*The Annals of Applied Statistics*
2024, Vol. 18, No. 3, 2342–2358
https://doi.org/10.1214/24-AOAS1884


# ASSESSING AQUATIC TOXICITY ASSESSMENT VIA A CLUSTERED VARIANCE MODEL

BY XIN WANG[1,a] AND JING ZHANG[2,b]

[1]*Department of Mathematics and Statistics, San Diego State University,* [a]*xwang14@sdsu.edu*
[2]*Department of Statistics, Miami University,* [b]*zhangj8@miamioh.edu*

Motivated by the need to assess consistency in the outcomes of aquatic toxicity tests conducted by different labs at different time points, we propose a clustering of variance method in linear mixed models. The proposed method, referred as CVM, is able to identify the cluster structure of the variances and estimate model parameters simultaneously. In our proposed method, a penalized approach based on pairwise penalties is proposed to identify the cluster structure. We construct an optimization problem and develop an algorithm based on the alternating direction method of multipliers. Simulation studies show that the proposed approach can identify the cluster structure well and outperforms traditional methods based on $k$-means. In the end, the proposed approach is applied to the aquatic toxicity assessment data, which gives a more reasonable cluster structure than the traditional methods.

**1. Introduction.** In aquatic toxicity tests, organisms are exposed to the chemical of interest and various biological endpoints, for example, hatching, survival, reproduction, or growth, are observed. These test outcomes are used to evaluate the potential negative impact of chemicals on different life stages of organisms (Bailer and Oris (1993)). The *C. dubia* reproduction test is a standard tool to assess the chronic impact of effluents discharged into freshwater systems (Amato et al. (1993), Bailey et al. (1996)). In a *C. dubia* reproduction test, organisms are randomly assigned to one of a few different concentration groups, including a zero-concentration control group. The number of young produced in a certain number of broods (usually three broods) or a specified study duration (typically, seven to eight days) is recorded. Multiple organisms (replicates) are observed at each concentration level. To conduct the whole effluent toxicity (WET) tests, a group exposed to the toxicant will be compared to the control group. In order to assess the negative impact of chemicals with different concentration levels, the number of young of each organism is then modeled as a function of the concentration via a generalized linear regression model (Bailer and Oris (1997), Dobson and Barnett (2018)). The estimated regression coefficients can be used to estimate the concentration level associated with a specific level of reproduction inhibition ($RI_p$) relative to the control group (Bailer and Oris (1997)), which is a popular analysis endpoint used in the comparison across different chemicals for toxicity assessment and policy making. For example, Safer Choice is the EPA's label for products with safer chemical ingredients. Safer Choice products meet strict human health and environmental criteria, including safer chemical criteria for ingredients like surfactants, solvents, and chelants, and it requires toxicity test results as supporting facts.

High consistency in control group observations increases the probability that unacceptable toxicant is identified in a WET test. The precision of $RI_p$ estimation also relies highly on the variability of control group observations (Zhang et al. (2022)). Therefore, in order to assess the toxicity of a chemical, it is essential to improve the consistency of control group

Received November 2023.
*Key words and phrases.* Aquatic toxicity assessment, heterogeneous variance, linear mixed models, penalty functions.

observations. This would help ensure that the potency estimates are comparable when they are produced in different batches of experiments conducted by different labs. When the control observations are highly variable and not comparable from batch to batch or from lab to lab, the resulting potency estimates or WET test findings would be misleading. Therefore, the consistency of the control observations plays an important role in toxicity assessment using the *C. dubia* reproduction tests.

1.1. *The* C. dubia *reproduction test data.* In this work we consider a study evaluating the consistency in *C. dubia* reproduction tests participated by 17 labs in California from August 2013 to July 2021. Due to the delay of experiments in one lab, the data set we study consists of *C. dubia* reproduction test outcomes from 16 labs. In the raw data, there are 1013 batches of experiments, among which 551 batches were control-only tests. In a single batch of control tests, 10 organisms (replicates) were exposed to the lab water and monitored for seven to eight days until the third brood of young was reproduced. The resulting total number of young was the experimental outcome of interest.

In addition to the *C. dubia* reproduction outcomes, water chemistry, and other experiment condition measurements were also collected, including alkalinity, conductivity, dissolved oxygen (unit: mg/L), hardness and pH of the water used in the tests, air temperature, light intensity, age of organisms when the tests started, etc. The objective of analyzing these test information altogether is to identify factors that impact the consistency of *C. dubia* reproduction test outcomes. As part of the early phase effort in the process, we are interested in grouping labs according to the variability of their experimental outcomes. In particular, the number of young produced in three broods in the control group is the response of interest; the water chemistry and experimental condition variables serve as covariates. The water chemistry variables impact the living environment of the organisms and hence are likely important factors that affect their reproduction. In addition, other covariates that could impact reproduction are also considered, including air temperature, light intensity as well as the age of organisms when these tests began. Due to the high percentage of missing information in many of these variables, we only considered a few water chemistry variables (conductivity, dissolved oxygen, and pH) and the air temperature of the labs as candidate covariates in the present study.

How the consistency of control observations are impacted by the water, light, and other experimental conditions can be evaluated through the proposed method, which utilizes a model that relates these control observations with the water chemistry measures and other experimental conditions. Grouping the variability or consistency of the control observations from different labs would allow us to evaluate the reliability of their potency estimates and conduct a quality check of the tests conducted in these labs.

The experimental outcomes were summarized for each batch as the average number of young of the 10 replicates used in the test. Let $y_{ih}$ be the logarithm of the average number of young in the $h$th batch conducted by the $i$th lab, and $\boldsymbol{x}_{ih}$ be the corresponding vector of covariates, including the intercept, for $h = 1, \ldots, m_i$ and $i = 1, \ldots, n$. In this data set, $n = 16$, and $m_i$'s take values from 7 to 80. In Figure 1 the distributions of the response variable are shown. The $x$-axis is the lab name along with the number of observations in the parenthesis. Different labs have different mean of the logarithm of the average number of young and different variabilities.

We first consider a linear mixed model (Stroup (2013)) relating the mean number of young and the covariates as follows:

$$y_{ih} = \boldsymbol{x}_{ih}^T \boldsymbol{\beta} + v_i + \epsilon_{ih}, \tag{1}$$

where $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ is the random effect of lab $i$, and $\epsilon_{ih} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ is the random error for $h = 1, \ldots, m_i$ and $i = 1, \ldots, n$. The covariates are conductivity, dissolved oxygen, pH of

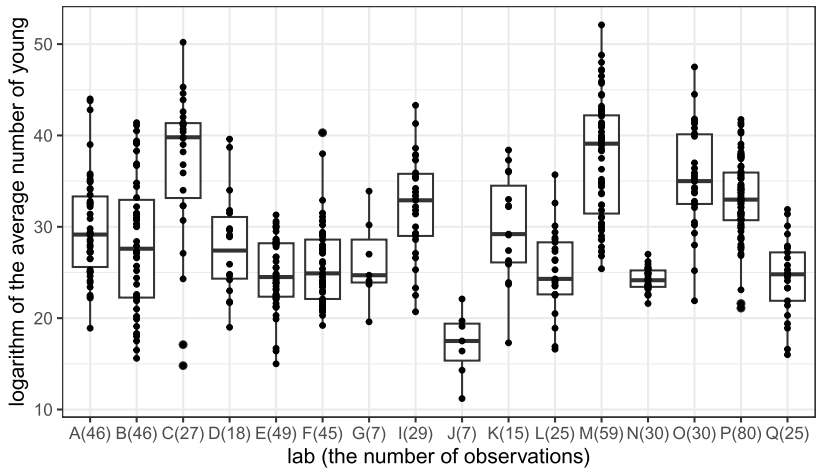



FIG. 1. *Boxplots of the logarithm of the average number of young in all* 16 *labs. The number of batches in each lab is in parentheses at x-axis. Dots represent the raw values of the observations.*

water used in the experiments, and the air temperature when the experiments were conducted. Based on this model, the density plot of raw conditional residuals for observations in each lab is shown in Figure 2, where these residuals are conditional on the random effects. It can be seen that some labs have higher consistency (lab $N$) in these residuals, while a few labs appear to have quite different spreads in the residuals. This indicates that the assumption of a constant variance of $\epsilon_{ih}$ in a regular linear mixed model is not reasonable, or equivalently, the reproduction outcomes from these labs are not of the same consistency. From the preliminary exploratory data analysis, we can see that lab $N$ has smaller variability than other labs. From the residuals plot in Figure 2, after fitting the linear mixed model, lab $B$ potentially has the largest variability.

A natural follow-up question is how these labs differ in terms of the consistency of these reproduction outcomes. Given that the control group outcomes are used as the baseline of toxicity assessment, labs that produce more consistent reproduction outcomes in their control group of the toxicology tests will estimate toxicity test endpoints with higher accuracy and precision. Thus, our statistical problem becomes finding clusters of labs based on the variability of random errors.

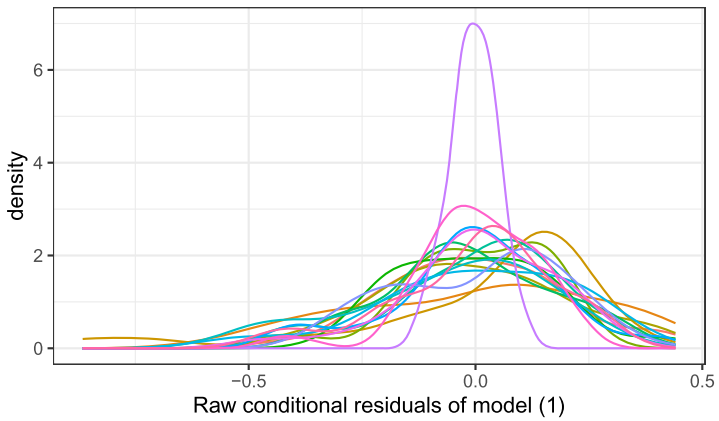



FIG. 2. *Density of raw conditional residuals based on the linear mixed model in* (1) *for each lab. Different lines represent different labs.*

1.2. *Literature review.* Clustering analysis is one of the most fundamental problems of understanding data sets (Jain (2010)); $k$-means and hierarchical clustering are the two most commonly used clustering methods. These two methods are model free, and they can only be used for clustering observations directly. If we want to find clusters of labs according to their variances in the test outcomes using $k$-means, we would need to have estimates of the lab-specific variances first. However, the accuracy of such variance estimates depends on the number of observations in these labs, and hence the accuracy varies as the number of tests available differs among labs, making the clustering results sensitive to the number of tests available from these labs. Also, it does not use any candidate covariates in the process and could not quantify the consistency of test outcomes after adjusting for the impact of the covariates on the responses. Thus, we are motivated to develop a new approach that can identify the clusters of observations based on their variances while simultaneously estimating regression coefficients and random effects. A popular model-based clustering method is the Gaussian finite mixture model (Fraley and Raftery (2002)). However, Gaussian finite mixture models often suffer from the convergence issue brought by outliers in the data (Archambeau, Lee and Verleysen (2003)). Moreover, when component likelihoods are misspecified, finite mixture models tend to overestimate the number of clusters and lead to unreliable conclusions (Cai, Campbell and Broderick (2021)). Frühwirth-Schnatter, Malsiner-Walli and Grün (2021) and Malsiner-Walli, Frühwirth-Schnatter and Grün (2016) proposed sparse finite mixture models using sparse priors for multivariate observations. However, these approaches assumed that different observations have the same dimension, which is different from our situation. Besides these, if the heterogeneity is due to the variability instead of the mean structure, including heterogeneity of the mean structure in the model will lead to bias. We applied the finite mixture model to our motivated data in Section 4, which shows that the heterogeneity of the variances is not identified properly.

Recently, there has been some work using optimization approaches and penalty functions to find clusters. In these approaches, penalty functions are imposed on the differences of regression coefficients, and optimization algorithms are developed to estimate parameters and find cluster structures simultaneously. For example, Ma and Huang (2017) used a concave fusion approach to find clustered intercepts in linear regression models. Wang, Zhu and Zhang (2023) considered finding clustered regression coefficients for spatial areal data with repeated measures based on pairwise penalties. These ideas were also applied to different nonlinear models (Hu et al. (2021), Miljkovic and Wang (2021), Wang, Zhang and Zhu (2023)). However, these models did not consider random effects. Wang and Zhu (2019) considered linear mixed models in small area estimation using pairwise penalties to find clusters of regression coefficients. Wang (2024) used linear mixed models and B-spline models to find clusters of functional data based on pairwise penalties. Zhou et al. (2022) used the approach based on pairwise penalties in linear mixed models to improve the initial values in their EM algorithm. All these models considered clustering means of different models rather than variances and cannot be applied directly to our motivating data set, which desires clustering outcomes according to variances.

In this work we propose a clustered variance model (CVM) with clusters of variances in linear mixed models. We construct an optimization problem based on the likelihood function and pairwise penalty functions on variances. An efficient algorithm is developed based on the alternating direction method of multipliers (ADMM) (Boyd et al. (2011)) to estimate regression coefficients, random effects, and the cluster structure of variances together. Simulation studies are conducted to compare our proposed approach with $k$-means based methods and finite mixture models. The results show that CVM recovers the cluster structure better than the traditional $k$-means based methods and finite mixture models.

This article is organized as follows. In Section 2 the proposed clustered variance model and the proposed algorithm are described in detail. The simulation study is conducted in Section 3

under several scenarios to compare the performances of our proposed method with competing approaches. We apply the proposed approach to the motivating dataset in Section 4. Finally, some discussions are given in Section 5.

## 2. Methodology.

2.1. *The model.* Recall the linear mixed model in (1). Instead of assuming a common variance for random error $\epsilon_{ih}$, we assume that the model has unit-specific random error variances $\sigma^2_{i,\epsilon}$, that is, $\epsilon_{ih} \overset{iid}{\sim} N(0, \sigma^2_{i,\epsilon})$ for $h = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Assume that there are $K$ heterogeneous variance groups, that is, the $n$ units can be partitioned into $K$ groups based on their values of variances. We denote the partition as $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$ such that $\sigma^2_{i,\epsilon} = \sigma^2_{j,\epsilon} = e^2_{k,\epsilon}$ if $i, j \in \mathcal{G}_k$, where $e^2_{k,\epsilon}$ is the random error variance of group $k$. However, neither the group structure $\mathcal{G}$ nor the number of groups $K$ is known. The goal is to cluster units (labs) based on their variances, while accounting for the intrabatch and interbatch uncertainty, and get the number of groups $\hat{K}$ and the estimates of variances based on observed data. To achieve the goal, we will construct an optimization problem based on log-likelihood function and pairwise functions. We consider the following log-likelihood of the linear mixed model in (1) with unit-specific random error variances,

$$l(\boldsymbol{\beta}, \sigma^2_v, \sigma^2_{1,\epsilon}, \ldots, \sigma^2_{n,\epsilon}) = \frac{1}{2}\sum_{i=1}^{n} \log|\Sigma_i| + \frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^T \Sigma_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}), \tag{2}$$

where $\boldsymbol{y}_i = (y_1, \ldots, y_{m_i})^T$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,m_i})^T$. The covariance matrix $\Sigma_i$ and its inverse, $\Sigma_i^{-1}$, have the following forms:

$$\Sigma_i = \mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\sigma^2_v + \boldsymbol{I}_{m_i}\sigma^2_{i,\epsilon},$$

$$\Sigma_i^{-1} = (\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\sigma^2_v + \boldsymbol{I}_{m_i}\sigma^2_{i,\epsilon})^{-1} = \frac{1}{\sigma^2_{i,\epsilon}}\left(\boldsymbol{I}_{m_i} - \mathbf{1}_{m_i}\mathbf{1}_{m_i}^T\frac{\sigma^2_v}{\sigma^2_{i,\epsilon} + m_i\sigma^2_v}\right).$$

In order to find group structure of $\sigma^2_{i,\epsilon}$, we impose pairwise penalties on the difference between $\sigma^2_{i,\epsilon}$ and $\sigma^2_{j,\epsilon}$. Similar approaches are used to find clusters of regression coefficients in the literature (Ma and Huang (2017), Wang, Zhu and Zhang (2023)). However, if the penalty functions are imposed to variances directly, extra constraints of the parameters are needed such that $\sigma^2_{i,\epsilon} > 0$. To get rid of the positive constraints of $\sigma^2_{i,\epsilon}$, we implement the following reparameterizations. Let $\tau_i = \log(\frac{\sigma^2_v}{\sigma^2_{i,\epsilon}})$ for $i = 1, 2, \ldots, n$; then we know that $(\sigma^2_v, \tau_1, \ldots, \tau_n)$ and $(\sigma^2_v, \sigma^2_{1,\epsilon}, \ldots, \sigma^2_{n,\epsilon})$ has a one-to-one mapping. This implies that estimating $(\sigma^2_v, \tau_1, \ldots, \tau_n)$ is equivalent to estimating $(\sigma^2_v, \sigma^2_{1,\epsilon}, \ldots, \sigma^2_{n,\epsilon})$. Then we can write the log-likelihood function in (2) as a function of $\boldsymbol{\beta}$, $\sigma^2_v$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^T$ as below:

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2_v, \boldsymbol{\tau}) = {} & \frac{1}{2}\sum_{i=1}^{n} m_i \log\sigma^2_v - \frac{1}{2}\sum_{i=1}^{n} m_i\tau_i + \frac{1}{2}\sum_{i=1}^{n}\log(1 + m_i\exp(\tau_i)) \\ & + \frac{1}{2}\sum_{i=1}^{n}\frac{\exp(\tau_i)}{\sigma^2_v}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^T(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \\ & - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma^2_v}\frac{\exp(2\tau_i)}{1 + m_i\exp(\tau_i)}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^T\mathbf{1}_{m_i}\mathbf{1}_{m_i}^T(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}). \end{aligned} \tag{3}$$

In (3), each unit has its own unit-specific variance, which is represented by $\tau_i$. To obtain the cluster structure, we apply a pairwise penalty on the difference between $\tau_i$ and $\tau_j$, and construct the following optimization problem: minimize the loglikelihood subject to a pairwise penalty,

$$Q(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}) = l(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}) + \sum_{1 \le i < j \le n} p(|\tau_i - \tau_j|; \gamma, \lambda), \tag{4}$$

where $p(\cdot; \gamma, \lambda)$ is a penalty function, $\lambda \ge 0$ is tuning parameter, and $\gamma > 0$ is a built-in constant in the penalty function. $L_1$ penalty (Tibshirani, Walther and Hastie (2001)), smoothly clipped absolute deviation penalty (SCAD) (Fan and Li (2012)), and the minimax concave penalty (MCP) (Zhang (2010)) are three popular used penalty functions. Ma and Huang (2017) explored the properties of these three penalties when using pairwise penalties for clustering intercepts in linear regression models. They showed that $L_1$ tended to result in more groups in the simulation, while SCAD and MCP performed similarly. In this work we use MCP. SCAD can be implemented in a similar way. $\lambda$ will be selected based on data-driven criteria, and $\gamma$ is fixed at 3, as in Ma and Huang (2017). The MCP is defined as follows:

$$p(t; \gamma, \lambda) = \begin{cases} \lambda |t| - \dfrac{t^2}{2\gamma} & |t| \le \gamma\lambda, \\ \dfrac{1}{2}\gamma\lambda^2 & |t| > \gamma\lambda. \end{cases}$$

As $\lambda$ increases, some pairs of $\tau_i - \tau_j$ will be shrunk to zeros, then the corresponding group structure will be found.

2.2. *The algorithm.* To solve the minimization problem based on the objective function in (4) to obtain estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\tau}}$ and $\hat{\sigma_v^2}$, an algorithm based on the ADMM algorithm is developed.

First, slack variables are introduced for all pairs, $\delta_{ij} = \tau_i - \tau_j$ for $1 \le 1 < j \le n$. Then the problem is equivalent to minimizing the following objective function with regard to $(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}, \boldsymbol{\delta})$, as in the ADMM algorithm:

$$L_0(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}, \boldsymbol{\delta}) = l(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}) + \sum_{1 \le i < j \le n} p(|\delta_{ij}|; \gamma, \lambda),$$

$$\text{subject to} \quad \tau_i - \tau_j - \delta_{ij} = 0,$$

where $\boldsymbol{\delta} = (\delta_{ij}, 1 \le i < j \le n)$. In the ADMM algorithm, the augmented Lagrangian is considered below:

$$L(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}, \boldsymbol{\delta}, \boldsymbol{v}) = L_0(\boldsymbol{\beta}, \sigma_v^2, \boldsymbol{\tau}, \boldsymbol{\delta}) + \sum_{i<j} v_{ij}(\tau_i - \tau_j - \delta_{ij}) + \frac{\rho}{2} \sum_{i<j} |\tau_i - \tau_j - \delta_{ij}|^2,$$

where $\boldsymbol{v} = (v_{ij}, 1 \le i < j \le n)$ are Lagrange multipliers and $\rho > 0$ is the penalty parameter. We fix it as $\rho = 1$ as in the references (Ma and Huang (2017), Wang, Zhu and Zhang (2023)).

Then parameters $\boldsymbol{\beta}$, $\sigma_v^2$, $\boldsymbol{\tau}$, $\boldsymbol{\delta}$, $\boldsymbol{v}$ will be updated iteratively. At the $(m+1)$th iteration, given their current values $(\boldsymbol{\beta}^{(m)}, \sigma_v^{2(m)}, \boldsymbol{\tau}^{(m)}, \boldsymbol{v}^{(m)})$, the updates of $\boldsymbol{\beta}$, $\sigma_v^2$, $\boldsymbol{\tau}$, $\boldsymbol{v}$ are given by as

follows:

$$
\begin{aligned}
\boldsymbol{\beta}^{(m+1)} &= \arg\min_{\boldsymbol{\beta}} L\big(\boldsymbol{\beta}, \sigma_v^{2(m)}, \boldsymbol{\tau}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{v}^{(m)}\big),\\
\sigma_v^{2(m+1)} &= \arg\min_{\sigma_v^2} L\big(\boldsymbol{\beta}^{(m+1)}, \sigma_v^2, \boldsymbol{\tau}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{v}^{(m)}\big),\\
\boldsymbol{\tau}^{(m+1)} &= \arg\min_{\boldsymbol{\tau}} L\big(\boldsymbol{\beta}^{(m+1)}, \sigma_v^{2(m+1)}, \boldsymbol{\tau}, \boldsymbol{\delta}^{(m)}, \boldsymbol{v}^{(m)}\big),\\
\boldsymbol{\delta}^{m+1} &= \arg\min_{\boldsymbol{\delta}} L\big(\boldsymbol{\beta}^{(m+1)}, \sigma_v^{2(m+1)}, \boldsymbol{\tau}^{(m+1)}, \boldsymbol{\delta}, \boldsymbol{v}^{(m)}\big),\\
v_{ij}^{(m+1)} &= v_{ij}^{(m)} + \rho\big(\tau_i^{(m+1)} - \tau_j^{(m+1)} - \delta_{ij}^{(m+1)}\big).
\end{aligned} \tag{5}
$$

The detailed algorithm in each update is provided in Supplement A in the Supplementary Material (Wang and Zhang (2024)). R code is provided in Supplementary Material (Wang and Zhang (2024)), which is also the available on Github https://github.com/wangx23/CluRER/.

REMARK 1. $\boldsymbol{\beta}^{(0)}$ are obtained by fitting a linear mixed model in (1). Based on the result of the linear mixed model, the conditional residuals can be calculated, that is, $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(0)} - \hat{v}_i$, where $\hat{v}_i$ is the best linear unbiased prediction (BLUP) of the random effect. Then we can calculate and $\bar{e}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} e_{ij}$. Based on $\boldsymbol{e}_i$, $\sigma_{i,\epsilon}^2$ can be initialized as $\sigma_{i,\epsilon}^{2(0)} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(e_{ij} - \bar{e}_i)^2$. $\sigma_v^{2(0)}$ can be initialized from the fitted model. Then $\boldsymbol{\tau}^{(0)} = \log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$, and $\delta_{ij}^{(0)} = \tau_i^{(0)} - \tau_j^{(0)}$. And $\boldsymbol{v}^{(0)} = \mathbf{0}$.

REMARK 2. The convergence criterion used is the same as Ma and Huang (2017), which is based on the primal residual $\boldsymbol{r}^{(m+1)} = \boldsymbol{D}\boldsymbol{\tau}^{(m+1)} - \boldsymbol{\delta}^{(m+1)}$. The algorithm is stopped if $|\boldsymbol{r}^{m+1}\| \le \epsilon$, where $\epsilon$ is a small positive value. Here we use $\epsilon = 0.001$.

We need to select the tuning parameter $\lambda$ in the proposed algorithm. In this paper we use the modified Bayes Information Criterion (BIC) (Ma and Huang (2017), Wang, Li and Tsai (2007)) to determine the best tuning parameter, which is used in other works using pairwise penalties, such as Ma and Huang (2017), Ma et al. (2020), and Wang, Zhu and Zhang (2023). In particular, we have

$$
\mathrm{BIC}_\lambda = -2l\big(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\boldsymbol{\tau}}\big) + \log\big(\log(n)\big)\log(N)\hat{K}, \tag{6}
$$

where $l(\cdot)$ is the log-likelihood function, $\hat{K}$ is the estimated number of groups, and $N = \sum_{i=1}^{n} m_i$. In real data analysis, we can also explore different group sizes, which is used in Miljkovic and Wang (2021).

**3. Simulation.** In this section we use several examples to evaluate the performance of the proposed CVM approach. In Section 3.1 and Section 3.2, we consider scenarios with multiple groups in the variances and compare it to $k$-means based methods and the finite mixture model. In Section 3.3 we consider a scenario without clustered variance structure to evaluate the performance of the proposed method when subgrouping does not exist. In Section 3.4 we discussed the initial values of the proposed algorithm.

$k$-means is the most widely used clustering approach to cluster observations directly. We will compare our proposed approach to $k$-means based approaches. Recall that we calculate initial values $\sigma_{i,\epsilon}^{2(0)}$ and $\sigma_v^{2(0)}$ and construct the initial values of $\tau_i^0$, as $\log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$. $\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)}$ represents the variabilities in different labs. Thus, we will use $\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)}$ as observations in $k$-means. Besides these, we also consider $\tau_i^0 = \log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$ as observations

in $k$-means since $\tau_i$'s are used as parameters to represent labs variabilities in our algorithm. In $k$-means, Gap statistic is widely used to select the number of clusters. Here we consider two Gaps statistics; one is proposed in Tibshirani, Walther and Hastie (2001), and the other one is proposed in Dudoit and Fridlyand (2021). Thus there are four $k$-means based approaches. In particular, "$k$-means$_{\text{Gap1}}$" and "$k$-means$_{\text{Gap2}}$" represent approaches using $\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)}$ and two Gap statistics, respectively. "$k$-means$^{\log}_{\text{Gap1}}$" and "$k$-means$^{\log}_{\text{Gap2}}$" represent approaches using $\tau_i^0 = \log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$ based on two Gap statistics, respectively. Besides these, we also consider the finite mixture model, denoted as "FMM." We use R package *flexmix* (Leisch (2004)) for fitting the finite mixture model, allowing differences both in the mean structure and the variances. BIC is used to select the number of clusters in FMM. And "CVM" represents our proposed method. We use modified BIC to select tuning parameters in (6).

To evaluate the clustering performance of CVM, we report the average estimated group number $\hat{K}$, average adjusted Rand index (ARI) (Rand (1971), Hubert and Arabie (1985), Vinh, Epps and Bailey (2010)) over 200 simulations. ARI is used to measure the degree of agreement between two partitions, with the largest value of 1. The larger the ARI value is, the more agreement between the two partitions. We also report the standard deviation values of different measures across 200 simulations in parentheses.

3.1. *Heterogeneous groups*. The data are simulated from model (1) in Section 2 with a clustered structure in $\sigma_{i,\epsilon}^2$. The covariates vector $\boldsymbol{x}_{ih}$ includes an intercept and two other components that are drawn from the standard normal distribution. The true values of parameters are set as follows: $\boldsymbol{\beta} = (3, 0.2, -0.2)$, and $\sigma_v = 0.18$; $m_i$'s are randomly drawn from (10, 20), (30, 40), (50, 60), (70, 80), and (90, 100). Motivated by the lab data, we will consider a cluster structure that includes a cluster with a single unit. Assume there are $K = 3$ true groups, $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$. $\mathcal{G}_1$ has one unit with $\sigma_{i,\epsilon} = 0.05$. For the other two groups, $\tau_i = -0.7$ if $i \in \mathcal{G}_2$ and $\tau_i = 0.5$ if $i \in \mathcal{G}_3$, which correspond to $\sigma_{i,\epsilon} = 0.2554$ and $\sigma_{i,\epsilon} = 0.1402$. Different units are randomly assigned to $\mathcal{G}_2$ and $\mathcal{G}_3$ with probabilities 4/5 and 1/5.

Table 1 and Table 2 show the results for ARI and the estimated number of groups $\hat{K}$ when $n = 15$, $n = 20$ and $n = 30$. We observe that, as local sample sizes increase, CVM can recover the group structure better. Regardless of the local sample sizes, the proposed method appears to be much more stable in terms of the classification agreement than all the $k$-means based approaches. The proposed method does overestimate the number of groups a little bit, regardless of the local sample size and the number of units in the data. It still performs better than $k$-means$_{\text{Gap1}}$ based on Gap statistics in Tibshirani, Walther and Hastie (2001), which overestimates by a lot more, or $k$-means$_{\text{Gap2}}$ based on the Gap statistics in Dudoit and Fridlyand (2021), which underestimates severely. It arguably outperforms $k$-means$_{\text{Gap1}}$, which has comparable performance when the local sample size is small while overestimating a lot more when the local sample size becomes 50 or higher. Even though the number of groups estimated by $k$-means$_{\text{Gap2}}$ was, on average, the closest to the true value, the ARI values suggest that this method does not achieve the same level of consistent grouping outcomes as our proposed method. Since our data are simulated from a model with common regression coefficients, FMM cannot identify the variance cluster structure well.

We also evaluate the estimation performance of $\sigma_{i,\epsilon}$ by calculating the root mean square error (RMSE), which is defined as $\text{RMSE} = \sqrt{\frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}(\hat{\sigma}_{i,\epsilon}^{(b)} - \sigma_{i,\epsilon})^2}$, where $n$ is the number of units (labs), $\hat{\sigma}_{i,\epsilon}^{(b)}$ is the estimate of $\sigma_{i,\epsilon}$ for the $i$th unit in the $b$th simulation. We compute the RMSE across 50 different group structures. The results are shown in Figure 3. It can be seen that as $m_i$ increases, we can have better estimates of $\sigma_{i,\epsilon}$.

TABLE 1

*Average ARI for different values of $m_i$ when $n = 15, 20, 30$ for different approaches. $m_i$'s are uniformly sampled from* (10, 20), (30, 40), (50, 60), (70, 80) *and* (90, 100)

| | Method | 10–20 | 30–40 | 50–60 | 70–80 | 90–100 |
|---|---|---|---|---|---|---|
| $n = 15$ | CVM | 0.60(0.28) | 0.82(0.22) | 0.91(0.17) | 0.92(0.17) | 0.95(0.14) |
| | $k$-means$_{\text{Gap1}}$ | 0.32(0.22) | 0.42(0.32) | 0.49(0.38) | 0.49(0.37) | 0.43(0.37) |
| | $k$-means$_{\text{Gap2}}$ | 0.49(0.32) | 0.77(0.30) | 0.84(0.28) | 0.85(0.28) | 0.86(0.27) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 0.13(0.18) | 0.26(0.25) | 0.33(0.27) | 0.40(0.29) | 0.41(0.31) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 0.01(0.10) | 0.04(0.19) | 0.07(0.25) | 0.12(0.31) | 0.19(0.38) |
| | FMM | 0.13(0.22) | 0.25(0.16) | 0.25(0.13) | 0.25(0.12) | 0.25(0.12) |
| $n = 20$ | CVM | 0.60(0.26) | 0.83(0.19) | 0.94(0.12) | 0.96(0.13) | 0.96(0.13) |
| | $k$-means$_{\text{Gap1}}$ | 0.32(0.20) | 0.45(0.31) | 0.51(0.34) | 0.53(0.37) | 0.56(0.37) |
| | $k$-means$_{\text{Gap2}}$ | 0.37(0.32) | 0.75(0.32) | 0.82(0.30) | 0.87(0.28) | 0.88(0.25) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 0.09(0.14) | 0.23(0.21) | 0.36(0.26) | 0.37(0.26) | 0.35(0.24) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 0.01(0.07) | 0.03(0.18) | 0.16(0.35) | 0.32(0.45) | 0.44(0.48) |
| | FMM | 0.11(0.18) | 0.20(0.12) | 0.21(0.11) | 0.21(0.11) | 0.22(0.10) |
| $n = 30$ | CVM | 0.61(0.23) | 0.84(0.16) | 0.96(0.07) | 0.97(0.09) | 0.99(0.05) |
| | $k$-means$_{\text{Gap1}}$ | 0.35(0.20) | 0.46(0.29) | 0.53(0.34) | 0.60(0.37) | 0.60(0.36) |
| | $k$-means$_{\text{Gap2}}$ | 0.30(0.32) | 0.57(0.42) | 0.69(0.42) | 0.78(0.37) | 0.78(0.38) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 0.05(0.14) | 0.19(0.24) | 0.36(0.28) | 0.37(0.28) | 0.37(0.25) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 0.02(0.13) | 0.09(0.28) | 0.30(0.45) | 0.54(0.48) | 0.70(0.44) |
| | FMM | 0.08(0.13) | 0.19(0.11) | 0.21(0.10) | 0.20(0.08) | 0.19(0.08) |

TABLE 2

*Average $\hat{K}$ for different values of $m_i$ when $n = 15, 20, 30$ for different approaches. $m_i$'s are uniformly sampled from* (10, 20), (30, 40), (50, 60), (70, 80) *and* (90, 100)

| | Method | 10–20 | 30–40 | 50–60 | 70–80 | 90–100 |
|---|---|---|---|---|---|---|
| $n = 15$ | CVM | 3.95(1.12) | 3.79(0.90) | 3.37(0.61) | 3.30(0.53) | 3.25(0.52) |
| | $k$-means$_{\text{Gap1}}$ | 5.88(2.02) | 5.92(2.24) | 5.69(2.38) | 5.61(2.39) | 6.23(2.59) |
| | $k$-means$_{\text{Gap2}}$ | 2.61(1.04) | 3.10(0.87) | 3.23(0.74) | 3.29(0.60) | 3.26(0.63) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 3.97(3.05) | 5.31(2.76) | 5.78(2.39) | 5.54(2.09) | 5.80(2.13) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 1.02(0.12) | 1.06(0.27) | 1.14(0.51) | 1.17(0.50) | 1.30(0.63) |
| | FMM | 2.67(0.72) | 4.17(0.80) | 4.84(0.91) | 5.16(0.82) | 5.28(0.81) |
| $n = 20$ | CVM | 3.83(1.12) | 4.06(1.04) | 3.53(0.79) | 3.31(0.61) | 3.31(0.61) |
| | $k$-means$_{\text{Gap1}}$ | 5.77(1.94) | 5.85(2.20) | 5.56(2.31) | 5.66(2.50) | 5.49(2.52) |
| | $k$-means$_{\text{Gap2}}$ | 2.46(1.26) | 3.13(0.85) | 3.22(0.67) | 3.25(0.62) | 3.22(0.53) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 3.65(3.25) | 5.17(2.88) | 5.43(2.17) | 5.79(2.10) | 6.08(2.05) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 1.01(0.10) | 1.04(0.22) | 1.22(0.53) | 1.46(0.72) | 1.65(0.78) |
| | FMM | 2.56(0.62) | 4.21(0.84) | 4.92(0.89) | 5.26(0.80) | 5.40(0.80) |
| $n = 30$ | CVM | 3.86(0.99) | 4.74(1.43) | 3.66(0.85) | 3.41(0.70) | 3.22(0.51) |
| | $k$-means$_{\text{Gap1}}$ | 5.91(1.89) | 5.83(2.06) | 5.67(2.35) | 5.38(2.56) | 5.22(2.43) |
| | $k$-means$_{\text{Gap2}}$ | 2.26(1.32) | 2.67(1.23) | 2.76(0.98) | 2.86(0.88) | 2.89(0.83) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 2.30(2.68) | 4.08(3.22) | 5.04(2.42) | 5.83(2.40) | 5.82(2.27) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 1.02(0.16) | 1.09(0.29) | 1.32(0.50) | 1.61(0.58) | 1.85(0.62) |
| | FMM | 2.83(0.64) | 4.40(0.80) | 5.26(0.90) | 5.66(0.86) | 5.82(0.79) |

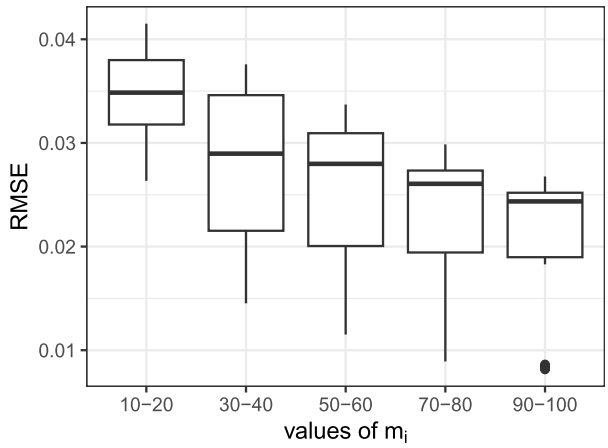



FIG. 3. *RMSE of $\hat{\sigma}_{i,\epsilon}$ for different values of $m_i$. $m_i$'s are uniformly generated from* (10, 20),(30, 40),(50, 60),(70, 80) *and* (90, 100).

3.2. *A different setup for $m_i$.* In this section, we consider a setup where $m_i$ has a larger range compared to those in Section 3.1. Table 3 shows the average ARI and average $\hat{K}$ when $m_i$'s are uniformly sampled from (10, 85), (20, 85), (30, 85) and (40, 85) across 200 simulations when $n = 15$. Among these cases, the case with (10, 85) is very close to the range of $m_i$ in the real data. We observe that as the lower bound of $m_i$ increases, we can expect larger ARI and $\hat{K}$ close to the true number of clusters 3. Among other approaches, $k$-means$_{\text{Gap2}}$ also gives reasonable results. FMM still cannot recover cluster structure since there is no mean cluster structure in the true model.

3.3. *Homogeneous groups.* In this section we consider a homogeneous case; that is, all units have the same random error variance. The overestimation in the number of groups in the previous section motivated us to consider this new scenario and investigate the performance of our model when there is no variance grouping structure in the data. A successful grouping method is expected to recover the fact that there is a single group when fitted to such a data set.

TABLE 3
*Average ARI and $\hat{K}$ for different values of $m_i$ when $n = 15$ for different approaches. $m_i$'s are uniformly sampled from* (10, 85), (20, 85), (30, 85) *and* (40, 85)

| | Method | 10–85 | 20–85 | 30–85 | 40–85 |
|---|---|---|---|---|---|
| ARI | CVM | 0.71(0.25) | 0.76(0.25) | 0.83(0.19) | 0.87(0.20) |
| | $k$-means$_{\text{Gap1}}$ | 0.42(0.32) | 0.42(0.33) | 0.45(0.35) | 0.45(0.36) |
| | $k$-means$_{\text{Gap2}}$ | 0.74(0.30) | 0.82(0.27) | 0.80(0.30) | 0.83(0.27) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 0.30(0.28) | 0.29(0.27) | 0.35(0.28) | 0.38(0.28) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 0.04(0.18) | 0.05(0.21) | 0.09(0.27) | 0.12(0.31) |
| | FMM | 0.21(0.15) | 0.25(0.15) | 0.24(0.14) | 0.26(0.14) |
| $\hat{K}$ | CVM | 4.32(1.18) | 4.21(1.16) | 3.97(0.97) | 3.76(0.91) |
| | $k$-means$_{\text{Gap1}}$ | 5.91(2.35) | 5.99(2.23) | 5.96(2.44) | 5.89(2.49) |
| | $k$-means$_{\text{Gap2}}$ | 3.32(0.83) | 3.29(0.70) | 3.24(0.76) | 3.28(0.58) |
| | $k$-means$^{\log}_{\text{Gap1}}$ | 4.91(2.51) | 5.44(2.72) | 5.78(2.36) | 5.35(2.13) |
| | $k$-means$^{\log}_{\text{Gap2}}$ | 1.06(0.31) | 1.06(0.28) | 1.12(0.40) | 1.19(0.54) |
| | FMM | 4.30(0.96) | 4.44(0.91) | 4.85(0.88) | 4.78(0.91) |

TABLE 4
*Summary of $\hat{K}$ results for different values of $m_i$ under the homogeneous setup. $m_i$'s are uniformly sampled from* $(10, 20)$, $(30, 40)$, $(50, 60)$, $(70, 80)$ *and* $(40, 85)$

| | | 10–20 | 30–40 | 50–60 | 70–80 | 90–100 |
|---|---|---|---|---|---|---|
| $n = 15$ | average | 1.36(0.59) | 1.54(0.76) | 1.44(0.72) | 1.39(0.65) | 1.34(0.58) |
| | per | 0.70 | 0.62 | 0.68 | 0.70 | 0.71 |
| $n = 20$ | average | 1.28(0.52) | 1.35(0.74) | 1.43(0.72) | 1.33(0.60) | 1.32(0.61) |
| | per | 0.76 | 0.77 | 0.69 | 0.73 | 0.75 |
| $n = 30$ | average | 1.18(0.41) | 1.10(0.36) | 1.23(0.64) | 1.28(0.64) | 1.26(0.64) |
| | per | 0.83 | 0.93 | 0.85 | 0.80 | 0.82 |

In this simulation we set $\sigma_{i,\epsilon} = 0.18$ for all $i = 1, \ldots, n$. Similar to the previous section, the average estimated number of groups ("average") and the percentage of correctly identifying a single group ("per") over 200 simulations are reported in Table 4, for $n = 15$, $n = 20$, and $n = 30$ under different values of local sample size $m_i$. It can be seen that, most of the time, by using the BIC defined in (6), we can identify the correct group structure.

3.4. *Discussions on initial values.* Initial values are important in the ADMM algorithm. In the work of using the algorithm based on pairwise penalties for clustering regression coefficients, most of them used fixed initial values, such as Ma et al. (2020), Zhu and Qu (2018), Lv et al. (2020), and Fang et al. (2022). In Wang (2024), they also had a discussion on the initial values. In the discussion they used different initial values and compared the results. Our algorithm focuses on estimating the heterogeneity of variances. In this section we conduct a similar study to evaluate the initial values setup.

For each simulated data set in Section 3.2 with $m_i$ uniformly generally from $(40, 85)$, 50 different initial values are generated by adding random noises to the proposed initial value $\boldsymbol{\tau}^{(0)}$. These random noises are drawn independently from a normal distribution with mean 0 and standard deviation 2. We calculate the differences of $\hat{l}$ values in (2), the differences of BIC values and ARI values between the results based on initial values generated from random noises and the results based on the original initial values. For each simulated data set, we calculate the minimum and maximum values of these differences across 50 different initial values. Out of the 100 different data sets, we observe that 99 of them have zero differences, which means that these initial values produce the same results as the results based on the original initial values. And only one has slightly different values. Table 5 shows the minimum and maximum of these differences for this particular data set. Among these 50 different initial values, 39 of them have the same results as the original initial values. Others have slightly worse ARI.

From the above results, we can see that the proposed algorithm is pretty robust to the initial values. And the proposed initial values in Remark 1 work well.

TABLE 5
*Differences of $\hat{l}$, ARI, and BIC compared to the original initial values*

| | $\hat{l}$ | ARI | BIC |
|---|---|---|---|
| minimum difference | −0.3367 | −0.0781 | 0 |
| maximum difference | 0 | 0 | 6.0776 |
| number of zero differences | 39 | 39 | 39 |

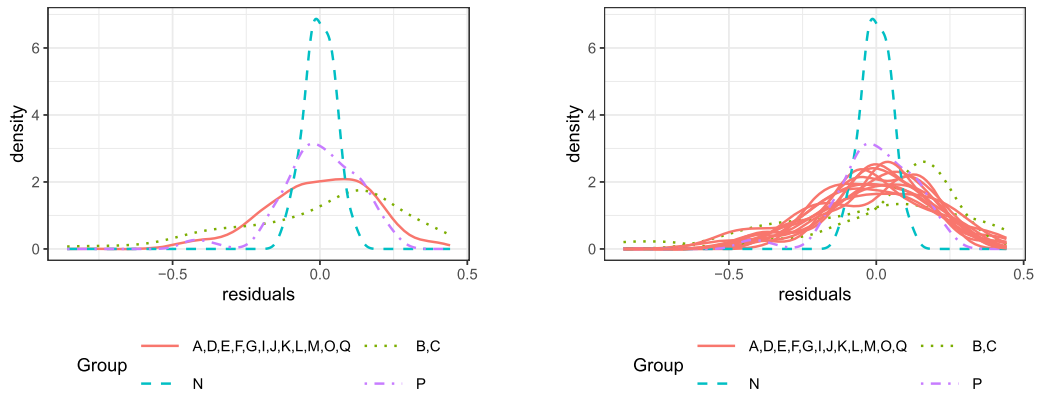
TABLE 6
*Estimated group structure and estimated standard deviation values in each estimated group*

| Group | Labs | $\hat{\sigma}_{k,\epsilon}$ |
|---|---|---|
| 1 | $A, D, E, F, G, I, J, K, L, M, O, Q$ | 0.1808 |
| 2 | $B, C$ | 0.2755 |
| 3 | $N$ | 0.0519 |
| 4 | $P$ | 0.1335 |

**4. Real data analysis.** In this section we apply our proposed method to the motivating aquatic toxicity test data in Section 4.1 and compare our results to the results based on $k$-means in Section 4.2 and the finite mixture model in Section 4.3. We consider the control data only. Recall that the response variable is the logarithm of the average number of young in each batch of the experiment, and the covariates are conductivity, dissolved oxygen level, pH of the water samples used, and the room air temperature where the experiment was conducted. In this data set, there are 16 labs, and the number of batches in each lab is between seven and 80.

4.1. *Results based on CVM.* The BIC values for two groups, three groups, four groups and five groups are $-1205.352$, $-1206.841$, $-1242.476$, and $-1225.143$. Based on BIC values, the selected number of groups is four. The estimated standard deviation of the lab effect is $\hat{\sigma}_v = 0.1921$. Table 6 shows the estimated group structure, corresponding labs, and the estimated standard deviation values. It can be seen that lab $N$ and lab $P$ have relatively smaller variances compared to other labs. And lab $B$ and lab $C$ have relatively large variances compared to the majority group.

Figure 4 shows the estimated density functions of raw conditional residuals for each group (left) and each lab (right), respectively. These raw conditional residuals are defined as the difference between the observed value $y_{ih}$ and the predicted value $\boldsymbol{x}_{ih}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$, where $\hat{v}_i$ is the BLUP for the random effects. These groups are associated with density functions of very different shapes and hence have different variability. The group consisting of lab $N$ only is associated with a density plot that is clearly of the highest consistency, suggesting the lowest variability of control outcomes from this lab. The density function of the group with lab $P$




(a) Estimated density functions of raw conditional residuals for four groups based on CVM.

(b) Estimated individual density functions of raw conditional residuals for four groups based on CVM.

FIG. 4. *Estimated density functions of raw conditional residuals for four groups based on CVM.*

TABLE 7
*Estimated group structure and estimated standard deviation values in each estimated group based on the data set without lab N*

| Group | Labs | $\hat{\sigma}_{k,\epsilon}$ |
|---|---|---|
| 1 | $A, D, E, F, G, I, J, K, L, M, O, Q$ | 0.1809 |
| 2 | $B, C$ | 0.2758 |
| 4 | $P$ | 0.1332 |

only appears to be more variable than the one of lab $N$ but more consistent than the other two groups. The density function of the group with labs $B$ and $C$ is left skewed compared to that of the biggest group with all the remaining labs $A, D, E, F, G, I, J, K, L, M, O, Q$.

From the preliminary data analysis and the results in Table 6 and Figure 4, we can see that lab $N$ has the smallest variability. We also explore the results when lab $N$ is removed from the original data set. When lab $N$ is removed, we implement our algorithm to other 15 labs. Table 7 shows the result. We have the same group structure for other 15 labs, which indicates that our proposed approach is also robust.

4.2. *Results based on k-means*. As a comparison, we also explore the results using $k$-means to find the clusters of variance based on $\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)}$ and $\log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$, as in the simulation study. Recall that, we use Gap statistics in $k$-means, there will be four different comparisons here. The approach based on $\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)}$, using the Gap statistics in Tibshirani, Walther and Hastie (2001) ("$k$-means$_{\text{Gap1}}$"), identifies two groups, and the approach based on $\log(\sigma_v^{2(0)}/\sigma_{i,\epsilon}^{2(0)})$, using the same Gap statistics ("$k$-means$_{\text{Gap1}}^{\log}$"), identifies one group. When the Gap statistics in Dudoit and Fridlyand (2021) are used in the two $k$-means methods ("$k$-means$_{\text{Gap2}}$" and "$k$-means$_{\text{Gap2}}^{\log}$"), the same group structure with six groups is identified.

More specifically, in "$k$-means$_{\text{Gap1}}$," lab $N$ is separated from the other groups, while it grouped lab $P$ with the remaining labs as a big group. As shown in Figure 4, based on the proposed approach, lab $P$ has a smaller variability than other labs in the other groups, and hence a homogeneous variance structure is not reasonable to describe the variability of this big group. Also, $k$-means$_{\text{Gap1}}^{\log}$ failed to find any groups in the data and ignored the big difference in variability among these labs. On the other hand, "$k$-means$_{\text{Gap2}}$" and "$k$-means$_{\text{Gap2}}^{\log}$" did identify the same six groups. Table 8 shows estimated standard deviation values using $k$-means when $\hat{K} = 6$. Figure 5 shows the estimated density functions of raw conditional residuals for each group. In the group structure with six groups, labs $D, J, K$ and labs $E, F, O$ are separated from the majority group. We can also observe that the shape of the density function of the group with labs $E, F, O$ is much similar to that with labs $A, G,$

TABLE 8
*Estimated group structure and estimated standard deviation values in each estimated group based on k-means*

| Group | Labs | $\hat{\sigma}_{k,\epsilon}$ |
|---|---|---|
| 1 | $A, G, I, L, M, Q$ | 0.1833 |
| 2 | $B, C$ | 0.2753 |
| 3 | $N$ | 0.0518 |
| 4 | $P$ | 0.1337 |
| 5 | $D, J, K$ | 0.2104 |
| 6 | $E, F, O$ | 0.1668 |

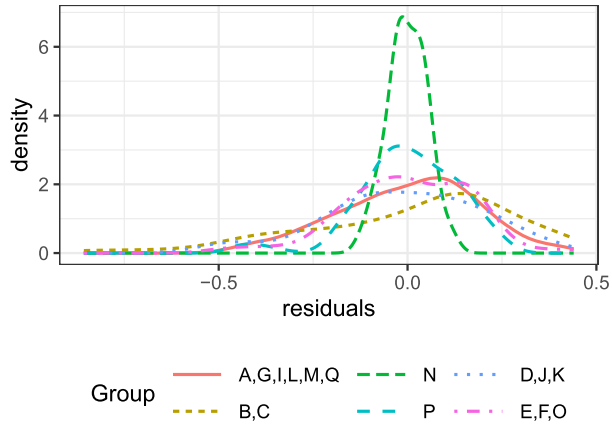



FIG. 5. *Density functions of raw conditional residuals based on k-means with six groups. Each line represents one group.*

$I$, $L$, $M$, $Q$ (Figure 5). Thus, it is more reasonable to merge labs $D$, $J$, $K$ and labs $E$, $F$, $O$ to the majority group, as in the grouping findings of our proposed method.

How stable are the grouping findings? As a followup comparison, we also explore the cluster structure obtained by using $k$-means$_{\text{Gap2}}$ and $k$-means$^{\log}_{\text{Gap2}}$ with a fixed number of groups at four. We also find that our proposed approach gives more reasonable results. The detailed results are provided in Supplement B in Wang and Zhang (2024).

4.3. *Results based on the finite mixture model.* In this section we apply the finite mixture model to the motivated data. BIC is used to select the number of groups. In Table 9 and Figure 6, we show the results based on the data set with lab $N$ and the data set without lab $N$. When lab $N$ is included, lab $N$ is grouped in the same group with other labs. However, if we check the residuals in Figure 6, we can see that lab $N$ has very different variability compared to the labs in the same group. If lab $N$ is not included in the analysis, we observe that the residuals of some labs are shifted away from 0, and some similar curves are not in the same group. This could be because FMM assumes that the heterogeneity of variances is along with the mean structure differences. However, the toxicity data we analyzed here may not have different effects of covaraites on the response. In this scenario, considering both mean structures and variance structures together, the heterogeneity of variances may not be identified properly.

**5. Conclusion and discussions.** In this work we consider a problem of clustering of variances and develop a new method for estimating parameters and identifying the heterogeneity of random error variances in linear mixed models. Our simulation studies have shown

TABLE 9
*Estimated group structure and estimated standard deviation values in each estimated group based on FMM*

| | With lab $N$ | | Without lab $N$ | |
|---|---|---|---|---|
| Group | Labs | $\hat{\sigma}_{k,\epsilon}$ | Labs | $\hat{\sigma}_{k,\epsilon}$ |
| 1 | $A, C, D, G, I, K, M, N, Q$ | 0.1890 | $A, B, D, E, F, G, J, K, L, Q$ | 0.2090 |
| 2 | $B, E, F, J, L$ | 0.2096 | $C, I, M, O, P$ | 0.1788 |
| 3 | $O, P$ | 0.1360 | | |

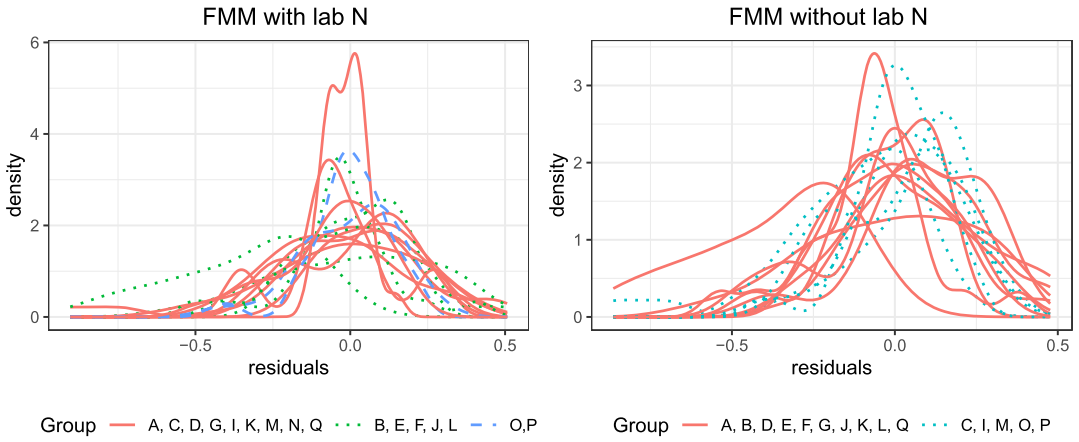



FIG. 6. *Density functions of raw conditional residuals based on the FMM model. Each line presents the density function of one lab. Different line types represent different groups. The left figure shows the density functions of residuals based on all labs. The right figure shows the density functions of residuals based on labs without lab N.*

that the proposed clustered variance approach could successfully detect groups of units according to their variability of observations while estimating the regression coefficients simultaneously. In the context of toxicity assessment, it allows us to find clusters of labs that are similar in the consistency of their control experiment outcomes as well as to identify the factors that impact the control experimental outcomes. In the application study, we compared the group structure found by the proposed method, $k$-means based methods and the finite mixture model according to the variability of the observations. The proposed methods identified more reasonable group structure compared to the $k$-means methods and the finite mixture model based approach.

There are a few possible extensions of our current study. The mean number of young in each batch of control test is used as the response in our proposed methods; it is possible to extend the methodology to group labs according to the original test results of individual organisms rather than aggregated batch/test level results. Based upon the individual-organism level clustering, it is also possible to extend the methodology to group labs according to the dose-response relationship when we include the tests using water with varying concentration levels of the reference toxicant in addition to the tests with control outcomes only. These questions would lead us to find the best practice in toxicity tests, but it is not a trivial application. Instead, it requires a novel statistical methodology due to the hierarchical structure of the data.

It is also possible to consider other experiment endpoints; for example, we could possibly cluster the labs according to both survival and reproduction test results. This leads to the application of our approaches in other types of toxicity tests, for example, acute fish tests, fish growth tests, etc. (Burden et al. (2017)).

Grouping the labs is the first step. The ultimate goal in the motivating data is to identify the combination of water chemistry, temperature, and other test condition variables that would help improve the consistency of reproduction outcomes in *C. dubia* tests and find an "optimal" combination of experimental conditions. This would require modeling both the mean and variability of the experimental outcomes as a function of these experimental conditions.

**Acknowledgments.** Many thanks to Darrin Greenstein, David Gillett, Alvine C. Mehinto, and Ken Schiff for providing the data and introducing of the background knowledge. We thank the Editor, the Associate Editor, and two anonymous reviewers for their helpful comments and suggestions that led to a substantially improved revision of the paper.

**Funding.** The work of Xin Wang is supported in part by the National Science Foundation grant NSF SES-2316353.

This project is supported by the Southern California Coastal Water Research Project.

## SUPPLEMENTARY MATERIAL

**Supplement** (DOI: 10.1214/24-AOAS1884SUPPA; .pdf). The detailed algorithm in each update and detailed results.

**R code** (DOI: 10.1214/24-AOAS1884SUPPB; .zip). Clustered variances in LMM.

## REFERENCES

AMATO, J. R., LUKASEWYCZ, M. T., ROBERT, E. D., MOUNT, D. I., DURHAN, E. J. and GERALD, T. A. (1993). An example of the identification of diazinon as a primary toxicant in an effluent. *Environ. Toxicol. Chem.* **11** 209–216.

ARCHAMBEAU, C., LEE, J. and VERLEYSEN, M. (2003). On convergence problems of the EM algorithm for finite Gaussian mixtures. In *European Symposium on Artificial Neural Networks* (*ESANN*'2003) 99–104, Bruges.

BAILER, A. J. and ORIS, J. T. (1993). Modeling reproductive toxicity in Ceriodaphnia tests. *Environ. Toxicol. Chem.* **12** 787–791.

BAILER, A. J. and ORIS, J. T. (1997). Estimating inhibition concentrations for different response scales using generalized linear models. *Environ. Toxicol. Chem.* **16** 1554–1559.

BAILEY, H. C., DIGIORGIO, C., KROLL, K., HINTON, D. E., MILLER, J. L. and STARRETT, G. (1996). Development of procedures for identifying pesticide toxicity in ambient waters: Carbofuran, diazinon and chlorpyrifos. *Environ. Toxicol. Chem.* **15** 837–845.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B., ECKSTEIN, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

BURDEN, N., GELLATLY, N., BENSTEAD, R., BENYON, K., BLICKLEY, T. M., CLOOK, M., DOYLE, I., EDWARDS, P., HANDLEY, J. et al. (2017). Reducing repetition of regulatory vertebrate ecotoxicology studies. *Integr. Environ. Assess. Manag.* **13** 955–957. https://doi.org/10.1002/ieam.1934

CAI, D., CAMPBELL, T. and BRODERICK, T. (2021). Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning* 1158–1169.

DOBSON, A. J. and BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*, 4th ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. For the third edition see [MR2459739]. MR3890007

DUDOIT, S. and FRIDLYAND, J. (2021). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **3** 1–21.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* **40** 2043–2068. MR3059076 https://doi.org/10.1214/12-AOS1028

FANG, K., CHEN, Y., MA, S. and ZHANG, Q. (2022). Biclustering analysis of functionals via penalized fusion. *J. Multivariate Anal.* **189** Paper No. 104874, 20. MR4384116 https://doi.org/10.1016/j.jmva.2021.104874

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

FRÜHWIRTH-SCHNATTER, S., MALSINER-WALLI, G. and GRÜN, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Anal.* **16** 1279–1307. MR4381135 https://doi.org/10.1214/21-BA1294

HU, X., HUANG, J., LIU, L., SUN, D. and ZHAO, X. (2021). Subgroup analysis in the heterogeneous Cox model. *Stat. Med.* **40** 739–757. MR4198442 https://doi.org/10.1002/sim.8800

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JAIN, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* **31** 651–666.

LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* **11** 1–18. https://doi.org/10.18637/jss.v011.i08

LV, Y., ZHU, X., ZHU, Z. and QU, A. (2020). Nonparametric cluster analysis on multiple outcomes of longitudinal data. *Statist. Sinica* **30** 1829–1856. MR4260746 https://doi.org/10.5705/ss.202018.0032

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MA, S., HUANG, J., ZHANG, Z. and LIU, M. (2020). Exploration of heterogeneous treatment effects via concave fusion. *Int. J. Biostat.* **16**.

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. MR3439375 https://doi.org/10.1007/s11222-014-9500-2

MILJKOVIC, T. and WANG, X. (2021). Identifying subgroups of age and cohort effects in obesity prevalence. *Biom. J.* **63** 168–186. MR4204907 https://doi.org/10.1002/bimj.201900287

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

STROUP, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Texts in Statistical Science Series*. CRC Press. MR2977489

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

VINH, N. X., EPPS, J. and BAILEY, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11** 2837–2854. MR2738784

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008 https://doi.org/10.1093/biomet/asm053

WANG, X. (2024). Clustering of longitudinal curves via a penalized method and EM algorithm. *Comput. Statist.* **39** 1485–1512. MR4730670

WANG, X. and ZHANG, J. (2024). Supplement to "Assessing aquatic toxicity assessment via a clustered variance model." https://doi.org/10.1214/24-AOAS1884SUPPA, https://doi.org/10.1214/24-AOAS1884SUPPB

WANG, X., ZHANG, X. and ZHU, Z. (2023). Clustered coefficient regression models for Poisson process with an application to seasonal warranty claim data. *Technometrics* **65** 514–523. MR4662685 https://doi.org/10.1080/00401706.2023.2190779

WANG, X. and ZHU, Z. (2019). Small area estimation with subgroup analysis. *Stat. Theory Relat. Fields* **3** 129–135. MR4028311 https://doi.org/10.1080/24754269.2019.1659097

WANG, X., ZHU, Z. and ZHANG, H. H. (2023). Spatial heterogeneity automatic detection and estimation. *Comput. Statist. Data Anal.* **180** Paper No. 107667, 23. MR4519305 https://doi.org/10.1016/j.csda.2022.107667

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, J., KONG, Y., BAILER, A. J., ZHU, Z. and SMUCKER, B. (2022). Incorporating historical data when determining sample size requirements for aquatic toxicity experiments. *J. Agric. Biol. Environ. Stat.* **27** 544–561. MR4459080 https://doi.org/10.1007/s13253-022-00496-0

ZHOU, L., SUN, S., FU, H. and SONG, P. X.-K. (2022). Subgroup-effects models for the analysis of personal treatment effects. *Ann. Appl. Stat.* **16** 80–103. MR4400504 https://doi.org/10.1214/21-aoas1503

ZHU, X. and QU, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electron. J. Stat.* **12** 171–193. MR3756096 https://doi.org/10.1214/17-EJS1389