Clustered data are ubiquitous in a variety of scientific fields. In this article, we propose a flexible and interpretable modeling approach, called grouped heterogeneous mixture modeling, for clustered data, which models cluster-wise conditional distributions by mixtures of latent conditional distributions common to all the clusters. In the model, we assume that clusters are divided into a finite number of groups and mixing proportions are the same within the same group. We provide a simple generalized EM algorithm for computing the maximum likelihood estimator, and an information criterion to select the numbers of groups and latent distributions. We also propose structured grouping strategies by introducing penalties on grouping parameters in the likelihood function. Under the settings where both the number of clusters and cluster sizes tend to infinity, we present asymptotic properties of the maximum likelihood estimator and the information criterion. We demonstrate the proposed method through simulation studies and an application to crime risk modeling in Tokyo.

1. Introduction

Data having clustered structures such as postal area, school, individual, and species is called clustered data, and appears in a variety of scientific fields. The main goal of statistical analysis of clustered data would be modeling the response variable as a function of covariates or auxiliary variables while accounting for heterogeneity among clusters. In this context, mixed models or function of clustered data would be model based on generalized estimating equation, but the primary interest in these works is estimation of the component distributions in the mixture by taking account of correlations within clusters rather than between-cluster heterogeneity. Moreover, as nonparametric Bayesian approaches, Teh et al. (2006) and Rodriguez, Dunson, and Gelfand (2008) might have similar philosophy to the mixture modeling (1) in that they assume sharing mixing components among clusters, while they require Markov chain Monte Carlo algorithm to

\[ f_i(y|x) = \sum_{k=1}^{L} \pi_i \theta_k(y|x), \quad i = 1, \ldots, m, \quad (1) \]

where \( m \) is the number of clusters, \( \theta_k(y|x) \) is a latent conditional distribution with parameter \( \phi_k \). In the model (1), the cluster-wise distributions share the common latent distributions \( \theta_k \), and between-cluster heterogeneity is captured only through the cluster-wise mixing proportions \( \pi_{i1}, \ldots, \pi_{iL} \) which play a similar role to random effects. This approach was considered in Sun, Rosen, and Sampson (2007) and Sugasawa, Kobayashi, and Kawakubo (2019) which employ logistic mixed models and Dirichlet distributions for modeling the mixing proportions, respectively. However, both methods still require complicated numerical algorithms to estimate model parameters. Apart from parametric modeling of cluster-wise conditional distributions, Rosen, Jiang, and Tanner (2000) and Tang and Qu (2016) proposed a mixture modeling based on the generalized estimating equation, but the primary interest in these works is estimation of the component distributions in the mixture by taking account of correlations within clusters rather than between-cluster heterogeneity. Moreover, as nonparametric Bayesian approaches, Teh et al. (2006) and Rodriguez, Dunson, and Gelfand (2008) might have similar philosophy to the mixture modeling (1) in that they assume sharing mixing components among clusters, while they require Markov chain Monte Carlo algorithm to

CONTACT Shonosuke Sugasawa sugasawa@csis.u-tokyo.ac.jp Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 2778568 Japan.
fit the models. Therefore, existing approaches for modeling cluster-wise conditional distributions are computationally intensive and may have poor interpretability, so that effective approaches to estimating cluster-wise conditional distributions while accounting for between-cluster heterogeneity are still limited.

In this article, we seek an effective mixture modeling that is flexible, interpretable, and computationally tractable. We work on the model formulation (1) and propose a grouping strategy for cluster-wise mixing proportions, that is, we assume that all the clusters are divided into a finite number of groups and mixing proportions within the same groups share the same mixing proportions, which we call grouped heterogeneous mixture (GHM) modeling. We introduce unknown group membership parameters, and estimate them together with unknown mixing proportions and parameters in latent distributions via the maximum likelihood method. Owing to such a simple structure of the mixing proportion, the maximum likelihood estimator can be easily obtained via the generalized EM algorithm (Meng and Rubin 1993), in which all the steps are simple and do not require any computationally intensive methods such as numerical integrations. We propose selecting the numbers of groups and latent distributions using an information criterion. Under the framework that both the number of clusters and within-cluster sample sizes go to infinity, we show that the grouping assignment and unknown model parameters can be consistently estimated. We also show that the numbers of groups and latent distributions can be consistently estimated via the information criterion under the same asymptotic framework.

The estimation of group membership parameters is closely related to k-means algorithm (Forgy 1965), and the idea of grouping parameters has been recently considered in the context of panel data analysis (e.g., Hahn and Moon 2010; Bonhomme and Manresa 2015; Ando and Bai 2016), but this article is the first one to adopt the idea for modeling cluster-wise mixing proportions. Moreover, in this article, we also propose a new approach to estimating grouping parameters, referred to structured grouping, which incorporates auxiliary information of clusters into grouping assignment to obtain more interpretable results. We realize this strategy via simple penalization on grouping parameters, motivated from stochastic processes on discrete spaces, and demonstrate this strategy with spatial information in our application given in Section 4.3.

The rest of this article is organized as follows. In Section 2, we introduce the proposed model, describe the generalized EM algorithm, the information criterion for selecting numbers of groups and latent distributions, and then discuss structured grouping via penalization. In Section 3, we illustrate the asymptotic properties of the maximum likelihood estimator and the information criterion. In Section 4, we carry out simulation studies to demonstrate the proposed method compared with some existing methods, and apply the proposed method to crime risk modeling in 23 wards of Tokyo by adopting mixture models with zero-inflated Poisson regression models. Finally, some discussions are given in Section 5. All the proofs are given in Appendix A, and R code implementing the proposed method is available at Github repository (https://github.com/sshonosuke/GHM).

## 2. Grouped Heterogeneous Mixture Modeling

### 2.1. The Proposed Model

Let $y_{ij}$ denote the $j$th measurement in the $i$th cluster for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where $m$ is the number of clusters and $n_i$ is the number of within-cluster samples or cluster size which could be unequal over clusters. Let $x_{ij}$ be a $p$-dimensional vector of covariates associated with $y_{ij}$, where we assume $p$ is fixed. We are interested in the conditional density or probability mass function of $y_{ij}$ given $x_{ij}$, denoted by $f_i(y_{ij}|x_{ij})$, which could be heterogeneous across clusters. For modeling $f_i$, we propose the following GHM model:

$$f_i(y_{ij}|x_{ij}) = \sum_{k=1}^{L} \pi_{gik} h_k(y_{ij}|x_{ij}, \phi_k), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i, \quad (2)$$

where $h_k$’s are latent conditional distributions common to all the clusters, $g_i \in \{1, \ldots, G\}$ denotes an unknown group membership parameter and $\phi_k$ is a vector of unknown parameters in the $k$th latent distributions. As mentioned in Section 1, the important feature of the GHM model (2) is the structure that $m$ clusters are divided into $G$ groups, and clusters within the same groups share the same mixing proportions, which results in the same conditional distributions in the same groups. Therefore, the GHM model would be easy to interpret while it is still flexible to account for the between-cluster heterogeneity. The latent conditional distribution $h_k$ could be specified depending on the type of response variable $y_{ij}$. For instance, we may use a normal distribution with mean $x^T \beta_k$ and variance $\sigma_k^2$ when the response variable is continuous, and a Poisson distribution with rate parameter $\exp(x^T \beta_k)$ when the response variable is count valued. Note that all the $h_k$ does not have to be the same family. In fact, we adopt combinations of a zero-inflation component and Poisson regression models as latent distributions in the application to crime data given in Section 4.3.

### 2.2. Generalized EM Algorithm

We estimate the grouping parameter $g_i$ and unknown parameters $\phi_k$ and $\pi_{gik}$ simultaneously via the maximum likelihood method. Define $\Theta = (\gamma, \pi, \phi)$ be the set of unknown parameters, where $\gamma, \pi, \phi$ are collections of $g_i$, $\pi_{gik}$, and $\phi_k$, respectively. We here assume that the numbers of groups $G$ and latent distributions $L$ are known, but the estimation will be discussed later. Given the data, the log-likelihood function of $\Theta$ is given by

$$Q(\Theta) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \left( \sum_{k=1}^{L} \pi_{gik} h_k(y_{ij}|x_{ij}, \phi_k) \right).$$

The maximum likelihood estimator $\hat{\Theta}$ of $\Theta$ is defined as the maximizer of $Q(\Theta)$. We develop an iterative method for maximizing $Q(\Theta)$ using the generalized EM algorithm (Meng and Rubin 1993). To this end, we introduce latent variables $z_{ij} \in \{1, \ldots, L\}$ and consider a hierarchical expression of the model (2) given by $y_{ij}|(z_{ij} = k) \sim h_k(y_{ij}|x_{ij}, \phi_k)$ and $P(z_{ij} = k) = \pi_{gik}$. Given $z_{ij}$’s, the complete log-likelihood is given by

$$L'(\Theta) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{L} 1(z_{ij} = k) \left\{ \log h_k(y_{ij}|x_{ij}, \phi_k) + \log \pi_{gik} \right\}. $$
In the E-step, we compute the conditional probabilities being \( z_{ij} = k \) for \( k = 1, \ldots, L \), given the data and current parameter values \( \Theta^{(r)} \), which are denoted by \( w_{ijk}^{(r)} \). The detailed expression of \( w_{ijk}^{(r)} \) will be given in Algorithm 1. Then, the objective function in the M-step, which is the expectation of the complete log-likelihood, is given by

\[
Q(\Theta | \Theta^{(r)}) = \sum_{k=1}^{L} \sum_{j=1}^{m} \sum_{i=1}^{n_k} w_{ijk}^{(r)} \log h_k(y_{ij}|x_{ij}; \phi_k) + \sum_{i=1}^{m} \sum_{k=1}^{L} \log \pi_{g,k} \sum_{j=1}^{n_i} w_{ijk}^{(r)}
\]

\[
= Q_1(\phi | \Theta^{(r)}) + Q_2(\gamma, \pi | \Theta^{(r)}).
\]

It is observed that the maximization of \( Q_1(\phi | \Theta^{(r)}) \) with respect to \( \phi \) can be divided into \( L \) maximization problems and \( \phi_1, \ldots, \phi_L \) can be separately updated. On the other hand, maximizing \( Q_2(\gamma, \pi | \Theta^{(r)}) \) includes discrete optimization of \( \gamma \) on the space \([1, \ldots, G]^m \), and brute-force search over the space would be infeasible. Instead of simultaneous maximization, we first maximize \( Q_2(\gamma, \pi | \Theta^{(r)}) \) with respect to \( \pi \) and we set \( \pi^{(r+1)} \) to the maximizer, and then, maximize \( Q_2(\gamma, \pi^{(r+1)} | \Theta^{(r)}) \) with respect to \( \gamma \) to get \( \gamma^{(r+1)} \), which allows separate updating for each element of \( \gamma \). This updating process guarantees that monotone increasing of the objective function, that is, \( Q_2(\gamma^{(r)}, \pi^{(r)} | \Theta^{(r)}) \leq Q_2(\gamma^{(r+1)}, \pi^{(r+1)} | \Theta^{(r)}) \). The proposed GEM algorithm is summarized in what follows.

**Algorithm 1 (GEM algorithm).** Starting from the initial values \( \Theta^{(0)} \) and \( r = 0 \), repeat the following procedure until the algorithm converges:

- **(E-step)** Compute the following weights:
  \[
  w_{ijk}^{(r)} = \frac{\pi_{g,k}^{(r)} h_k(y_{ij}|x_{ij}; \phi_k^{(r)})}{\sum_{k=1}^{L} \pi_{g,k}^{(r)} h_k(y_{ij}|x_{ij}; \phi_k^{(r)})},
  \]
  for \( j = 1, \ldots, n_i \) and \( i = 1, \ldots, m \).

- **(M-step)** Update unknown parameters as follows:
  \[
  \phi_k^{(r+1)} = \text{argmax} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ijk}^{(r)} \log h_k(y_{ij}|x_{ij}; \phi_k) \right\},
  \]
  \[
  \pi_{g,k}^{(r+1)} = \left( \sum_{\ell \in g} \pi_{g,k}^{(r)} \sum_{j=1}^{L} w_{ijk}^{(r)} \right)^{-1} \sum_{\ell \in g} \pi_{g,k}^{(r)} \sum_{j=1}^{n_i} w_{ijk}^{(r)},
  \]
  \[
  g_i^{(r+1)} = \text{argmax}_{g=1, \ldots, G} \left( \sum_{k=1}^{L} \log \pi_{g,k}^{(r+1)} \sum_{j=1}^{n_i} w_{ijk}^{(r)} \right),
  \]
  for \( i = 1, \ldots, m, g = 1, \ldots, G, \) and \( k = 1, \ldots, L \).

Note that updating \( \phi_k \) requires maximizing the weighted log-likelihood of \( k \)-th latent distribution, which can be readily carried out as long as the latent density \( h_k \) is familiar regression models such as normal linear regression or Poisson regression. Moreover, for updating \( g_i \), we simply compute all the values of the objective function for \( g = 1, \ldots, G \) and select the maximizer. Algorithm 1 can be seen as a combination of the classical EM algorithm for mixture models and a modified \( k \)-means algorithm.

The above algorithm could be sensitive to the choice of initial estimators of \( \Theta \), and we propose the following strategy to obtain reasonable initial estimators. For each cluster, we firstly fit different finite mixture models with \( L \) components to get estimates of cluster-wise mixing proportions \( \pi_{is}^{(0)} = (\pi_{is}^{(0)}_1, \ldots, \pi_{is}^{(0)}_{L-1}) \) as well as parameters \( \phi_{ik}^{(0)} \) in the mixing distributions. Then, we apply the standard \( k \)-means algorithm with \( G \) groups to \( \pi_{is}^{(0)}, \ldots, \pi_{ms}^{(0)} \) and adopt the estimated grouping and centroid as the initial estimators of \( \gamma \) and \( \pi \), respectively. The initial estimators of \( \phi_k \) could be obtained via element-wise means or medians of \( \{\phi_{ik}^{(0)}, \ldots, \phi_{mk}^{(0)}\} \) for \( k = 1, \ldots, L \).

### 2.3. Selection of the Numbers of Groups and Latent Distributions

The proposed mixture model has two tuning parameters, \( G \) (the number of groups) and \( L \) (the number of latent distributions), which would be critically related to performance and interpretability of the proposed method. Using unnecessarily small values of \( G \) or \( L \) would result in failure to capture the between-cluster heterogeneity or loss of flexibility for modeling the conditional distributions, whereas the use of unnecessarily large values of \( G \) or \( L \) may lead to redundant modeling which does not hold meaningful interpretations. Therefore, we suggest selecting adequate values of \( G \) and \( L \) by minimizing the following information criterion:

\[
\text{IC}(G, L) = -2Q(\hat{\Theta}_{GL}) + \log N \left\{ G(L - 1) + m + \sum_{k=1}^{L} \dim(\phi_k) \right\},
\]

where \( N = \sum_{i=1}^{m} n_i \) is the total sample size, \( \Theta_{GL} \) is the set of unknown parameters under \( G \) groups and \( L \) latent distributions, and \( \dim(\phi_k) \) is the dimension of the component-specific parameter \( \phi_k \). Clearly, the above information criterion is motivated from Bayesian information criterion (Schwarz 1978), which is known to provide consistent estimation of the number of components in the standard finite mixture models (e.g., Keribin 2000). In Section 3, we show that the information criterion (4) enables us to obtain consistent estimators of \( L \) as well as \( G \). The minimizer can be easily searched over all the combinations of prespecified candidates of \( G \) and \( L \), namely \( G \in \{G_{\min}, \ldots, G_{\max}\} \) and \( L \in \{L_{\min}, \ldots, L_{\max}\} \). Typically, we adopt \( G_{\min} = 1 \) and \( L_{\min} = 1 \), and set \( G_{\max} \) and \( L_{\max} \) to large and moderate values, respectively. Since the computation of \( \Theta_{GL} \) for each choice of \((G, L)\) can be easily computed via Algorithm 1, the brute-force search would be feasible in practice.

### 2.4. Structured Grouping

In the proposed method, the grouping membership is determined in a model-dependent way. On the other hand, we may also incorporate auxiliary or external cluster-level information into grouping. Let \( s_{ij} \) be some similarity measure between clusters \( i \) and \( i' \), satisfying \( s_{ij} \in [0, 1] \). For example, if the cluster-
level covariate \( z_i \) is available, we may define \( s_{ir} = \exp(-b\|z_i - z_r\|^2) \) with a scaling constant \( b \). In our application in Section 4.3, we consider incorporating adjacent information of areas (clusters) into grouping assignment. Specifically, we adopt the indicator of adjacency between \( i \)th and \( i' \)th areas for \( s_{i'i} \), that is, \( s_{i'i} = 1 \) if two areas are adjacent and \( s_{i'i} = 0 \) otherwise. Under the settings, it would be reasonable to make two clusters having larger similarity \( s_{i'i} \) be more likely to be classified in the same group, which would make the results more interpretable. To incorporate such information in the estimation process, we consider maximizing the following penalized likelihood function:

\[
Q(\Theta) + \xi \sum_{i < i'} s_{i'i} I(g_i = g_{i'}),
\]

(5)

where \( \xi \geq 0 \) is a tuning parameter that controls the strength of the penalty. The penalty term in (5) is motivated from a stochastic process on the space of grouping parameters, that is, \( \{1, \ldots, G\}^m \), called Potts model (Potts 1952), where its probability function is proportional to \( \exp(\xi \sum_{i < i'} s_{i'i} I(g_i = g_{i'})) \). Hence, the estimation strategy has a similar philosophy to the use of spatially correlated random effects in generalized linear mixed models. Since \( Q(\Theta) \) is the log-likelihood function, maximizing the penalized likelihood function (5) can be seen as the maximum a posterior probability estimation under the prior stochastic process. The maximization of (5) can be still easily carried out by a similar algorithm to Algorithm 1, where the details are given as follows:

**Algorithm 2 (Modified GEM algorithm under penalization).** The E-step and M-steps for \( \pi_k^r \) and \( \pi_{ik}^r \) are the same as those in Algorithm 1, but the M-step for \( g_i \) is given by

\[
g_i^{(r+1)} = \arg \max_{g = 1, \ldots, G} \left( \sum_{k=1}^L \log \pi_{ik}^{(r+1)} + n_i^{(r)} \right) + \xi \sum_{i' \neq i} s_{i'i} I(g_i = g_{i'}^{(r)}),
\]

where \( i = 1, \ldots, m \).

We note that the updating process for \( g_i \) can be separately carried out for \( i = 1, \ldots, m \), thereby we can simply calculate the values of the objective function for \( g = 1, \ldots, G \), and choose the maximizer. Regarding the value of \( \xi \), we here recommend simply using small values such as \( \xi = 0.2 \) or \( \xi = 0.5 \), as adopted in Section 4.3. If we adopt a too large value of \( \xi \), the penalty term will be too strong and the information from the likelihood would be ignored in the estimation of \( g_i \), which is quite unreasonable. Although it might be able to select \( \xi \) via some data-dependent manner, we do not consider the details further since it would extend the scope of this article. We also note that all the theoretical results given in Section 3 would be valid for the penalized likelihood (5) as long as the penalty term is asymptotically negligible.

### 3. Asymptotic Properties

We investigate the large-sample properties of the maximum likelihood estimator \( \hat{\Theta} \) and selection performance of the information criterion (4) when both the number of clusters and cluster sizes tend to infinity. Without loss of generality, we suppose that the cluster labels are assigned such that the cluster size increases with the cluster index, that is, \( n_1 \leq n_2 \leq \cdots \leq n_m \). We assume that \( m \rightarrow \infty \) as well as \( n_1 \rightarrow \infty \) as considered in literatures (e.g., Vonesh et al. 2002; Hui, Muller and Welsh 2017).

We also adopt the setting that \( m/n_i^0 \rightarrow 0 \) for some \( v > 0 \), that is, the cluster sizes do not have to grow at the same rate as \( m \), and assume that \( n_m = O(n_i^0) \) for some \( c > 1 \), that is, the largest cluster size \( n_m \) is the polynomial order of the smallest cluster size \( n_1 \). In Theorem 1, we further assume that \( G \) and \( L \) are known. We then obtain the following asymptotic results for the maximum likelihood estimator.

**Theorem 1.** Under regularity conditions given in Appendix A and \( m, n_1 \rightarrow \infty \) and \( m/n_i^0 \rightarrow 0 \) for some \( v > 0 \), it holds that

\[
\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^L (\pi_{ik} - \pi_{ik}^0) = O_p\left( \frac{1}{m} \right),
\]

(6)

\[
\sqrt{N} F_m(\tilde{\psi})^{1/2} (\tilde{\psi} - \psi^0) \rightarrow N(0, I_{\dim(\psi)}),
\]

(7)

where \( F_m(\Theta) \) is the Fisher information matrix defined in Appendix A, and \( \hat{\psi}_0, \pi_{ik}^0 \) and \( \psi^0 \) are the true parameters.

Although the number of cluster-wise mixing proportions grows at the number of clusters \( m \), (6) shows that they are consistently estimated and the convergence rate depends on \( m \). Moreover, (7) shows that the structural parameters \( \Phi \) in latent distributions and group-wise mixing proportions \( \pi \) whose dimensions are fixed are \( \sqrt{N} \)-consistent. For estimating the standard errors of \( \hat{\psi} \), we employ the generalized clustered bootstrap (Field et al. 2010). The method works by replacing the objective function with weighted version given by

\[
Q_n(\Theta) = \sum_{i=1}^m \sum_{j=1}^{n_i} \log \left( \sum_{k=1}^L \pi_{ik} h_k(y_{ij}|x_{ij}; \Phi_k) \right),
\]

where \( u_i^n \) is a cluster-specific random weight. We here use \( m \times \text{Dir}(1, \ldots, 1) \) for the distribution of \( u^n = (u_1^n, \ldots, u_m^n) \), so the expectation of \( u_i^n \) is 1. We generate \( u^n \) for \( B \) times to get \( B \) estimators of \( \hat{\psi} \), where \( B \) is the number of bootstrap samples. Note that the weighted objective function \( Q_n(\Theta) \) can be easily maximized by a similar GEM algorithm obtained by replacing \( w_{jk}^{(r)} \) with \( u_i^n w_{jk}^{(r)} \) in Algorithm 1.

We next consider asymptotic properties of the information criterion (4) to select \( G \) and \( L \). We obtain the following result.

**Theorem 2.** Let \( (\hat{G}, \hat{L}) \) be the minimizer of the information criterion (4), and \( (G^0, L^0) \) be the true values. Under regularity conditions given in Appendix A and \( m \rightarrow \infty \) and \( m/n_i^0 \rightarrow 0 \) for some \( v > 0 \), it holds that \( P(\hat{G}, \hat{L}) = (G^0, L^0) \rightarrow 1 \).

Theorem 2 shows that the information criterion (4) holds selection consistency for both \( G \) and \( L \). This can be seen as an extension of the results regarding selection consistency of Bayesian information criterion in the standard mixture model (Keribin 2000).
4. Numerical Studies

4.1. Simulation Study: Continuous Response

We investigate the finite sample performance of the proposed method together with some existing methods. For simplicity, we set the cluster sizes to be the same in all the clusters, that is, $n_i = n$ for $i = 1, \ldots, m$, where $m$ is the number of clusters. We first generated two covariates $x_{ij}$ and $x_{2ij}$ from the uniform distribution on $(0, 0.8)$ and Bernoulli distribution with success probability 0.5, respectively. We considered five scenarios for the true structures of cluster-wise conditional distributions written in the following form:

$$
\begin{align*}
    f_i(y_{ij}|x_{1ij}, x_{2ij}) &= \pi_i \phi(y_{ij}; \beta_{10} + \beta_{11}x_{1ij} + \beta_{12}x_{2ij}, \sigma_i^2) \\
    &\quad + (1-\pi_i) \phi(y_{ij}; \beta_{20} + \beta_{21}x_{1ij} + \beta_{22}x_{2ij}, \sigma_i^2),
\end{align*}
$$

where $\phi(\cdot; a, b)$ denotes a normal density function with mean $a$ and variance $b$. In the first three scenarios, we set $\beta_{20} = \beta_{22} = -\beta_{10} = -\beta_{12} = 0.5, \beta_{11} = -\beta_{21} = 1$ for all $i$, and set $\sigma_1 = 0.2$ and $\sigma_2 = 0.5$, thereby all the clusters share the same latent distributions and variations of cluster-wise conditional distributions are determined by the cluster-specific mixing proportion $\pi_i$. The generating processes of $\pi_i$ are given by (I) Beta(2,1), (II) TP(0.1,0.9), and (III) TP(0.1,0.9) + U(0-1,0.1), where TP(a,b) denotes a two point distribution on $a$ and $b$ with equal probabilities. We note that scenario (I) produces different mixing proportions over all the clusters, while $\pi_i$ takes only two values in scenario (II), that is, all the clusters can be divided into two groups. Scenario (III) corresponds to the case where the clusters can be approximately clustered into two groups.

In scenario (IV), we generated the regression coefficients from $\beta_{10}, \beta_{12} \sim N(-0.5, (0.3)^2), \beta_{20}, \beta_{22} \sim N(0.5, (0.3)^2), \beta_{11} \sim N(1, (0.3)^2)$, and $\beta_{21} \sim N(-1, (0.3)^2)$ and set $\sigma_1 = 0.2$ and $\sigma_2 = 0.5$. We also generated the mixing proportion $\pi$ from Beta(2,1). In this scenario, both mixing proportions and distributions are different in each cluster. Finally, we considered a more simplified structure as scenario (V), that is, we set $\pi_i = 1$ for all $i$, and adopted $\beta_{10}, \beta_{11}, \beta_{12} \sim N(0, (0.5)^2)$ and $\sigma = 0.3$. Regarding the settings of $m$ (number of clusters) and $n$ (number of within-cluster sample size), we considered five settings, namely, $(m, n) = (20, 80), (20, 160), (40, 80), (40, 160)$, and $(60, 80)$.

For the simulated dataset, we applied the proposed GHM model with normal linear regression models as latent conditional distributions in which $G$ (the numbers of groups) and $L$ (the number of latent distributions) are selected by the information criterion (4) among combinations of $G \in \{1, 2, \ldots, 10\}$ and $L \in \{1, 2, 3, 4\}$. We also applied the proposed method with $L = 2$ and the largest number of groups, $G = 10$, denoted by fGHM. For comparison, we firstly consider two logical options that directly apply the standard finite mixture of normal linear regression models, which we call global mixture (GM) and local mixture (LM). The GM method applies the finite mixture model to all the data by ignoring the nature of clustered data, while the LM method applies the mixture model separately to the clusters. The number of components in both GM and LM were selected from $\{1, \ldots, 4\}$ by Bayesian information criterion. As a competitor from random effects models, we adopted random coefficient (RC) models using R package “lme4” (Bates et al. 2016). Moreover, as a competitor from an advanced mixture methods for clustered data, we applied the latent Dirichlet mixture (LDM) approach with normal linear regression models as latent conditional distributions, developed in Sugawara, Kobayashi, and Kawakubo (2019), which models the cluster-wise mixing proportions via the Dirichlet distribution. It should be noted that the computational cost of LDM is quite intensive since it requires a Monte Carlo EM algorithm including Markov chain Monte Carlo algorithm in its E-step, thereby we set the number of component to 2 in all the cases.

Based on these methods, we computed estimators of cluster-wise conditional distributions $f_i$, and compute the following mean integrated squared errors (MISE):

$$
\text{MISE} = \frac{1}{m} \sum_{i=1}^{m} \sum_{x_{1i} \in \{0,1\}} \int_{-0.2}^{0.2} \int_{-\infty}^{\infty} \left| \frac{f_i(y|x_1, x_2)}{f_i(x_1, x_2)} - f_i(y|x_1, x_2) \right|^2 \ dy dx_1,
$$

where $\Delta_{x_1}$ and $\Delta_x$ are sets of equally-spaced points from $-0.2$ to $0.8$ with $b_{x_1}$ increment and from $-5$ to $5$ with $b_y$ increment, respectively, and we set $b_{x_1} = 0.02$ and $b_y = 0.1$. In Table 1, we reported the MISE values averaged over 300 replications, for five scenarios of data generating mechanism with five combinations of $(m, n)$ and $L$ (within-cluster sample sizes).

| Scenario | $(m, n)$ | GHM | fGHM | GM | LM | RC | LDM |
|----------|----------|-----|------|----|----|----|-----|
| (I)      | (20, 80) | 3.95 | 3.90 | 2.00 | 5.80 | 6.00 | 5.60 |
|          | (20, 160)| 3.99 | 3.91 | 2.67 | 6.17 | 6.01 | 5.70 |
|          | (40, 80) | 3.91 | 3.94 | 2.83 | 7.12 | 6.92 | 6.68 |
|          | (40, 160)| 3.90 | 3.91 | 2.80 | 7.05 | 6.76 | 6.50 |
|          | (60, 80) | 3.94 | 3.97 | 2.95 | 8.02 | 7.00 | 6.80 |
| (II)     | (20, 80) | 4.52 | 4.50 | 2.70 | 23.14 | 4.76 | 4.53 |
|          | (20, 160)| 4.51 | 4.50 | 2.70 | 23.20 | 4.73 | 4.52 |
|          | (40, 80) | 4.50 | 4.50 | 2.70 | 23.50 | 4.73 | 4.52 |
|          | (40, 160)| 4.50 | 4.50 | 2.70 | 24.25 | 4.75 | 4.53 |
|          | (60, 80) | 4.50 | 4.50 | 2.70 | 24.00 | 4.75 | 4.53 |
| (III)    | (20, 80) | 4.57 | 4.57 | 2.70 | 21.44 | 4.76 | 4.53 |
|          | (20, 160)| 4.55 | 4.55 | 2.70 | 21.49 | 4.75 | 4.54 |
|          | (40, 80) | 4.55 | 4.55 | 2.70 | 21.94 | 4.79 | 4.56 |
|          | (40, 160)| 4.54 | 4.54 | 2.70 | 21.92 | 4.76 | 4.56 |
|          | (60, 80) | 4.54 | 4.54 | 2.70 | 21.78 | 4.78 | 4.54 |
| (IV)     | (20, 80) | 4.75 | 5.08 | 2.44 | 22.67 | 5.25 | 5.08 |
|          | (20, 160)| 4.66 | 5.04 | 2.43 | 22.79 | 5.19 | 5.04 |
|          | (40, 80) | 4.81 | 5.11 | 2.41 | 22.55 | 5.24 | 5.10 |
|          | (40, 160)| 4.76 | 5.08 | 2.45 | 22.72 | 5.20 | 5.07 |
|          | (60, 80) | 4.84 | 5.13 | 2.41 | 22.45 | 5.23 | 5.11 |
| (V)      | (20, 80) | 5.85 | 6.09 | 2.01 | 23.71 | 5.88 | 6.05 |
|          | (20, 160)| 5.80 | 6.14 | 2.00 | 23.60 | 5.80 | 6.10 |
|          | (40, 80) | 5.76 | 6.17 | 2.00 | 19.68 | 5.79 | 6.10 |
|          | (40, 160)| 5.80 | 6.20 | 2.00 | 20.24 | 5.85 | 6.13 |
|          | (60, 80) | 5.82 | 6.23 | 2.00 | 21.76 | 5.84 | 6.14 |

NOTE: The values are multiplied by 10.
of \((m, n)\). It is observed that the LM and GM approaches do not provide preferable performance compared with the proposed methods possibly due to the lack of model flexibility in GM and instability of estimation in LM under moderate sample sizes such as \(n = 80\) and 160. Although RC works quite well when the model is correctly specified as in scenario (V), its performance is not necessarily satisfactory under the other scenarios. On the other hand, the LDM approach works reasonably well in all the scenarios by flexibly modeling the cluster-wise mixing proportions via the Dirichlet distribution. However, overall the proposed GHM performs as well as or better than the LDM method. Moreover, it is interesting to see that the performance of GHM is quite comparable with LDM even under scenario (I) in which mixing proportions are completely different over clusters, while GHM imposes clustering structures. In scenario (I), both mixing proportions and distributions are different, thereby both GHM and LDM are misspecified, but GHM provides better estimation performance than LDM. In Table 2, we reported average values of selected numbers of \(G\) and \(L\) in GHM. In scenarios (I)–(III), the average values of selected \(L\) are almost equal to the true value 2, and the average value of \(G\) tends to increase in the order of scenarios (II), (III), and (I), as the variations in mixing proportions increase. We note that the true data generating process is consistent with the proposed GHM in scenario (II), and the average values of the selected \(G\) is almost equal to 2, which would support the theoretical result given in Theorem 2. In these scenarios, the performance of GHM and fGHM are comparable, so that the effect of selecting smaller numbers of \(G\) on estimation performance could be limited. In scenarios (IV) and (V), the average numbers of \(G\) and \(L\) are larger than those in the other scenarios possibly because the data generating processes are much more complicated than scenarios (I)–(III). In these cases, selecting suitable values of \(G\) as well as \(L\) in GHM leads to the improvement of estimation performance compared with fGHM.

### 4.2. Simulation Study: Count Response

We next investigate the performance under count responses by adopting the following structures of cluster-wise conditional distributions:

\[
f(y|y_1y_2; \lambda) = \pi_1y_1e^{(\beta_{10} + \beta_{11}y_1 + \beta_{12}y_2)} + (1 - \pi_1)y_2e^{(\beta_{20} + \beta_{21}y_1 + \beta_{22}y_2)},
\]

where \(\pi_1(\cdot; \lambda)\) denotes the probability mass function of Poisson distribution with rate parameter \(\lambda\). We considered the same five scenarios for the mixing proportion \(\pi\) as in Section 4.1. Regarding the regression coefficients, we set \(\beta_{10} = -\beta_{12} = -0.25, \beta_{11} = -\beta_{21} = 0.5\) for all \(i\), in scenarios (I), (II), and (III), and generated from \(\beta_{10}, \beta_{12} \sim N(-0.25, (0.3)^2)\), \(\beta_{20}, \beta_{22} \sim N(0.25, (0.3)^2)\), \(\beta_{11} \sim N(0.5, (0.3)^2)\), and \(\beta_{21} \sim N(-0.5, (0.3)^2)\) in scenario (IV). In scenario (V), we employed the same structure as in Scenario 4.1. Also we considered the same five combinations of \((m, n)\).

For the simulated dataset, we applied the GHM, FGHM, GM, LM, and LDM methods with Poisson regression models as latent distributions and the other settings equivalent to the previous simulation study in Section 4.1. Moreover, we applied a Poisson mixed model with random intercept (RI), by using R package "lme4." The estimated cluster-wise conditional distributions were evaluated by the MISE (8) with \(\Delta_y = \{0, 1, \ldots, 15\}\). In Table 3, we reported the MISE values averaged over 300 replications. Since the true generating process for \(\pi\) is the same as the assumed model in LDM in scenario (I), it is quite reasonable that the LDM performs the best in this scenario. Although the proposed GHM and aGHM work slightly worse than LDM, they perform much better than the other approaches. This is because GHM pursues computational tractability and interpretable results by grouping clusters at the cost of some loss of flexibility compared with LDM. In this scenario, FGHM using \(G = 10\) performs better than GHM, possibly because the true mixing proportions do not admit grouping structures and the large number of groups could be necessary for modeling such complicated structures. The GHM method performs the best in scenario (II) since the mixing proportion \(\pi\) admits the grouped structure in this case. In the other scenarios, the assumed models for \(\pi\) in both GHM and LDM are misspecified, and the results show that the two methods

| Scenario | (m, n) | GHM | FGHM | GM | LM | RI | LDM |
|----------|-------|-----|------|----|----|----|-----|
| (I)      |       |     |      |    |    |    |     |
| (II)     |       |     |      |    |    |    |     |
| (III)    |       |     |      |    |    |    |     |
| (IV)     |       |     |      |    |    |    |     |
| (V)      |       |     |      |    |    |    |     |

NOTE: The values are multiplied by 100.
are quite comparable, thereby the proposed method would be more appealing in that the proposed method is computationally less intensive and is more interpretable than LDM. We also note that the two logical approaches, GM and LM, do not perform well in all the scenarios due to the same reasons given in Section 4.1. Although the RI method performs quite well in scenario (V) since the true data generating process of Poisson regressions and a zero-inflated component: given the covariates, we employed the following mixture model (9), the cluster-specific mixtures act as ward-specific effects to address the heterogeneity among 23 wards, we define the adjacent indicator Gm, which shows that the selected G may adaptively change depending on the complexity of the true structure of πi. In particular, the average numbers of G and L in scenario (IV) are larger than the other scenarios since the cluster-specific distributions in scenario (IV) hold quite complicated between-cluster heterogeneity.

### 4.3. Example: Crime Risk Modeling in 23 Wards of Tokyo

We here apply the proposed method to crime risk modeling in 23 wards of Tokyo, Japan. We used a dataset of number of police-recorded crime in Tokyo metropolitan area, provided by University of Tsukuba and publicly available online ("GIS database of number of police-recorded crime at O-aza-chome in Tokyo, 2009-2017," available at https://commons.sk.tsukuba.ac.jp/data_en). In this study, we focus on the number of violent crimes in m = 23 wards in Tokyo metropolitan area in 2015, and wards are recognized as clusters. For i = 1, . . . , m, each ward contains ni local towns, and ni ranges from 51 to 275, and the number of total local towns (sample sizes) is 2855. The wards are indexed as i = 1, . . . , m and local towns in the ith ward are indexed as j = 1, . . . , ni. For auxiliary information for each town, area (km²), population densities in noon and night, density of foreign people, percentage (%) of unemployment, single-person household and university graduation and average duration of residence are available in each local town.

Let yij be the number of crimes in the jth local town in the ith ward, and we found that the ratios of zero counts in yij is relatively large. For modeling the conditional distribution of yij given the covariates, we employed the following mixture model of Poisson regressions and a zero-inflated component:

\[ y_{ij} | x_{ij} \sim \pi_{G1} \delta_0(y_{ij}) + \sum_{k=2}^{L} \pi_{Gk} \text{Po}(y_{ij} | a_{ij} \exp(x_{ij}' \beta_k)), \]  

where \( \delta_0(\cdot) \) denotes the one-point distribution on the origin which corresponds to the zero-inflation component, \( a_{ij} \) is the area, \( x_{ij} \) is a vector of 7 covariates described above and an intercept term, and \( \pi_{Gk} \in [1, \ldots, G] \) are group parameters. In the model (9), \( L \) is the number of components including the zero-inflated one, so we set \( L \geq 2 \). Note that \( \log a_{ij} \) can be regarded as an offset term, and \( \exp(x_{ij}' \beta_k) \) can be interpreted as the crime risk per square kilometer. In the model (9), the cluster-specific mixing proportions act as ward-specific effects to address the heterogeneity among the wards. Based on the adjacent information among 23 wards, we define the adjacent indicator \( s_{ij}/s_{i'j'} \) for \( i, i' \in [1, \ldots, m] \), that is, \( s_{ij} = 1 \) if the \( i \)th ward and the \( i' \)th ward are adjacent and \( s_{ij} = 0 \) otherwise, and adopted the penalized likelihood (5) with \( \xi = 0.2 \) to address potential spatial similarity. We also tried \( \xi = 0.3 \) and \( \xi = 0.5 \), but the following results did not change much. Based on the information criterion (4), we estimated \( G \) and \( L \) from candidates \( \{1, 2, \ldots, 8\} \) and \( \{2, 3, 4\} \), respectively, and \( G = 4 \) and \( L = 3 \) were selected. For comparison, we also applied the LDM model with \( L = 3 \) in which the first component is the zero-inflation component and the other two components are Poisson regression models.

In Table 5, we reported the estimates of regression coefficients in the latent two Poisson regression models in both GHM and LDM. We also reported standard errors using 1000 bootstrap samples for GHM. The results show that the coefficients in GHM and LDM are not much different since the modeling frameworks of GHM and LDM are quite similar. Comparing the regression coefficients in the two components, they seems quite different in some covariates. In particular, the regression coefficients of “percentage of single-person household” and “percentage of university graduation” are not significant in the 2nd component, but they are significant in the 3rd component. This fact would indicate the importance of using mixture modeling in this application.
with group-wise mixing proportions in GHM for the second and third components. We colored the estimates of cluster-wise mixing proportions according to the grouping assignment in GHM. It is noted that the estimated prior probabilities in LDM are 0.14 and 0.74 for the second and third components, respectively, thereby the estimated mixing proportions in LDM are scattered around these values. From the figure, it is observed that the variations of the mixing proportions in LDM seems less than the group-wise mixing proportions in GHM, possibly because the use of the parametric Dirichlet distribution for the mixing proportions in LDM may restrict variations in the estimates. On the other hand, since the mixing proportions in the four groups are relatively different each other, we may easily interpret the characteristics of the groups, which could be one of practical advantages of the proposed method compared with LDM. We also note that the mixing proportions of the first component (zero inflated term) are 0.11, 0.23, 0.00, 0.13 in groups 1–4, respectively, so that the zero-inflation component would also be a meaningful component to distinguish the four groups.

We next investigated the predictive performance of GHM and LDM. We randomly omitted $h\%$ of samples, where $h \in \{5, 10, 15, 20\}$, to act as test data, and set the rest of the data as training data. We then estimated GHM and LDM based on the training data, where $G$ and $L$ in GHM were selected via the information criterion (4) from the same candidates and $L = 3$ was adopted in LDM, and predict the response values (number of crimes) using the covariates and offset in the test data. For measuring the performance of the prediction, we calculated the negative predictive log-likelihood (PLL) and mean squared error (MSE) that measures the prediction performance of GHM and LDM under $h\%$ of samples are set to test data.

$$PLL = -\sum_{i=1}^{m} \sum_{j \in r_i} \log \left\{ \hat{\pi}_{i1} \delta_0(y_j) + \sum_{k=2}^{L} \hat{\pi}_{ik} \text{Pois}(y_j; a_j \exp(x'_j \hat{\beta}_k)) \right\},$$

$$MSE = \sum_{i=1}^{m} \sum_{j \in r_i} \left\{ \log(\hat{\mu}_{ij} + 1) - \log(y_{ij} + 1) \right\}^2,$$

where $r_i$ denotes the index set of the test data in the $i$th cluster. Note that smaller values indicate better prediction performance in both criterion, and PLL and MSE are concerned with performance in terms of distribution prediction and point prediction, respectively. We repeated the procedure for 100 times, and reported the mean and median values in Table 6. It shows that the proposed GHM approach outperforms the LDM method in terms of both distribution prediction and point prediction, and the amount of improvement tends to increase as the percentage of test data increases. This may be because the characteristics of GHM that provides parsimonious modeling and is numerically stable compared with LDM results in good prediction especially when the training data used in the estimation is limited.

Finally, we assessed computation time of the proposed method and existing approaches. Here we also considered a Poisson mixed model with random coefficients as a potential candidate from generalized linear mixed models (GLMM), using R package "lme4" (Bates et al. 2016). We randomly sampled $\lfloor aN \rfloor$ units from the original dataset, where $\lfloor \cdot \rfloor$ is the floor function, $N = 2855$ is the original sample size and $a$ is a scaling constant. We considered five cases for $a$, namely $a \in \{0.6, 0.8, 1.0, 1.2, 1.4\}$ to investigate the potential impact of the sample size on the computation time. For each

**Table 6.** Mean and median values of 100 replicated values of negative predictive log-likelihood (PLL) and mean squared error (MSE) that measures the prediction performance of GHM and LDM under $h\%$ of samples are set to test data.

| $h$ | PLL (mean) | PLL (median) | MSE (mean) | MSE (median) |
|-----|------------|--------------|------------|--------------|
|     | GHM | LDM | GHM | LDM | GHM | LDM | GHM | LDM |
| 5   | 210.1 | 213.4 | 209.2 | 212.5 | 54.1 | 62.7 | 50.9 | 51.2 |
| 10  | 417.4 | 424.9 | 417.5 | 423.6 | 120.5 | 143.1 | 104.8 | 107.4 |
| 15  | 625.9 | 635.0 | 623.6 | 634.0 | 166.3 | 186.0 | 155.0 | 161.7 |
| 20  | 834.3 | 844.0 | 834.3 | 843.9 | 231.2 | 264.6 | 213.1 | 253.3 |
easy interpretations of the estimation results, we may lose a certain degree of flexibility when the data does not admit any grouping structures. To overcome the difficulty, we may consider "soft grouping" version of the proposed method by considering pseudo-probability of grouping assignment instead of deterministic grouping in the estimation algorithm. This direction of research could also be an interesting future work.

Appendix A

A.1. Proof of Theorem 1

We require the following regularity conditions:

(C1) The density (probability) function \( g(Y|X; \psi) \equiv \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_i(y_{ij}|x_{ij}; g_i, \psi) \) is identifiable in \( \Theta \) up to the permutation of the component and grouping labels.

(C2) The Fisher information matrix

\[
F_n(\Theta) = \mathbb{E} \left( \frac{\partial}{\partial \psi} \log g(Y|X; \psi) \right) \left( \frac{\partial}{\partial \psi} \log g(Y|X; \psi) \right)^T
\]

is finite and positive definite at \( \psi = \psi^0 \) and \( \gamma = \gamma^0 \).

(C3) There exists an open subset \( \Omega \) containing true parameters \( \psi \) such that there exists functions \( M_k(x, y), k = 1, 2, 3, \) with

\[
f_i(y|x; \psi) < M_1(x, y), \quad \frac{\partial}{\partial \psi} \log f_i(y|x; \psi) < M_2(x, y), \quad \frac{\partial^2}{\partial \psi_s \partial \psi_t} \log f_i(y|x; \psi) < M_3_{st}(x, y),
\]

for arbitrary \( \gamma \) and \( \gamma^1 \). Moreover, \( E[M_1(x, y)^2] < \infty, E[M_2(x, y)^2] < \infty \) and \( E[M_3_{st}(x, y)^2] < \infty \).

(C4) \( \psi^0 \) is an interior point in the compact set \( \Omega \subset \mathbb{R}^{\dim(\psi)} \).

We define a set of neighborhood of \( \psi^0 \) as \( N_\eta = \{ \psi^0 + \eta u; \| u \| = 1 \} \), where \( \eta \) is a scalar, \( u \in \mathbb{R}^{\dim(\psi)} \), and \( \| \cdot \| \) denotes the \( L^2 \)-norm. We first show the following lemma.

Lemma A1. For any \( \delta > 0 \) and \( \psi \in N_\eta \), it holds that

\[
\frac{1}{m} \sum_{i=1}^{m} I(\mathcal{G}_i(\psi) \neq g_0) = o_P(n_1^{-\delta}).
\]

Proof. From Markov’s inequality, it follows that

\[
P \left( \frac{1}{m} \sum_{i=1}^{m} I(\mathcal{G}_i(\psi) \neq g_0) > \epsilon n_1^{-\delta} \right) \leq \frac{n_1}{m} \sum_{i=1}^{m} P(\mathcal{G}_i(\psi) \neq g_0) \quad (A.1)
\]

for arbitrary \( \epsilon > 0 \) and \( \psi \in N_\eta \). Note that

\[
P(\mathcal{G}_i(\psi) \neq g_0) \leq \sum_{g \neq g_0} \left( \sum_{j=1}^{n_i} \log \frac{f_i(y_{ij}|g_i; \psi)}{f_i(y_{ij}|g_0; \psi)} \right) \geq \sum_{j=1}^{n_i} \log \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} \geq \sum_{j=1}^{n_i} \log \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} 
\]

\[
= \sum_{j=1}^{n_i} \log \left( \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} \right) \geq 0.
\]

Let \( v_{ij} = \log(\frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)}) \). Then, we have

\[
P \left( \sum_{j=1}^{n_i} v_{ij} \geq 0 \right) = P \left( \exp \left( \sum_{j=1}^{n_i} v_{ij} \right) \geq 1 \right) \leq \mathbb{E} \left( \exp \left( \sum_{j=1}^{n_i} v_{ij} \right) \right)
\]

\[
= \mathbb{E} \left( \exp \left( \sum_{j=1}^{n_i} \log \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} \right) \right) \leq \mathbb{E} \left( \exp \left( \sum_{j=1}^{n_i} \log \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} \right) \right)^{n_i}
\]

\[
= \exp \left( n_i \log \left( \frac{f_i(y_{ij}|g; \psi)}{f_i(y_{ij}|g_0; \psi)} \right) \right) \leq \text{const}.
\]
The second inequality follows from the general form of Markov inequality with increasing function \( \sqrt{x} \), and the third equality uses independence among \( v_j \) for \( j = 1, \ldots, n \). Note that

\[
\sqrt{\frac{1}{f_i(y_{ij}; g_i^0, \psi)}} = \sqrt{\frac{1}{f_i(y_{ij}; g_i^0, \psi^0)}} + f_i(y_{ij}; g_i^0, \psi^0)^{-3/2}(\psi - \psi^0)^{\frac{3}{2}} \left( \frac{\partial}{\partial \psi} f_i(y_{ij}; g_i^0, \psi) \right) |_{\psi = \psi^0}.
\]

Under Condition (C1), \( \lim_{\psi \to \psi^0} f_i(y_{ij}; g_i^0, \psi) = 0 \) when \( g_i \neq g_i^0 \) so that for sufficiently small \( \eta \), there exists a constant \( c > 0 \) such that

\[
\sup_{\psi \in \mathcal{N}_0} \log \left\{ 1 - H(f(\cdot; g, \psi), f(\cdot; g_i^0, \psi^0)) + Cn \right\} < -c.
\]

Hence, we have

\[
P \left( \sum_{j=1}^{n_i} \frac{\log f_i(y_{ij}; g, \psi)}{f_i(y_{ij}; g_i^0, \psi)} \geq 0 \right) \leq \exp(-cn_i) = o(n_i^{-\delta}),
\]

together the right side of (A.1) is \( o(1) \), which completes the proof.

We next show the consistency of \( \hat{\Theta} \) as given in the following lemma:

**Lemma A2.** As \( m \to \infty \) and \( n_1 \to \infty \), it holds that \( \hat{\Theta} \to \Theta_0 \).

**Proof.** Define \( u \in \mathbb{R}^{\dim(\psi)} \) such that \( \|u\| = 1 \) and \( y_1^1 = (g_{11}, \ldots, g_{1m}) \in \{1, \ldots, G\}^m \) such that \( d(y_1^1, \gamma^0) = \sum_{i=1}^m 1(g_{is} \neq g_{is}^0) = s \). Further, we define

\[
R(\psi, \gamma) = \frac{1}{N} \sum_{i=1}^m R_i(\psi, g_i) = \frac{1}{N} Q(\Theta).
\]

Note that

\[
\left| R(\psi) - R(\psi^0, \gamma^0) \right| = \left| R(\psi^0 + \nu u, \gamma^0) - R(\psi^0, \gamma^0) \right| + \left| R(\psi^0 + \nu u, \gamma^0) - R(\psi^0 + \nu u, \gamma^0) \right| = I_1 + I_2.
\]

Under Conditions (C2) and (C3), using the similar argument given in the proof of Theorem 1 in Hui, Warton, and Foster (2015), it holds that \( I_1 = o(1) \). Regarding \( I_2 \), it is noted that

\[
I_2 \leq \frac{1}{N} \sum_{i=1}^m \left| R_i(\psi^0 + \nu u, g_i) - R_i(\psi^0, g_i) \right| \leq \frac{1}{N} \sum_{i=1}^m I(g_{is} \neq g_{is}^0) |R_i(\psi^0 + \nu u, g_i) - R_i(\psi^0, g_i)| \leq aC,
\]

for some \( C > 0 \). Therefore, for any \( s \), there exists a local maximum inside the set \( \{ \psi^0 + u \varphi : \|u\| = 1 \} \cup \{ \gamma^1, \gamma^0 \} \), so that the consistency follows.

We define the following two objective functions:

\[
\hat{Q}(\psi) = \frac{1}{N} \sum_{i=1}^m \log f_i(y_{ij}; \psi) \equiv \frac{1}{N} \sum_{i=1}^m \log \left( \sum_{k=1}^{L} \pi_{\Theta}(\psi) h_k(y_{ij}|x_{ik}, \phi_k) \right),
\]

\[
\hat{Q}(\psi) = \frac{1}{N} \sum_{i=1}^m \log f_i^0(y_{ij}; \psi) \equiv \frac{1}{N} \sum_{i=1}^m \log \left( \sum_{k=1}^{L} \pi_{\Theta}(\psi) h_k(y_{ij}|x_{ik}, \phi_k) \right),
\]

and we define \( \hat{\psi} \) and \( \hat{\psi}^0 \) as the maximizer of \( \hat{Q}(\psi) \) and \( \hat{Q}(\psi) \), respectively. Note that \( \sqrt{N} \hat{Q}(\psi) - \hat{Q}(\psi) \sim N(0, \text{I}_{\dim(\psi)}) \). It holds that

\[
\left| \frac{1}{N} \hat{Q}(\psi) - \frac{1}{N} \hat{Q}(\psi) \right| \leq \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\log f_i(y_{ij}; \psi) - \log f_i^0(y_{ij}; \psi)}{2} \leq \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \log \frac{f_i(y_{ij}; \psi)}{f_i^0(y_{ij}; \psi)} \right)^2 \leq \frac{m}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \log \frac{f_i(y_{ij}; \psi)}{f_i^0(y_{ij}; \psi)} \right)^2,
\]

from Lemma A1 and the same argument used in (A.3). Hence, (6) follows under \( m/n_1^2 \to 0 \).
A.2. Proof of Theorem 2

In this proof, let $Q^0$ be the true parameter. Moreover, we use $\hat{Q}_{GL}$ and $Q_{GL}$ to denote $\hat{Q}$ and $Q$ in (A.2) to address the dependency on $G$ and $L$. Here we require an additional condition given below.

(C5) There exists a constant $R$ such that $E[Q_{GL}(\theta^0) - Q_{GL}(\hat{\theta}_{GL})] \geq R > 0$ for all models with $G < G_0$ or $L < L^0$, where $Q_{GL}(\theta_{GL})$ denotes the log-likelihood of a single observation, and $\theta_{GL}$ denotes the pseudo-true parameters (e.g., White 1982) which maximizes $E[Q_0(\theta_{GL})]$ with respect to $\theta_{GL}$.

Note that when $G < G_0$ or $L < L^0$, the model is under-specified, thereby the above condition requires that the Kullback–Leibler distance between any under-specified models with pseudo-true parameter and the true model with $G = G_0$ and $L = L^0$ is positive. Conditions similar to (C5) are typically imposed in theoretical development on consistency of model selection or variable selection (e.g., Keribin 2000; Hui, Muller and Welsh 2017).

We show that $P(1) = \frac{1}{L}$ for each $(G, L) \neq (G_0, L^0)$. We evaluate the probability separately in two cases, namely, $G < G^0$ or $L < L^0$ and $G > G_0$ and $L > L^0$.

We first consider the under-specified case. It holds that

$$\frac{1}{N}IC(G, L) - \frac{1}{N}IC(G_0, L^0) = \frac{2}{N}Q(\hat{\theta}_{GL}) - \frac{2}{N}Q(\hat{\theta}_{GL}) + P\log N,$$

where $P = G(L - 1) - G^0(L^0 - 1) - \sum_{k=1}^{\text{max}(L, L^0)} \text{dim}(\phi_k)$ is the difference of the numbers of parameters, which is $O(1)$. From the definition of $\hat{Q}$ and $\hat{Q}$ in (A.2) and the property given in (A.3), for sufficiently large $N$, it holds that

$$\frac{1}{N}Q(\hat{\theta}(\phi)) = \frac{1}{N}Q(\hat{\theta}(\phi)) + o_p(n^{-2}).$$

For sufficiently large $\delta^s$. For the under-specified choice, we can use a similar argument to the proof of Theorem 1 to show that $\hat{Q}_{GL}$ and $Q_{GL}$ hold the same property as (A.3), so that a similar argument given above leads to $N^{-1}Q(\hat{\theta}_{GL}) = N^{-1}Q(\hat{\theta}_{GL}) + o_p(1)$, where $\theta_{GL}$ is the pseudo-true parameter. Moreover, by the weak law of large numbers, we have $N^{-1}Q(\theta^0) \to E[Q_0(\theta^0)]$ and $Q(\hat{\theta}_{GL}) \to E[Q_0(\hat{\theta}_{GL})]$ in probability. Therefore, it follow that

$$N^{-1}Q(\hat{\theta}_{GL}) - Q(\hat{\theta}_{GL}) = \frac{1}{N}Q(\hat{\theta}_{GL}) - Q(\hat{\theta}_{GL}) + o_p(1) \geq R + o_p(1).$$

Then, (A.4) is strictly positive with probability tending to 1.

We next consider the over-specified case, namely, $G > G^0$ and $L > L^0$. Again, we let $P = G(L - 1) - G^0(L^0 - 1) - \sum_{k=1}^{L} \text{dim}(\phi_k)$ be the difference of the numbers of parameters, which is $O(1)$ and $P > 0$.

Note that

$$IC(G, L) - IC(G^0, L^0) = 2 \{Q(\hat{\theta}_{GL}) - Q(\hat{\theta}_{GL})\} + N\log N,$$

and

$$Q(\hat{\theta}_{GL}) - Q(\hat{\theta}_{GL}) = \{Q(\hat{\theta}_{GL}) - Q(\theta^0)\} - \{Q(\hat{\theta}_{GL}) - Q(\theta^0)\} = \{Q(\hat{\theta}_{GL}) - Q(\theta^0)\} \geq -\{Q(\hat{\theta}_{GL}) - Q(\theta^0)\} + \{Q(\hat{\theta}_{GL}) - Q(\theta^0)\}.$$

From a similar argument to (A.5), we can obtain that the first term is $o_p(n_1^{-2})$ for sufficiently large $\delta^s$. Regarding the second term, it holds that

$$Q(\hat{\theta}_{GL}) - Q(\hat{\theta}_{GL}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(\hat{\theta}_{GL} = \theta^0) \log \frac{f_{ij}(\hat{\theta}_{GL})}{f_{ij}(\theta^0)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(\hat{\theta}_{GL} = \theta^0) \log \frac{f_{ij}(\hat{\theta}_{GL})}{f_{ij}(\theta^0)} \leq 1 + I_2.$$

From a similar argument to Lemma A1 and the derivation of (A.3), it follows that $N^{-1}I_2 = o_p(n_1^{-2})$. On the other hand, we have

$$I_1 = \sum_{g=1}^{G} \sum_{j=1}^{m} I(\hat{\theta}_{GL} = \theta^0) \log \frac{f_{ij}(\hat{\theta}_{GL})}{f_{ij}(\theta^0)} = \sum_{g=1}^{G} \ell_g,$$

where $\ell_g$ is the likelihood ratio statistic for testing the number of components $L_g$ against $L$ in the mixture model fitted only to the data classified $g$th group. From (C4) and Dacunha-Castelle and Gassiat (1999), $\ell_g$ converges to a random variable characterized by the supremum of a stochastic process, whereby we obtain $I_1 \to O_p(1)$. Finally, we have

$$IC(G, L) - IC(G^0, L^0) = -2 \times O_p(1) + o_p(1) + P\log N,$$

so that the right-hand side tends to infinity as $N \to \infty$, and the difference is guaranteed to be positive with probability tending to 1.

Funding

The authors gratefully acknowledge JSPS KAKENHI grant numbers 16H07406 and 18K12757.

References

Ando, T., and Bai, J. (2016), "Panel Data Models With Grouped Factor Structure Under Unknown Group Membership," Journal of Applied Econometrics, 31, 163–191. [1000]

Bates, D., Machler, M., Bolker, B., and Walker, S. (2016), "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software, 67, 1–48. [1003,1006]

Bonhomme, S., and Manresa, E. (2015), "Grouped Pattern of Heterogeneity in Panel Data," Econometrica, 83, 1147–1184. [1000]

Chung, Y., and Dunson, D. B. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection," Journal of the American Statistical Association, 104, 1646–1660. [999]

Dacunha-Castelle, D., and Gassiat, E. (1999), "Testing the Order of a Model Using Locally Conic Parametrization: Population Mixtures and Stationary ARMA Processes," The Annals of Statistics, 27, 1178–1209. [1009]

Demidenko, E. (2013), Mixed Models: Theory and Applications With R, New York: Wiley. [999]

Field, C. A., Pang, Z., and Welsh, A. H. (2010), "Bootstrapping Robust Estimators for Clustered Data," Journal of the American Statistical Association, 105, 1606–1616. [1002]

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., and Kelderman, H. (2018), "Detecting Treatment-Subgroup Interactions in Clustered Data With Generalized Linear Mixed-Effects Model Trees," Behavior Research Methods, 50, 2016–2034. [999]

Forgy, E. W. (1965), "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications," Biometrics, 21, 768–769. [1000]

Geweke, J., and Keane, M. (2007), "Smoothly Mixing Regressions," Journal of Econometrics, 138, 252–290. [999]
Hahn, J., and Moon, H. (2010), "Panel Data Models With Finite Number of Multiple Equilibria," *Econometric Theory*, 26, 863–881. [1000]

Hajjem, A., Bellavance, F., and Larocque, D. (2011), "Mixed Effects Regression Trees for Clustered Data," *Statistics & Probability Letters*, 81, 451–459. [999]

Hui, F. K. C., Warton, D. L., and Foster, A. D. (2015), "Multi-Species Distribution Modeling Using Penalized Mixture of Regressions," *The Annals of Applied Statistics*, 9, 866–882. [1007,1008]

Hui, F. K. C., Muller, S., and Welch, A. H. (2017), "Joint selection in mixed models using regularized PQL," *Journal of the American Statistical Association*, 112, 1323–1333. [1002,1009]

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87. [999]

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, New York: Springer. [999]

Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhya*, 62, 49–66. [1001,1002,1009]

Khalili, A., and Chen, J. (2007), "Variable Selection in Finite Mixture of Regression Models," *Journal of the American Statistical Association*, 102, 1025–1038. [1007]

Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Models by Using Smoothing Splines," *Journal of the Royal Statistical Society*, Series B, 61, 381–400. [999]

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [999]

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278. [1000]

Ng, S. K., and McLachlan, G. J. (2014), "Mixture Models for Clustering Multilevel Growth Trajectories," *Computational Statistics & Data Analysis*, 71, 43–51. [999]

Nguyen, H. D., and McLachlan, G. J. (2016), "Laplace Mixture of Linear Experts," *Computational Statistics & Data Analysis*, 93, 177–191. [999]

Potts, R. B. (1952), "Some Generalized Order-Disorder Transformations," *Mathematical Proceedings of the Cambridge Philosophical Society*, 48, 106–109. [1002]

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008), "The Nested Dirichlet Process," *Journal of the American Statistical Association*, 103, 1131–1154. [999]

Rosen, O., Jiang, W., and Tanner, M. A. (2000), "Mixtures of Marginal Models," *Biometrika*, 87, 391–404. [999]

Rubin, D. B., and Wu, Y. (1997), "Modeling Schizophrenic Behavior Using General Mixture Components," *Biometrics*, 53, 243–261. [999]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1001]

Sugasawa, S., Kobayashi, G., and Kawakubo, Y. (2019), "Latent Mixture Modeling for Clustered Data," *Statistics and Computing*, 29, 537–548. [999,1003]

Sun, Z., Rosen, O., and Sampson, A. R. (2007), "Multivariate Bernoulli Mixture Models With Application to Postmortem Tissue Studies in Schizophrenia," *Biometrics*, 63, 901–909. [999]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [999]

Villani, M., Kohn, R., and Giordani, P. (2009), "Regression Density Estimation Using Smooth Adaptive Gaussian Mixtures," *Journal of Econometrics*, 153, 155–173. [999]

Villani, M., Kohn, R., and Nott, D. J. (2012), "Generalized Smooth Finite Mixtures," *Journal of Econometrics*, 171, 121–133. [999]

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002), "Conditional Second-Order Generalized Estimating Equations for Generalized Linear and Nonlinear Mixed Effects Models," *Journal of the American Statistical Association*, 97, 271–283. [1002]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–26. [1009]