Computational Statistics and Data Analysis 194 (2024) 107942

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sequential estimation for mixture of regression models for heterogeneous population

Na You [a,b], Hongsheng Dai [b,c,*], Xueqin Wang [d], Qingyun Yu [a]

[a] School of Mathematics, Sun Yat-sen University, Guangdong, 510275, China
[b] School of Mathematics, Statistics and Actuarial Science, University of Essex, Colchester, CO4 3SQ, UK
[c] School of Mathematics, Statistics and Physics, Newcastle University, NE1 7RU, UK
[d] School of Management, University of Science and Technology of China, Anhui, China

ARTICLE INFO

Keywords:
EM algorithm
Heterogeneous population
Mixture model
Sub-type

ABSTRACT

Heterogeneity among patients commonly exists in clinical studies and leads to challenges in medical research. It is widely accepted that there exist various sub-types in the population and they are distinct from each other. The approach of identifying the sub-types and thus tailoring disease prevention and treatment is known as precision medicine. The mixture model is a classical statistical model to cluster the heterogeneous population into homogeneous sub-populations. However, for the highly heterogeneous population with multiple components, its parameter estimation and clustering results may be ambiguous due to the dependence of the EM algorithm on the initial values. For sub-typing purposes, the finite mixture of regression models with concomitant variables is considered and a novel statistical method is proposed to identify the main components with large proportions in the mixture sequentially. Compared to existing typical statistical inferences, the new method not only requires no pre-specification on the number of components for model fitting, but also provides more reliable parameter estimation and clustering results. Simulation studies demonstrated the superiority of the proposed method. Real data analysis on the drug response prediction illustrated its reliability in the parameter estimation and capability to identify the important subgroup.

1. Motivation

The heterogeneity of patients in response to treatment is a common problem in practice, which challenges precision medicine and motivates more appropriate clinical strategies. For some tumor diseases, it is recognized that chemotherapy treatment may not be effective for every patient. However, since there are typically no standard criteria to clearly distinguish the patients who are not responsive to the treatment, these patients also have to suffer the chemotherapy pains and bear high medical costs even though they can not benefit from the treatment. The underlying reason that causes this fact is the complexity of tumor disease. In clinical practice, cancer patients are classified based on their tumor characteristics such as histology and morphology. However, more and more evidences indicate that one perceived disease may have various sub-types. The patients with different sub-types are distinct in etiology and pathogenesis, and therefore prognosis to systematic treatment (Curtis et al., 2012; Schlicker et al., 2012; Punt et al., 2017; Fan et al., 2014). Assigning patients into different sub-populations to apply corresponding effective treatments is precision

* Corresponding author at: School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
E-mail address: hongsheng.dai@newcastle.ac.uk (H. Dai).

https://doi.org/10.1016/j.csda.2024.107942
Received 13 November 2022; Received in revised form 15 February 2024; Accepted 16 February 2024

Available online 23 February 2024

medicine's main task. In order to stratify the cancer patients according to their pharmacologic response to the drug intervention, the Cancer Cell Line EncyclopFedia (CCLE) project (https://portals.broadinstitute.org/ccle) acquired the responses of 24 chemical compounds on 504 cancer cell lines, along with their genetic profiles, including DNA mutations, DNA methylations, RNAseq gene expressions, mRNA expressions, etc. The drug responses show high heterogeneity among cell lines. It is of great interest to identify the subgroups of cell lines within which they show similar responsive patterns to the drug intervention.

In classical statistics, the mixture model is commonly used to handle the data heterogeneity due to subgroups among the population (McLachlan and Peel, 2000; Fruhwirth-Schnatter, 2006). For the interesting outcome y and its related risk factors $\boldsymbol{x}$, the finite mixture of regression models (FMR) with K components is denoted by

$$f(y;\boldsymbol{x}) = \sum_{k=1}^{K} \rho_k g_k(y;\boldsymbol{x}), \tag{1}$$

where $g_k(y;\boldsymbol{x})$ in different components may belong to the same density family but with different parameters $\boldsymbol{\theta}_k$, and thereafter $g_k(y;\boldsymbol{x}) = g(y;\boldsymbol{\theta}_k,\boldsymbol{x})$. For instance, in the mixture of Gaussian regression models, $g(y;\boldsymbol{\theta}_k,\boldsymbol{x})$ is the normal density with mean $\boldsymbol{x}^T\boldsymbol{\beta}_k$ and standard deviation σ_k, where the superscript T indicates the transpose, and $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k)^T$. Although this type of mixture model is extensively studied (Grün and Leisch, 2008; Benaglia et al., 2009; Balakrishnan et al., 2017), the previous works usually put more effort into revealing the heterogeneous relationship between the explanatory variables $\boldsymbol{x}$ and the response y, but neglect the inference on partitioning criteria for subgroups. With the fitted model (1), if we want to stratify a new incoming patient into any subgroup, then we have to incorporate the response to calculate the posterior classification probability. However, this is not feasible for practical use since the response can not be known in advance in most cases. One of the research aims of the CCLE project is to classify a new cell line into a subgroup to predict how it would respond to different drugs and therefore help to choose the most appropriate treatment.

A more general type of FMR incorporates the patient's individual characterization as the concomitant variable $\boldsymbol{z}$ to model the mixing probabilities, that is

$$f(y;\boldsymbol{x},\boldsymbol{z}) = \sum_{k=1}^{K} \rho_k(\boldsymbol{z}) g(y;\boldsymbol{\theta}_k,\boldsymbol{x}), \tag{2}$$

where $\rho_k(\cdot)$ are unknown functions to be estimated (Grün and Leisch, 2008; Benaglia et al., 2009; Huang and Yao, 2012; Huang et al., 2013, 2018). The covariates in $\boldsymbol{z}$ may be the same, different, or overlapped with that in the explanatory variable $\boldsymbol{x}$. It is remarkable that model (2) is crucially important to classify the heterogeneous population into subgroups. With the aid of the concomitant variable such as the expression of a set of genes, one can establish sub-typing criteria to assign the new patient to the appropriate subgroup, and then predict the response using the regression model of the corresponding component. Without the subgroup identification, the prediction can not be properly derived, no matter how precisely the component models are described. This motivates us to work on the finite mixture regression model with concomitant variables (general finite mixture regression, gFMR), i.e. model (2), rather than (1) in this paper.

EM algorithm (Dempster et al., 1977) is widely employed to calculate the maximum likelihood estimate (MLE) of the parameter in the mixture model (McLachlan and Peel, 2000; Fruhwirth-Schnatter, 2006; Grün and Leisch, 2008; Huang et al., 2013, 2018). The main issue about the EM-type iterative algorithms is that their computation results may depend on the setting of initial values (Biernacki et al., 2003; Balakrishnan et al., 2017), due to the fact that the objective function to be maximized, i.e., the likelihood of the mixture model, may not be concave. It is usually suggested to repeat the EM algorithm from multiple initial values and take the maximum of the converged likelihoods as the global maximum. In order to answer how many different initial points should be tried to achieve the global maximum, Jin et al. (2016) found the probability that the EM algorithm starting from a random initial value converges to the global maximum decreases exponentially as K increases.

The number of components K in the mixture model affects the computation complexity to a great extent. In addition to the label-switching problem being commonly addressed in the literature (Papastamoulis, 2016), multiple components in the mixture model contribute to the non-concavity of the likelihood function with bad local maxima owing to the flexibility in the formulation of the FMR. Suppose that the data are from a mixture of three components $\rho_A g_A + \rho_B g_B + \rho_C g_C$. Even though we use the mixture model with the true value $K = 3$, say $\rho_1 g_1 + \rho_2 g_2 + \rho_3 g_3$, to fit the data, the parameter value corresponding to $g_1 = g_2 = g_A$, $\rho_1 + \rho_2 = \rho_A$ and $\rho_3 g_3 = \rho_B g_B + \rho_C g_C$ contributes a bad local maximum of the likelihood function. Furthermore, because of $g_1 = g_2$, the mixture $(\rho_1 - c)g_1 + (\rho_2 + c)g_2$ with any constant c such that both $\rho_1 - c$ and $\rho_2 + c$ fall in $[0, \rho_A]$ provides equal fitting to $\rho_1 g_1 + \rho_2 g_2$. These bad local maxima form a wave on the surface of the likelihood function. As K increases, the waves become much more intensive and challenge the iterative algorithm to achieve the global maxima.

Since the definitions of sub-types are unclear, K is usually unknown in reality. There have been many proposed methods for the determination of K in the literature, such as AIC, BIC, and the likelihood ratio test (McLachlan and Peel, 2000), extended K-means approach Pelleg and Moore (2000) and some Bayesian approaches (Stephens, 2000a; Miller and Harrison, 2018). They consider all possible values of K and choose the best one under certain criteria. Baudry and Celeux (2015) proposed a recursive algorithm to determine K and considered different approaches to split a mixture component into two components to avoid irrelevant parameter estimation. Similar approaches of splitting a component or merging mixture components were also used in the Bayesian framework via reversible-jump MCMC (Richardson and Green, 1997). However, the problematic model fitting results can mislead the choice of K. Ho et al. (2019) and Dwivedi et al. (2020) studied the singularity behaviours of EM algorithm with over-specified K. Moreover, as aforementioned, a reasonably large K setting can not avoid the model fitting from bad local maxima and presents

initial-dependent results. When the study population is highly heterogeneous, it is appealing to develop a statistical method for the parameter estimation which is less initial-dependent to ensure stable inference.

In recent years, although there are substantive progresses on the global performance and convergence rate of EM algorithm for solving the mixture models, the conclusions were mainly drawn on those with two components (Xu et al., 2016; Balakrishnan et al., 2017; Klusowski et al., 2019; Kwon et al., 2019; Ho et al., 2019; Dwivedi et al., 2020; Kwon et al., 2021). In this paper, for the gFMR (2) with large K, we propose a sequential procedure based on a mixture model with two components for the parameter estimation and component identification. The basic 2-component mixture model is composed of one interested parametric component, and the mixture of all other components modelled via a nonparametric component. Therefore, it can deal with the problems with $K > 2$, but still has the simplicity of a two-component mixture model. For a heterogeneous population, the proposed procedure runs sequentially. At each time, a set of subjects are classified into the interested component (the parametric component), and the rest are left for further partitioning (the non-parametric component). As more subjects are separated out, the algorithm stops until either no more interested component is identified or no sufficient samples remain. Therefore, we need not pre-specify K to fit the gFMR. Meanwhile, the 2-component mixture model used for each partitioning simplifies the gFMR at most efforts, providing a stabler parameter estimation than the directly fitting model (2). In summary, our method offers a novel alternative statistical strategy for the gFMR estimation to deal with the highly heterogeneous population. Unlike Baudry and Celeux (2015) and Richardson and Green (1997) which need to decide which component to be split, our recursive algorithm always splits the non-parametric component in every step into a combination of a Gaussian component and a non-parametric component. Therefore, our algorithm is simpler and bypasses the step to decide which component to split comparing to Baudry and Celeux (2015) and Richardson and Green (1997). Since our algorithm always split the non-parametric component, it means that any parametric components, once identified, will not change in later steps of iterations. Therefore, the samples identified from the non-parametric component will become smaller and smaller as the algorithm iterates. This will naturally provide a stopping criterion for our algorithm, e.g., it will stop when the non-parametric component does not have enough samples to be split further to identify another parametric component.

The remainder of this paper is structured as follows. In Section 2, we present the methodology of statistical inferences on the basic 2-component mixture model. In Section 3, we introduce the sequential analysis procedure for the gFMR with $K > 2$. The simulations are described in Section 4 and the analysis on a real data set is illustrated in Section 5. A short discussion is given in Section 6.

2. Two-component mixture model with concomitant variables

First, we consider a mixture regression model with two components. The response y is associated with the explanatory variables $\boldsymbol{x}$ and the concomitant variable $\boldsymbol{z}$ via

$$f(y;\boldsymbol{\Theta}, f_1, \boldsymbol{x}, \boldsymbol{z}) = \pi_0(\boldsymbol{\alpha};\boldsymbol{z}) f_0(y;\boldsymbol{\theta},\boldsymbol{x}) + \pi_1(\boldsymbol{\alpha};\boldsymbol{z}) f_1(y;\boldsymbol{x},\boldsymbol{z}), \tag{3}$$

where $f_0(y;\boldsymbol{\theta},\boldsymbol{x})$ is our main research interest and is in some parametric form with unknown parameter $\boldsymbol{\theta}$, and $f_1(y;\boldsymbol{x},\boldsymbol{z})$ is unspecified, representing the mixture of other components. For notation simplicity, we drop unknown component parameters from $f_1(y;\boldsymbol{x},\boldsymbol{z})$ since they are not of the main research interests in this section. The mixing probabilities such that $\pi_0(\boldsymbol{\alpha};\boldsymbol{z}) + \pi_1(\boldsymbol{\alpha};\boldsymbol{z}) = 1$ are determined by $\boldsymbol{z}$ via a function with unknown parameter $\boldsymbol{\alpha}$, for instance the logit function where $\pi_0(\boldsymbol{\alpha};\boldsymbol{z}) = \exp(\boldsymbol{z}^T\boldsymbol{\alpha})/\big(1+\exp(\boldsymbol{z}^T\boldsymbol{\alpha})\big)$, and $\boldsymbol{\Theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)^T$. Note that the gFMR (2) can be written in the formula of (3) by setting $\pi_0(\boldsymbol{\alpha};\boldsymbol{z}) = \rho_1(\boldsymbol{z})$, $f_0(y;\boldsymbol{\theta},\boldsymbol{x}) = g(y;\boldsymbol{\theta}_1,\boldsymbol{x})$, and $f_1(y;\boldsymbol{x},\boldsymbol{z}) = \sum_{k=2}^{K} \rho_k(\boldsymbol{z}) g(y;\boldsymbol{\theta}_k,\boldsymbol{x})/(1-\rho_1(\boldsymbol{z}))$.

Let $\mathcal{D} = \{\boldsymbol{y}, \mathcal{X}, \mathcal{Z}\}$ denote the observed data of n i.i.d. samples, with $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathcal{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ and $\mathcal{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$. The log-likelihood function of model (3) is

$$\mathcal{L}(\boldsymbol{\Theta};\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \log\Big\{\pi_0(\boldsymbol{\alpha};\boldsymbol{z}_i) f_0(y_i;\boldsymbol{\theta},\boldsymbol{x}_i) + \pi_1(\boldsymbol{\alpha};\boldsymbol{z}_i) f_1(y_i;\boldsymbol{x}_i,\boldsymbol{z}_i)\Big\}. \tag{4}$$

Since $f_1(y;\boldsymbol{x},\boldsymbol{z})$ is unknown, (4) is not ready for optimization. We consider formulating it as

$$\tilde{f}_1(y;\boldsymbol{\omega}) = \frac{1}{nh_n}\sum_{j=1}^{n} \omega_j \mathbb{K}\left(\frac{y_j - y}{h_n}\right), \tag{5}$$

where $\mathbb{K}(z)$ is a kernel function that satisfies the following Condition 2.1, h_n is the bandwidth, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ are nuisance parameters.

Condition 2.1. *The kernel function* $\mathbb{K}(z)$ *is such that* $\int \mathbb{K}(z)\mathrm{d}z = 1$, $\int \mathbb{K}(z) z \mathrm{d}z = 0$ *and* $\mu_{\mathbb{K}} = \int \mathbb{K}(z) z^2 \mathrm{d}z < \infty$.

Note that the value of ω_j in (5) depends on the data point $(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$, which should be denoted by $\omega_j(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$ exactly. However, in model (3), our main interest is to identify the parametric component $f_0(\cdot)$. It is not our target to describe how the response y depends on $\boldsymbol{x}$ in $f_1(\cdot)$, so we drop off this dependency in (5) and regard ω_j, $j = 1, \ldots, n$ as nuisance parameters to simplify the inference. In particular, (5) can be viewed as a weighted kernel density estimation of $f_1(\cdot)$, which was utilized in different manner from existing literature (Wang and Wang, 2007). Instead of using the random samples from f_1 to estimate the density function f_1, we are using the samples from the mixture f to estimate f_1, and we show that $\tilde{f}_1$ in (5) is asymptotically unbiased to f_1. Such large sample properties are provided in the Supplementary Material.

Replacing $f_1(y;\boldsymbol{x})$ in (4) by (5), it is natural to consider the following working log-likelihood function for the estimation of the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\omega}$, i.e.,

$$\widetilde{\mathcal{L}}(\boldsymbol{\Theta},\boldsymbol{\omega};\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\pi_0(\boldsymbol{\alpha};\boldsymbol{z}_i)f_0(y_i;\boldsymbol{\theta},\boldsymbol{x}_i) + \pi_1(\boldsymbol{\alpha};\boldsymbol{z}_i)\tilde{f}_1(y_i;\boldsymbol{\omega})\right\}. \tag{6}$$

Denoted $\boldsymbol{\Theta}^*$ as the truth of $\boldsymbol{\Theta}$ and f_1^* as the truth of f_1, we have the following theorem under mild conditions.

Condition 2.2.

1. *Functions $\pi_0(\boldsymbol{\alpha};\boldsymbol{z}), \pi_1(\boldsymbol{\alpha};\boldsymbol{z})$ and $f_0(y;\boldsymbol{\theta},\boldsymbol{x})$ are continuous and have second derivatives with respect to $\boldsymbol{\theta},\boldsymbol{\alpha}$.*
2. *There exists a constant $\delta \in (0,1)$ such that*

$$\sup_{\boldsymbol{\alpha},\boldsymbol{z}} \frac{\pi_1(\boldsymbol{\alpha};\boldsymbol{z})}{\pi_0(\boldsymbol{\alpha};\boldsymbol{z})} \in (\delta,\delta^{-1}).$$

3. *The second derivative of $f_0(y;\boldsymbol{\theta}^*,\boldsymbol{x})$ and $f_1^*(y;\boldsymbol{x},\boldsymbol{z})$ with respect to y exists such that for some $M>0$*

$$\sup_{y,\boldsymbol{x}} \frac{\mathrm{d}^2}{\mathrm{d}y^2} f_1^*(y;\boldsymbol{x},\boldsymbol{z}) \le M, \quad \sup_{y,\boldsymbol{x}} \frac{\mathrm{d}^2}{\mathrm{d}y^2} f_0^*(y;\boldsymbol{\theta}^*,\boldsymbol{x}) \le M.$$

$\int y^2 f_0(y;\boldsymbol{\theta}^*,\boldsymbol{x})\mathrm{d}y < \infty$, $\int y^2 f_1^*(y;\boldsymbol{x},\boldsymbol{z})\mathrm{d}y < \infty$, *and* $\sup_{y,\boldsymbol{\theta},\boldsymbol{x}} f_0(y;\boldsymbol{\theta},\boldsymbol{x}) < \infty$, $\sup_{y,\boldsymbol{x}} f_1^*(y;\boldsymbol{x},\boldsymbol{z}) < \infty$.

Theorem 2.1. *If Condition 2.1 holds and $h_n \to 0, nh_n \to \infty$ as $n \to \infty$, under mild Conditions 2.2 and A.1 in the Supplementary Material, there exists $\widehat{\boldsymbol{\omega}}$ such that the maximum $\widehat{\boldsymbol{\Theta}}$ of $\widetilde{\mathcal{L}}(\boldsymbol{\Theta},\widehat{\boldsymbol{\omega}};\mathcal{D})$ in (6) converges to $\boldsymbol{\Theta}^*$ in probability.*

Remark 2.1. Note that in a neighbourhood $\mathcal{A}$, the identifiability of $\widehat{\boldsymbol{\Theta}}$ is guaranteed by the conditions stated in the above theorem. Condition 2.2 says the mixing probability $\pi_0(\boldsymbol{\alpha};\boldsymbol{z})$ is bounded up and below for all $\boldsymbol{z}$ and $\boldsymbol{\Theta} \in \mathcal{A}$, which means we can only identify the component that is large enough, and the component with very small weight can not be identified. The concavity in Condition A.1 guarantees the uniqueness of the parameter estimate.

Theorem 2.2. *Under the same conditions as Theorem 2.1 and $\sqrt{n}h_n^2 \to 0$ as $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, with the detailed formula and its estimator provided in the Supplementary Material.*

In order to obtain the maximum estimate $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\omega}}$, we implement an EM iterative algorithm. Let $\boldsymbol{\Delta} = (\Delta_1,\ldots,\Delta_n)^T$, where $\Delta_i = 0$ or 1 indicates whether the data point $(\boldsymbol{x}_i,\boldsymbol{z}_i,y_i)$ belongs to the parametric component $f_0(\cdot)$ or the nonparametric component $f_1(\cdot)$, and $P(\Delta_i = k|\boldsymbol{z}_i) = \pi_k(\boldsymbol{\alpha};\boldsymbol{z}_i)$, $k=0,1$. The complete working log-likelihood of $(\mathcal{D},\boldsymbol{\Delta})$ is

$$\begin{aligned}\widetilde{\mathcal{L}}^c(\boldsymbol{\Theta},\boldsymbol{\omega};\mathcal{D},\boldsymbol{\Delta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{I}(\Delta_i=0)\log\pi_0(\boldsymbol{\alpha};\boldsymbol{z}_i) + \mathbb{I}(\Delta_i=1)\log\pi_1(\boldsymbol{\alpha};\boldsymbol{z}_i)\right\} \\ + \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{I}(\Delta_i=0)\log f_0(y_i;\boldsymbol{\theta},\boldsymbol{x}_i) + \mathbb{I}(\Delta_i=1)\log\tilde{f}_1(y_i;\boldsymbol{\omega})\right\}.\end{aligned} \tag{7}$$

Given the parameter estimates $\boldsymbol{\omega}^{(m)}$, $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\alpha}^{(m)}$ from the mth iteration, the E-step calculates the posterior probability that each sample belongs to the parametric component $f_0(\cdot)$,

$$u_{0,i}^{(m+1)} = \frac{\pi_0(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_i)f_0(y_i;\boldsymbol{\theta}^{(m)},\boldsymbol{x}_i)}{\pi_0(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_i)f_0(y_i;\boldsymbol{\theta}^{(m)},\boldsymbol{x}_i) + \pi_1(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_i)\tilde{f}_1(y_i;\boldsymbol{\omega}^{(m)})}. \tag{8}$$

Substituting $\mathbb{I}(\Delta_i=0)$ in (7) by $u_{0,i}^{(m+1)}$ and $\mathbb{I}(\Delta_i=1)$ by $1-u_{0,i}^{(m+1)}$, the maximization of the expected complete working log-likelihood in the M-step can be solved separately, i.e.,

$$\omega_j^{(m+1)} = n\cdot\frac{\left(1-u_{0,j}^{(m+1)}\right)\pi_1(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_j)^{-1}}{\sum_{i=1}^{n}\left(1-u_{0,i}^{(m+1)}\right)\pi_1(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_i)^{-1}}, \tag{9}$$

$$\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^{n}\left\{u_{0,i}^{(m+1)}\log f_0(y_i;\boldsymbol{\theta},\boldsymbol{x}_i)\right\}, \tag{10}$$

$$\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}}\frac{1}{n}\sum_{i=1}^{n}\left\{u_{0,i}^{(m+1)}\log\pi_0(\boldsymbol{\alpha};\boldsymbol{z}_i) + \left(1-u_{0,i}^{(m+1)}\right)\log\pi_1(\boldsymbol{\alpha};\boldsymbol{z}_i)\right\}. \tag{11}$$

The updating equation (9) is derived based on (8) and equation (A13) in the Supplementary Material, from which we know $\omega_j^{(m+1)} \propto \left(1-u_{0,j}^{(m+1)}\right)\pi_1(\boldsymbol{\alpha}^{(m)};\boldsymbol{z}_j)^{-1}$ and (9) is the normalised version.

Given initial assignments among two groups, the above E-step and M-step run iteratively until convergence. The whole procedure is summarized in Algorithm 1 as follows. To tackle bad local maxima, in practice we may rerun Algorithm 1 multiple times from

Algorithm 1: The algorithm for $K=2$.

Result: the converged parameter estimates $\widehat{\boldsymbol{\Theta}}$;
Randomly divide the samples into two groups by assigning $u_{0,i}^{(0)} \sim Binom(1,0.5)$, and calculate $\boldsymbol{\omega}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\alpha}^{(0)}$ according to equation (9), (10) and (11) respectively;
Compute the observed working log-likelihood (6);
repeat
  E-step: Update the posterior probabilities $u_{0,i}^{(m+1)}$ by equation (8);
  M-step: renew the parameter estimates according to (9), (10) and (11);
  Compute the observed working log-likelihood (6);
until *the observed working likelihood converges*;

different initial values and take the parameter estimates corresponding to the largest likelihood as $\widehat{\boldsymbol{\Theta}}$. It is worth noting that, since there are only two components being considered in the mixture model (3), the probability that our algorithm is trapped by the bad local maxima is greatly decreased. Meanwhile, the inclusion of the nonparametric component aids its applicability in dealing with heterogeneous population.

3. Sequential partitioning for gFMR with $K > 2$

In real applications, there usually are multiple interested components in the gFMR (2), and even the number of components K is unknown. In order to avoid the maximization of the likelihood function with plenty of bad local maxima and pre-specification of K, we propose a sequential EM (seqEM) procedure based on Algorithm 1 to identify the components $g(y;\boldsymbol{\theta}_k,\boldsymbol{x})$ and its associated mixing probabilities $\rho_k(\boldsymbol{z})$ one by one.

First, the gFMR (2) can be written in the formula of the basic 2-component mixture model,

$$f(y;\boldsymbol{x},\boldsymbol{z}) = \pi_{0;1}(\boldsymbol{\alpha}_1;\boldsymbol{z})g(y;\boldsymbol{\theta}_1,\boldsymbol{x}) + \pi_{1;1}(\boldsymbol{\alpha}_1;\boldsymbol{z})f_{1;1}(y;\boldsymbol{x},\boldsymbol{z}),$$

where $f_{1;1}(y;\boldsymbol{x},\boldsymbol{z}) = \sum_{k=2}^{K}\rho_k(\boldsymbol{z})g(y;\boldsymbol{\theta}_k,\boldsymbol{x})/(1-\rho_1(\boldsymbol{z}))$, $\pi_{0;1}(\boldsymbol{\alpha}_1;\boldsymbol{z}) = \rho_1(\boldsymbol{z})$, and $\pi_{1;1}(\boldsymbol{\alpha}_1;\boldsymbol{z}) = 1-\pi_{0;1}(\boldsymbol{\alpha}_1;\boldsymbol{z})$. Note that $f_{1;1}(y;\boldsymbol{x},\boldsymbol{z})$ can be further formulated as a 2-component mixture model, and this separation can be done recursively, i.e.,

$$f_{1;\tau-1}(y;\boldsymbol{x},\boldsymbol{z}) = \pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})g(y;\boldsymbol{\theta}_\tau,\boldsymbol{x}) + \pi_{1;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})f_{1;\tau}(y;\boldsymbol{x},\boldsymbol{z}), \quad \tau = 1,2,\ldots,K, \tag{12}$$

where $\pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z}) = \rho_\tau(\boldsymbol{z})/\left(1-\sum_{k<\tau}\rho_k(\boldsymbol{z})\right)$, $\pi_{1;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z}) = 1-\pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})$ and $f_{1;\tau}(y;\boldsymbol{x},\boldsymbol{z}) = \sum_{k>\tau}\rho_k(\boldsymbol{z})g(y;\boldsymbol{\theta}_k,\boldsymbol{x})/\left(1-\sum_{k\le\tau}\rho_k(\boldsymbol{z})\right)$ by defining $f_{1;0}(y;\boldsymbol{x},\boldsymbol{z}) = f(y;\boldsymbol{x},\boldsymbol{z})$ and $f_{1;K}(y;\boldsymbol{x},\boldsymbol{z}) = 0$. Then the log-likelihood of the gFMR (2) can be expressed as

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\sum_{k<\tau}\rho_k(\boldsymbol{z}_i)g(y_i;\boldsymbol{\theta}_k,\boldsymbol{x}_i) + \left(1-\sum_{k<\tau}\rho_k(\boldsymbol{z}_i)\right)f_{1;\tau-1}(y;\boldsymbol{x}_i,\boldsymbol{z}_i)\right\} \tag{13}$$

for any $\tau = 1,\ldots,K$. Suppose there are $\tau-1$ components being identified, say the first $\tau-1$ ones without loss of generality, with their parameter estimates $\widehat{\boldsymbol{\theta}}_k$ and the corresponding mixing probabilities $\widehat{\rho}_k(\boldsymbol{z})$, $k=1,2,\cdots,\tau-1$. Reformulating $f_{1;\tau}(\cdot)$ in (12) by (5) and then substituting it into (13) yields the working log-likelihood to identify the τth component,

$$\widetilde{\mathcal{L}}_{\mathcal{G}}(\boldsymbol{\Theta}_\tau,\boldsymbol{\omega}_\tau;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\sum_{k<\tau}\widehat{\rho}_k(\boldsymbol{z}_i)g(y_i;\widehat{\boldsymbol{\theta}}_k,\boldsymbol{x}_i) + \left(1-\sum_{k<\tau}\widehat{\rho}_k(\boldsymbol{z}_i)\right)\left[\pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})g(y;\boldsymbol{\theta}_\tau,\boldsymbol{x}) + \pi_{1;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})\tilde{f}_{1;\tau}(y;\boldsymbol{\omega}_\tau)\right]\right\}, \tag{14}$$

where $\pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z}) = \rho_\tau(\boldsymbol{z})/\left(1-\sum_{k<\tau}\widehat{\rho}_k(\boldsymbol{z})\right)$. We still use the notation $\pi_{0;\tau}(\boldsymbol{\alpha}_\tau;\boldsymbol{z})$ as that in (12) without cause of any confusion. The parameter estimate $\widehat{\boldsymbol{\Theta}}_\tau$ for $\boldsymbol{\Theta}_\tau$ is obtained by maximizing (14). As τ goes from 1 to K, a series of components $g(y;\widehat{\boldsymbol{\theta}}_\tau,\boldsymbol{x})$ and their associated mixing probabilities $\widehat{\rho}_\tau(\boldsymbol{z}) = \left(1-\sum_{k<\tau}\widehat{\rho}_k(\boldsymbol{z})\right)\pi_{0;\tau}(\widehat{\boldsymbol{\alpha}}_\tau;\boldsymbol{z})$ can be identified sequentially. Denoted by $\boldsymbol{\Theta}^* = (\boldsymbol{\Theta}_1^*,\cdots,\boldsymbol{\Theta}_K^*)$ the truth of $\boldsymbol{\Theta}$ in (12), we can show that the parameter estimates $\widehat{\boldsymbol{\Theta}}_\tau$, $\tau = 1,2,..K$, enjoy the following statistical properties.

Corollary 3.1. *Suppose we already have the consistent parameter estimates for the first $\tau-1$ components, $\widehat{\boldsymbol{\Theta}}_k, k=1,\cdots,\tau-1$. If Condition 2.1 holds and $h_n \to 0, nh_n \to \infty$ as $n \to \infty$, then under mild Conditions C.1 and C.2 in the Supplementary Material, there exists $\widehat{\boldsymbol{\omega}}_\tau$ such that the maximum $\widehat{\boldsymbol{\Theta}}_\tau$ of the working log-likelihood function $\widetilde{\mathcal{L}}_{\mathcal{G}}(\boldsymbol{\Theta}_\tau,\widehat{\boldsymbol{\omega}}_\tau;\mathcal{D})$ in (14) converges to $\boldsymbol{\Theta}_\tau^*$ in probability.*

Remark 3.1. The identification of the τth component can be viewed as the separation of the parametric part from a basic 2-component model, therefore the uniqueness of the MLE is guaranteed similarly as in Section 2. However, when there are multiple

components to be separated, the order of $\widehat{\boldsymbol{\Theta}}_1, \cdots, \widehat{\boldsymbol{\Theta}}_K$ may alter due to possible label-switching, thus we only consider $\widehat{\boldsymbol{\Theta}}_\tau$ in the neighbourhood $\mathcal{A}_\tau$ of $\boldsymbol{\Theta}_\tau^*$ as Condition C.2 states in Corollary 3.1. In real applications, the initial value of the algorithm only affects the order of the components to be identified, but not their parameter estimation.

Corollary 3.2. *Under the same conditions as Corollary 3.1 and* $\sqrt{n}h_n^2 \to 0$ *as* $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\Theta}}_\tau - \boldsymbol{\Theta}_\tau^*)$ *is asymptotically normal with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}_\tau$*, with the detailed formula being given in the Supplementary Material.*

In order to maximize (14) to get $\widehat{\boldsymbol{\Theta}}_\tau$, we utilize the EM iterations. Let $\Delta_i = k$ indicate that the ith data point $(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i)$ belongs to the kth component, $k = 1, 2, \ldots, \tau$, and $\Delta_i = \tau + 1$ to stand for that it belongs to one of the latter $(K - \tau)$ components. The complete working likelihood is

$$\begin{aligned}\widetilde{\mathcal{L}}_{\mathcal{G}}^c(\boldsymbol{\Theta}_\tau, \boldsymbol{\omega}_\tau; \mathcal{D}, \boldsymbol{\Delta}) = \prod_{i=1}^n \Bigg\{ \prod_{k<\tau} \Big(\widehat{\rho}_k(\boldsymbol{z}_i) g(y_i; \widehat{\boldsymbol{\theta}}_k, \boldsymbol{x}_i)\Big)^{\mathbb{I}(\Delta_i = k)} \Bigg[\Big(1 - \sum_{k<\tau} \widehat{\rho}_k(\boldsymbol{z}_i)\Big) \\ \cdot \Big(\pi_{0;\tau}(\boldsymbol{\alpha}_\tau; \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}_\tau, \boldsymbol{x}_i) + \pi_{1;\tau}(\boldsymbol{\alpha}_\tau; \boldsymbol{z}_i) \tilde{f}_{1;\tau}(y_i; \boldsymbol{\omega}_\tau)\Big)\Bigg]^{\mathbb{I}(\Delta_i \geq \tau)} \Bigg\} \\ \propto \prod_{i=1}^n \Big[\Big(\pi_{0;\tau}(\boldsymbol{\alpha}_\tau; \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}_\tau, \boldsymbol{x}_i) + \pi_{1,\tau}(\boldsymbol{\alpha}_\tau, \boldsymbol{z}_i) \tilde{f}_{1,\tau}(y_i; \boldsymbol{\omega}_\tau)\Big)\Big]^{\mathbb{I}(\Delta_i \geq \tau)},\end{aligned}$$

given the parameter estimates for the first $\tau - 1$ components, $\widehat{\boldsymbol{\Theta}}_k$, $k = 1, \ldots, \tau - 1$. Let

$$\begin{aligned}\varpi_{i;\tau-1} &= P\big(\Delta_i \geq \tau | \mathcal{D}, \widehat{\boldsymbol{\Theta}}_k, \widehat{\boldsymbol{\omega}}_k; k = 1, \ldots, \tau - 1\big) = \prod_{k=1}^{\tau-1} P\Big(\Delta_i \geq k+1 | \Delta_i \geq k, \mathcal{D}, \widehat{\rho}_k, \widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\omega}}_k\Big) \\ &= \prod_{k=1}^{\tau-1} \frac{\pi_{1;k}(\widehat{\boldsymbol{\alpha}}_k; \boldsymbol{z}_i) \tilde{f}_{1;k}(y_i; \widehat{\boldsymbol{\omega}}_k)}{\pi_{0;k}(\widehat{\boldsymbol{\alpha}}_k; \boldsymbol{z}_i) g(y_i; \widehat{\boldsymbol{\theta}}_k, \boldsymbol{x}_i) + \pi_{1;k}(\widehat{\boldsymbol{\alpha}}_k; \boldsymbol{z}_i) \tilde{f}_{1;k}(y_i; \widehat{\boldsymbol{\omega}}_k)}. \end{aligned} \tag{15}$$

The conditional expectation of $\log \widetilde{\mathcal{L}}_{\mathcal{G}}^c(\boldsymbol{\Theta}_\tau, \boldsymbol{\omega}_\tau; \mathcal{D}, \boldsymbol{\Delta})$, denoted by

$$Q(\boldsymbol{\Theta}_\tau, \boldsymbol{\omega}_\tau; \mathcal{D}, \boldsymbol{\Delta}) = \sum_{i=1}^n \varpi_{i;\tau-1} \log \Big[\Big(\pi_{0;\tau}(\boldsymbol{\alpha}_\tau, \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}_\tau, \boldsymbol{x}_i) + \pi_{1;\tau}(\boldsymbol{\alpha}_\tau, \boldsymbol{z}_i) \tilde{f}_{1,\tau}(y_i; \boldsymbol{\omega}_\tau)\Big)\Big],$$

is the weighted log-likelihood of the basic 2-component mixture model. Therefore, the parameters $\boldsymbol{\Theta}_\tau$ and $\boldsymbol{\omega}_\tau$ can be estimated using Algorithm 1 by assigning the weight $\varpi_{i;\tau-1}$ on the ith data point, $i = 1, 2, \ldots, n$.

The sequential procedure is concluded as Algorithm 2. It runs until $\sum_{i=1}^n \varpi_{i;\tau-1} \leq \kappa$, where κ is a tuning parameter to threshold the effective sample size for the τth partitioning. We set $\kappa = 10$ as default, i.e. there are at least 10 observations being left to fit the 2-component mixture model for the identification of more components. This prevents the model from being over-fitted.

Remark 3.2. In the above algorithm, the weight ω_j will be updated in each iteration, which can be seen from (16). The multiplicative factor $\varpi_{j;\tau-1}$ in (16) is a value always less than one and will become smaller and smaller (see the definition in equation (15) and the last step of Algorithm 2). The factor $\varpi_{j;\tau-1}$ shows the probability that a sample does not belong to clusters $1, \cdots, \tau - 1$. Therefore, $\sum_{j=1}^n \varpi_{j;\tau-1}$ can be viewed as the total number of samples that belong to clusters $\tau, \tau + 1, \cdots$, which will be estimated by the non-parametric kernel density. If this number is too small, then we should stop the algorithm since there are not enough data to further split this non-parametric cluster into a smaller parametric cluster and a smaller non-parametric cluster.

In practice, we can simply use a classical heuristic argument that from any parameter estimation, we will need at least 20 observations. Therefore, we can stop the algorithm, if $\sum_{j=1}^n \varpi_{j;\tau-1} \leq \kappa = 20$. In this paper, we set a smaller threshold value $\kappa = 10$, for the purpose to obtain more small clusters and showing that the small clusters identified are not needed. Note that, setting $\kappa = 10$ or $\kappa = 20$ does not affect the large clusters identified in the beginning. This is because once the algorithm identified a cluster, it will be fixed and will not change when identifying clusters in a later stage. For example, with $\kappa = 20$ the algorithm identified clusters 1, 2, and 3, plus a non-parametric component. If we change $\kappa = 10$, we will still have exactly the same clusters 1, 2, and 3 as if we are using $\kappa = 20$, however, we may identify two smaller clusters 4 and 5, plus a smaller non-parametric cluster.

Please see Section F in the Supplementary Material for more simulations studies about the choice of κ, to demonstrate that changing the value of κ will not affect the clusters we have identified.

4. Simulation studies

The subgroups to be identified by the parametric components using our method can be modelled with different density families. Considering the popularity of Gaussian models and their easy implementation using existing computational tools, we illustrate the proposed seqEM algorithm under the framework of the Gaussian model and compare its performance to that of the FlexMix (Grün and Leisch, 2008). For a given K, Grün and Leisch (2008) started the EM algorithm by separating the samples into K classes evenly at random to estimate the model parameters. Besides, they provided ready-for-use procedures to run EM algorithm repeatedly from

Algorithm 2: The sequential EM (seqEM) algorithm for gFMR with $K > 2$.

Initialise κ, set $\tau = 1$ and $\boldsymbol{\varpi}_{\tau-1} = (\varpi_{1;0}, \cdots, \varpi_{n;0}) = (1, \cdots, 1)^T$;
while $\sum_i \varpi_{i;\tau-1} > \kappa$ **do**
  Randomly divide the samples into two groups by assigning $u_{0,i}^{(0)} \sim Binom(1, 0.5)$;
  Set $m = 0$, $c^{(0)} = 1$, $\widetilde{\mathcal{L}}_\tau^{(0)} = 0$;
  while $c^{(m)} > \epsilon$ **do**
    $m = m + 1$;
    M-step: Calculate

$$\omega_j^{(m)} = n \frac{\varpi_{j;\tau-1}(1 - u_{0,j}^{(m)})\pi_{1;\tau}(\boldsymbol{\alpha}_\tau^{(m)}; \boldsymbol{z}_j)^{-1}}{\sum_{i=1}^n \varpi_{i;\tau-1}(1 - u_{0,i}^{(m)})\pi_{1;\tau}(\boldsymbol{\alpha}_\tau^{(m)}; \boldsymbol{z}_i)^{-1}}, \quad j = 1, 2, \ldots, n, \tag{16}$$

$$\boldsymbol{\theta}^{(m)} = \arg\max_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \left\{ \varpi_{i;\tau} u_{0,i}^{(m)} \log g(y_i; \boldsymbol{\theta}, \boldsymbol{x}_i) \right\},$$

$$\boldsymbol{\alpha}^{(m)} = \arg\max_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{i=1}^n \varpi_{i;\tau} \left\{ u_{0,i}^{(m)} \log \pi_{0;\tau}(\boldsymbol{\alpha}, \boldsymbol{z}_i) + (1 - u_{0,i}^{(m)}) \log \pi_{1;\tau}(\boldsymbol{\alpha}, \boldsymbol{z}_i) \right\},$$

$$\widetilde{\mathcal{L}}_\tau^{(m)} = \frac{1}{n} \sum_{i=1}^n \varpi_{i;\tau-1} \log \left\{ \pi_{0;\tau}(\boldsymbol{\alpha}^{(m)}; \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}^{(m)}, \boldsymbol{z}_i) + \pi_{1;\tau}(\boldsymbol{\alpha}^{(m)}; \boldsymbol{z}_i) \tilde{f}_{1;\tau}(y_i; \boldsymbol{\omega}^{(m)}) \right\};$$

    and update $c^{(m)} = \widetilde{\mathcal{L}}_\tau^{(m)} - \widetilde{\mathcal{L}}_\tau^{(m-1)}$;
    E-step: compute the posterior probability of each sample

$$u_{0,i}^{(m+1)} = \frac{\pi_{0;\tau}(\boldsymbol{\alpha}^{(m)}; \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}^{(m)}, \boldsymbol{x}_i)}{\pi_{0;\tau}(\boldsymbol{\alpha}^{(m)}; \boldsymbol{z}_i) g(y_i; \boldsymbol{\theta}^{(m)}, \boldsymbol{x}_i) + \pi_{1;\tau}(\boldsymbol{\alpha}^{(m)}; \boldsymbol{z}_i) \tilde{f}_{1;\tau}(y_i; \boldsymbol{\omega}^{(m)})};$$

  end
  Output $\widehat{\boldsymbol{\Theta}}_\tau = (\boldsymbol{\alpha}^{(m)}, \boldsymbol{\theta}^{(m)})$, $\widehat{\boldsymbol{\omega}}_\tau = \boldsymbol{\omega}^{(m)}$, update $\varpi_{i;\tau} = \varpi_{i;\tau-1}(1 - u_{0,i}^{(m+1)})$, $i = 1, 2, \ldots, n$ and set $\tau = \tau + 1$;
end

different random initial values and for different numbers of components to pursue the global maximum of the likelihood as the MLE and determine a suitable K.

We simulate the data according to the gFMR with K components, i.e., $g(y; \boldsymbol{\theta}_k, \boldsymbol{x})$ in (2) corresponds to the density of the Gaussian regression model, with $\boldsymbol{\theta}_k' = (\boldsymbol{\beta}_k', \sigma_k)$ consisting of the regression parameter $\boldsymbol{\beta}_k$ and the residual standard deviation σ_k. The mixing probabilities are given by a multinomial logistic function $\rho_k(\boldsymbol{z}) = \exp(\boldsymbol{z}\boldsymbol{\gamma}_k) / \left(1 + \sum_{j=1}^{K-1} \exp(\boldsymbol{z}\boldsymbol{\gamma}_j)\right)$, with the parameters $\boldsymbol{\gamma}_k$, $k = 1, \ldots, K-1$. Since we are going to evaluate the parameter estimation of a mixture model, for the easiness of visualization, we set both $\boldsymbol{x}$ and $\boldsymbol{z}$ as 2-dimensional vectors, of which one element is discrete from the binomial distribution Binom $(1, 0.5)$ and the other is continuous from the standard normal $N(0, 1)$. To investigate the impact of K on the parameter estimation, we fixed a dataset and fit the simpler FMR model (1) by both FlexMix and our seqEM. As shown by the numerical results in Section D of the Supplementary Material, our seqEM performs more stably than the FlexMix in the parameter estimation with random initial values. In this section, we focus on the simulation studies and comparisons of our proposed seqEM and the FlexMix under gFMR (2).

To make things clear, we refer to '*component*' as the true component in the mixture model, and '*cluster*' as the component being identified using FlexMix or the proposed seqEM. Note that our seqEM identifies the clusters one by one. When we talk about cluster 1, it means the first cluster being identified, which may represent one of the other components rather than component 1. For the FlexMix, we include both FlexMix(K) whose number of components K is given as the truth, and FlexMix(BIC) whose K is selected from 2 to 10 according to BIC.

In order to mimic the real data analysis, we simulate the data from the gFMR with unevenly weighted components, where the main components with leading mixing probabilities are of more interest for treatment evaluation. More studies, such as the results of our method on the gFMR with equally weighted components, are provided in Section E of the supplementary Material. In the following, we start from the gFMR with a single main component and then move on to that with multiple main components. Through the simulation studies, we will show the advantages of seqEM, compared to FlexMix, in four aspects. The seqEM

a) is more reliable to identify the main components;
b) does not need to pre-specify K;
c) is more reliable to handle cases with outliers.
d) is more reliable to the choice of the initial values of the EM algorithms (see Section D in the Supplementary Material).

4.1. gFMR with single main component

We set $K = 5$ and the random sample of size $n = 200$ is generated by setting the parameters

Table 1
Frequencies of numbers of clusters being identified by seqEM and FlexMix in 500 simulations of Section 4.1 and 4.2.

# of clusters		2	3	4	5	6	7	8
Simulation in Section 4.1	seqEM	5	72	151	147	113	11	1
	FlexMix(BIC)	118	156	139	57	25	5	
Simulation in Section 4.2	seqEM		12	46	156	192	94	
	FlexMix(BIC)		4	284	193	19		

$$\begin{cases} \boldsymbol{\gamma}_1 = (-1,-1,1)', \\ \boldsymbol{\gamma}_2 = (2,1,-1)', \\ \boldsymbol{\gamma}_3 = (-0.5,1,-1)', \\ \boldsymbol{\gamma}_4 = (0,-1,-1)', \end{cases} \quad \begin{cases} \boldsymbol{\beta}_1 = (1,-1,2)',\ \boldsymbol{\beta}_2 = (1,1,1)', \\ \boldsymbol{\beta}_3 = (1,2,0)',\ \boldsymbol{\beta}_4 = (1,0,-2)', \\ \boldsymbol{\beta}_5 = (1,-2,-1)'. \\ \sigma_1 = \sigma_2 = \cdots = \sigma_5 = 0.1. \end{cases}$$

Such a setting leads to the samples from different components being well separated, and their proportions are about $0.06, 0.74, 0.06$, 0.05, and 0.08, respectively. The second component dominates all the others. Due to the limited sample size, other components 1, 3, 4, and 5 only have a few observations and may be taken as outliers. This simulation is to mimic the real data analysis and it is more meaningful to distinguish the largest component from those negligible outliers than to identify every component.

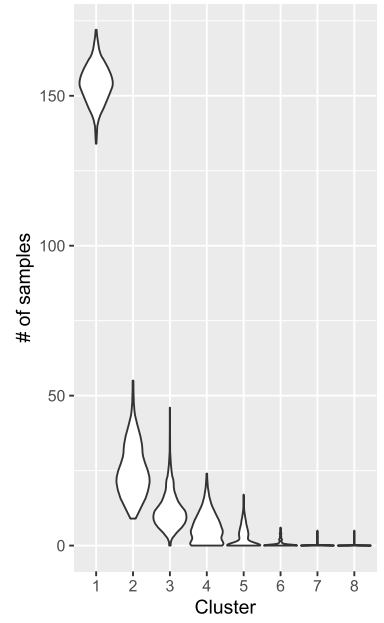
The numbers of clusters being identified using our method in 500 simulations are summarized by their frequencies in Table 1. From the results in Table 1, we can see that seqEM is more likely to identify 3 to 6 clusters, however, FlexMix(BIC) has more than 20% probability (118 out of 500 simulations) to identify 2 clusters, while the true number of components $K = 5$. Thus seqEM is much better than FlexMix(BIC) for estimating K.

In Fig. 1a, we show the violin plot of the number of samples being classified into each cluster in 500 simulations. There are around 154 samples being classified into cluster 1 in each repetition of the simulation, whose size is greatly larger than those of the others. The median proportion 0.77 of cluster 1 is very close to the proportion 0.74 of component 2 in the population. We calculated the proportion of samples in cluster 1 that come from component 2, and found that it ranges from 0.88 up to 1, with the median 0.96. Besides, the regression parameter estimates of the first 2 clusters from our method in 500 simulations are presented in Fig. 1b. To best visualise the results, we only show the estimates for slope parameters in $\widehat{\boldsymbol{\beta}}_i$, $i = 1, 2$. All parameter estimates of cluster 1 clearly gather around the true parameter of component 2. It is demonstrated that component 2 with the largest mixing probability is always firstly identified as cluster 1 by our method.

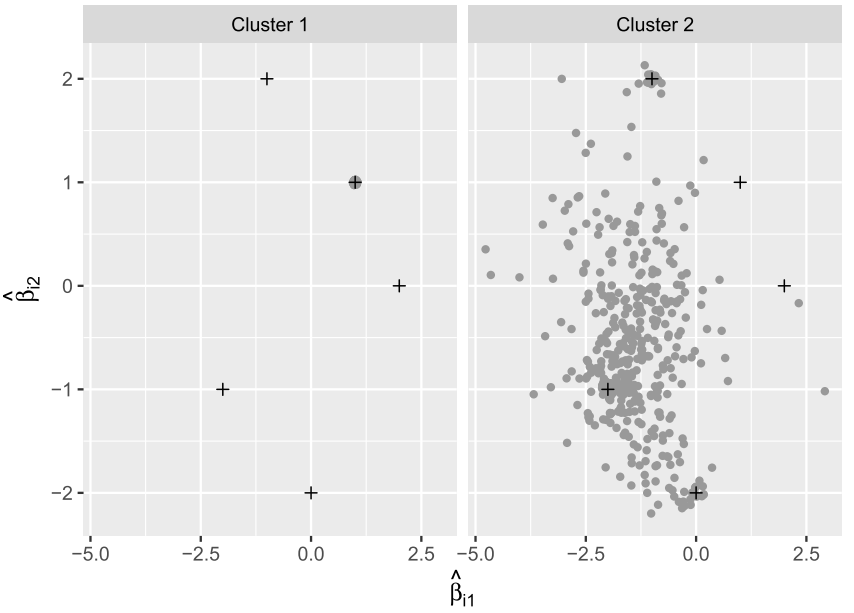
To evaluate the performance of seqEM in the component identification and compare it to the FlexMix, we search the permuted labels of the identified clusters for an optimal matching to the true components such that there are as many samples as possible being correctly classified into the matched pairs of clusters and components (Stephens, 2000b). We calculate the proportion of total samples in the matched pairs (PMS), as well as the well-known adjusted Rand index (ARI, (Hubert and Arabie, 1985)) and normalized variation of information (NVI, (Meilă, 2007)) for reference. Moreover, for each component, we define the classification accuracy rate (CAR) as the proportion of the samples in the best-matched cluster that come from this component, and the detection rate (DR) as the proportion of the samples in the component that are included by its best-matched cluster. The ARI, NVI, PMS, CARs, and DRs from different methods in 500 simulations are presented in Fig. 1c. FlexMix(BIC) presents significantly higher ARI and PMS, and lower NVI than FlexMix(K), indicating that K should be determined by the data, rather than arbitrarily assigned, even though it is given as the truth. Our seqEM presents a higher ARI or PMS than FlexMix(BIC), but a slightly higher NVI than FlexMix(BIC). In addition, seqEM efficiently avoids those lower outliers in ARI and PMS results from FlexMix(BIC).

The boxplots of CAR and DR for each component reveal the different performance between our seqEM and the FlexMix. For the main component 2, our method achieves significantly higher DRs than FlexMix(K), and avoids those extremely smaller values from FlexMix(BIC), while their CARs remain at similar levels. Those lower DRs appear in the results of FlexMix(K) and FlexMix(BIC) because component 2 is split into two or more clusters by FlexMix, i.e., $g(y; \boldsymbol{\theta}_2, \boldsymbol{x})$ itself was fitted by a gFMR. However, our method is less likely to split the main component. For the others, due to limited sample size, all methods present variable CARs and DRs. Compared to the FlexMix, our seqEM provides more reliable clustering results with higher DR for the main component, leading to the advantage a).

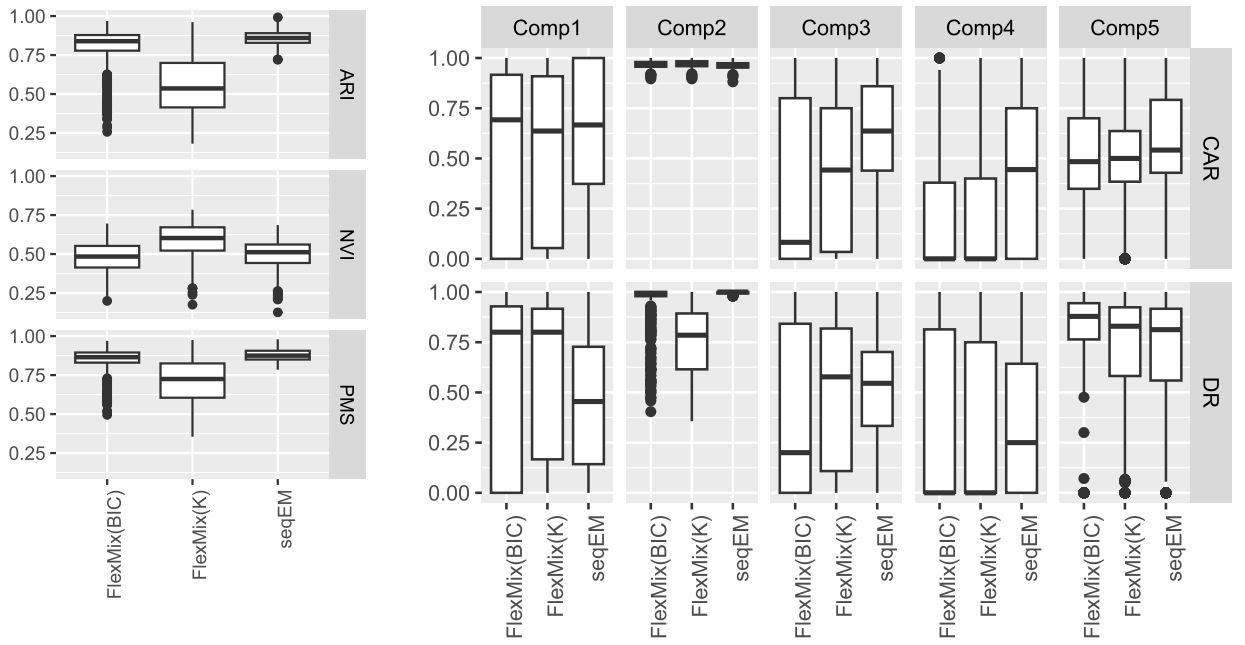
Although multiple clusters are identified from our method, the research aim of this paper is not to estimate the total number of components in the mixture. When the population is highly heterogeneous, it may be infeasible to identify all components with a limited sample size. Instead, we propose to focus on the identification of the main component, which has sufficient data support. As illustrated in this simulation, the main component is identified by sequentially fitting the basic 2-component mixture model, with no need to specify K, yielding the advantage b) of our method. In our opinion, the available sample size determines whether or not the partitioning continues, which is controlled by a tuning parameter κ. In the beginning, there is a full sample, and the sample size decreases as more and more clusters are identified. It is shown that the main component is always identified as the first cluster, so that the setting of κ does not affect the clusters being identified in the beginning, but only those at the end. In Section F of the Supplementary Material, we present the clusters' sizes and parameter estimation of cluster 1 using different κ. While κ is set to be 5, 10, or 20, the size of cluster 1 does not change and its parameter estimation always points to component 2. This performance also demonstrates the reliability of our method to outliers. As mentioned at the beginning of this subsection, the components except component 1 can be taken as outliers. Our method identifies the main component 2 more accurately with these non-negligible outliers, clarifying the advantage c).



(a) Cluster size



(b) Parameter estimation



(c) The overall PMS, CAR and DR for each component

Fig. 1. (a) The numbers of samples being clustered into different clusters using seqEM in Section 4.1. Each violin plot is based on the results of 500 replicated simulation experiments. (b) Estimates for the regression parameters of clusters 1 and 2 from seqEM, where the grey dots indicate the estimates and plus signs (+) locate the true parameter values of components 1-5. (c) The obtained ARI, NVI, PMS, CARs, and DRs using seqEM, FlexMix($K = 5$) and FlexMix(BIC).

4.2. gFMR with multiple main components

In this scenario, we set the parameters $\boldsymbol{\beta}_k$ and σ_k, $k = 1, 2, \ldots, K$ the same as that in section 4.1, but $\boldsymbol{\gamma}_1 = (-1, -1, 1)'$, $\boldsymbol{\gamma}_2 = (1, 1, -1)'$, $\boldsymbol{\gamma}_3 = (0, -1, -1)'$, and $\boldsymbol{\gamma}_4 = (0, 1, 1)'$ to alter the mixing probabilities among different components. The average proportions

of components 1-5 in this scenario are $0.05, 0.5, 0.09, 0.26$, and 0.1 respectively, where there are more than one main component with leading mixing probabilities, the second and the fourth, and the others may be taken as outliers.

We present the frequencies of numbers of clusters being identified using seqEM and FlexMix(BIC) in 500 simulations in Table 1. Our seqEM identifies 4 to 7 clusters with very large probabilities. Among the 500 simulations, seqEM identifies 5 or 6 clusters 348 times. On the other hand, FlexMix(BIC) identifies 4 to 6 clusters with a very large probability. Among the 500 simulations, FlexMix(BIC) identifies 4 clusters 284 times, and 5 clusters 193 times. FlexMix(BIC) tends to identify less number of clusters, by merging small clusters into large clusters, which is why the parameter estimation from FlexMix(BIC) is not good. On the contrary, our seqEM tends to identify more clusters, but the parameter estimation for the main clusters is not distorted by the small clusters being identified (see Remark 3.2 and Section F in the Supplementary Material). Therefore seqEM is more reliable to handle the outliers from the small components.

The violin plot in Fig. 2a indicates that, the numbers of samples classified as cluster 1 from our seqEM have two modes, ones with the median 106 and the others with the median 62. That is to say, in all the simulations, there is either a sub-dataset of about 106 samples being classified into cluster 1, or another sub-dataset of about 62 samples being classified into cluster 1. The proportions of these two sub-datasets, 0.53 and 0.31, are very close to the proportions of components 2 and 4 in the population, indicating that cluster 1 may correspond to components 2 or 4 in different simulation experiments. The numbers of samples being classified into cluster 2 also gather around two modes, ones with median 94.5 and the others with median 51, and their corresponding proportions 0.47 and 0.26 are close to that of components 2 and 4, too. Besides, we present the estimates for the slope parameters in the regression models of the first 4 clusters from our method in 500 simulations in Fig. 2b. It is demonstrated that the first two clusters being identified by our method point to either component 2 or 4, with possible permutations in their orders in different simulation experiments. Subsequently, as the cluster sizes decrease, the parameter estimates for clusters 3 and 4 become more variable.

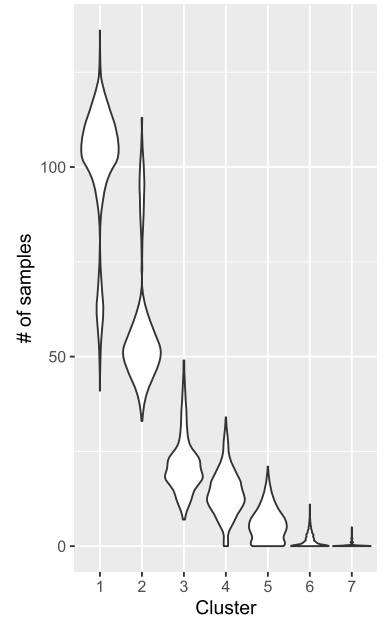
Our method shows a similar result compared to FlexMix in ARI or PMS, but a slightly higher NVI. A significant advantage of our method is that it avoids yielding extremely small ARI and PMS values, as shown in Fig. 2c. Comparisons for each component can also be seen in the last plot of Fig. 2c. For the main components 2 and 4, our method maintains higher DRs, while both FlexMix(K) and FlexMix(BIC) may produce much lower ones, with their CARs at a similar level. The vanishing extremely lower DRs indicate that our method can fully identify the main components and avoid further splitting them into more clusters, resulting in the advantage c). On the other hand, both FlexMix(K) and FlexMix(BIC) are affected by the small components (the outliers) and their components 2 and 4 have a very long tail in the boxplots. Since our procedure runs sequentially, the samples belonging to the overlapped components are more likely to be captured by the former cluster. This prevents us from over-splitting components and distinguishes us from FlexMix. In summary, our method provides more accurate and reliable clustering results for the main components than the FlexMix, referred to as advantage a). As the same as that in Section 4.1, our method identifies clusters sequentially according to the 2-component mixture model, which has nothing to do with K and leads to the advantage b).

When the available sample size of a component is small, it may have to be treated as an outlier, even if there is a component. None of the existing methods can identify a component with a very small sample from it. This is also the exact idea that motivates us to set up the stopping criteria for our sequential algorithm. When the sample size is large, our method will identify components sequentially until there are not enough available data, for instance, less than 10 observations with $\kappa = 10$ (see Remark 3.2 for more details). In the situation that all the components' weights are equal, they are not potentially classified as main or outlier components, but identified one by one as a sequence, with possible order switching as indicated by components 2 and 4 in section 4.2. Please refer to Section E of Supplementary Material for detailed results.

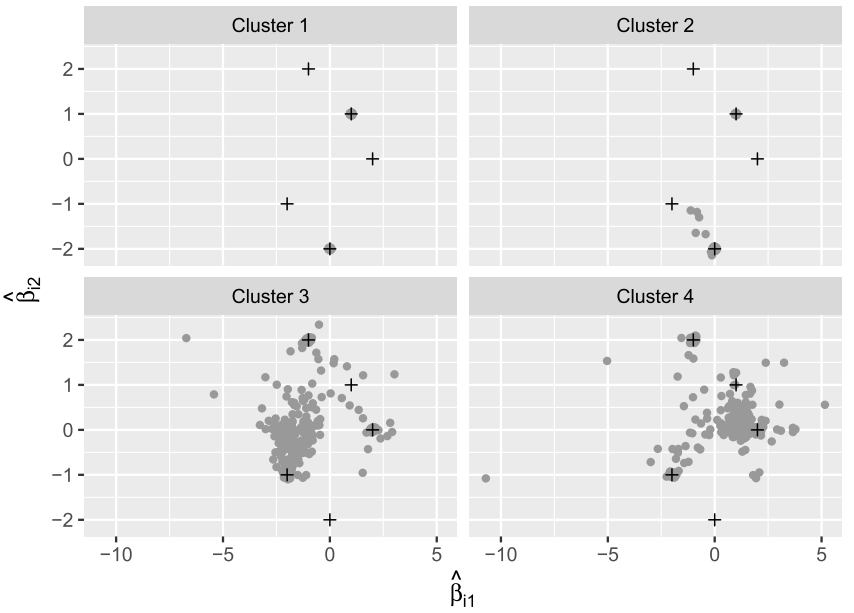
5. A cancer cell line study

In the treatment of tumor diseases, many drugs show highly heterogeneous responses due to the complexity of the disease. Certain treatments may significantly benefit some patients, but cause serious side effects and provide little benefit to some others. It is greatly valuable in clinical practice to identify patients who would best benefit from a specific treatment and exclude those who are not benefited, based on our increasing understanding of the molecular mechanism of tumors. For this purpose, the CCLE project generated the genomic profiles of 504 cancer cell lines and tested the drug responses of 24 chemical compounds on them. The heterogeneity of tumor disease makes the drug responses vary among cell lines. There are some more responsive, some less responsive, and others not responsive at all, which implies that the patient with a particular genetic feature will either benefit more or less from treatment or not benefit at all. Our research aim is to quantify how responsive a cell line is to a particular treatment and identify such a heterogeneous structure among the cell line population, to classify patients who potentially benefit from the treatment, with little side effects; or those who are not benefited from the treatment but have serious side effects; or those belonging to other subgroups.

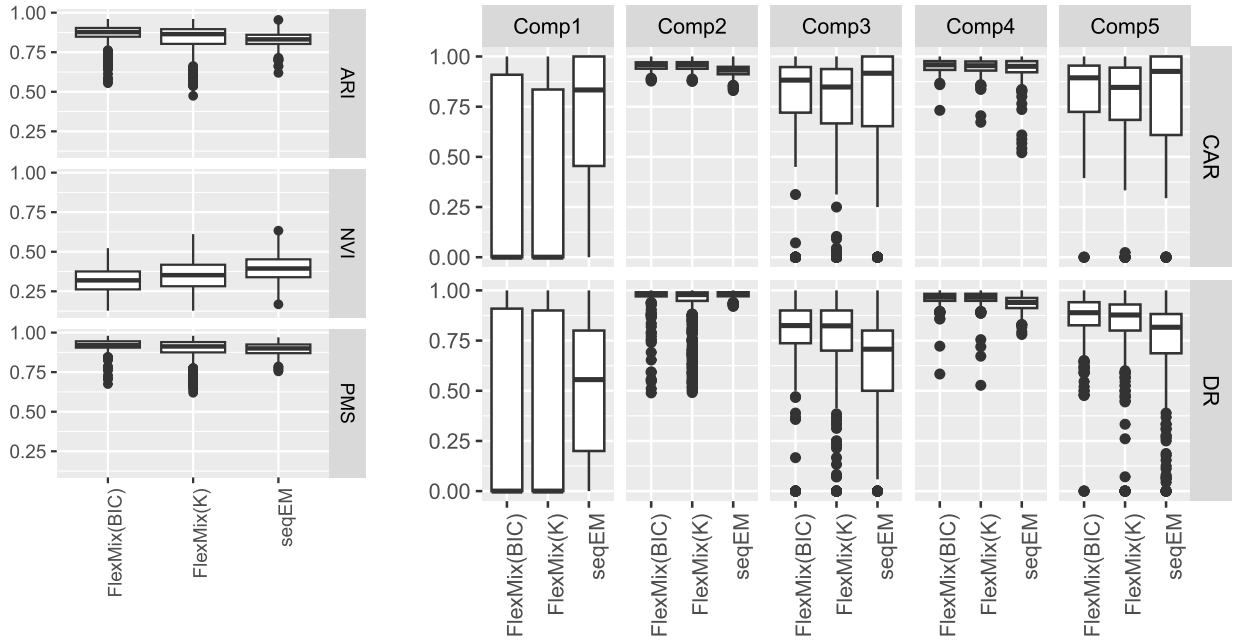
The mixture of regression models is employed to account for the heterogeneous dependency of pharmacologic response on the genetic profiles of different cell lines. Li et al. (2019) presented a feature selection procedure using model (1) to choose the regression covariates from tens of thousands of genes for drug sensitivity prediction of the heterogeneous population. To illustrate the possible uncertainty of FlexMix due to initial values, for each drug, we took their selected genes as the explanatory variables and fitted model (1) by both seqEM and FlexMix 500 times from random initial values. Since the true value of K is unknown, it is chosen from $2, 3, \ldots, 9$ by BIC in the FlexMix. To eliminate the impact of label-switching and summarize the parameter estimation for comparison, we have an investigation on $\max_{k=1}^{K} \widehat{\rho}_k$, which is the estimate for the largest component weight. The histograms of $\max_{k=1}^{K} \widehat{\rho}_k$, that were obtained from both methods in 500 repetitions with random initial values, are presented in Fig. 3, with one compound *PD-0332991* being excluded because its selected genes given by Li et al. (2019) were not successfully matched in the database. For a



(a) Cluster size



(b) Parameter estimation



(c) The overall PMS, CAR and DR for each component

Fig. 2. (a) The numbers of samples being clustered into different clusters using seqEM in Section 4.2. Each violin plot is based on the results of 500 replicated simulation experiments. (b) Estimates for the regression parameters of clusters 1-4 from seqEM, where the grey dots indicate the estimates and plus signs (+) locate the true parameter values of components 1-5. (c) The obtained ARI, NVI, PMS, CARs, and DRs using seqEM, FlexMix($K = 5$) and FlexMix(BIC).

number of compounds, both methods can provide consistent estimates $\max_{k=1}^{K} \widehat{\rho}_k$, although their results may be distinct. Besides, there are many others such as *Nutlin-3, PF2341066, PLX4720, Sorafenib* and *TKI258*, for which the estimates $\max_{k=1}^{K} \widehat{\rho}_k$ from the FlexMix with different initial values become more scattered than that from the seqEM, which means seqEM gives more reliable estimate for the largest component weight in these compounds, compared to the FlexMix.

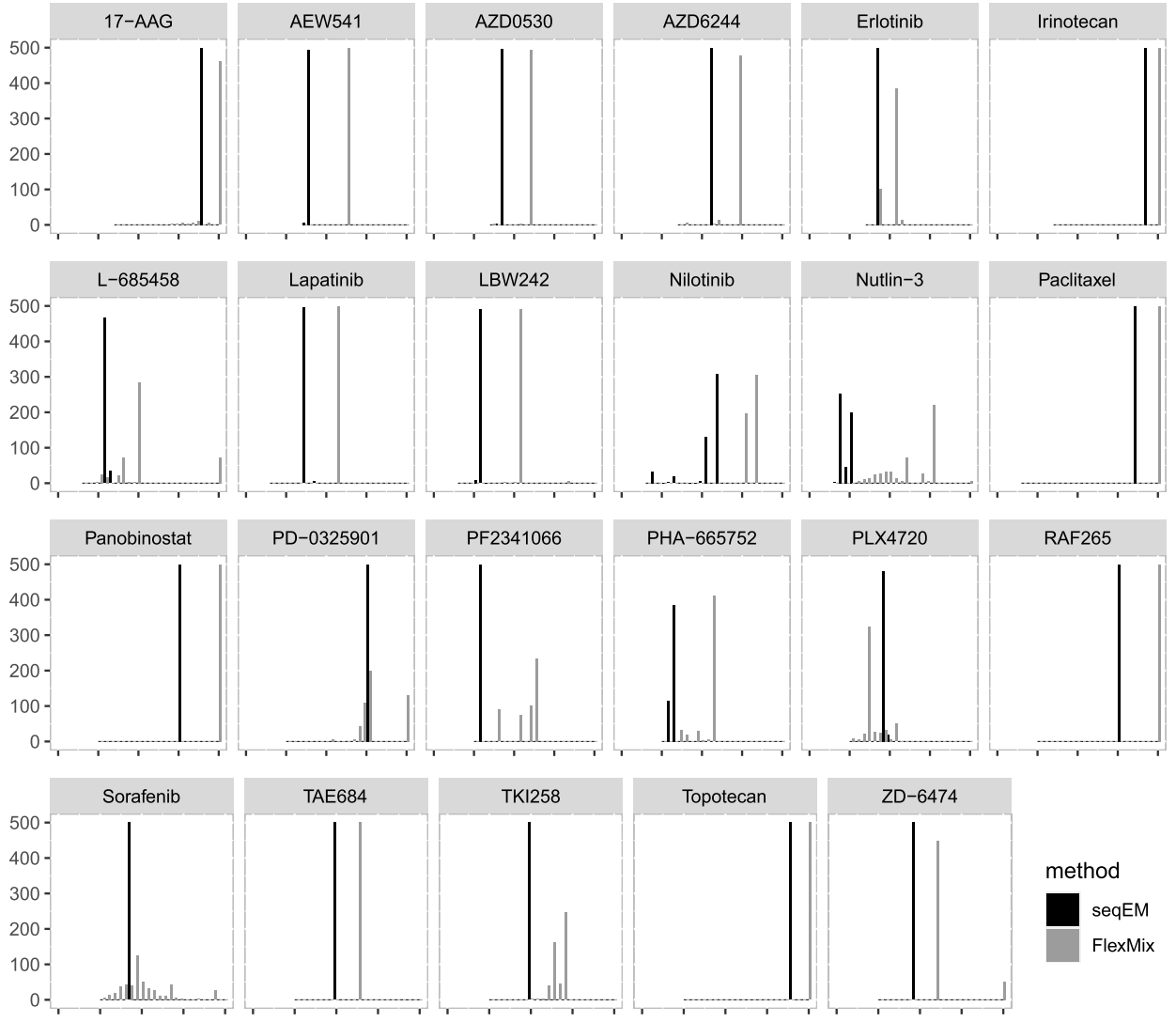
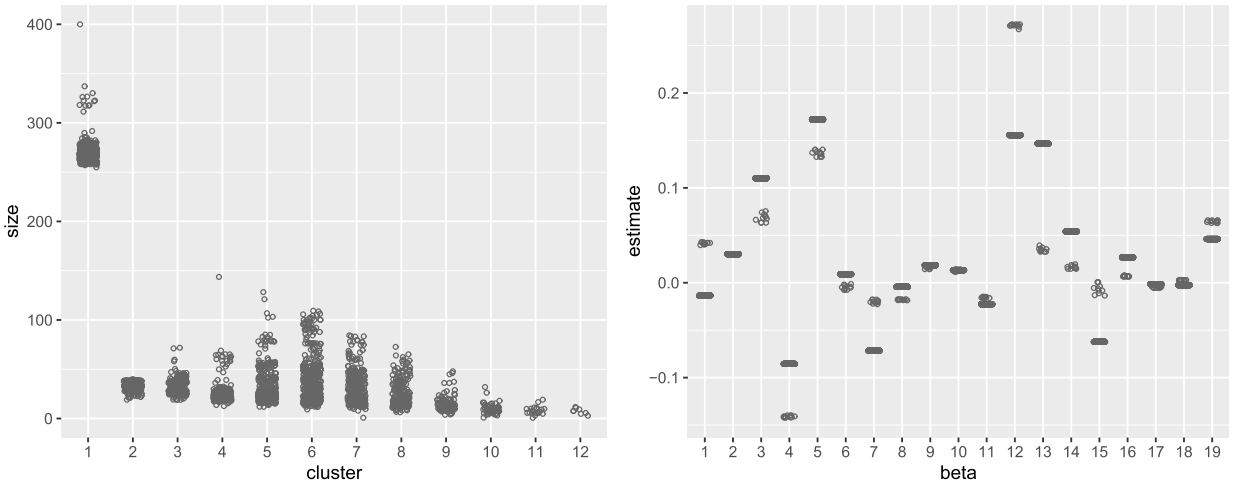


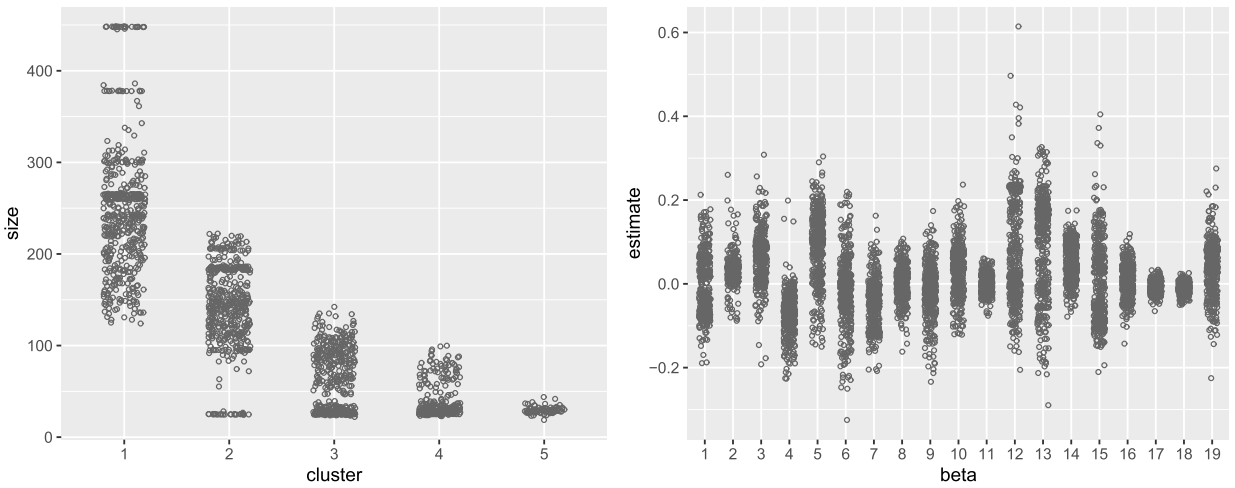
Fig. 3. Histogram of $\max_{k=1}^{K} \widehat{\rho}_k$ from our seqEM and FlexMix with random initial values for 23 compounds of CCLE database, where all the x-axes range from 0 to 1.

To illustrate the advantage of seqEM on the working model (2), we choose the data set of *Sorafenib* for further analysis. First, since there are no concomitant variables reported in the original analysis, we took the clusters identified by our seqEM using model (1) to label the samples into subgroups, and then apply the screening and random forest method to select the concomitant variables. In addition to 19 explanatory genes given by Li et al. (2019), five genes PPAN, STEAP1, KLRG1, DENND4A and JAG1 were chosen as the concomitant variables. With the selected explanatory and concomitant variables, we run our proposed algorithm 500 times with random initial values. The numbers of samples that were classified into different clusters in 500 repetitions were summarized in Fig. 4a. It is shown that cluster 1 is significantly larger than the others and the cluster sizes after the second one are particularly smaller. Considering there are 19 explanatory variables being involved, the parameter estimation for those clusters with limited sample sizes may be inaccurate and cluster 1 is the only important component. In all 500 repetitions of our method with random initial values, the estimates for the regression parameter $\boldsymbol{\beta}_1$ remain consistent, only with a few exceptions, as shown by the jitter plots in Fig. 4a. For comparison, we also repeated the FlexMix 500 times with random initial values. Although there were fewer numbers of clusters being identified, the clustering results from the FlexMix are very sensitive to the initial values. We also rank the identified clusters in each repetition by their sizes from the largest to the smallest, and present the cluster sizes in 500 repetitions in Fig. 4b, as well as the regression parameter estimates for cluster 1. It can be seen that both the cluster sizes and the parameter estimates from the FlexMix fluctuate in a significantly larger span than that from our method, demonstrating the robustness of our method to the initial settings in the model estimation.

To evaluate the accuracy of parameter estimation from our method, we conducted 5-fold cross-validation. With the parameter estimates obtained by fitting the training data, the drug responses in the test data are predicted by $\widehat{y}_i = \sum_{k=1}^{K} \widehat{\rho}_k(\boldsymbol{z}_i)\widehat{f}_k(\boldsymbol{x}_i)$, where $\widehat{f}_k(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_k$ for the identified clusters with the regression parameter estimates $\widehat{\boldsymbol{\beta}}_k$, but $\sum_{j=1}^{n} y_i\widehat{\omega}_j$ for the last cluster of nonpara-



(a) seqEM



(b) FlexMix

Fig. 4. The jitter plots of the cluster sizes (left) that were obtained using (a) seqEM and (b) FlexMix in 500 repetitions with random initial values, and their parameter estimates (right) for the largest component in the analysis of compound *Sorafenib*.

metric component. The cluster that a sample in the test data belongs to is determined by $\arg\max_k \widehat{\rho}_k(\boldsymbol{z}_i)$. In different folds, distinctive clusters may be produced using different training data, especially for the components with limited samples. It may be not feasible to match all the clusters from different folds. Our method focuses on the identification of important clusters, thus we summarize the results by clusters. As aforementioned, cluster 1 is the only important component, therefore we only present the averaged mean square error (MSE) and correlation between the prediction and the truth $(\widehat{y}_i, y_i)$ of cluster 1 in Table 2, together with that of the largest clusters obtained by the FlexMix. It is shown that the important cluster identified using our method involves more samples than the largest cluster from the FlexMix, with similar MSEs and correlations being achieved. Our method avoids the splitting of the important cluster and separates the largest subgroup. The results of FlexMix using model (1) were also included in Table 2. By utilizing the concomitant variables to partition the samples into appropriate clusters for prediction, the working model (2) consistently presented smaller MSE and higher correlation than model (1), no matter either the proposed or FlexMix method was used for the model estimation.

In summary, our approach can identify a very significant cluster and the cell lines in this cluster show a common responsive pattern to the chemical compound *Sorafenib*. On the contrary, the existing FlexMix cannot identify a single cluster, since its results are not reliable in the sense that they highly depend on the initial values of the EM algorithm. The concomitant variables in our model can help doctors to identify this cluster before they apply treatment. For patients not belonging to this cluster, doctors may look for alternative treatments.

Table 2
MSE and correlation between the prediction and the truth in the largest cluster of the test data in 5-fold cross-validation.

	seqEM	FlexMix of model (2)	FlexMix of model (1)
MSE	0.18	0.17	0.25
Correlation	0.33	0.34	0.29
Cluster proportion	0.58	0.49	0.56

6. Discussion

Heterogeneity commonly exists in this era of big data, which changed the typical statistical analysis from the model foundation and challenges all the following inferences. In this paper, we contribute a novel statistical method to identify the components in the gFMR sequentially, providing an alternative way to estimate the gFMR. Compared to directly fitting the gFMR, our method not only avoids the need to pre-specify the number of components K in the mixture but also can avoid local maxima in the parameter estimation and therefore provide more reliable clustering results. For the regression model of each component, we used the Gaussian model for illustration. The proposed method can be easily adapted to other parametric families. The convergence rate of the EM algorithm discussed in peer literature might be extended to our method but beyond the scope of this paper.

Identifiability is a common issue being widely discussed for mixture models. The gFMR (2) is quite general, and its identifiability has been thoroughly discussed by Wang et al. (2014) in parametric, semi-parametric and non-parametric settings. In this paper, we consider $g(y;\boldsymbol{\theta}_k,\boldsymbol{x})$ in some known parametric form, and (2) is not subject to the identifiability issue. Note that the two-component mixture model (3) is nonidentifiable due to the fact that any arbitrary part excluding a parametric component can be regarded as a non-parametric component. However, the sequence of identified components from our algorithm is identifiable without consideration of label-switching due to the identifiability of gFMR (2).

In terms of computational complexity, our seqEM actually searches a small number of K only, since it will stop automatically once enough clusters are identified. On the contrary, existing FlexMix EM algorithms for mixture models will need to specify a very large value for the maximum possible value of K, then do a full search of the space and run EM algorithms for every possible value of K. Thus FlexMix needs to run more EM algorithms in practice. However, in terms of real computational time, our seqEM does not show advantages, due to the computational cost in the kernel density estimation, for instance in the analysis of *Sorafenib* data set, it took about one second for each FlexMix EM. Thus FlexMix will take about half a minute if we do a deep search for K up to 30, as suggested by Richardson and Green (1997). On the other hand, seqEM took about 6 minutes when we choose $\kappa = 100$ to identify 8 components. For these highly heterogeneous real data with possible outliers, we may have to run the EM algorithms much more times than the expected number of components in the population, when we use FlexMix. However, when these small outlier components are not of research interests, seqEM bypasses this challenge by setting a large κ and the computational burden will be significantly reduced. We would consider improving the computational efficiency of seqEM in future work.

Another common issue about big data analysis is the high dimensionality. Both the concomitant variable and the explanatory variable may be of high dimension. However, to focus on the discussions on the essential problems in solving the complex mixture model, we did not consider variable selection in this work. With more reliable model estimation, the proposed method can be extended to implement variable selection for both the concomitant and explanatory variables in each partitioning of the sequential procedure. For ultra-high dimensional data, appropriate univariate screening methods should be incorporated in advance to reduce their dimensions to a manageable size. We leave this to future research work.

Acknowledgements

You's research is partially supported by the National Natural Science Foundation of China (12126610), and Guangdong Basic and Applied Basic Research Foundation (2023A1515012254). Dai's research is partially supported by the EPSRC PINCODE grant, EP/X027872/1, and the UKRI OCEAN grant, EP/Y014650/1. Wang's research is partially supported by the National Natural Science Foundation of China (12231017, 72171216, 71921001, 71991474), and the Science and Technology Program of Guangzhou, China (202002030129). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2024.107942.

Appendix B. Computational codes

The R package `SEMRflexmix` implementing the proposed sequential analysis procedure was developed on the basis of `flexmix` (Grün and Leisch, 2008) and publicly available at https://github.com/scrcss/SEMRflexmix.

References

Balakrishnan, S., Wainwright, M.J., Yu, B., 2017. Statistical guarantees for the EM algorithm: from population to sample-based analysis. Ann. Stat. 45 (1), 77–120.

Baudry, J.-P., Celeux, G., 2015. Em for mixtures: initialization requires special care. Stat. Comput. 25, 713–726.

Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S., 2009. mixtools: an R package for analyzing mixture models. J. Stat. Softw. 32 (6), 1–29.

Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Stat. Data Anal. 41 (3–4), 561–575.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., et al., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. B 39, 1–38.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M.J., Jordan, M.I., Yu, B., 2020. Singularity, misspecification and the convergence rate of EM. Ann. Stat. 48 (6), 3161–3182.

Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Nat. Sci. Rev. 1, 293–314.

Fruhwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York.

Grün, B., Leisch, F., 2008. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28 (4), 1–35.

Ho, N., Yang, C.-Y., Jordan, M.I., 2019. Convergence rates for Gaussian mixtures of experts. ArXiv. arXiv:1907.04377 [abs].

Huang, M., Yao, W., 2012. Mixture of regression models with varying mixing proportions: a semiparametric approach. J. Am. Stat. Assoc. 107 (498), 711–724.

Huang, M., Li, R., Wang, S., 2013. Nonparametric mixture of regression models. J. Am. Stat. Assoc. 108 (503), 929–941.

Huang, M., Yao, W., Wang, S., Chen, Y., 2018. Statistical inference and applications of mixture of varying coefficient models. Scand. J. Stat. 45 (3), 618–643.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M.J., Jordan, M.I., 2016. Local maxima in the likelihood of Gaussian mixture models: structural results and algorithmic consequences. Adv. Neural Inf. Process. Syst. 29, 4116–4124.

Klusowski, J.M., Yang, D., Brinda, W., 2019. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. IEEE Trans. Inf. Theory 65 (6), 3515–3524.

Kwon, J., Qian, W., Caramanis, C., Chen, Y., Davis, D., 2019. Global convergence of the em algorithm for mixtures of two component linear regression. In: Conference on Learning Theory. PMLR, pp. 2055–2110.

Kwon, J., Ho, N., Caramanis, C., 2021. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1405–1413.

Li, Q., Shi, R., Liang, F., 2019. Drug sensitivity prediction with high-dimensional mixture regression. PLoS ONE 14 (2), e0212108.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Meilă, M., 2007. Comparing clusterings—an information based distance. J. Multivar. Anal. 98 (5), 873–895.

Miller, J.W., Harrison, M.T., 2018. Mixture models with a prior on the number of components. J. Am. Stat. Assoc. 113 (521), 340–356.

Papastamoulis, P., 2016. label.switching: an R package for dealing with the label switching problem in mcmc outputs. J. Stat. Softw., Code Snippets 69 (1), 1–24.

Pelleg, D., Moore, A.W., 2000. X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conference on Machine Learning, pp. 727–734.

Punt, C.J.A., Koopman, M., Vermeulen, L., 2017. From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nat. Rev. Clin. Oncol. 14, 235–246.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. B 59, 731–792.

Schlicker, A., Beran, G., Chresta, C.M., McWalter, G., Pritchard, A., Weston, S., et al., 2012. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med. Genom. 5, 66.

Stephens, M., 2000a. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Ann. Stat., 40–74.

Stephens, M., 2000b. Dealing with label switching in mixture models. J. R. Stat. Soc. B, 795–809.

Wang, B., Wang, X., 2007. Bandwidth selection for weighted kernel density estimation. arXiv preprint. arXiv:0709.1616.

Wang, S., Yao, W., Huang, M., 2014. A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. Stat. Probab. Lett. 93, 41–45.

Xu, J., Hsu, D.J., Maleki, A., 2016. Global analysis of expectation maximization for mixtures of two Gaussians. Adv. Neural Inf. Process. Syst. 29.