilon_{ik}$$

with $i = 1, 2, \ldots, n$ and $k = 1, \ldots, K$ is the cluster index. The ϵ_{ik} values follow the normal distribution with mean 0 and variance σ_k^2.

$$\begin{array}{ll} \text{Case 1:} & \text{subject specific random effects } \ \eta_i \sim N(0, 0.01) \ \ i = 1, \ldots, n \\ & \text{cluster 1: } \ w_1 = 0.4, \ \boldsymbol{\beta}_1 = (0, 0, 4), \qquad \sigma_1^2 = 0.5 \\ & \text{cluster 2: } \ w_2 = 0.4, \ \boldsymbol{\beta}_2 = (-1, 0, -2), \ \ \sigma_2^2 = 0.2 \\ & \text{cluster 3: } \ w_3 = 0.2, \ \boldsymbol{\beta}_2 = (1, 1, 0), \qquad \sigma_3^2 = 0.1 \\ \text{Case 2:} & \text{subject specific random effects } \ \eta_i \sim N\,(0, 0.01) \ \ i = 1, \ldots, n \\ & \text{cluster 1: } \ w_1 = 0.3, \ \boldsymbol{\beta}_1 = (0, 0, 4), \qquad \sigma_1^2 = 1 \\ & \text{cluster 2: } \ w_2 = 0.5, \ \boldsymbol{\beta}_2 = (-1, 0, -2), \ \ \sigma_2^2 = 1 \\ & \text{cluster 3: } \ w_3 = 0.2, \ \boldsymbol{\beta}_2 = (1, 1, 0), \qquad \sigma_3^2 = 1 \\ \text{Case 3:} & \text{cluster specific random effects } \ \eta_k \sim N\,(0, 0.25) \ \ k = 1, \ldots, K \\ & \text{cluster 1: } \ w_1 = 0.3, \ \boldsymbol{\beta}_1 = (0, 0, 2), \qquad \sigma_1^2 = 0.5 \\ & \text{cluster 2: } \ w_2 = 0.3, \ \boldsymbol{\beta}_2 = (-1, 0, -2), \ \ \sigma_2^2 = 0.2 \\ & \text{cluster 3: } \ w_3 = 0.4, \ \boldsymbol{\beta}_2 = (1, 1, 0), \qquad \sigma_3^2 = 0.1 \end{array}$$

Case 4: cluster specific random effects $\eta_k \sim N(0,\ 0.25)\ \ k = 1, \dots, K$

cluster 1: $w_1 = 0.3,\ \boldsymbol{\beta}_1 = (0, 0, 4),\qquad \sigma_1^2 = 0.5$

cluster 2: $w_2 = 0.5,\ \boldsymbol{\beta}_2 = (-1, 0, -2),\ \sigma_2^2 = 0.2$

cluster 3: $w_3 = 0.2,\ \boldsymbol{\beta}_2 = (1, 1, 0),\qquad \sigma_3^2 = 0.1$

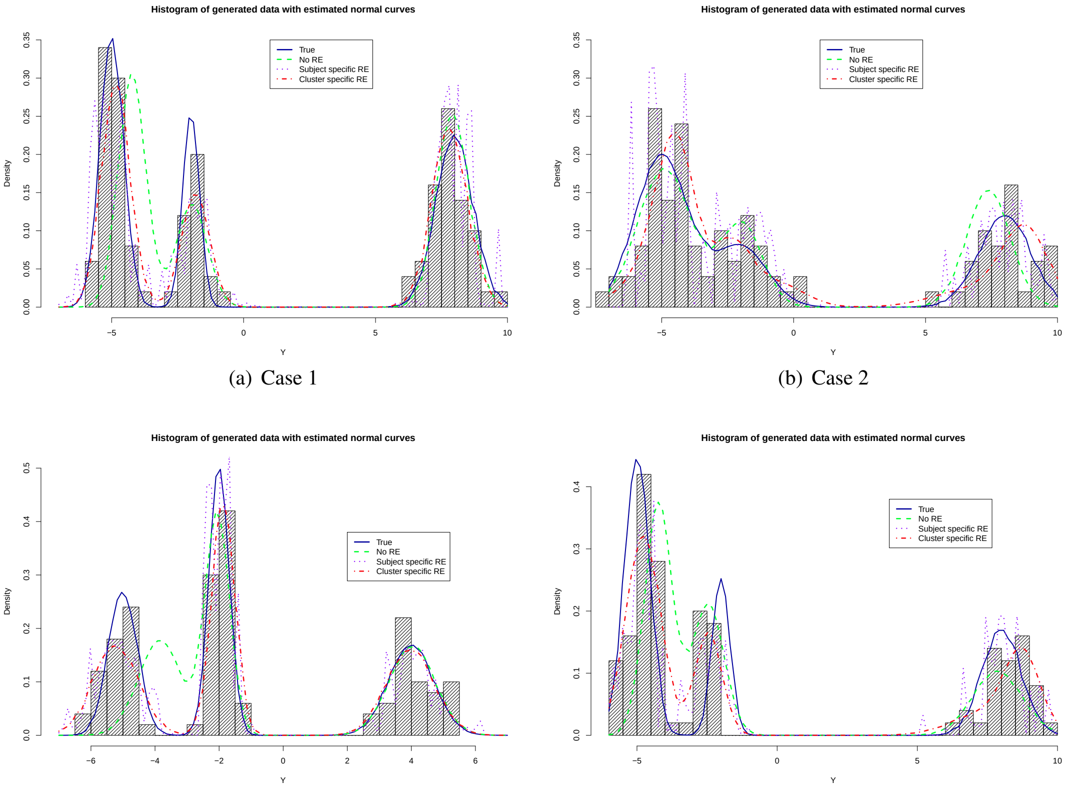
We consider three different models for each set of generated data: DP linear mixture with no random effect, DP linear mixture with subject specific random effects, and DP linear mixture with cluster specific random effects. The Gibbs sampler was iterated 20,000 times to get values of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ as well as to get the cluster index $\mathbf{z}$: consequently, 10,000 draws were then saved as simulations from the posterior. For the cluster index of the Bayesian normal mixture model, we computed the posterior likelihood after repeating three steps of the sampling procedure described in Section 3. The transaction plot of the posterior likelihood has been used as a convergence tool, and it showed that the proposed Gibbs sampler converges very fast (plot was omitted).

To report the outcome of our proposed model, the cluster index $\mathbf{z}$ of the highest posterior likelihood value was chosen as the allocation of the clusters from the Bayesian DP mixture model. Label switching problem needs to be considered adequately if we compute the posterior probability that an observation belongs to the each cluster based on after burn-in 10,000 iterations. However, we use the results from the settings with the highest posterior likelihood in our work and we do not consider here.

Figure 1 shows the estimated curves of three components mixture data based on the cluster classifications of the Bayesian DP linear mixture models with no random effect, subject specific random effects, and cluster specific random effects. From the figures, we observe that all of the curves with the cluster specific random effects are well classified into each true partition. There might exist a higher chance that these two clusters might be detected as one big cluster or might be detected as three small clusters if the two true means are close to each other because samples with subject specific random effects diffuse more compared to samples generated closely around the true mean functions.

For Case 1 and Case 2, data sets are generated based on three quite separated means with subject specific random effects. The true curve of three component normal mixture (bold line) and histogram of the generated data do not perfectly match because of the noise (subject specific random effects) on each observation. We might not be able to estimate the true curve perfectly because the estimation is based on the generated data. The estimated curves of the Bayesian DP linear mixture models are on the first low in Figure 1. For Case 1, normal mixture of linear model, no random effect model, (long dash line) seems to underestimate the mean of the first high peak because of the neighboring distance of two components. However, DP mixtures of linear mixed models with cluster specific random effects (dot dash line) and subject specific random effects (short dash line) adequately estimate the true mean and the true curve. The estimated line of the DP mixture with subject specific random effects wiggles due the subject specific random effects. For Case 2, two components are very close to each other. No random effect model and the cluster specific random effects model seem to estimate each means closed to each other in fully neighboring area. However, DP mixture with subject specific random effects model adequately estimate the true means.

For Case 3 and Case 4, data sets are generated based on three separated means with cluster specific random effects. Because of the cluster specific random effects, the true curve and the histogram. The estimated curves of the Bayesian DP linear mixture models are on the second low in Figure 1. Normal mixture of linear model, no random effect model, seems to underestimate the means and the mixture of subject specific random effects seems to overestimate the means of each components. For the neighboring area of two components, normal mixture of linear model adequately estimate each



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 1: *Estimated curves of the Bayesian Dirichlet process linear mixture models.*

means of components close to each other. The DP mixture of subject specific random effects models estimates the curve wiggly because of the model structure. However, the DP mixture of cluster specific random effects model adequately estimate the true curve. Models do not estimate the true cover well because of the distance between the true curve and the generated data; however, the DP mixture of cluster specific random effects model adequately compromise the true curve and the generated data.

Table 1 lists the estimate cluster probabilities of the Bayesian DP mixture of linear mixed regression. The proposed Bayesian DP mixture models with cluster specific random effects seem to choose more clusters if the neighboring area is hard to detected. This might be the reason for the ambiguous edge structure or the smoothness that results in the highly sensitive detection of the differences from the mean curve for the clustering. The DP mixture model with cluster specific random effects choose dominant mean curves and combine samples to be detected as a variation of the non-trend mean for the widely spread samples.

The proposed Bayesian DP mixture models with cluster specific random effects choose more clusters if the neighboring area cannot be detectable easily. However, in the sense of curve estimation, the DP mixture of linear regression models estimate the true curve and the generated data poorly, and the DP mixture of linear mixed models with subject specific random effects over-estimate the true curve and the generated data. As we discussed above, the DP mixture with cluster specific random effects seem to estimate mean curves adequately based on the generated data.

Table 1: Estimated component probabilities of the Bayesian Dirichlet process mixture of the linear mixed regression

Cluster ID	Case 1			Case 2					
	1	2	3	1	2	3	4	5	6
True	0.4	0.4	0.2	0.3		0.5		0.2	
No random effect (RE)	0.4	0.4	0.2	0.32		0.22	0.25	0.21	
Subject specific RE	0.4	0.4	0.2	0.3		0.46		0.24	
Cluster specific RE	0.4	0.4	0.2	0.21	0.9	0.35	0.12	0.16	0.7
Cluster ID	Case 3			Case 4					
	1	2	3	1	2	3	4		
True	0.3	0.3	0.4	0.3		0.5	0.2		
No random effect (RE)	0.3	0.3	0.4	0.3		0.47	0.23		
Subject specific RE	0.3	0.3	0.4	0.3		0.51	0.19		
Cluster specific RE	0.3	0.3	0.4	0.19	0.11	0.5	0.2		

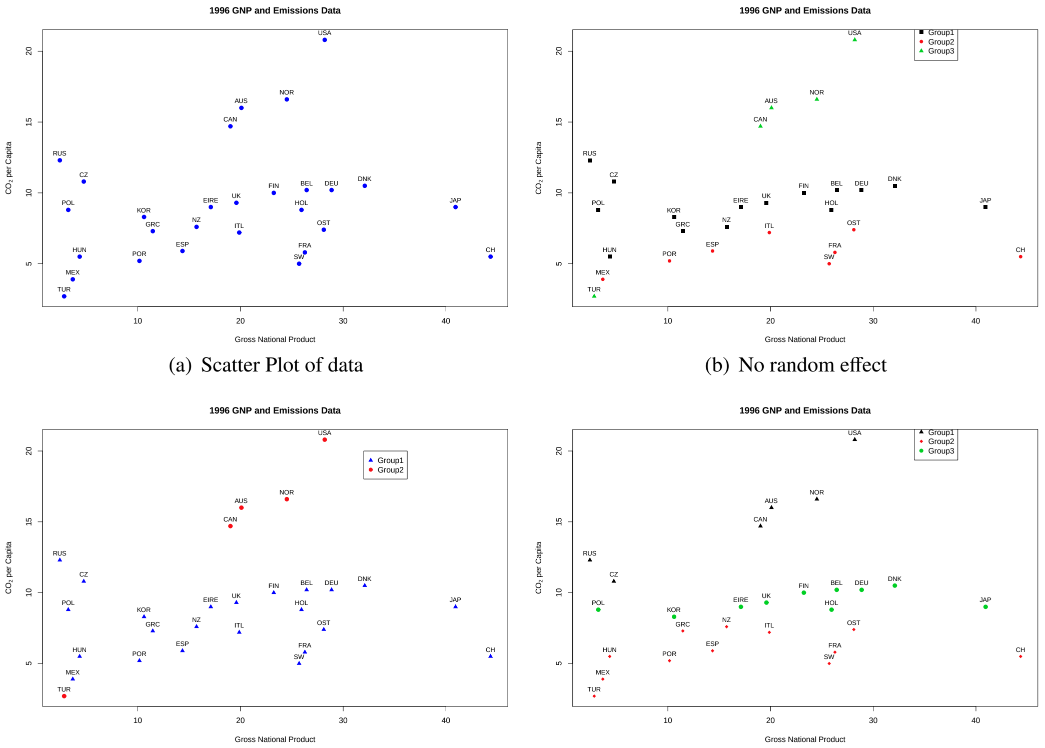
5. Data Analysis

We applied our method of clustering to CO_2 data set from Hurn *et al.* (2003). This data set contains the gross national product (GNP) per capita in 1996 for 28 countries and their estimated carbon dioxide (CO_2) emission per capita for the same year. An abbreviation pertaining to the country measured such as GRC = Greece and CH = Switzerland. Originally, to identify groups and the corresponding linear models was of interest for low GNP countries as it may help clarify which development path they are embarking on.

In Hurn *et al.* (2003), the CO_2 dataset has been analyzed with the conclusion that $k = 2$ was the solution most favored by the data. It is also noted that, in the case $k = 3$, there was a stable third line appearing, but the weight associated with the third regression line was estimated by 0.0069 and was negligible.

We compare out proposed Bayesian DP mixture models for the representation of the allocations of the observations to components: DP linear mixture with no random effect, DP linear mixture with subject specific random effects, and DP linear mixture with cluster specific random effects. After 10,000 burn-in iterations, 10,000 Gibbs samplers of the Bayesian DP mixture models were saved for the posterior inference. For the Bayesian DP mixture output, the clustering structure of the largest value of the posterior log likelihood was chosen among 10,000 Gibbs samplers. The proposed sampler converges fast. The trace plots and other convergence check tools are not included.

The proposed DP normal mixture of linear mixed regression models shows different output from Hurn *et al.* (2003). Figure 2 is a comparison of the representation of the allocations of the Bayesian DP linear mixture models to components. The DP mixture of linear regression (no random effect) model (in the first low of the second column in Figure 2) allocated observations into three groups. With the no random effect normal mixture model, `USA`, `NOR`, `AUS`, `CAN` and `TUR` are chosen as one group. The DP mixture of linear mixed model with subject specific random effects in the second low of the first column in Figure 2, allocated observations into two groups. As the same results of the no random effect mixture model, `USA`, `NOR`, `AUS`, `CAN` and `TUR` are chosen as one group. These are similar to the conclusion of Hurn *et al.* (2003). However, the DP mixture of linear mixed model with cluster specific random effects, in the second low of the second column in Figure 2, allocated observations into three groups. The allocations seems based on the CO_2 level. The estimated slop of the linear regression for each cluster is 0.31, 0.01, and 0.09, respectively. With estimated slops numerically smaller compared to other models.



(a) Scatter Plot of data

(b) No random effect

(c) Subject specific random effects

(d) Cluster specific random effects

Figure 2: *Representation of the allocations of the Bayesian Dirichlet process linear mixture models to components.*

6. Concluding Remarks

A Bayesian normal mixture model with cluster specific random effects based on DP as a prior on clustering structure has been discussed. It is a generalization of the normal mixture models with random effects. The proposed model carries out inferences on a range of plausible values of the number of clusters according to the data with the mean trajectories. Our approach provides simultaneous partitioning and parameter estimation with the computation of the classification probabilities, unlike its counterparts.

We note that the Bayesian DP linear mixture models with cluster specific random effects consider the hidden structure in the simultaneous detection of multimodality. The cluster-specific random effects are assumed to be independent between clusters and can be seen to control of all unobserved group characteristics that are shared by group members. We compare the proposed model to mixtures of linear mixed models with subject-specific random effects that account for individual heterogeneity.

Based on the simulation studies, we observed that the proposed DP mixture model with cluster specific random effects that detect clusters sensitively by combining vague edges into different clusters. It might be due to the proposed model detected the clustering structure fully according to data. In the sense of curve estimation, the DP mixture of linear regression models estimate the true curve and the generated data poorly, and the DP mixture of linear mixed models with subject specific random effects over-estimate the true curve and the generated data. As we discussed above, the DP mixture with cluster specific random effects seem to estimate mean curves adequately based on generated data.

We have demonstrated that our proposed model can provide further insight into our understanding of CO_2 data. Each partition is characterized by a set of parameters in our enhanced model. The proposed DP mixture with cluster specific random effects shows a different conclusion from other models and Hurn *et al.* (2003). This might be the reason for the cluster specific random effects.

The number of clusters seem numerically larger than the numbers of clusters based on other clustering algorithms. However, disregarding hidden variation because of the computational simplicity might not be a good idea even for the interpretation of the data itself. Thus, we expect that the Bayesian DP mixture model with cluster specific random effects might construct clusters based on meaningful structures according to information from the data.

The validation of the partitioning should be based on scientific investigation with plausible interpretation. The statistical data analysis provides numerical support and direction for searches that can improve and enhance research.

Acknowledgement

The authors would like to thank the referees and the associate editor for providing informative feedback which led to an improved version of a previous draft.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes, *Annals of Statistics* ,**1**, 353–355.

Booth, J. G., Casella, G. and Hobert, J. P. (2008). Clustering using objective functions and stochastic search, *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 119–139.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model, *Journal of the American Statistical Association*, **91**, 633–642.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-bases clustering, *Journal of the American Statistical Association*, **93**, 294–302.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–38.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *Annals of Statistics*, **1**, 209–230.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.

Hurn, M., Justel, A. and Robert, C. P. (2003). Estimating mixtures of regressions, *Journal of Computational and Graphical Statistics*, **12**, 55–79.

Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model, *Econometrica*, **46**, 427–434.

Kyung, M., Gill, J. and Casella G. (2010). Estimation in Dirichlet random effects models, *Annals of*

Statistics, **38**, 979–1009.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process model, *Journal of Computational and Graphical Statistics*, **7**, 223–238.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, **9**, 249–265

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association*, **53**, 873–880.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association*, **73**, 730–738.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society: Series B (Methodological)*, **59**, 731–792.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, **81**, 82–86.

Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models, *Journal of the American Statistical Association*, **93**, 249–261.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Received August 12, 2015; Revised October 9, 2015; Accepted October 10, 2015