Latent mixture modeling for clustered data

Shonosuke Sugawasa1 · Genya Kobayashi2 · Yuki Kawakubo2

Received: 14 May 2017 / Accepted: 22 June 2018 / Published online: 26 June 2018 © Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract
This article proposes a mixture modeling approach to estimating cluster-wise conditional distributions in clustered (grouped) data. We adapt the mixture-of-experts model to the latent distributions, and propose a model in which each cluster-wise density is represented as a mixture of latent experts with cluster-wise mixing proportions distributed as Dirichlet distribution. The model parameters are estimated by maximizing the marginal likelihood function using a newly developed Monte Carlo Expectation–Maximization algorithm. We also extend the model such that the distribution of cluster-wise mixing proportions depends on some cluster-level covariates. The finite sample performance of the proposed model is compared with some existing mixture modeling approaches as well as mixed effects models through the simulation studies. The proposed model is also illustrated with the posted land price data in Japan.

Keywords Conditional distribution · Monte Carlo EM algorithm · Hierarchical model · Mixture modeling · Random effect

1 Introduction
We suppose that data is divided into several known clusters. For example, when posted land prices are observed, the dataset might be divided according to their closest station. Throughout this paper, we call this type of data clustered data, which often arises in many scientific fields such as econometrics, epidemiology, and genetics. Although mixed effects models (Demidenko 2004) have been widely used for such data, it fundamentally aims at modeling conditional means in each cluster, which could be inappropriate if the data distribution is skewed or multimodal. As an alternative modeling strategy, the finite mixture model (McLachlan and Peel 2000) has been extensively applied for its flexibility to capture heterogeneity in the data. For modeling independent data, the mixture model with covariates was originally proposed in Jacobs et al. (1991), known as mixture-of-experts. To date, a large body of literature has been concerned with flexible modeling of the conditional density for independent data. For example, see Jordan and Jacobs (1994), Hurn et al. (2003), Geweke and Keane (2007), Villani et al. (2009), Villani et al. (2012) and Nguyen and McLachlan (2016).

However, the existing models for independent data are not suitable for estimating cluster-wise conditional distributions. If we globally apply the mixture models to a whole dataset ignoring the clustering labels (we call global mixture modeling), the estimated conditional distributions are the same over all clusters, which is clearly inappropriate in clustered data analysis. On the other hand, applying the mixture models independently to each cluster in order to capture the cluster heterogeneity (we call local mixture modeling) leads to unstable results since the within-cluster samples sizes are usually not large in practice. Hence, another flexible modeling strategy for clustered data is desired. Up to now, several methods have been proposed for modeling cluster-wise distributions. Rubin and Wu (1997) and Ng and McLachlan (2007) proposed mixtures of mixed effects models. Sun et al. (2007), Ng and McLachlan (2007), Vermunt (2007) developed a mixture of experts with cluster-dependent mixing proportions. Rosen et al. (2000) and Tang and Qu (2016) used the generalized estimating equation approach to estimate the component distributions by incorporating the correlations within clusters.

In this article, we propose a compromised model between the global and local mixture modeling. Note that the local mixture model can be expressed as
\[ f_i(y|x) = \sum_{k=1}^{K} \pi_{ik} h_k(y|x), \]

where \( y \) is the response variable, \( x \) is the vector of covariates, and \( h_k \) is the component distribution for the \( k \)-th component of the \( i \)-th cluster with the mixing proportion \( \pi_{ik} \) satisfying \( \sum_{k=1}^{K} \pi_{ik} = 1 \). Since the within-cluster sample size is usually small in practice, \( h_k(y|x) \) would not be stably estimated. Hence, we restrict \( h_k(y|x) = h_k(z|x) \), that is, the component distributions are the same over all the clusters as in the global modeling. Then the model reduces to

\[ f_i(y|x) = \sum_{k=1}^{K} \pi_{ik} h_k(y|x), \]

which can be interpreted as there exists \( K \) latent distributions and each cluster-wise distribution \( f_i(y|x) \) is expressed by these distributions with cluster-wise mixing proportions \( \pi_{ik} \). Hence, as long as \( K \) is a moderate number, one can estimate \( K \) component distributions with reasonable accuracy. On the other hand, estimating unstructured \( \pi_{ik} \) is not feasible since the number of \( \pi_{ik} \)'s grows as the number of clusters increases. To overcome this difficulty, we assume that the vector of proportions \( \pi_i = (\pi_{i1}, \ldots, \pi_{iK})' \) that characterizes the conditional distribution of the \( i \)-th cluster, is a realization from a multivariate distribution. Therefore, \( \pi_i \) plays a similar role to the mixed effects in the context of the mixed-effects model and it can make observations within the same cluster be correlated. When \( h_k(y|x) \) is a normal distribution, the model is similar to Sun et al. (2007) who used a logistic mixed model for modeling \( \pi_i \), but we here directly model \( \pi_i \) by adopting a Dirichlet distribution, which allows us to develop a tractable estimating method for model parameters.

In this article, the model parameters are estimated based on a likelihood-based approach. The model can be viewed as a three-stage hierarchical model, where the first stage consists of the model for the response variable, the second stage consists of the latent variables which assign the latent distribution, and the third stage consists of the model for the mixing proportions. We develop a Monte Carlo Expectation–Maximization (MCEM) algorithm (Dempster et al. 1977; Wei and Tanner 1990) for parameter estimation of which the E-step consists of a simple Gibbs sampling scheme for imputing the latent variables. Since the number of latent distributions \( K \) is generally unknown, we consider selecting \( K \) based on the Akaike information criteria (AIC) or Bayesian information criteria (BIC), where the maximum log-marginal likelihood can be easily computed from a simple Monte Carlo approximation.

The rest of the paper is organized as follows: Sect. 2 describes the proposed model in detail and develops the MCEM algorithm for maximizing the marginal likelihood. In Sect. 3, the performance of the proposed method is demonstrated along with some existing methods through simulation studies. An application to the real data set is also presented. In Sect. 4, some discussion is provided.

## 2 Latent mixture model

### 2.1 Model setup

Suppose that we have the clustered (grouped) observations \( y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i \), with an associated \( p \)-dimensional vector of covariates \( x_{ij} \). Let \( f_i(y|x) \) be a density or probability mass function of \( y_{ij} \) given \( x_{ij} \), which are the same within clusters but different across clusters. Our aim is to estimate the cluster-wise conditional density \( f_i(y|x) \) from the data set \( \{y_{ij}, x_{ij}\} \). To this end, we consider the following latent mixture model:

\[ f_i(y|\pi_i, x, \phi) = \sum_{k=1}^{K} \pi_{ik} h_k(y|x, \phi_k), \]

where \( \pi_{ik} \) is the weight for the \( k \)-th component in the \( i \)-th cluster, \( h_k(|\cdot, \phi_k) \), \( k = 1, \ldots, K \) are the latent conditional densities characterized by the parameter \( \phi_k \), and \( K \) is the unknown number of latent densities. Moreover, we assume that the mixing proportions \( \pi_i \)'s are independent realizations from the Dirichlet distribution with the density

\[ p(\pi_i|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k) \Gamma(\alpha_{i1}) \cdots \Gamma(\alpha_{iK})}{\prod_{k=1}^{K} \Gamma(\alpha_k) \prod_{k=1}^{K} \pi_{ik}^{\alpha_{ik} - 1}} \]

for \( i = 1, \ldots, m \), where \( \Gamma(\cdot) \) denotes the gamma function and \( \alpha = (\alpha_1, \ldots, \alpha_K)' \) is a vector of unknown parameters.

In this article, we let (1) and (2) together denote the latent mixture model. The unknown model parameters to be estimated are \( \phi_1, \ldots, \phi_K \) in latent distributions and \( \alpha \) in the Dirichlet distribution. Under the setting (1) and (2), taking expectation of \( \pi_{ik} \) with respect to \( \text{Dir}(\alpha) \), we have

\[ f_i(y|x, \alpha, \phi) = \sum_{k=1}^{K} p_k h_k(y|x, \phi_k), \]

\[ p_k = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}, \]

which is referred to the marginal model, and is common over all the clusters.

As often done in estimating mixture models, by introducing the latent component indicator \( z_{ij} \in \{1, \ldots, K\} \), the proposed model (1) and (2) can be expressed in the three-stage hierarchical model:
1st stage: $y_{ij} | x_{ij}, (z_{ij} = k) \sim F_k(x_{ij}, \phi_k)$,
2nd stage: $z_{ij} | \pi_i \sim \text{Cat}(K, \pi_i)$,
3rd stage: $\pi_i \sim \text{Dir}(\alpha)$,

where $F_k$ is the distribution having density $h_k$, and $\text{Cat}(K, \pi_i)$ is the categorical distribution on $\{1, \ldots, K\}$ with the probability vector $\pi_i$. In hierarchy (4), $z_{ij}$ and $\pi_i$ are the latent variables. The latent density $h_k$ is determined by the user and the generalized linear model is an attractive choice. For example, we may use a normal distribution with mean $x_{ij} \beta_k$ and variance $\sigma^2$ for $F_k$ when $y_{ij}$ is a continuous variable, and a Poisson distribution with mean $\exp(x_{ij} \beta_k)$ when $y_{ij}$ is a counting variable.

### 2.2 Monte Carlo EM algorithm for parameter estimation

We here consider estimating unknown model parameters

$$\theta = \{\phi_1, \ldots, \phi_K, \alpha\}$$

based on the data. Under the hierarchical formulation (4), the marginal likelihood function $L(\theta)$ is expressed as

$$L(\theta) = \prod_{i=1}^n L_i(\theta)$$

$$L_i(\theta) = \int \left\{ \prod_{j=1}^{n_i} \sum_{k=1}^K \pi_{ik} h_k(y_{ij} | x_{ij}, \phi_k) \right\} p(\pi_i | \alpha) d\pi_i$$

$$= \sum_{z_i} \left\{ \prod_{j=1}^{n_i} \sum_{k=1}^K \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{1}{z_i} \prod_{j=1}^{n_i} h_k(y_{ij} | x_{ij}, \phi_k)^{w_{ijk}} \right\} p(\pi_i | \alpha) d\pi_i$$

$$= \frac{\prod_{k=1}^K \Gamma(\sum_{j=1}^{n_i} \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \sum_{z_i} \prod_{j=1}^{n_i} h_k(y_{ij} | x_{ij}, \phi_k)^{w_{ijk}}$$

$$\times \prod_{k=1}^K \frac{\Gamma(\sum_{j=1}^{n_i} \alpha_k)}{\Gamma(n_i + \sum_{k=1}^K \alpha_k)}$$

where $w_{ijk} = I(z_{ij} = k)$ and $\sum_{z_i}$ denotes the summation over all the combination of $z_i \in \{1, \ldots, K\}^{n_i}$. Hence, a direct maximization of the marginal likelihood is not feasible since evaluation of the likelihood function $L(\theta)$ requires the summation over $K^{n_i}$ elements for each $i$, which is computationally prohibitive even for small $K$.

Instead, we exploit the hierarchical representation (4) and develop the EM algorithm (Dempster et al. 1977) which indirectly and iteratively maximizes $L(\theta)$. Let $\pi = \{\pi_1, \ldots, \pi_m\}$ and $z = \{z_1, \ldots, z_m\}$. Then, the complete log-likelihood function $\ell^c$ of (4) is given by

$$\ell^c(\theta, z, \pi) = \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{k=1}^K I(z_{ij} = k) \log h_k(y_{ij} | x_{ij}, \phi_k)$$

$$+ \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{k=1}^K I(z_{ij} = k) \log \pi_{ik} + \sum_{i=1}^m \log p(\pi_i | \alpha),$$

where $p(\pi_i | \alpha)$ denotes the density function of $\text{Dir}(\alpha)$. Then, given the value of $\theta$ in the $i$th iteration denoted by $\theta^{(i)}$, the E-step entails the imputation of the latent variables $z$ and $\pi$ by taking expectation

$$Q(\theta | \theta^{(i)}) = \mathbb{E}[\ell^c(\theta, z, \pi) | Y, \theta^{(i)}],$$

where the expectation is taken with respect to the posterior distribution of $(z, \pi)$ given all the response variables $Y$. Since $Q(\theta | \theta^{(i)})$ contains the summation of $z_i$ as in $L(\theta)$, an analytical evaluation of $Q(\theta | \theta^{(i)})$ would be infeasible. Therefore, we consider Monte Carlo approximation of $Q(\theta | \theta^{(i)})$ as

$$Q(\theta | \theta^{(i)}) \approx \frac{1}{L} \sum_{l=1}^L \ell^c(\theta, z^{(i)}, \pi^{(i)}),$$

where $L$ is a sufficiently large number, and $z^{(i)}$ and $\pi^{(i)}$ are the $l$th random sample generated from the posterior distribution of $(z, \pi)$ given $Y$ with $\theta = \theta^{(i)}$. Under the hierarchy (4), the marginal posterior distributions of $z$ and $\pi$ are not in simple forms, but the full conditional distributions of $z | \pi, Y$ and $\pi | z, Y$ are the following familiar distributions:

$$z_{ij} | \pi_i, Y \sim \text{Cat}(K, \tilde{p}_{ij}),$$

$$\pi_i | z, Y \sim \text{Dir}(\tilde{\alpha}_i),$$

where $\tilde{p}_{ij} = (\tilde{p}_{ij1}, \ldots, \tilde{p}_{ijk})'$ and $\tilde{\alpha}_i = (\tilde{\alpha}_{i1}, \ldots, \tilde{\alpha}_{iK})'$ with

$$\tilde{p}_{ijk} = \pi_{ik} h_k(y_{ij} | x_{ij}, \phi^{(i)}_k) \sum_{l=1}^K \pi_{il} h_l(y_{ij} | x_{ij}, \phi^{(i)}_l)$$

$$\tilde{\alpha}_k = \alpha^{(i)}_k + \sum_{j=1}^{n_i} I(z_{ij} = k).$$

Then we can use a Gibbs sampler for generating random samples of the posterior distribution of $(z, \pi)$.

The M-step maximizes $Q(\theta | \theta^{(i)})$ obtained from the E-step, noting that

$$Q(\theta | \theta^{(i)}) = C + \sum_{i=1}^m \sum_{j=1}^{n_i} \sum_{k=1}^K z_{ijk} \log h_k(y_{ij} | x_{ij}, \phi_k)$$

$$+ \sum_{i=1}^m \mathbb{E}[\log p(\pi_i | \alpha) | Y, \theta^{(i)}].$$
where $C$ is a constant independent of $\theta$ and $z_{ikj}^* = E[I(z_{ij} = k)|Y, \theta^{(r)}]$ computed from the E-step. Therefore, the maximization problem of $Q(\theta|\theta^{(r)})$ can be divided into the following:

$$\hat{\phi}_k = \arg\max_{\phi_k} \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ijk}^* \log h_k(y_{ij}|x_{ij}, \phi_k),$$

$$\hat{\alpha} = \arg\max_{\alpha} \left\{ \log \Gamma \left( \sum_{k=1}^{K} \alpha_k \right) - m \sum_{k=1}^{K} \log \Gamma(\alpha_k) \right\} + \sum_{k=1}^{K} \alpha_k \sum_{i=1}^{m} \log(\pi_{ik})^*, \quad \text{for } (6)$$

where $(\log \pi_{ik})^* = E[\log \pi_{ik}|Y, \theta^{(r)}]$. It is noted that the maximization with respect to each $\phi_k$ is identical to maximizing the weighted log-likelihood function of the latent conditional distributions, which can be easily carried out by using, for example, the Newton–Raphson algorithm. Similarly, the maximization with respect to $\alpha$ is similar to performing the maximum likelihood method in the Dirichlet distribution.

The whole procedure of the proposed MCEM algorithm is summarized as follows.

**Algorithm 1 (MCEM algorithm) Iterative:**

1. **Set the initial values** $\theta^{(0)}$ and $t = 0$.
2. **Draw a large number of samples** $\pi$ and $\mathbf{z}$ by Gibbs sampling with the full conditionals (5), and compute $z_{ij}^* = E[I(z_{ij} = k)|Y, \theta^{(r)}]$ and $(\log \pi_{ik})^* = E[\log \pi_{ik}|Y, \theta^{(r)}]$.
3. **Solve the maximization problem** (6) and set $\phi_k^{(t+1)} = \hat{\phi}_k$ and $\alpha^{(t+1)} = \hat{\alpha}$.
4. **If the algorithm has converged,** the algorithm is terminated. Otherwise, set $t = t + 1$ and go back to Step 2.

In the case of the normal linear regression model as the latent model, namely $F_k(x_{ij}, \phi_k) = N(x_{ij}^T \beta_k, \sigma_k^2)$ in (4), the M-step for $\phi_k = (\beta_k, \sigma_k^2)$ in (6) can be obtained analytically:

$$\hat{\beta}_k = \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ijk}^* x_{ij} x_{ij}^T \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ijk}^* x_{ij} y_{ij},$$

$$\hat{\sigma}_k^2 = \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ijk}^* \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} z_{ijk}^* (y_{ij} - x_{ij}^T \hat{\beta}_k)^2.$$  

for $k = 1, \ldots, K$.

Following Shi and Copas (2002), the convergence of the proposed MCEM algorithm is monitored by using the batch mean $\hat{\theta}^{(t)} = H^{-1} \sum_{h=0}^{H-1} \theta^{(t-h)}$, after the $H$th iteration. The algorithm is terminated when the relative difference $||\hat{\theta}^{(t)} - \hat{\theta}^{(t-d)}|| / (||\hat{\theta}^{(t-d)}|| + \delta)$, is smaller than some predetermined (small) $\varepsilon$. Here, $H, d, \varepsilon$ and $\delta$ are specified by the user, and we use $H = 30, d = 5, \varepsilon = \delta = 0.0001$ as default choices. For the E-step, $L = 500$ is used as the default choice and this choice appears to work well in the numerical examples in Sect. 3.

For selecting the number of latent distributions, $K$, we use the Akaike information criteria (AIC) or the Bayesian information criteria (BIC) based on the log-marginal likelihood, without any theoretical justifications. When $\phi_k$ is $p$-dimensional, the number of parameters included in the model (4) is $pK + K$. Then the formulations of AIC and BIC are given by

$$\text{AIC} = -2 \sum_{i=1}^{m} \log f_i^{m}(y_i|x_i, \hat{\theta}) + 2(pK + K),$$

$$\text{BIC} = -2 \sum_{i=1}^{m} \log f_i^{m}(y_i|x_i, \hat{\theta}) + (pK + K) \log N,$$

where $N = \sum_{i=1}^{m} n_i$ is the total number of observations and $f_i^{m}(y_i|x_i, \hat{\theta}) = f_i^{m}(y_i|x_i, \hat{\theta}) \text{ s.t. } h_k(y_{ij}|x_{ij}, \hat{\phi}_k)$ is the maximum marginal likelihood. As noted in Sect. 2.2, since the direct evaluation of the marginal likelihood is computationally prohibitive, the maximum marginal likelihood is evaluated by the Monte Carlo integration. Let $\pi_{ik}^*(\pi_{11}^*, \ldots, \pi_{KK}^*)$ be the random vector generated from $\text{Dir}(\alpha)$. Then, the Monte Carlo approximation of (7) is

$$f_i^{m}(y_i|x_i, \hat{\theta}) \approx \frac{1}{B} \sum_{b=1}^{B} \left\{ \prod_{j=1}^{n_i} \sum_{k=1}^{K} \pi_{ik}^{(b)} h_k(y_{ij}|x_{ij}, \hat{\phi}_k) \right\},$$

for a large $B$, where $(\pi_{11}^{(b)}, \ldots, \pi_{KK}^{(b)})'$ is the $b$th draw from $\text{Dir}(\alpha)$. As demonstrated in Sect. 3.1, BIC works better than AIC, so that we recommend to adopt BIC rather than AIC for selecting the number of components.

Let $K^*$ be the selected number of latent distributions based on AIC or BIC. Then, the cluster-wise estimated conditional density is given by

$$\hat{f}_i(y|x) = \sum_{k=1}^{K^*} \pi_{ik} h_k(y|x, \hat{\phi}_k),$$

where $\pi_{ik} = E[\pi_{ik}|Y]$ evaluated at $\theta = \hat{\theta}$, which can be computed via the Gibbs sampler with the full conditionals.
(5) with $\theta = \hat{\theta}$. Under (1) and (2), response variables in different clusters are mutually independent, so that it holds $E[\pi_i | Y] = E[\pi_i | Y_i]$ with $Y_i = \{y_{i1}, \ldots, y_{in_i}\}$, which means that the cluster-wise mixing proportions $\pi_{ik}$ are determined by the observation in the cluster. Generally speaking, $\pi_{ik}$ tends close to the estimates of marginal means $p_k$ in (3) if the cluster-specific sample size $n_i$ is small, so that the estimated conditional density would be close to the estimated marginal model (3). On the other hand, in clusters with relatively large $n_i$, the estimated conditional density might vary from the marginal model (3), depending on the information of $Y_i$.

2.3 Flexible modeling of mixing proportions

One possible criticism for the formulation of the proposed latent mixture model (1) is its simplicity in the relationship between the response variable $y$ and the covariate vector $x$. The flexibility of the model (1) can be extended by incorporating the covariate vector $w$ in the mixing proportions. Then, we here consider the following structure of the mixing proportions:

$$
\pi_i \sim \text{Dir}(\alpha_i), \quad \alpha_i = (\alpha_{i1}, \ldots, \alpha_{iK})', \\
\alpha_{ik} = \exp(w_i'y_k), \quad k = 1, \ldots, K,
$$

(8)

where $w_i$ is the $q$-dimensional vector of the cluster-specific covariates and $y_k$ is the corresponding coefficient. One can take, for example, $w_i = \tilde{x}_i^{(s)}$ where $\tilde{x}_i^{(s)} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$ and $x_{ij}$ is the subvector of $x_{ij}$. Under this setting, it holds that

$$
E[\pi_{ik}] = \frac{\exp(w_i'y_k)}{\sum_{k=1}^{K} \exp(w_i'y_k)}.
$$

the MCEM algorithm developed in Sect. 2.2 can be easily modified to estimate the model with (8). Specifically, in the E-step $\tilde{a}_{ik}$ appeared in the full conditional distribution of $\pi_i | w, Y$ in (5) is replaced with

$$
\tilde{a}_{ik} = \exp(w_i'y_k) + \sum_{j=1}^{n_i} I(z_{ij} = k),
$$

and the M-step for $\alpha$ in (6) is replaced with the maximizing

$$
Q(\alpha) = \sum_{i=1}^{m} \sum_{k=1}^{K} \log \Gamma(\alpha_k) + \sum_{i=1}^{m} \sum_{k=1}^{K} \log \Gamma(\exp(w_i'y_k)) \exp(w_i'y_k) (\log \pi_{ik})^*,
$$

with $\gamma > 0$ and $\Gamma(\gamma) = \int_0^{\infty} t^\gamma \exp(-t) dt$, and $\pi_{ik} = \pi_{ik}(x)$ is a differentiable function with respect to $\gamma$. Finally, it is noted that the number of parameters under (8) is $K(p + q)$, so that the penalty terms in AIC and BIC used for selecting $K$ should be changed accordingly.

3 Numerical studies

3.1 Simulation studies (continuous response)

The finite sample performance of the proposed latent mixture model was investigated together with some existing methods. We considered $m = 50$ clusters and two cases of within-cluster sample sizes $n_1 = 30$ and $n_2 = 50$ for $i = 1, \ldots, m$. In the following simulations, the covariate $x_{ij}$ is generated from the standard normal distribution $N(0, I)$.

We first consider the following true conditional density in the $i$th cluster:

$$
f_i(y|x) = \pi_i \phi(y; -1 + x, 1) + (1 - \pi_i) \phi(y; 1 - x, 1),
$$

(I)

where $\phi(\cdot; a, b)$ denotes the density function of the normal distribution $N(a, b)$. The true density corresponds to the proposed latent mixture regression (LMR) model in which normal linear regression models are used as latent models. In this case, the true number of latent components is 2.

We first investigated the difference in performance between AIC and BIC for selecting the number of components in LMR models. To this end, we fitted the LMR model with normal linear regression models to the simulated data with $K \in \{2, 3, 4\}$ and recorded the selected number of components by AIC and BIC over 100 data replications. The MCEM algorithm was implemented using the setting described in Sect. 2.2. For the initial values, we used dispersed values for the constant term in the latent mixture regressions, $\beta_{k1}$, namely, $(\beta_{11}, \beta_{21}) = (-3, 3)$, $(\beta_{11}, \beta_{21}, \beta_{31}) = (-4, 0, 4)$ and $(\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}) = (-4.5, -1.5, 1.5, 4.5)$, and set $\sigma_k = 1, \alpha_k = 1$ for $k = 1, \ldots, K$. We have tried other choices for $\beta_{k1}$, but they also worked reasonably well.

Table 1 presents the frequency of the selected component numbers, which shows that BIC selected the true number of components all the time whereas AIC selected larger number of components in some replications. Moreover, we computed the cluster-wise mean integrated squared error (MISE) defined as

$$
\text{MISE}_i(x) = \frac{1}{R} \sum_{r=1}^{R} \int \left[ \hat{f}_i^{(r)}(t|x) - f_i^{(r)}(t|x) \right]^2 dt,
$$

(9)
where \( R = 100 \) and \( \hat{\pi}_{i}(r|x) \) is the estimated conditional density obtained from the \( r \)th replication. Since \( \text{MISE}_i(x) \) depends on \( x \), we considered averaged MISE over equally spaced 41 points of \( x \) from -2 and 2. The boxplots of the averaged MISE are shown in Fig. 1, which reveals that the use of BIC leads to slightly smaller MISEs than AIC. Therefore, we adopt BIC for selecting the number of components \( K \) for the rest of the paper.

We next considered the estimation accuracy of the mixing proportion \( \pi_i \) through the LMR model. For comparison, we also fitted the local mixture (LM) model where the mixture of normal linear regressions was fitted to each cluster separately and global mixture (GM) model where the single mixture of normal linear regressions was fitted to the whole data ignoring the cluster heterogeneity. Note again that GM ignores the clustering structure and produces the same conditional densities over all the clusters. To focus only on the estimation of \( \pi_i \), we fix the number of components of the three models to 2, the true number of components. Based on \( R = 100 \) data replications, we computed the following mean squared error (MSE) of \( \pi_i \) for each model:

\[
\text{MSE}_i = \frac{1}{R} \sum_{r=1}^{R} (\hat{\pi}_{i}(r) - \pi_i)^2,
\]

where \( \hat{\pi}_{i}(r) \) is an estimate of \( \pi_i \) in the \( r \)th replication. Figure 2 presents the boxplots of log-MSEs. From Fig. 2, we can observe that LMR can estimate the mixing proportion \( \pi_i \) with higher accuracy than LM and GM since GM cannot account for the heterogeneity of \( \pi_i \) and LM is unstable due to the moderate within-cluster sample sizes.

Furthermore, we evaluated the performance of the LMR model together with some existing models in terms of MISE (9) with \( R = 100 \). In addition to LM and GM, we also fitted a random intercept (RI) model as a competitor from mixed effects models. For applying LMR, LM and GM, the number of components was selected based on BIC. The MISE at \( x = -1.5, 0, 1.5 \) as well as the averaged MISE over the 41 points of \( x \) are presented in Fig. 3, which reveals the better performance of LMR than other models.

As the second scenario of the true cluster-wise conditional density, we considered the following structure:

\[
f_i(y|x) = (1 - \pi_{i1} - \pi_{i2})\phi(y; -x, 1.5^2) \tag{II}
+ \pi_{i1}\phi(y; -1 + 2x, 0.5^2) + \pi_{i2}\phi(y; 1.5 + x, 1),
\]

where \( \pi_{i1} = I(1 \leq i \leq 15), \ \pi_{i2} = I(16 \leq i \leq 30). \)

Note that the above cluster-wise densities cannot be expressed by the LMR model, so that this scenario enables us to evaluate the performance of LMR under a misspecified setting. We again fitted the four models, LMR, LM, GM and RI, to the simulated data, and computed MISE (9) over 100 data replications. In applying LMR, LM, GM, we used BIC to select the number of components from \( K \in \{2, 3, 4\} \). Figure 4 presents the simulated values of MISE at \( x = -1.5, 0, 1.5 \), and averaged MISE over the 41 points of \( x \). We can observe that the proposed LMR again outperforms the other models. Moreover, we computed BIC values of LMR, GM and RI, and found that BIC of the LMR model was the smallest among the models in every simulation run.

We finally investigated the efficacy of the modeling the distribution of the mixing proportion in terms of some covariates as introduced in Sect. 2.3. To this end, we considered the following true conditional density:

\[
f_i(y|x) = \pi_i\phi(y; -1 + x, 1) + (1 - \pi_i)\phi(y; 1 - x, 1), \tag{III}
\]

\( \pi_i \sim \text{Beta}(\alpha_{i1}, \alpha_{i1}), \)

\( \alpha_{i1} = \exp(1 + 0.6w_i), \quad \alpha_{i2} = \exp(1 - 0.5w_i), \)

\( w_i \sim \text{Ber}(0.4). \)
We considered four cases of \( m, m = 50, 100, 150 \) and \( 200 \). In each case, clusters were divided into 50 groups and each group \( g = 1, \ldots, 50 \) includes the same number of clusters. We set \( n_i = g \) if the \( i \)th cluster is included in the group \( g \). The latent mixture regression model with covariate-dependent structure of mixing proportions (LMR-CD) and the latent mixture regression model (LMR) are fitted to the simulated data. For the MCEM algorithm, the initial values for \((\beta_k, \sigma_k)\) are set to those described above and we set \((\gamma_{k1}, \ldots, \gamma_{kq})^f = (0, \ldots, 0)^f\).
We first evaluated the estimation accuracy of $\alpha_{i1}$ and $\alpha_{i2}$ as well as cluster-wise mixing proportion $\pi_i$. To this end, MSEs (10) of $\hat{\pi}_i$ were calculated based on $R = 100$ data replications and those of $\hat{\alpha}_{i1}$ and $\hat{\alpha}_{i2}$ were calculated in the same way. Figure 5 presents the boxplots of log MSE ratios of LMR-CD to LMR, in which value smaller than 0 means LMR-CD provides better estimates than LMR. The figure shows that the estimation accuracy of $\alpha_{i1}$ and $\alpha_{i2}$ improves as $m$ becomes larger. More importantly, the estimation accuracy of $\pi_i$ is directly related to the accuracy of conditional density estimation, and LMR-CD can provide better estimates of $\pi_i$ than LMR regardless of $m$. We next evaluated the MISE (9) of LMR-CD and LMR, for which the number of components was selected by using BIC in both models. In Fig. 6, we show log MISE ratios of LMR-CD to LMR, in which the value smaller than 0 means LMR-CD provides better conditional density estimates than LMR. Figure 6 reveals that LMR-CD performs better than LMR in terms of conditional density estimation and these results would be consistent with the results of estimation accuracy of $\pi_i$ in Fig. 5. It is also observed that LMR-CD outperforms LMR especially in the clusters with the smaller within-cluster sample sizes $n_i$ where the estimated densities are close to the marginal model.

### 3.2 Simulation studies (counting response)

In this section, we investigated the performance of the proposed LMR model in the case where the response values are counting values. To this end, we considered the following cluster-wise probability functions:

\[
f_i(y|x) = \pi_i\text{Po}(y; 0.5 + 0.5x) + (1 - \pi_i)\text{Po}(y; 2 - x)
\]

\[
\pi_i \sim \text{Beta}(\alpha_{i1}, \alpha_{i2}),
\]

\[
\alpha_{i1} = \exp(1 + 0.6w_i), \quad \alpha_{i2} = \exp(1 - 0.5w_i),
\]

\[
w_i \sim \text{Ber}(0.4),
\]

where $\text{Po}(y; \lambda)$ denotes a Poisson probability function with mean $\exp(\lambda)$. We considered two cases of $m$, $m = 50$ and 100, and the within-cluster sample sizes were determined as in Scenario (III). As in the previous section, the covariates $x_{ij}$’s were generated from $N(0, 1)$. We applied both LMR-CD and LMR models in which Poisson regression models were used as latent regression models and the number of components were selected by BIC from $K \in \{2, 3, 4\}$. Moreover, we also fitted the GM model of Poisson regression models and selected the number of components by using BIC. Based on $R = 100$ simulation runs, we computed MISE (9) for each model and the results are reported in Fig. 7. It is revealed that both LMR-CD and LMR outperform GM and as in Scenario (III) LMR-CD performs better than LMR especially in the clusters with the smaller within-cluster sample sizes.

---

Figure 5: Ratios of log-MSEs of estimators of $\alpha_{i1}, \alpha_{i2}$ and $\pi_i$ in LMR-CD to those in LMR under Scenario (III) in the four cases of $m$.

Figure 6: Ratios of log-MISEs of conditional density estimates from LMR-CD to those from LMR evaluated at $x = -1.5, 0, 1.5$ and averaged log-MISE values over 41 points of $x$ in Scenario (III) with four cases of $m$, $m = 50, 100, 150$ and 200.
m (PLP) data in Tokyo and the surrounding four prefectures (Chiba, Saitama, Kanagawa and Ibaraki) in 2001. The data units (locations) are clustered with respect to the nearest station. The number of clusters is \( m = 295 \) and the total number of units is \( N = 2363 \). The number of within-cluster samples \( n_i \) are ranging from 1 to 45, and the histogram of \( n_i \) is provided in the left panel in Fig. 8. We note that there are 221 clusters with \( n_i \) smaller than 10 and 25 clusters with \( n_i = 1 \). The response variable \( y_{ij} \) is the PLP which is measured in 100,000 yen per squared meter. In each \( j \)th unit (location) in \( i \)th cluster (station), \( y_{ij} \) is observed with the floor area ratio (\% ) \( F_{ij} \) and amount of time \( A_{ij} \) (second) to station \( i \) on foot. Moreover, as cluster level information, the amount of time \( T_i \) from Tokyo station by train and the prefecture to which the station belongs are available. We use four dummy variables \( D_{1i}, D_{2i}, D_{3i}, \) and \( D_{4i} \) for Chiba, Saitama, Kanagawa, and Ibaraki, respectively, which take value one if the station \( i \) belongs to the corresponding prefecture and zero otherwise. The values of \( y_{ij} \) range from 0.158 to 20.3. The right panel of Fig. 8 shows that the histogram of \( y_{ij} \) for \( y_{ij} < 8 \). Note that the number of samples with \( y_{ij} \geq 8 \) is only 20 which is \(<1\%\) of the total number of observations. The cluster-wise conditional density of PLP would be valuable since it enables us to see the distributional relationships between PLP and available covariates. Moreover, based on the cluster-wise conditional density, we can compute cluster-wise characteristics, for example, mean, median, quantile and more complex quantity like expectile. Since the distribution of PLP seems skewed as observed from Fig. 8, a mixture of normal linear regression would be suitable rather than a single normal linear regression for modeling PLP.

Let \( x_{ij} = (1, F_{ij}, A_{ij}, T_i, D_{1i}, \ldots, D_{4i})^T \). We consider the following latent mixture regression model with covariate dependent structure in mixing proportions (LMR-CD):

\[
 f_i(y_{ij} | x_{ij}) = \sum_{k=1}^{K} \pi_{ik} \phi(y_{ij}; x_{ij}^T \beta_k, \sigma_k^2),
\]

\[
(\pi_{1}, \ldots, \pi_{K})^T \sim \text{Dir}(\alpha_1, \ldots, \alpha_K),
\]

\[
\alpha_k = \exp(\gamma_{i1} + \gamma_{2k} T_i^* + \gamma_{3k} F_{ij}^* + \gamma_{4k} A_{ij}^* + \gamma_{5k} D_{ij}),
\]

\[
k = 1, \ldots, K.
\]

### 3.3 Example: posted land price data in Japan

To demonstrate the proposed method in a practical situation, we apply the latent mixture model to the posted land price (PLP) data in Tokyo and the surrounding four prefectures.
with \( j = 1, \ldots, n_i, \ i = 1, \ldots, m \), where \( \beta_k = (\beta_{k1}, \ldots, \beta_{k8})' \) and \( T_i^*, F_i^*, A_i^* \) are the standardized version of \( T_i, F_i, A_i \), respectively. The MCEM algorithm was implemented using the setting described in Sect. 2.2 with the initial values for \( \beta_k \) are set to the estimates obtained from the global mixture (GM) model and we set \( \alpha_{ik} = 1 \) and \( y_\gamma = 0 \) for \( k = 1, \ldots, K \). To determine the number of components in the latent mixture and assess the adequacy of the model for \( \alpha_{ik} \) in (11), we calculated BIC for in total of \( 8 \times 2^4 = 128 \) models, considering all the combination of the four variables and \( T_i^*, F_i^*, A_i^* \) and \( D_i \) and the number of components \( K \in \{1, 2, \ldots, 8\} \). The best model minimizing BIC was the model including only \( T_i^* \) with \( K = 6 \).

It is noted that the marginal model (3) is given by

\[
f_i(y_{ij}) = \sum_{k=1}^{K} p_{ik} \phi(y_{ij}; x'_{ij} \beta_k, \sigma_k^2),
\]

and the cluster-wise estimated density is

\[
f_i(y) = \sum_{k=1}^{K} E[\pi_{ik}|Y_i] \phi(y; x' \beta_k, \sigma_k^2),
\]

where \( Y_i = \{y_1, \ldots, y_{ni}\} \) and \( E[\pi_{ik}|Y_i] \) can be computed from the Gibbs sampling (5). Moreover, based on BIC, the number of latent components was selected to be \( K = 6 \) from \( \{1, \ldots, 8\} \). We also doubled the number of Gibbs draws in the E-step, but the same result was obtained.

For comparison with (11), we also applied the global mixture (GM) model with \( K_a \) components:

\[
f(y) = \sum_{k=1}^{K_a} p_k \phi(y; x' \beta_k, \sigma_k^2),
\]

where \( \sum_{k=1}^{K_a} p_k = 1 \). It is expected that the estimated GM is similar to the marginal model in LMR. Based on BIC \( K_a = 5 \) was selected. We also fitted the LMR model obtained by putting \( y_\gamma = y_\gamma = y_\gamma = 0 \) and \( y_5h = 0 \) in (11), denoted by LMR, and \( K = 7 \) was selected by BIC. The resulting BIC values of LMR-CD, LMR and GM are as follows:
so that the LMR-C model was the best model among the three models. To see the behavior of the proposed Monte Carlo EM algorithm, we give the updating processes of $\beta_k^4$ and $\sigma_k^4$ for $k = 1, \ldots, K (= 6)$ in Fig. 9. We can observe that these parameter values reasonably converge to their final estimates.

To visualize the estimated conditional density in each cluster, we fixed the covariate vector $x$ at $(1, 100, 600, T_i, D_{i1}, \ldots, D_{i4})^T$, in which $f_i(y|x)$ corresponds to the density function of the PLP of each cluster when the floor area ratio is 100 and the location is 10 minutes’ walk from the nearest station. Figure 10 presents the estimated density under LMR-C, the marginal model of LMR-C (mLMR-C) in (12), and GM for the selected stations with small $n_i$. The figure shows that the cluster-wise estimated densities under LMR-C are close to those under mLMR-C when $n_i$ is small. This is because the small $n_i$ values leads to a small difference between the prior mean $\pi_{ik}$ and posterior mean $E[\pi_{ik}|Y_i]$ of $\pi_{ik}$, so that the estimated densities in such clusters are automatically close to those under mLMR-C which can be stably estimated from the data. Figure 11 presents the estimated densities for the selected stations with relatively large $n_i$. Contrary to Fig. 10, the estimated densities under LMR-C deviate from mLMR-C in some clusters. The result implies that mLMR-C is adjusted by the observed data in these clusters. We finally point out that mLMR-C and GM are similar in most cases since their modeling strategies are similar in the sense that they aim at estimating the global density by ignoring the clustering structure.

Finally, Figs. 12 and 13 present the posterior means $E[\pi_{ik}|Y_i]$ and prior means $p_{ik}$ of the mixing proportions for the stations arranged by $n_i$, respectively. The figures indicate that the distributions of PLP for the stations with larger $n_i$ are represented with fewer latent-distributions, while $E[\pi_{ik}|Y_i]$ with smaller $n_i$ appear to be more uniform and are close to $p_{ik}$, supporting the aforementioned conjecture. While it is not the main purpose of the proposed modeling approach, by rearranging the result, for example, by railway routes, the clustering of the stations can also be visualized as presented in Fig. 14. The figure shows that the stations on the same route tend to result in similar distributions of PLP.
4 Conclusion and discussion

We have proposed the latent mixture model for estimating the cluster-wise conditional distributions. The model parameters are estimated by using the simple Monte Carlo EM algorithm instead of the brute force maximization of the marginal likelihood. Through the simulation and empirical studies, the proposed method is found to be useful for flexible modeling of clustered data.

In this article, we selected the number of components by using AIC and BIC. However, the use of AIC or BIC is not necessarily justified in this case since the mixture model might be a singular model and observations within the same clusters are not independent. The detailed investigation of selecting the number of latent components with theoretical validity would be a valuable future work. Moreover, one possible extension of the proposed model is to use mixed effects models as latent distributions. When the response values are continuous, one may use linear mixed models as latent distributions, the estimation algorithm would be computationally intensive. The detailed investigation is left to a future study.

Acknowledgements This work was supported by JSPS KAKENHI Grant Nos. [16H07406, 15K17036, 16K17101, 18K12754, 18K12757]. The computational results were obtained using Ox version 6.21 (Doornik 2007).

References

Demidenko, E.: Mixed Models: Theory and Applications. Wiley, New York (2004)
Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B 39, 1–38 (1977)
Doornik, J.: Ox: Object Oriented Matrix Programming. Timberlake Consultants Press, London (2007)
Geweke, J., Keane, M.: Smoothly mixing regressions. J. Econom. 138, 252–290 (2007)
Hurn, M., Justel, A., Robert, C.P.: Estimating mixtures of regressions. J. Comput. Graph. Stat. 12, 55–79 (2003)
Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Comput. 3, 79–87 (1991)
Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Comput. 214, 181–214 (1994)
McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
Ng, S.K., McLachlan, G.J.: Extension of mixture-of-experts networks for binary classification of hierarchical data. Artif. Intell. Med. 41, 57–67 (2007)
Ng, S.K., McLachlan, G.J.: Mixture models for clustering multilevel growth trajectories. Comput. Stat. Data Anal. 71, 43–51 (2014)
Nguyen, H.D., McLachlan, G.J.: Laplace mixture of linear experts. Comput. Stat. Data Anal. 93, 177–191 (2016)
Rosen, O., Jiang, W., Tanner, M.A.: Mixtures of marginal models. Biometrika 87, 391–404 (2000)
Rubin, D.B., Wu, Y.: Modeling schizophrenic behavior using general mixture components. Biometrics 53, 243–261 (1997)
Shi, J.Q., Copas, J.: Publication bias and meta-analysis for 2×2 tables: an average Markov chain Monte Carlo EM algorithm. J. R. Stat. Soc. B 64, 221–236 (2002)
Sun, Z., Rosen, O., Sampson, A.R.: Multivariate Bernoulli mixture models with application to postmortem tissue studies in schizophrenia. Biometrics 63, 901–909 (2007)
Tang, X., Qu, A.: Mixture modeling for longitudinal data. J. Comput. Graph. Stat. 25, 1117–1137 (2016)
Vermunt, J.K.: A hierarchical mixture model for clustering three-way data sets. Comput. Stat. Data Anal. 51, 5368–5376 (2007)
Villani, M., Kohn, R., Giordani, P.: Regression density estimation using smooth adaptive Gaussian mixtures. J. Econom. 153, 155–173 (2009)
Villani, M., Kohn, R., Nott, D.J.: Generalized smooth finite mixtures. J. Econom. 171, 121–133 (2012)
Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithm. J. Am. Stat. Assoc. 85, 699–704 (1990)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.