Article
http://justc.ustc.edu.cn
Received: February 13, 2025; Accepted: May 14, 2025

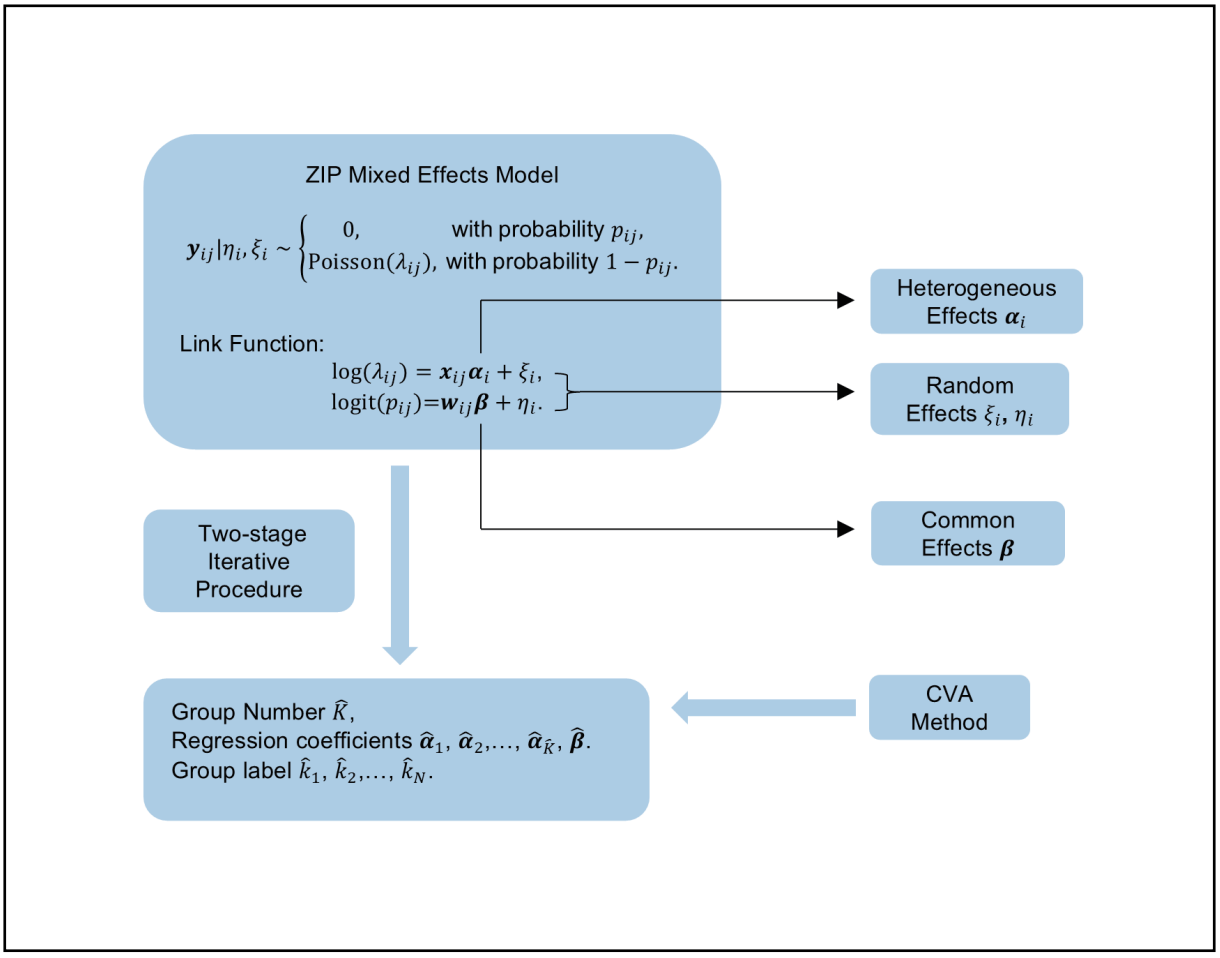
Subgroup analysis for zero-inflated count panel data

Yiyan Pan, Chao Han ✉, and Weiping Zhang

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Chao Han, E-mail: ustchanchao@mail.ustc.edu.cn

Graphical abstract



The main framework of the proposed method.

Public summary

- We propose a zero-inflated Poisson mixed effects model to account for the correlation among repeated measurements and identify subgroups in heterogeneous populations.
- Based on the K-means algorithm, we develop an iterative algorithm to estimate subject partitions and group-wise coefficients, with the number of groups determined by cross-validation using the averaging method.
- Compared with existing methods, our method is easier to implement and has a lower computational burden. The proposed estimator is consistent and asymptotically normal under mild conditions.

Citation: Pan Y Y, Han C, Zhang W P. Subgroup analysis for zero-inflated count panel data. *JUSTC*, **2025**, 55(0): . CSTR: 32290.14.JUSTC-2025-0016. DOI: 10.52396/JUSTC-2025-0016

JUSTC

Article

http://justc.ustc.edu.cn

Received: February 13, 2025; Accepted: May 14, 2025

Subgroup analysis for zero-inflated count panel data

Yiyan Pan, Chao Han ✉, and Weiping Zhang

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Chao Han, E-mail: ustchanchao@mail.ustc.edu.cn

Cite This: *JUSTC*, **2025**, 55(X): (13pp)

Read Online

Abstract: Zero-inflated Poisson (ZIP) regression model is widely used for count data with excess zeros. In this paper, we propose a zero-inflated Poisson mixed effects model to account for the correlation among repeated measurements and identify subgroups in heterogeneous populations. We develop an iterative algorithm to estimate the subgroup assignments and corresponding regression coefficients for all individuals, with the number of subgroups determined by the averaging cross-validation (CVA) method. Under mild assumptions, the proposed estimator exhibits consistency and asymptotically follows a normal distribution. The effectiveness of our method is demonstrated through extensive simulation studies and an application to a real dataset.

Keywords: count panel data; heterogeneity; mixed effects; subgroup analysis

CLC number: O213.9 **Document code:** A

2020 Mathematics Subject Classification: 62J99

1 Introduction

Zero-inflated count panel data are frequently encountered in clinical and public health research, with examples including alcohol consumption[1], blood product units administered post-surgery[2] and spatial epidemiology[3]. These measurements and studies offer valuable insights into the progression of targeted diseases and the long-term effects of interventions. The zero-inflated Poisson (ZIP) regression proposed by Lambert[4] is commonly used to handle an excessive number of zeros for count data, where the count response variables are assumed independently and identically follow a mixture of a Poisson distribution and a distribution with a point mass of one at zero. In a repeated measures design, where the assumption of independence and identical distribution is clearly violated, Hall[5] proposed a ZIP model with random effects and demonstrated via the whitefly data that ignoring the correlation among repeated observations on the same subject can seriously affect the validity of statistical inference. Later, Zhu et al.[6] considered potential heterogeneity and proposed a zero-inflated count model with heterogeneous random effects, modeling their variance as a function of covariates to account for random effects heterogeneity. Ghosh et al.[7] extended the ZIP models to include a broad class of distributions (e.g. power series distributions) and presented a Bayesian estimation approach. Considerable attention has been given to modeling correlations among repeated zero-inflated count data[1,8,9].

In recent decades, subgroup analysis has attracted considerable attention because of its ability to uncover meaningful subgroups within diverse populations and thus enhance predictive accuracy. By using a Bayesian random partition model[10] to inform the subgroup structure, Barcella et al.[11] proposed a Dirichlet process (DP) mixture of the ZIP distribution. However, they assumed that the subjects in the same subgroup follow identical ZIP distributions and that Markov Chain Monte Carlo (MCMC) algorithms are adopted for estimation. Moreover, their method requires specifying the number of mixture components and the underlying distribution of the data[12,13], which is rather complicated in practice. Commonly used methods in subgroup analysis, such as K-means clustering, hierarchical clustering, and spectral clustering with various linkage methods[14], are also effective and popular, but are typically applied only in low-dimensional scenarios. Previously, Ma and Huang[15] proposed a concave pairwise fusion approach to estimate group-wise regression coefficients and the number of groups simultaneously by imposing penalties. They utilized the alternating direction method of multipliers (ADMM)[16] algorithm to optimize the objective function. Chen et al.[17] applied the pairwise fusion approach to identify group-specific intercepts in the Poisson part, with the mixing probability not depending on covariates. However, this method cannot be extended to a ZIP model with random effects straightforward because of the introduction of integrals within the likelihood function, which would significantly increase the computational complexity of the ADMM algorithm.

In this paper, we propose a ZIP mixed effects model for the zero-inflated count panel data, accounting for the correlation among repeated measurements and simultaneously partitioning the subjects into a finite number of groups, with subjects in the same group sharing the same regression coefficients. Since the introduction of random effects leads to integrals in the calculation, the pairwise fusion approach is computationally intensive. Therefore, we present an iterative algorithm to obtain our estimator, referring to Ref. [18]. Given the grouping parameters, the remaining parameters are convenient for

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

estimation via the maximum likelihood estimate (MLE). On the other hand, given parameters related to the distribution, we estimate the grouping parameters based on the log-likelihood of each individual. To determine the number of clusters, we apply cross-validation with the averaging (CVA) method proposed by Wang[19], which has been shown to be consistent in selecting the number of clusters when they are properly separated into subgroups. Compared with the ADMM algorithm, the proposed algorithm has a much lower computational burden, and can directly utilize the R package **glmmTMB**[20] to solve the MLE. We further establish the statistical properties of our estimator in an asymptotic framework where N (the number of subjects) and T (the number of repeated measurements) tend to infinity simultaneously, but it is allowed for N to grow substantially faster than T. The estimators are shown to be consistent and asymptotically normal under the condition that $N = o(T^v)$ for some $v > 0$ and some additional assumptions. Unlike many large-T panel studies, such as Refs. [21, 22], we do not impose identical distributions or stationarity over the time series dimension, conditional on the unobserved effects. Furthermore, as illustrated in the simulation studies, the homogeneous regression coefficient estimates are consistent without strong restrictions on the growth rate of N relative to T.

The rest of the paper is organized as follows. In Section 2, we introduce the ZIP mixed effects model for zero-inflated count panel data. An iterative algorithm is developed to estimate regression coefficients and a CVA criterion is adopted for selecting the number of groups. In Section 3, we study the asymptotic properties of the proposed estimators. In Section 4, we evaluate the performance of the proposed estimators through various simulation settings. In Section 5, we apply ZIP mixed effects models to a real panel dataset. Some concluding remarks are given in Section 6. In the Appendix, we present the details of the technical proof in Section 3, along with additional results for Sections 4 and 5.

2 Methodology

2.1 Model

For panel data, let y_{ij} be the response of interest, $\boldsymbol{x}_{ij}$, $\boldsymbol{w}_{ij}$ be p-dimensional and q-dimensional vectors containing covariate information for subject i at time j, respectively, where $i = 1, \cdots, N$ and $j = 1, \cdots, T$. Assuming that, conditional on the random effects (η_i, ξ_i) we have

$$y_{ij}|\eta_i, \xi_i \sim \begin{cases} 0, & \text{with probability } p_{ij}; \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } 1 - p_{ij}. \end{cases} \tag{1}$$

The above model can be interpreted as a combination of a point mass of one at zero and a Poisson distribution with mean λ_{ij}. The probability p_{ij} allows the model to account for extra zeros in the data, and thus captures the zero-inflated phenomenon more effectively.

Suppose that the covariates associated with λ_{ij} are $\boldsymbol{x}_{ij}$ and the covariates associated with p_{ij} are $\boldsymbol{w}_{ij}$, where $\boldsymbol{x}_{ij}$ and $\boldsymbol{w}_{ij}$ can be identical or partially identical. Let the models for λ_{ij} and p_{ij} be specified as follows:

$$\begin{aligned} \log(\lambda_{ij}) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_i + \xi_i, \\ \text{logit}(p_{ij}) &= \boldsymbol{w}_{ij}^\top \boldsymbol{\beta} + \eta_i, \end{aligned} \tag{2}$$

where $\text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij}))$, $\boldsymbol{\alpha}_i$ is the regression parameter for the Poisson part that can be heterogeneous among subjects while parameter $\boldsymbol{\beta}$ is common to all subjects. To avoid potential identifiability issues[17,23] and simplify the problem, this paper only considers the heterogeneity of $\boldsymbol{\alpha}_i$. Random effects $(\eta_i, \xi_i)^\top$ are assumed to be independent and identically follow a bivariate normal distribution $N(0, 0, a, b, \rho)$, with $a > 0$, $b > 0$ and $-1 < \rho < 1$, which are incorporated to account for the correlation among repeated measurements[5]. Let $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, a, b, \rho)^\top$ represents the vector of common parameters, which does not change over i. We posit that subjects can be categorized into K distinct groups, with subjects sharing a common group parameter within the same group (i.e., $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ only when subjects i and j belong to the same group). To facilitate this classification, we introduce the label $k_i \in \{1, \cdots, K\}$, which signifies that subject i belongs to group k when $k_i = k$. Consequently, we denote the grouping parameter $\boldsymbol{\gamma} = \{k_1, \cdots, k_N\}$ and define $G_k = \{i : k_i = k\}$ as the set of subjects belonging to group k. Under such subgroup allocation $\boldsymbol{\gamma}$, we further define $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{k_i}$.

2.2 Estimation algorithm

Define $\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{iT})^\top$ as a T-dimensional response vector, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iT})^\top$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \cdots, \boldsymbol{w}_{iT})^\top$ as $T \times p$, $T \times q$ covariate matrices, respectively. The joint probability mass function of $\boldsymbol{Y}_i$ under (1) and (2) is

$$\begin{aligned} &P(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{\alpha}_{k_i}, \boldsymbol{\Theta}) = \\ &\int\int P(\boldsymbol{Y}_i, \eta_i, \xi_i|\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{\alpha}_{k_i}, \boldsymbol{\Theta}) \mathrm{d}\eta_i \mathrm{d}\xi_i = \\ &\int\int \prod_{j=1}^{T} \Big\{ \mathbb{I}\{y_{ij} = 0\} \frac{\exp(\boldsymbol{w}_{ij}^\top \boldsymbol{\beta} + \eta_i) + \exp(-\exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_{k_i} + \xi_i))}{1 + \exp(\boldsymbol{w}_{ij}^\top \boldsymbol{\beta} + \eta_i)} + \\ &\mathbb{I}\{y_{ij} > 0\} \frac{\exp(-\exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_{k_i} + \xi_i))}{1 + \exp(\boldsymbol{w}_{ij}^\top \boldsymbol{\beta} + \eta_i)} \frac{\exp(y_{ij}(\boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_{k_i} + \xi_i))}{y_{ij}!} \Big\} f(\eta_i, \xi_i|\boldsymbol{\Sigma}) \mathrm{d}\eta_i \mathrm{d}\xi_i, \end{aligned} \tag{3}$$

where $f(\eta_i, \xi_i|\boldsymbol{\Sigma})$ represents the probability density function of a bivariate normal distribution centered at the origin with covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} a^2 & ab\rho \\ ab\rho & b^2 \end{pmatrix}$.

Given K, the negative log-likelihood as the objective function is given as follows,

$$Q(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K, \boldsymbol{\Theta}, \boldsymbol{\gamma}) = -\sum_{i=1}^{N} \log(P(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{\alpha}_{k_i}, \boldsymbol{\Theta})). \tag{4}$$

The estimators are therefore defined as the solution of the following minimization problem:

$$(\hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_K, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\gamma}}) = \underset{(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K, \boldsymbol{\Theta}, \boldsymbol{\gamma})}{\arg\min}\, Q(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K, \boldsymbol{\Theta}, \boldsymbol{\gamma}), \tag{5}$$

where the minimum is taken over all possible groupings $\boldsymbol{\gamma} = \{k_1, \cdots, k_N\}$ of the N subjects into K groups, common parameters $\boldsymbol{\Theta}$, and group-specific parameters $\boldsymbol{\alpha}$.

We propose the following iterative algorithm (Algorithm 1) to solve (5) with a given K. In Step 3, the MLE is calculated by minimizing the negative log-likelihood function, with the

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

Algorithm 1. Two-stage estimation procedure.

1: Set $\boldsymbol{\gamma}^{(0)}=\{k_1^{(0)},\cdots,k_N^{(0)}\}$, $k_i^{(0)}\in\{1,2,\cdots,K\}$ as the initial estimate of $\boldsymbol{\gamma}$, and $t=0$.

2: **repeat**

3: For the given $\boldsymbol{\gamma}^{(t)}$, compute

$$(\boldsymbol{\alpha}_1^{(t+1)},\cdots,\boldsymbol{\alpha}_K^{(t+1)},\boldsymbol{\Theta}^{(t+1)})=\underset{(\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_K,\boldsymbol{\Theta})}{\arg\min}\ Q(\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_K,\boldsymbol{\Theta},\boldsymbol{\gamma}^{(t)}).$$

4: Compute for all $i\in\{1,\cdots,N\}$,

$$k_i^{(t+1)}=\underset{k\in\{1,2,\cdots,K\}}{\arg\min}\ -\log(P(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{W}_i,\boldsymbol{\alpha}_k^{(t+1)},\boldsymbol{\Theta}^{(t+1)})).$$

5: Set $t=t+1$.

6: **until** The stopping criteria of the algorithm are met.

random effects integrated via the Laplace approximation. We implement this calculation by using the R package **glmmTMB**[20]. To ease the computation, we treat $\boldsymbol{\Sigma}$ as a diagonal matrix in the applications in Section 4 and Section 5. For generality, we do not impose this assumption in the subsequent proofs. To obtain the initial cluster assignment, we first estimate $\boldsymbol{\alpha}_1^*,\cdots,\boldsymbol{\alpha}_N^*$ and $\boldsymbol{\Theta}^{(0)}$ from

$$(\boldsymbol{\alpha}_1^*,\cdots,\boldsymbol{\alpha}_N^*,\boldsymbol{\Theta}^{(0)})=\underset{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_N,\boldsymbol{\Theta}}{\arg\min}-\sum_{i=1}^{N}\log(P(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{W}_i,\boldsymbol{\alpha}_i,\boldsymbol{\Theta})).$$

Then we use K-means to divide $\boldsymbol{\alpha}_1^*,\cdots,\boldsymbol{\alpha}_N^*$ into K groups, with corresponding assignment $\boldsymbol{\gamma}^{(0)}=\{k_1^{(0)},\cdots,k_N^{(0)}\}$.

2.3 Determining the number of groups

The proposed algorithm proceeds with a prespecified number of groups, which is unknown in practice. In this paper, we adopt the cross-validation with averaging (CVA) method proposed by Wang[19], as its selection consistency has been proven. This method has also been applied by Zhang et al.[24] in the context of quantile regression for panel data, and Ito and Sugasawa[25] employed the same strategy to estimate the number of groups in grouped generalized estimating equations for longitudinal data.

The CVA focuses on the instability of clustering under given K. In the mth replication with $m=1,\cdots,M$, we randomly divide N subjects into three disjoint subsets Z_1, Z_2, and Z_3, where Z_1, Z_2 are regarded as training datasets, and each of them contains $n=[N/3]$ subjects. We first apply the proposed method to the two training datasets to obtain $\hat{\boldsymbol{\alpha}}_1^{(1)},\cdots,\hat{\boldsymbol{\alpha}}_K^{(1)},\hat{\boldsymbol{\Theta}}^{(1)}$; $\hat{\boldsymbol{\alpha}}_1^{(2)},\cdots,\hat{\boldsymbol{\alpha}}_K^{(2)},\hat{\boldsymbol{\Theta}}^{(2)}$. Then we obtain the optimal grouping assignment in Z_3:

$$\hat{k}_i^{(j)}=\underset{k\in\{1,2,\cdots,K\}}{\arg\min}-\log(P(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{W}_i,\hat{\boldsymbol{\alpha}}_k^{(j)},\hat{\boldsymbol{\Theta}}^{(j)}),\quad i\in Z_3,j\in\{1,2\}.$$

Based on two derived assignments, we introduce the function $\hat{s}_m(K)$ to measure the grouping instability,

$$\hat{s}_m(K)=\sum_{i,j\in Z_3}\mathbb{I}\{\mathbb{I}\{\hat{k}_i^{(1)}=\hat{k}_j^{(1)}\}+\mathbb{I}\{\hat{k}_i^{(2)}=\hat{k}_j^{(2)}\}=1\}.$$

If subjects i, j in the testing dataset are classified into the same group based on one training data, but are classified into different groups based on another training data, we have $\mathbb{I}\{\mathbb{I}\{\hat{k}_i^{(1)}=\hat{k}_j^{(1)}\}+\mathbb{I}\{\hat{k}_i^{(2)}=\hat{k}_j^{(2)}\}=1\}=1$, which implies that the corresponding grouping results are unstable. We calculate the average values of $\hat{s}_m(K)$ over M replications and denote it as $\hat{s}(K)$, i.e., $\hat{s}(K)=M^{-1}\sum_{i=1}^{M}\hat{s}_m(K)$. Then the minimizer of $\hat{s}(K)$ is selected as the number of groups.

3 Asymptotic properties

Let

$$l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})=(1/T)\log P(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{W}_i,\boldsymbol{\alpha},\boldsymbol{\Theta}),$$

$$\bar{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta})=\underset{\boldsymbol{\alpha}}{\arg\max}\sum_{i=1}^{N}\mathbb{I}\{k_i=k\}\mathbb{E}[l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})],$$

$$\hat{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta})=\underset{\boldsymbol{\alpha}}{\arg\max}\sum_{i=1}^{N}\mathbb{I}\{\hat{k}_i(\boldsymbol{\Theta})=k\}l_i(\boldsymbol{\alpha},\boldsymbol{\Theta}),$$

where k_i represents the true group label, $\hat{k}_i(\boldsymbol{\Theta})$ denotes the estimated group label of subject i under $\boldsymbol{\Theta}$. We denote the true value of the parameter for the kth group as $\boldsymbol{\Theta}^0$, $\boldsymbol{\alpha}_k^0$. Note that $\bar{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta}^0)=\boldsymbol{\alpha}_k^0$. Finally, for matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|$ denotes its spectral norm, and we use $\boldsymbol{M}>0$ to denote that $\boldsymbol{M}$ is positive definite. To state our first theorem, we make the following assumption.

Assumption 1.

(i) The parameter space $\boldsymbol{\Theta}$ for $\boldsymbol{\Theta}$ and parameter space $\mathcal{A}$ for $\boldsymbol{\alpha}_i$ are compact sets. N,T tend jointly to infinity and the number of groups K is finite.

(ii) $\max_i\sup_{\boldsymbol{\alpha},\boldsymbol{\Theta}}\|\mathbb{E}[l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})]\|=O(1)$, and similarly for the first three derivatives of l_i; $\max_i\sup_{\boldsymbol{\alpha},\boldsymbol{\Theta}}\mathbb{E}[\|\frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\alpha}^\top}\|^2]=O(1)$ and $\max_i\sup_{\boldsymbol{\alpha},\boldsymbol{\Theta}}\mathbb{E}[\|\frac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top}\|^2]=O(1)$; $\min_i\inf_{\boldsymbol{\alpha},\boldsymbol{\Theta}}\mathrm{mineig}(-\frac{\partial^2 l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top})\geq r+o_p(1)$, $r>0$, where $\mathrm{mineig}(\boldsymbol{M})$ is the minimum eigenvalue of $\boldsymbol{M}$; There exists

$$\delta>0$$

$$\text{s.t.}\quad \max_i\mathbb{E}[\|\frac{\partial l_i}{\partial\alpha}(\bar{\boldsymbol{\alpha}}(k_i,\Theta),\Theta)\|^4]=o(T^{-2\delta}),\quad \boldsymbol{\Theta}\in\boldsymbol{\Theta}.$$

In Assumption 1(i), we set the parameter spaces to be compact, which is also imposed in Refs. [22, 25]. Additionally, we assume that the number of groups, K, is finite as both N and T approach infinity. Assumption 1(ii) is similar to Assumption 3 in Ref. [22]. We require some moments to be bounded and strict concavity of the log-likelihood as a function of $\boldsymbol{\alpha}$. We assume that $\max_i\mathbb{E}[\|\frac{\partial l_i}{\partial\alpha}(\bar{\boldsymbol{\alpha}}(k_i,\Theta),\Theta)\|^4]=o(T^{-2\delta})$, which specifically holds if

$$\|\sum_{i=1}^{N}(\frac{\partial l_i}{\partial\boldsymbol{\alpha}}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})-\mathbb{E}[\frac{\partial l_i}{\partial\boldsymbol{\alpha}}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})])\|^2=o_p(NT^{-\delta})$$

and $\{(\boldsymbol{X}_i,\boldsymbol{W}_i):1\leq i\leq N\}$ are i.i.d. across i. For the consistency of $\hat{\boldsymbol{\theta}}$, we can relax this assumption and alternatively assume that $\mathbb{E}[\|\frac{\partial l_i}{\partial\boldsymbol{\alpha}}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\|^2]=o(1)$.

Theorem 1. Suppose Assumption 1 holds, we have

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}^0,$$

$$\frac{1}{N}\sum_{i=1}^{N}\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\theta}}),\hat{\boldsymbol{\theta}})-\boldsymbol{\alpha}_{k_i}^0\|^2 \xrightarrow{\mathrm{p}} 0.$$

The proof is provided in Appendix A.1. Theorem 1 shows the consistency of regression coefficient estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$. To obtain the asymptotic distributions of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\theta}}),\hat{\boldsymbol{\theta}})$, we need the following additional assumption.

Assumption 2.

(i) $N=o(T^{2\delta})$.

(ii) For all $k\in\{1,\cdots,K\}$, $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{k_i=k\}=\pi_k>0$, and there exists $C>0$ such that $\min_{k,\tilde{k}}\|\boldsymbol{\alpha}_k^0-\boldsymbol{\alpha}_{\tilde{k}}^0\|>C$ for all $k\neq\tilde{k}$, k, $\tilde{k}\in\{1,2,\cdots,K\}$.

(iii) Denote

$$\boldsymbol{s}_i=\frac{\partial l_i}{\partial\boldsymbol{\theta}}+\mathbb{E}\Big[\sum_{j\in G_{k_i}}\frac{\partial^2 l_j}{\partial\boldsymbol{\theta}\partial\boldsymbol{\alpha}^\top}\Big]\mathbb{E}\Big[\sum_{j\in G_{k_i}}\frac{\partial^2 l_j}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top}\Big]^{-1}\frac{\partial l_i}{\partial\boldsymbol{\alpha}},$$

$$\boldsymbol{H}_i=\mathbb{E}\Big[-\frac{\partial^2 l_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big]-\mathbb{E}\Big[\sum_{j\in G_{k_i}}\frac{\partial^2 l_j}{\partial\boldsymbol{\theta}\partial\boldsymbol{\alpha}^\top}\Big]\mathbb{E}\Big[\sum_{j\in G_{k_i}}\frac{\partial^2 l_j}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top}\Big]^{-1}\mathbb{E}\Big[\frac{\partial^2 l_i}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\theta}^\top}\Big],$$

$\boldsymbol{s}_i$ converges to the random variables $\boldsymbol{S}_i$ uniformly as T tends to infinity. And there exist

$$\boldsymbol{H}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{H}_i,\quad \boldsymbol{H}>0,$$

$$\boldsymbol{S}=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{s}_i\boldsymbol{s}_i^\top],$$

$$\boldsymbol{\Sigma}_k=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i\in G_k}\mathbb{E}[\frac{\partial^2 l_i}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top}],$$

$$\boldsymbol{R}_k=\lim_{N,T\to\infty}\frac{1}{N}\sum_{i\in G_k}\mathbb{E}[\frac{\partial^2 l_i}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\theta}^\top}].$$

Here we omit references to the evaluation points $\boldsymbol{\theta}^0$ and $\boldsymbol{\alpha}_{k_i}^0$.

Assumption 2(i) restricts the growth rate of N to ensure the classification consistency of the grouping variables, with δ defined in Assumption 1. We do not assume that N and T are of the same order as those in Refs. [21, 26], and that δ can be greater than 1. Assumption 2(ii) is typically imposed in the literature on the grouping approach in panel data models, e.g., Refs. [18, 25]. It ensures that the K groups are well separated so that the parameters $\boldsymbol{\alpha}_i^0$ and k_i can be identifiable. Assumption 2(iii) aims to provide several concrete asymptotic symbols.

Theorem 2. Supposing that Assumptions 1 and 2 hold, we have

$$P\Big(\sup_{i\in\{1,2,\cdots,N\}}|\hat{k}_i(\hat{\boldsymbol{\theta}})-k_i|>0\Big)=o_p(\sqrt{N}T^{-\delta}),$$

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\xrightarrow{\mathrm{d}}N(0,\boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}),$$

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}(k,\hat{\boldsymbol{\theta}})-\boldsymbol{\alpha}_k^0)\xrightarrow{\mathrm{d}}N(0,\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Omega}_k\boldsymbol{\Sigma}_k^{-1}),$$

where $\boldsymbol{\Omega}_k=\boldsymbol{R}_k\boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}\boldsymbol{R}_k^\top$.

The proof of Theorem 2 is provided in Appendix A.1. Under Assumption 2(i), we have $NT^{-2\delta}=o_p(1)$, which means that each subject can be classified into the right subgroup with a probability approaching 1.

4 Simulation studies

To further investigate the performance of the proposed method, we conduct two simulation experiments. As mentioned in Section 3, the proposed estimators possess the Oracle property and exhibit asymptotic normality under the stated assumptions. First, we design a simulation experiment to assess the performance of the estimators across varying N and T. Subsequently, we alter the distance among clusters to observe how the performance of our method is affected when the clusters are relatively close.

We set $p=q=3$, and the covariates $(x_{ij,1},x_{ij,2})^\top$ and $(w_{ij,1},w_{ij,2})^\top$ are generated independently from the bivariate standard normal distribution, let $\boldsymbol{x}_{ij}=(1,x_{ij,1},x_{ij,2})^\top$, $\boldsymbol{w}_{ij}=(1,w_{ij,1},w_{ij,2})^\top$. The true value covariance for random effects is set as $\boldsymbol{\Sigma}=\mathrm{diag}(1/8,1/8)$. For the sake of computational convenience, we fix ρ to be 0[27], while the remaining parameters are all unknown. The response variable is then generated according to the ZIP model (1) with

$$\log(\lambda_{ij})=\boldsymbol{x}_{ij}^\top\boldsymbol{\alpha}_{k_i}+\xi_i,$$
$$\mathrm{logit}(p_{ij})=\boldsymbol{w}_{ij}^\top\boldsymbol{\beta}+\eta_i.$$

We evaluate the performance of the estimation of $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}$ via the squared error loss defined as $SEL_\alpha=\sum_{k=1}^{K}\|\hat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^0\|^2$, $SEL_\beta=\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|^2$. The classification accuracy is assessed by $N^{-1}\sum_{i=1}^{N}\mathbb{I}\{\hat{k}_i=k_i\}$. We compare the proposed estimators with those obtained via the K-means method (treat $\boldsymbol{\alpha}$ as individual-specific and divide the groups based on their distance), and also present results based on the true group assignment (Oracle). Our method is denoted as G-ZIPm.

Example 1. In this example, we study the effect of sample size on the estimator. We set N=180 or 270 and T=20 or 30. For the true grouping assignment, we let $k_i=1$ for $i=1,\cdots,N/3$, $k_i=2$ for $i=N/3+1,\cdots,2N/3$ and $k_i=3$ for $i=2N/3+1,\cdots,N$. Here we set

$$\boldsymbol{\beta}^0=(1,-1,1)^\top,\ \boldsymbol{\alpha}_1^0=(1,-1,1)^\top,$$
$$\boldsymbol{\alpha}_2^0=(-1,-1,1)^\top,\ \boldsymbol{\alpha}_3^0=(0,1,-1)^\top.$$

Table 1 displays the selection probabilities of K. We ob-

Table 1. Selection probabilities (%) of the number of groups (K) obtained from the CVA based on 100 Monte Carlo replications for different methods.

(N,T)	Method	K				
		2	3	4	5	6
(180,20)	G-ZIPm	24	76	0	0	0
	K-means	2	0	3	13	82
(180,30)	G-ZIPm	20	80	0	0	0
	K-means	0	1	4	33	62
(270,20)	G-ZIPm	6	94	0	0	0
	K-means	0	6	5	25	64
(270,30)	G-ZIPm	5	95	0	0	0
	K-means	13	26	17	22	22

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

Table 2. Mean and variance (shown in parentheses) of SEL_α and SEL_β over 500 Monte Carlo replications and multiplied by 100. We compare our estimator (G-ZIPm) with the results obtained from K-means and those obtained with correctly specified groups (Oracle).

(N,T)	G-ZIPm		K-means		Oracle	
	SEL_α	SEL_β	SEL_α	SEL_β	SEL_α	SEL_β
(180,20)	3.223(0.086)	1.462(0.017)	53.869(147.474)	1.80(0.034)	2.762(0.051)	1.368(0.014)
(180,30)	1.949(0.019)	0.924(0.006)	7.045(1.333)	1.105(0.011)	1.843(0.020)	0.855(0.005)
(270,20)	2.066(0.028)	0.951(0.008)	16.232(3.485)	1.119(0.009)	1.857(0.025)	0.910(0.007)
(270,30)	1.256(0.009)	0.634(0.003)	2.587(0.079)	0.616(0.003)	1.238(0.009)	0.623(0.003)

Table 3. Average values of classification accuracy (%) over 500 Monte Carlo replications for different methods.

(N,T)	Method	
	G-ZIPm	K-means
(180,20)	95.90	82.38
(180,30)	98.75	94.94
(270,20)	96.01	87.59
(270,30)	98.75	96.14

serve that as the sample size increases, the probability of selecting the correct number of groups increases. Moreover, when the sample sizes are large, such as $(N,T)=(270,30)$, the adopted CVA strategy can select the true K with a probability of almost 1. In comparison, the K-means method struggles to identify the true number of groups accurately under these settings. Table 2 lists the average values of SEL_α and SEL_β on the basis of real group number, i.e. $K=3$. As N and T increase, both SEL_α and SEL_β decrease and approach the results obtained with the Oracle method. The SEL_α obtained via K-means only approaches our method's results when N and T are large, but it is much worse than our results. Table 3 presents the classification accuracy with $K=3$, showing that the accuracy of our estimator approaches 1 as N and T grows. Collectively, Tables 1–3 illustrate the good performance of the proposed method. These findings highlight that our method significantly outperforms K-means, particularly when N and T are not sufficiently large.

Under the settings of Example 1, we compared the performance of our method with that of the pairwise fusion method[15,17], focusing on the RMSE and the selection of the number of groups K. More details are presented in Appendix A.2.1.

Example 2. In this example, we study the performance of the proposed estimator when clusters are not well-separated. We let $(N,T)=(270,20)$ and $k_i=1$ for $i=1,\cdots,N/3$, $k_i=2$ for $i=N/3+1,\cdots,2N/3$ and $k_i=3$ for $i=2N/3+1,\cdots,N$. Then we set

$$\boldsymbol{\beta}^0=(1,-1,1)^\top,\ \boldsymbol{\alpha}_1^0=(0.5,-0.5,0.5)^\top\mu,$$
$$\boldsymbol{\alpha}_2^0=(-0.5,-0.5,0.5)^\top\mu,\ \boldsymbol{\alpha}_3^0=(0,0.5,-0.5)^\top\mu,$$

where $\mu=1,2,3$ for three different settings.

From Table 4, we observe that CVA performs poorly when $\mu=1$, indicating that identifying the true number of clusters becomes challenging when the clusters are not well-separated. When $\mu=2$ or 3, our method can select the true number of groups with a probability close to 1. In contrast, the CVA combined with the K-means method can identify K to some extent only when $\mu=3$. Tables 5 and 6 demonstrate that μ significantly impacts the estimation of regression coefficients and group classification accuracy when $K=3$. Specifically, when $\mu=1$, the SEL_α and SEL_β obtained from our method are much larger than those obtained when the assignments are correctly specified. However, when $\mu=3$, the accuracy approaches 1 and SEL_α, SEL_β approach to those of the Oracle method. Across all the settings, the K-means method performs worse than our method does, particularly when μ is small. These results highlight the flexibility and robustness of our method in handling varying levels of cluster separation.

In the aforementioned simulations, the subjects are alloc-

Table 4. Selection probabilities (%) of the number of groups (K) obtained from the CVA based on 100 Monte Carlo replications for different methods.

μ	Method	K				
		2	3	4	5	6
1	G-ZIPm	38	23	16	10	13
	K-means	17	12	19	14	38
2	G-ZIPm	6	94	0	0	0
	K-means	0	6	5	25	64
3	G-ZIPm	0	100	0	0	0
	K-means	17	83	0	0	0

Table 5. The comparative results between our estimator (G-ZIPm) and two other methods: K-means and the approach with correctly specified subgroups (Oracle). Mean and variance (shown in parentheses) of SEL_α and SEL_β over 500 Monte Carlo replications and multiplied by 100.

μ	G-ZIPm		K-means		Oracle	
	SEL_α	SEL_β	SEL_α	SEL_β	SEL_α	SEL_β
1	8.304(0.407)	1.675(0.018)	414.519(916.448)	6.527(0.056)	2.097(0.013)	1.117(0.010)
2	2.066(0.028)	0.951(0.008)	16.232(3.485)	1.119(0.009)	1.857(0.025)	0.910(0.007)
3	1.766(0.031)	0.885(0.006)	42.909(208.376)	1.044(0.010)	1.582(0.029)	0.884(0.005)

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

Table 6. Average values of accuracy of classification (%) over 500 Monte Carlo replications for different methods.

Method	μ		
	1	2	3
G-ZIPm	81.92	96.01	98.06
K-means	38.12	87.59	93.07

ated evenly into groups. To better assess the performance of CVA in selecting the number of groups, we conduct additional simulations under different group structures. The detailed results are documented in Appendix A.2.2.

5 Empirical example

We apply the proposed approach to the dataset on gender differences in alcohol-related problem across two years[28], which was measured by the Rutgers Alcohol Problem Index RAPI[29]. This dataset is derived from an intervention study aimed at a large sample of heavy-drinking college students. After filtering out incomplete data, we obtain a balanced dataset of $N = 561$ individuals, consisting of 213 men and 348 women, with $T = 5$ measurements. Each individual was measured at the same time point for 6 consecutive months. The histogram of RAPI outcomes shown in Fig. 1 reveals an obvious zero-inflated phenomenon.

According to Ref. [28], we apply our model (1) with the following links for λ_{ij} and p_{ij},

$$\begin{aligned}\log(\lambda_{ij}) &= \alpha_{k_i,0} + \alpha_{k_i,1}\cdot \text{Gender}_i + \alpha_{k_i,2}\cdot \text{Time}_j + \\ &\quad \alpha_{k_i,3}\cdot \text{Gender}_i \times \text{Time}_j + \xi_i, \\ \text{logit}(p_{ij}) &= \beta_0 + \beta_1\cdot \text{Gender}_i + \beta_2\cdot \text{Time}_j + \\ &\quad \beta_3\cdot \text{Gender}_i \times \text{Time}_j + \eta_i.\end{aligned}$$

Here $\text{Gender} = 1$ signifies male, whereas $\text{Gender} = 0$ signifies female. Using the CVA strategy, we initially select the number of groups, K, from the candidates $\{2, 3, \cdots, 6\}$. The CVA value for each K is shown in Fig. A1 in Appendix A.3, and the CVA value is minimized at $K = 5$. Consequently, we proceed with the proposed algorithm by setting $K = 5$ to estimate the regression coefficients for each group. The estimated common coefficients, $\hat{\boldsymbol{\beta}} = (-3.210, -0.523, 0.017, 0.016)^\top$, suggest that males are more likely to have a RAPI score greater than zero, and the likelihood of RAPI being zero increases over time, which is consistent with the conclusion in Ref. [8]. Additionally, we present the estimated $\boldsymbol{\alpha}$ of each group in Table 7.

The estimated regression coefficients in the five groups diff substantially from each other. To visualize the difference, we investigate the average predicted RAPI for each groupin comparison with the sample mean of the RAPI in Fig. 2. The average predicted RAPI in each group is computed from 1000 outcomes generated from the fitted model. We also derive maximum likelihood estimates (ignoring heterogeneity), substitute them into groups estimated by G-ZIPm, and obtain corresponding RAPI predictions from 1000 replications. These are added to the RAPI prediction graph in Fig. 2, using dotted lines. The grouping results obtained via pairwise methods are shown in Fig. A2 in Appendix A.3.

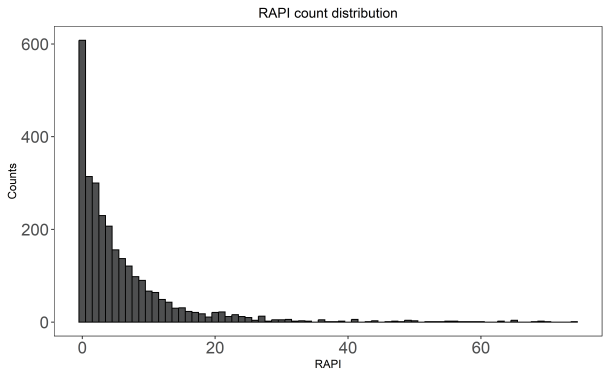


Fig. 1. Histogram of RAPI data.

Table 7. Group size and regression coefficients in each group for the Poisson part.

	Group				
	1	2	3	4	5
Group size	114	106	79	163	99
Intercept	2.448	0.799	1.530	0.918	2.222
Gender	0.240	0.652	1.104	0.239	–0.794
Time	–0.015	–0.129	0.051	0.005	–0.088
Gender×Time	–0.026	–0.020	–0.040	–0.013	0.141

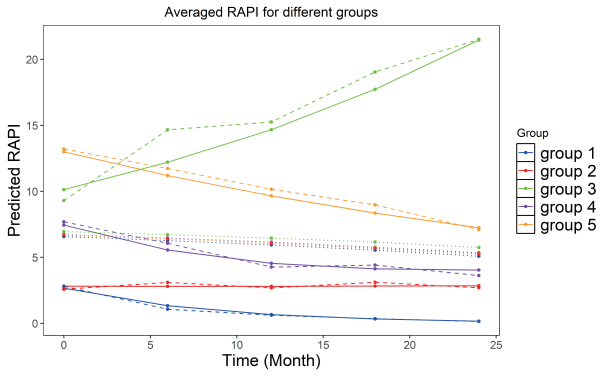


Fig. 2. Plot of average predicted RAPI obtained via GZIP-m (solid line), average predicted RAPI obtained by ignoring heterogeneity (dotted line), and sample mean of RAPI (dashed line) in each group.

In Table 7, we observe that Gender has a positive influence on RAPI within Groups 1-4, suggesting that males tend to have higher RAPI scores than females do. However, in Group 5, we find that $\alpha_{i1} < 0$, which contradicts common expectations. This highlights the importance of accounting for unobserved heterogeneity and conducting subgroup analysis to achieve personalized and accurate results. Furthermore, as shown in Fig. 2, the RAPI predictions derived without accounting for heterogeneity exhibit substantial deviations from the sample mean, thereby entirely violating the underlying distribution pattern. It indicates that subgroup analysis of this alcohol data is highly necessary. Our model provides a good fit, with five distinct groups clearly identified. The average predicted RAPI values align closely with the sample means in Groups 1, 2, 4, and 5. The suboptimal results in Group 3 may be attributed to either a small sample size or exceptionally

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

high RAPI values that deviate from the assumptions of the zero-inflated Poisson model. Fig. 2 presents the effects of the intervention on different groups of individuals. We note that as time progresses, the RAPI for Groups 1 and 2 decreases significantly, whereas the RAPI for Group 3 gradually increases, with no significant changes observed in the other groups.

6 Conclusions

In this paper, we propose a zero-inflated Poisson mixed effects model to account for potential heterogeneity among the subjects. We develop an iterative algorithm based on likelihood for partitioning individuals, and determine the number of subgroups via CVA. We also investigate the asymptotic properties of the obtained estimators. The proposed method performs well in both simulations and real data application.

The proposed method can be extended in several beneficial ways. For example, extending the proposed method to an unbalanced panel data setting is straightforward. In this context, we need to impose restrictions uniformly on the number of measurements for each individual (T_i) to ensure the consistency of the estimates. When the dimension of the covariates is large, it would be better to conduct variable selection by introducing a penalty function in the estimation equation[30]. Another future line of research is to relax the assumption that the number of groups (K) is finite.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (12171450).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yiyan Pan is currently a master's student at the Department of Statistics and Finance, University of Science and Technology of China. His research mainly focuses on subgroup analysis.

Chao Han is currently a Ph.D. candidate under the supervision of Prof. Weiping Zhang at the University of Science and Technology of China. His research mainly focuses on longitudinal data analysis.

Preprint statement

Methematics

References

[1] Buu A, Li R, Tan X, et al. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, **2012**, *31* (29): 4074–4086.

[2] DeSantis S M, Lazaridis C, Ji S, et al. Analyzing propensity matched zero-inflated count outcomes in observational studies. *Journal of Applied Statistics*, **2013**, *41* (1): 127–141.

[3] Lin P-S., Zhu J, Lin F-C. Iterative estimating equations for disease mapping with spatial zero-inflated poisson data. *Stat*, **2024**, *13* (1): e646.

[4] Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **1992**, *34* (1): 1–14.

[5] Hall D B. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, **2000**, *56* (4): 1030–1039.

[6] Zhu H, Luo S, DeSantis S M. Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Statistical Methods in Medical Research*, **2017**, *26* (4): 1774–1786.

[7] Ghosh S K, Mukhopadhyay P, Lu J-C. Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **2006**, *136* (4): 1360–1375.

[8] Zhang W, Wang J, Chen Y. A joint mean-correlation modeling approach for longitudinal zero-inflated count data. *Brazilian Journal of Probability and Statistics*, **2020**, *34* (1): 35–50.

[9] Jiang R, Zhan X, Wang T. A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data. *Journal of the American Statistical Association*, **2023**, *118* (542): 792–804.

[10] Lau J W, Green P J. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **2007**, *16* (3): 526–558.

[11] Barcella W, Iorio M D, Baio G, et al. A Bayesian nonparametric model for white blood cells in patients with lower urinary tract symptoms. *Electronic Journal of Statistics*, **2016**, *10* (2): 3287–3309.

[12] Everitt B, Hand D J. Finite Mixture Distributions. New York: Chapman & Hall, **1981**.

[13] Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, **1996**, *58* (1): 155–176.

[14] Romesburg H C. Cluster Analysis for Researchers. Belmont, CA, USA: Lifetime Learning Publications, **1984**.

[15] Ma S, Huang J. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, **2017**, *112* (517): 410–423.

[16] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **2011**, *3* (1): 1–122.

[17] Chen K, Huang R, Chan N H, et al. Subgroup analysis of zero-inflated poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics*, **2019**, *86*: 8–18.

[18] Bonhomme S, Manresa E. Grouped patterns of heterogeneity in panel data. *Econometrica*, **2015**, *83* (3): 1147–1184.

[19] Wang J. Consistent selection of the number of clusters via crossvalidation. *Biometrika*, **2010**, *97* (4): 893–904.

[20] Brooks M E, Kristensen K, van Benthem K J, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **2017**, *9* (2): 378–400.

[21] Hahn J, Kuersteiner G. Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, **2011**, *27* (6): 1152–1191.

[22] Bonhomme S, Lamadon T, Manresa E. Discretizing unobserved heterogeneity. *Econometrica*, **2022**, *90* (2): 625–643.

[23] Yip C H, Yau K W. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **2005**, *36* (2): 153–163.

[24] Zhang Y, Wang H J, Zhu Z. Quantile-regression-based clustering for panel data. *Journal of Econometrics*, **2019**, *213* (1): 54–67.

[25] Ito S, Sugasawa S. Grouped generalized estimating equations for longitudinal data analysis. *Biometrics*, **2023**, *79* (3): 1868–1879.

[26] Fernández-Val I, Weidner M. Individual and time effects in nonlinear panel models with large *N*, *T*. *Journal of Econometrics*, **2016**, *192* (1): 291–312.

[27] Liu J, Ma Y, Johnstone J. A goodness-of-fittest for zero-inflated Poisson mixed effects models in tree abundance studies. *Computational Statistics and Data Analysis*, **2020**, *144*: 106887.

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

[28] Neighbors C, Lewis M A, Atkins D C, et al. Efficacy of web-based personalized normative feedback: A two-year randomized controlled trial. *Journal of Consulting and Clinical Psychology*, **2010**, *78* (6): 898–911.
[29] White H R, Labouvie E W. Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, **1989**, *50* (1): 30–37.
[30] Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **2012**, *68* (2): 353–360.
[31] Wang H, Li R, Tsai C-L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **2007**, *94* (3): 553–568.
[32] Ma S, Huang J, Zhang Z et al. Exploration of heterogeneous treatment effects via concave fusion. *The International Journal of Biostatistics*, **2020**, *16* (1): 20180026.
[33] Shao L, Wu J, Zhang W et al. Integrated subgroup identification from multi-source data. *Computational Statistics and Data Analysis*, **2024**, *193*: 107918.
[34] Liu R, Shang Z, Zhang Y et al. Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, **2020**, *215* (2): 574–590.

Appendix

A.1 Proofs

Before the proof, we define several useful notations: $\boldsymbol{v}_i(\boldsymbol{\alpha},\boldsymbol{\Theta}) = \dfrac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}}$, $\boldsymbol{v}_i^{\alpha}(\boldsymbol{\alpha},\boldsymbol{\Theta}) = \dfrac{\partial^2 l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^{\top}}$, $\boldsymbol{\mu}_i(\boldsymbol{\alpha},\boldsymbol{\Theta}) = \dfrac{\partial l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$. $\hat{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta}) = \arg\max\limits_{\boldsymbol{\alpha}} \sum\limits_{i=1}^{N} \mathbb{I}\{\hat{k}_i(\boldsymbol{\Theta}) = k\} l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})$, $\bar{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta}) = \arg\max\limits_{\boldsymbol{\alpha}} \sum\limits_{i=1}^{N} \mathbb{I}\{k_i = k\} \mathbb{E}[l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})]$, where $\hat{k}_i(\boldsymbol{\Theta})$ is the estimated group label of subject i under $\boldsymbol{\Theta}$.

To show consistency of $\hat{\boldsymbol{\Theta}}$, we first establish the following technical lemma.

Lemma A1. Under the Assumption 1, we have

$$\frac{1}{N}\sum_{i=1}^{N} \|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})\|^2 = o_p\left(\frac{1}{\sqrt{N}T^{\delta}}\right),\ \forall \boldsymbol{\Theta} \in \boldsymbol{\Theta}, \tag{A1}$$

$$\sup_{\boldsymbol{\Theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})\|^2 = o_p(1). \tag{A2}$$

Proof. According to definition of $\hat{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta})$, we have $\sum\limits_i l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta}) \le \sum\limits_i l_i(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}),\boldsymbol{\Theta})$. Expand them to second order at $\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})$ and choose $\boldsymbol{a}_i(\boldsymbol{\Theta})$ between $\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})$ and $\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta})$ such that

$$\sum_i l_i(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}),\boldsymbol{\Theta}) =$$

$$\sum_i [l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta}) + \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^{\top}(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})) + \frac{1}{2}(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}))^{\top}\boldsymbol{v}_i^{\alpha}(\boldsymbol{a}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta})(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}))]$$

by Taylor expansion, then we have

$$\frac{1}{2N}\sum_i (\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}))^{\top}(-\boldsymbol{v}_i^{\alpha}(\boldsymbol{a}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}))(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})) \le \frac{1}{N}\sum_i \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^{\top}(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})). \tag{A3}$$

By Assumption 1(ii), there exists $r > 0$ such that

$$\min_i \inf_{\boldsymbol{\alpha},\boldsymbol{\Theta}} \min \operatorname{eig}[-\boldsymbol{v}_i^{\alpha}(\boldsymbol{\alpha},\boldsymbol{\Theta})] \ge r + o_p(1).$$

So $\frac{1}{2N}\sum_i (\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}))^{\top}(-\boldsymbol{v}_i^{\alpha}(\boldsymbol{a}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta})(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})) \ge \frac{r}{2}A$, where $A = \frac{1}{N}\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})\|^2$. By the Cauchy inequality,

$$A \le \frac{2}{rN}\sum_i \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^{\top}(\hat{\boldsymbol{\alpha}}(\hat{k}_i,\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})) \le 2\sqrt{\frac{1}{rN}\sum_i \|\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\|^2} \cdot \sqrt{A}. \tag{A4}$$

From the definition of $\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})$ and Assumption 1(ii), we know $\mathbb{E}[\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})] = \mathbf{0}$. By Assumption 1(ii), we see

$$\frac{1}{rN}\sum_i \|\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\|^2 = o_p\left(\frac{1}{\sqrt{N}T^{\delta}}\right). \tag{A5}$$

Then $A = o_p\left(\dfrac{1}{\sqrt{N}T^{\delta}}\right)$, that is (A1).

To prove (A2), we denote $Z(\boldsymbol{\Theta}) = \frac{1}{N}\sum_i \|\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\|^2$. We are going to show that if $\dfrac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = O_p(\sqrt{\sup_{\boldsymbol{\Theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta})})$, then

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta}) = o_p(1)$, as mentioned in Supplemental Material of Ref. [22]: For any $v>0$, $\epsilon>0$, there exists $M>0$ such that $P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\frac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\| > M\sqrt{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta})}\right) < \frac{\epsilon}{2}$. $\boldsymbol{\Theta}$ is a compact set, so there exists a finite cover of $\boldsymbol{\Theta} = B_1 \cup \cdots \cup B_r$, where B_i is an open ball with center $\boldsymbol{\Theta}_i$ and diameter $r_i \le \frac{1}{2M}\sqrt{v}$. Then, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta}) \le \max_i Z(\boldsymbol{\Theta}_i) + \sup_{\boldsymbol{\theta}} \|\frac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\| \frac{1}{2M}\sqrt{v}$. Since $a > v \Rightarrow a - \sqrt{a}\frac{1}{2}\sqrt{v} > \frac{v}{2}$, we can get $P(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta}) > v) \le \epsilon/2 + P(\max_i Z(\boldsymbol{\Theta}_i) > \frac{v}{2}) < \epsilon$ for N, T large enough.

Therefore it's sufficient to prove $\frac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = O_p(\sqrt{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta})})$.

$$\frac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \frac{2}{N}\sum_{i=1}^{N} \frac{\partial \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^\top}{\partial \boldsymbol{\Theta}} \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta}).$$

We have $\frac{\partial \boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^\top}{\partial \boldsymbol{\Theta}} = \frac{\partial \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})^\top}{\partial \boldsymbol{\Theta}} \boldsymbol{v}_i^{\alpha}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta}) + \boldsymbol{v}_i^{\theta}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})^\top$. From the definition of $\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})$, we have $\partial \mathbb{E}[\sum_{j\in G_{k_i}} l_j(\boldsymbol{\alpha},\boldsymbol{\Theta})]/\partial \boldsymbol{\alpha}|_{\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})} = \mathbf{0}$. Derivative of this function with respect to $\boldsymbol{\Theta}$, we obtain

$$\frac{\partial \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}^\top} = \mathbb{E}\left[\sum_{j\in G_{k_i}} -\boldsymbol{v}_j^{\alpha}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\right]^{-1} \mathbb{E}\left[\sum_{j\in G_{k_i}} \boldsymbol{v}_j^{\theta}(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\right]^\top. \tag{A6}$$

By Assumption 1(ii) and the Cauchy inequality, we have

$$\frac{\partial Z(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = O_p\left(\sqrt{\frac{1}{N}\sum_i \|\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})\|^2}\right) = O_p\left(\sqrt{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z(\boldsymbol{\Theta})}\right).$$

As we discussed above, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})\|^2 = o_p(1)$.

Proof of Theorem 1. From (A2), we then verify using a Taylor expansion that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\frac{1}{N}\sum_{i=1}^{N} l_i(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}),\boldsymbol{\Theta}),\boldsymbol{\Theta}) - \frac{1}{N}\sum_{i=1}^{N} l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})| = o_p(1).$$

By Assumption 1(ii), referencing the proof of the asymptotic properties of MLE, we have

$$\frac{1}{N}\sum_{i=1}^{N} l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}_r),\boldsymbol{\Theta}_r) \le \frac{1}{N}\sum_{i=1}^{N} l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0),\boldsymbol{\Theta}^0),$$

$\boldsymbol{\Theta}_r \in \boldsymbol{U}_r$, where $\boldsymbol{U}_r$ is a closed ball with center $\boldsymbol{\Theta}^0$ and diameter r. Then we have

$$\frac{1}{N}\sum_{i=1}^{N} l_i(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}_r),\boldsymbol{\Theta}_r),\boldsymbol{\Theta}_r) \le \frac{1}{N}\sum_{i=1}^{N} l_i(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\boldsymbol{\Theta}^0),\boldsymbol{\Theta}^0),\boldsymbol{\Theta}^0) + o_p(1).$$

Therefore, $\hat{\boldsymbol{\Theta}} \in \boldsymbol{U}_r$. Due to the arbitrariness of r, it is known that $\hat{\boldsymbol{\Theta}} \xrightarrow{\text{p}} \boldsymbol{\Theta}^0$.

Expanding $\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})$ at $\boldsymbol{\Theta}^0$, we obtain

$$\|\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}) - \bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)\|^2 = O_p(\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0\|^2) = o_p(1).$$

Combined with (A2), the second part of Theorem 1 is satisfied.

Denote $\tilde{\boldsymbol{\alpha}}(k,\boldsymbol{\Theta}) = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \mathbb{I}\{\hat{k}_i = k\} l_i(\boldsymbol{\alpha},\boldsymbol{\Theta})$, where $\hat{k}_i$ denotes the estimated group label under $\hat{\boldsymbol{\Theta}}$. $\tilde{\boldsymbol{\alpha}}$ is differentiable with respect to $\boldsymbol{\Theta}$. We note that $\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}) = \hat{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})$ and $\frac{1}{N}\sum_{i=1}^{N} \frac{\partial l_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}} = \mathbf{0}$.

Lemma A2. Under Assumption 1, we have

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial l_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}} - \frac{1}{N}\sum_{i=1}^{N} \frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}} = o_p\left(\sqrt[4]{\frac{1}{NT^{2\delta}}}\right). \tag{A7}$$

Proof. Under the conditions of Theorem 1, expanding $\boldsymbol{\mu}_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})$, around $\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})$, we obtain

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}=\frac{1}{N}\sum_{i=1}^{N}\{\boldsymbol{v}_i^{\boldsymbol{\theta}}(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})(\hat{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}))-\frac{\partial\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta})^{\top}}{\partial\boldsymbol{\Theta}}\boldsymbol{v}_i(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})\}+o_p\left(\sqrt[4]{\frac{1}{NT^{2\delta}}}\right)=o_p\left(\sqrt[4]{\frac{1}{NT^{2\delta}}}\right).$$

In the last equality, we have employed the expression $\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})=\hat{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})$, along with Assumption 1(2), Lemma A1, and (A5). It follows that (A7) is satisfied.

Finally, we provide the proof of Theorem 2 based on the above conclusions.

Proof of Theorem 2. To prove the theorem, we first show the asymptotic normality of $\hat{\boldsymbol{\Theta}}$, and then show the oracle property and asymptotic normality of $\hat{\boldsymbol{\alpha}}$.

(ⅰ) We have

$$\frac{\partial l_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}-\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}+\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}=\frac{\partial l_i(\tilde{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}=\mathbf{0}.$$

Then, according to Lemma A2 we expand $\frac{\partial l(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}}$ at $\boldsymbol{\Theta}^0$ to obtain the following equation,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0),\boldsymbol{\Theta}^0)}{\partial\boldsymbol{\Theta}}+\left(\frac{\partial}{\partial\boldsymbol{\Theta}^{\top}}\Big|_{\tilde{\boldsymbol{\Theta}}}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}\right)(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^0)=o_p\left(\sqrt[4]{\frac{1}{NT^{2\delta}}}\right),$$

where $\tilde{\boldsymbol{\Theta}}$ lies between $\boldsymbol{\Theta}^0$ and $\hat{\boldsymbol{\Theta}}$, and further expand $\frac{\partial}{\partial\boldsymbol{\Theta}^{\top}}\Big|_{\tilde{\boldsymbol{\Theta}}}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}),\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}$ around $\boldsymbol{\Theta}^0$ using that $\hat{\boldsymbol{\Theta}}$ is consistent. By (A6),

$$\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}^0+\boldsymbol{H}^{-1}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{s}_i+o_p\left(\sqrt[4]{\frac{1}{NT^{2\delta}}}\right).$$

Under Assumption 2(ⅲ), according to CLT, we get the asymptotic normality of $\hat{\boldsymbol{\Theta}}$.

(ⅱ) We construct inequality as

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbb{I}\{\hat{k}_i(\hat{\boldsymbol{\Theta}})\neq k_i\}\|\hat{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})\|^2]\leq\frac{1}{N}\sum_{i=1}^{N}\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})\|^2=o_p\left(\frac{1}{\sqrt{N}T^{\delta}}\right),\tag{A8}$$

where last equality we've used Lemma A1. Expanding $\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})$ at $\boldsymbol{\Theta}^0$, we have

$$\|\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)\|^2=O_p(\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^0\|^2)=o_p(1).$$

According to the classification of groups and the triangle inequality,

$$\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)\|\geq\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\boldsymbol{\Theta}^0)\|,$$

$$\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)\|\geq\|\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)-\bar{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\boldsymbol{\Theta}^0)\|-\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\boldsymbol{\Theta}^0)\|\geq\mathbb{I}\{\hat{k}_i(\hat{\boldsymbol{\Theta}})\neq k_i\}\cdot C-\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\boldsymbol{\Theta}^0)\|,$$

and summing the two inequalities show that $\|\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0)\|\geq\frac{1}{2}\mathbb{I}\{\hat{k}_i(\hat{\boldsymbol{\Theta}})\neq k_i\}\cdot C$. Therefore we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{\hat{k}_i(\hat{\boldsymbol{\Theta}})\neq k_i\}\|\hat{\boldsymbol{\alpha}}(\hat{k}_i,\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})\|^2\geq\frac{1}{N}\mathbb{I}\{\sup_{i\in\{1,\cdots,N\}}|\hat{k}_i(\hat{\boldsymbol{\Theta}})-k_i|>0\}\cdot\left(\frac{1}{4}C^2+o_p(1)\right).\tag{A9}$$

Combining (A8) and (A9), we obtain the first part of Theorem 2,

$$P(\sup_{i\in\{1,2,\cdots,N\}}|\hat{k}_i(\hat{\boldsymbol{\Theta}})-k_i|>0)=o(\sqrt{N}T^{-\delta}).$$

Define E_1 as the event $\{E_1:\hat{k}_i(\hat{\boldsymbol{\Theta}})=k_i,\forall i=1,\cdots,N\}$. From the above results, we can see that $P(E_1)=1$. Considering the consistency of $\hat{\boldsymbol{\Theta}}$, on E_1,we have

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0))=\sqrt{N}(\hat{\boldsymbol{\alpha}}(\hat{k}_i(\hat{\boldsymbol{\Theta}}),\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}}))+\sqrt{N}(\bar{\boldsymbol{\alpha}}(k_i,\hat{\boldsymbol{\Theta}})-\bar{\boldsymbol{\alpha}}(k_i,\boldsymbol{\Theta}^0))=o_p(\sqrt[4]{NT^{-2\delta}})+\frac{\partial\bar{\boldsymbol{\alpha}}(k_i,\tilde{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}^{\top}}\cdot\sqrt{N}(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^0),$$

where $\tilde{\boldsymbol{\Theta}}$ lies between $\boldsymbol{\Theta}^0$ and $\hat{\boldsymbol{\Theta}}$. By the asymptotic normality of $\hat{\boldsymbol{\Theta}}$ we've proved, we can get the asymptotic normality of $\hat{\boldsymbol{\alpha}}$ as mentioned in Theorem 2.

A.2 Additional results in Section 4

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

A.2.1 Comparison with the pairwise fusion method

Chen et al.[17] proposed a pairwise fusion method[15] for subgroup analysis of zero-inflated data. In this section, we apply this method in our context and compare it with our proposed approach (G-ZIPm). The objective function of the pairwise fusion method is given as

$$Q(\boldsymbol{\alpha},\boldsymbol{\Theta}) = -\sum_{i=1}^{N} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + \sum_{1\le i<j\le N} p_\kappa(\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|,\tau), \tag{A10}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top,\cdots,\boldsymbol{\alpha}_N^\top)^\top$, $l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) = (1/T)\log P(\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{W}_i,\boldsymbol{\alpha}_i,\boldsymbol{\Theta})$. And $p_\kappa(\cdot,\tau)$ is the SCAD penalty function with a tuning parameter $\tau > 0$[15, 17]. We minimize the modified Bayesian information criterion (mBIC) proposed by Ref. [31] to determine the tuning parameters.

$$\text{mBIC} = -\sum_{i=1}^{n} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + C_n\frac{\log n}{n}(\hat{K}p+q+3),$$

where $C_n = c\log(\log(np+q+3))$, and c is set to 50 in this simulation. We adopt the ADMM[16] algorithm to optimize the objective function. Specifically, we first recast (A10) as the equivalent constrained objective function:

$$\begin{aligned} &Q(\boldsymbol{\alpha},\boldsymbol{\Theta},\boldsymbol{\delta}) = -\sum_{i=1}^{N} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + \sum_{1\le i<j\le N} p_\kappa(\|\boldsymbol{\delta}_{ij}\|,\tau),\\ &\text{subject to } \boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j-\boldsymbol{\delta}_{ij} = \mathbf{0}, \end{aligned} \tag{A11}$$

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{ij}^\top, i<j\}^\top$. Thus the augmented Lagrangian is

$$L(\boldsymbol{\alpha},\boldsymbol{\Theta},\boldsymbol{\delta},\boldsymbol{\lambda}) = -\sum_{i=1}^{N} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + \sum_{1\le i<j\le N} p_\kappa(\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|,\tau) + \sum_{1\le i<j\le N}\boldsymbol{\lambda}_{ij}^\top(\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j-\boldsymbol{\delta}_{ij}) + \frac{\rho}{2}\sum_{1\le i<j\le N}\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j-\boldsymbol{\delta}_{ij}\|^2, \tag{A12}$$

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_{ij}^\top, i<j\}^\top$ are Lagrange multipliers and $\rho > 0$ is a pre-specified penalty parameter and Boyd et al.[16] showed that the ADMM algorithm converges for any arbitrary choice of ρ. In this simulation, we fix $\rho = 1$, and $\kappa = 3.7$. We update the estimates sequentially at the $(l+1)$th iteration step, as follows:

$$\begin{aligned} (\boldsymbol{\alpha}^{(l+1)},\boldsymbol{\Theta}^{(l+1)}) &= \arg\min_{\boldsymbol{\alpha},\boldsymbol{\Theta}} L(\boldsymbol{\alpha},\boldsymbol{\Theta},\boldsymbol{\delta}^{(l)},\boldsymbol{\lambda}^{(l)}),\\ \boldsymbol{\delta}^{(l+1)} &= \arg\min_{\boldsymbol{\delta}} L(\boldsymbol{\alpha}^{(l+1)},\boldsymbol{\Theta}^{(l+1)},\boldsymbol{\delta},\boldsymbol{\lambda}^{(l)}),\\ \boldsymbol{\lambda}_{ij}^{(l+1)} &= \boldsymbol{\lambda}_{ij}^{(l)} + \rho(\boldsymbol{\alpha}_i^{(l+1)}-\boldsymbol{\alpha}_j^{(l+1)}-\boldsymbol{\delta}_{ij}^{(l+1)}). \end{aligned} \tag{A13}$$

The Newton–Raphson method is used to solve the global minimizer of thefirst step of (A13). Define $ST(\boldsymbol{x},\tau) = (1-\tau/\|\boldsymbol{x}\|)_+\boldsymbol{x}$. For the second step in (A13), for $\kappa > 1/\rho+1$, $\boldsymbol{\delta}_{ij}^{(l+1)}$ can be updated via the following explicit solution:

$$\boldsymbol{\delta}_{ij}^{(l+1)} = \begin{cases} ST(\boldsymbol{u}_{ij}^{(l+1)},\tau/\rho) & \text{if } \|\boldsymbol{u}_{ij}^{(l+1)}\| \le \tau+\tau/\rho;\\ \dfrac{ST(\boldsymbol{u}_{ij}^{(l+1)},\kappa\tau/((\kappa-1)\rho))}{1-1/((\kappa-1)\rho)} & \text{if } \lambda+\lambda/\rho < \|\boldsymbol{u}_{ij}^{(l+1)}\| \le \kappa\lambda;\\ \boldsymbol{u}_{ij}^{(l+1)} & \text{if } \|\boldsymbol{u}_{ij}^{(l+1)}\| > \kappa\tau; \end{cases}$$

where $\boldsymbol{u}_{ij}^{(l+1)} = \boldsymbol{\alpha}_i^{(l+1)}-\boldsymbol{\alpha}_j^{(l+1)}+\rho^{-1}\boldsymbol{\lambda}_{ij}^{(l)}$. In the simulation, we found that the pairwise fusion method is highly sensitive to initial values in the context of our study. To find a resonable initial value, we consider the ridge fusion criterion given by

$$L_R(\boldsymbol{\alpha},\boldsymbol{\Theta}) = -\sum_{i=1}^{N} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + \frac{\lambda^*}{2}\sum_{1\le i<j\le N}\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|^2,$$

where λ^* is the tuning parameter having a small value and we set $\lambda^* = 0.001$ in simulstions. Ma et al.[32] selected initial values by adopting this strategy. To obtain better results, we further perform a simple grouping of $\boldsymbol{\alpha}_i$, dividing it into G groups ($G = \lceil\sqrt{N}\rceil$) via the k-means method, and the maximum likelihood estimate (MLE) $\boldsymbol{\alpha}^{(0)},\boldsymbol{\Theta}^{(0)}$ calculated based on this group structure as the initial values.

To accelerate the ADMM algorithm, we apply a k-nearest neighbors method to reduce the number of fusion terms as Ma et al.[32] and Shao et al.[33] did. Then the objective function can be written as:

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

$$Q(\boldsymbol{\alpha},\boldsymbol{\Theta}) = -\sum_{i=1}^{N} l_i(\boldsymbol{\alpha}_i,\boldsymbol{\Theta}) + \sum_{1\le i<j\le N} I_{ij}^{k} p_\kappa(\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|),$$

where I_{ij}^{k} indicates whether subject j is in i's k-nearest neighbors, and we set $k = 20$ ($k > \sqrt{N}$) in the simulations.

Under the settings of Example 1, we compared the performance of our method with that of the pairwise fusion method. The following tables report the mean and standard deviation (s.d.) of the RMSE, where $\mathrm{RMSE}(\boldsymbol{\alpha}) = \sqrt{\sum_{i=1}^{N}\|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{\alpha}_{k_i}^{0}\|^2/(Np)}$, $\mathrm{RMSE}(\boldsymbol{\beta}) = \sqrt{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\|^2/q}$, the selection probabilities of the number of groups and the average execution time per repetition (the running time of our method includes the time for selecting K using CVA). Table A1 presents the estimated values of K obtained using different methods. Our proposed method consistently achieves a higher probability of selecting the correct K across all experimental settings than does the pairwise fusion approach. The estimated values of K derived from the pairwise fusion method tend to be overestimated, which is likely due to the random effects incorporated into our model's framework. Table A2 presents the RMSE and average running time results. We observe that while the RMSE($\boldsymbol{\beta}$) of the pairwise fusion method are comparable to ours, our RMSE($\boldsymbol{\alpha}$) is lower than that of the pairwise fusion approach. Moreover, our method achieves a significantly shorter average running time than the pairwise fusion framework. Additionally, it should be noted that the pairwise method exhibits high sensitivity to initialization. To address this, its initial values are derived from ridge regression and k-means clustering[32, 33], which are closer to the true values compared with the initial values of our method. Despite employing acceleration strategies to expedite convergence, the pairwise method still underperforms our algorithm and incurs a longer average running time.

A.2.2 Performance of CVA

To better measure the performance of CVA in selecting group number K, we supplement a set of simulations under different group structures. In all the following scenarios, we set the frequency of observation $T = 20$, and $\boldsymbol{\beta}^0 = (1,-1,1)^\top$, $\boldsymbol{\alpha}_1^0 = (1,-1,1)^\top$, $\boldsymbol{\alpha}_2^0 = (-1,-1,1)^\top$, $\boldsymbol{\alpha}_3^0 = (0,1,-1)^\top$, $a = b = 1/8$, $\rho = 0$. The following scenarios are considered:

Scenario 1: $k_i = 1$ for $i = 1,\cdots,130$, $k_i = 2$ for $i = 131,\cdots,260$ and $k_i = 3$ for $i = 261,\cdots,270$.

Scenario 2: $k_i = 1$ for $i = 1,\cdots,260$, $k_i = 2$ for $i = 261,\cdots,520$ and $k_i = 3$ for $i = 521,\cdots,540$.

Scenario 3: $k_i = 1$ for $i = 1,\cdots,54$, $k_i = 2$ for $i = 55,\cdots,135$ and $k_i = 3$ for $i = 136,\cdots,270$.

Scenario 4: $k_i = 1$ for $i = 1,\cdots,90$, $k_i = 2$ for $i = 91,\cdots,180$ and $k_i = 3$ for $i = 181,\cdots,270$.

Except for Scenario 2, which has 540 observed individuals, all the other scenarios have 270 observed individuals. Scenario 1 represents a relatively unbalanced data where the size of group 3 is less than 5% of the total sample size. An extreme case in which there exists a group whose members are not assigned to a training set or testing set may occur in scenario 1 as the size of Group 3 is too small. In Scenario 3, the number of individuals in each group is moderate. Scenario 4 follows our previous setup, with all groups having an equal number of individuals. Based on 100 Monte Carlo replications, we obtained selection probabilities of each K, which are reported in Table A3.

From Table A3, we can observe that when each group has a moderate number of individuals, the CVA performs favorably, be-

Table A1. The estimated values of K obtained by different methods in 100 Monte Carlo replications.

(N,T)	Method	K					(N,T)	Method	K				
		2	3	4	5	6			2	3	4	5	6
(180,20)	G-ZIPm	24	76	0	0	0	(270,20)	G-ZIPm	6	94	0	0	0
	Pairwise	0	43	54	3	0		Pairwise	0	23	52	24	1
(180,30)	G-ZIPm	20	80	0	0	0	(270,30)	G-ZIPm	5	95	0	0	0
	Pairwise	0	82	18	0	0		Pairwise	0	72	25	3	0

Table A2. Average time per repetition, mean, and standard deviation (shown in parentheses) of the RMSE($\boldsymbol{\alpha}$) and RMSE($\boldsymbol{\beta}$) over 100 Monte Carlo replications.

(N,T)	Method	RMSE($\boldsymbol{\alpha}$)	RMSE($\boldsymbol{\beta}$)	Time (min)	(N,T)	Method	RMSE($\boldsymbol{\alpha}$)	RMSE($\boldsymbol{\beta}$)	Time (min)
(180,20)	G-ZIPm	0.148(0.036)	0.064(0.027)	16.512	(270,20)	G-ZIPm	0.147(0.026)	0.050(0.024)	13.982
	Pairwise	0.423(0.077)	0.075(0.036)	29.554		Pairwise	0.414(0.057)	0.054(0.028)	34.305
(180,30)	G-ZIPm	0.084(0.030)	0.052(0.022)	19.347	(270,30)	G-ZIPm	0.081(0.025)	0.042(0.018)	15.632
	Pairwise	0.280(0.053)	0.050(0.021)	24.435		Pairwise	0.279(0.047)	0.042(0.019)	40.355

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):

Table A3. Selection probabilities (%) of the number of groups (K) obtained from the CVA based on 100 Monte Carlo replications in different scenarios.

Scenarios	K					Scenarios	K				
	2	3	4	5	6		2	3	4	5	6
1	31	62	4	0	3	3	13	87	0	0	0
2	9	91	0	0	0	4	6	94	0	0	0

ing able to select the correct number of groups with a high probability. In Scenario 1, approximately one-third of the participants incorrectly identified the number of groups as 2. This is due to the extremely small size of the Group 3. However, in Scenario 2, where the group size proportions were identical to those in Scenario 1 and N was doubled compared with those in Scenario 1, our method surprisingly excelled by accurately identifying the correct K. This illustrates that when N is sufficiently large, our method can effectively avoid the occurrence of extreme situations and provide accurate estimates of K even for extremely unbalanced data. In special cases where the sample size N is small and the group assignment is imbalanced, CVA indeed has deficiencies in selecting K. In such circumstances, we can adopt the information criterion proposed by Liu et al.[34] to determine the number of groups, which can serve as a research direction for future studies.

A.3 Additional results in Section 5

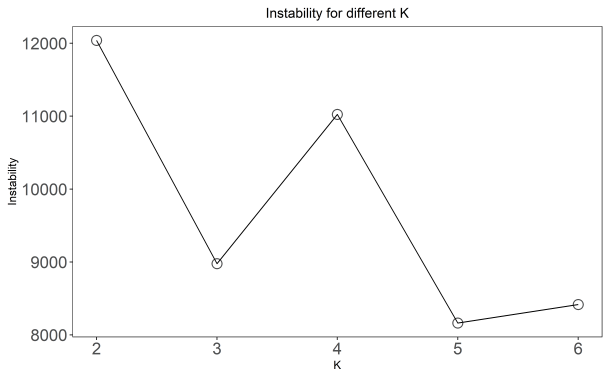


Fig. A1. $\hat{s}(K)$ values for different number of clusters.

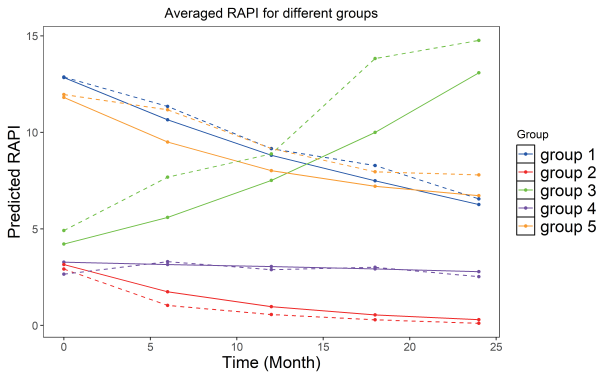


Fig. A2. Plot of average RAPI predictions derived from the pairwise fusion method (solid line) and sample mean of RAPI (dashed line).

In Fig. A1, we present the CVA values for the candidate values of K. Fig. A2 illustrates the grouping results of real data under the pairwise fusion approach. The average predicted RAPI in each group is computed from 1000 outcomes generated from the fitted model. The number of groups K is estimated as 5, and the group structures obtained from pairwise fusion and G-ZIPm are similar. Compared with the results obtained by G-ZIPm, the predicted RAPI values here exhibit a larger bias from the sample mean of RAPI in Group 3, with notably similar predictions between Group 1 and Group 5, indicating poor estimation performance.

CSTR: 32290.14.JUSTC-2025-0016
DOI: 10.52396/JUSTC-2025-0016
JUSTC, **2025**, 55(X):