

Post-Hoc Mixture Models of the Best Linear Unbiased Predictors (BLUP) from Linear Mixed Effects Models to Classify Longitudinal Data with Haphazardly Spaced Intervals: A Simulation Study

Md Jobayer Hossain<sup>1,2</sup>, Benjamin E. Leiby<sup>3</sup>

<sup>1</sup>Biostatistics Program, Department of Biomedical Research, Nemours/Alfred I. duPont Hospital for Children, 1600 Rockland Road, Wilmington, DE 19803

<sup>2</sup>Department of Applied Economics and Statistics, University of Delaware, 213 Townsend Hall, Newark, DE 19716

<sup>3</sup>Division of Biostatistics, Thomas Jefferson University, 1015 Chestnut Street, Philadelphia, PA 19107

**Corresponding:**

Md Jobayer Hossain

Biostatistics Core, Department of Biomedical Research, Nemours/Alfred I. duPont Hospital for Children, 1600 Rockland Road, Wilmington, DE 19803

Email: [jhossain@nemours.org](mailto:jhossain@nemours.org)

**Abstract**

In precision medicine, vast amounts of naturally occurring data are used to derive evidence-based knowledge. Recognizing distinct trajectories, identifying potential predictors, and making data-driven decisions are areas of growing interest. Mixture-based mixed effects models, implemented mainly in R and Mplus are useful for this. While theoretically sound, these approaches often fail to fully address the problem due to computational complexities. Linear mixed effects models ballpark curves of irregularly spaced longitudinal data with optimum precision. Random effects (REs), assumed to be distributed as multivariate normal (MVN), vary across individuals accounting for sources of heterogeneity. To identify presence of subgroups, we apply post-hoc Gaussian finite mixtures on empirical BLUPs assuming a mixture of MVN of REs to classify trajectories. Resulting ellipsoids can vary by center or by features- size, volume and orientation that are determined by eigenvalue decomposition of covariance matrix. This study used simulations to extend previous application of real data to evaluate the classification performance of this method compared with that of existing methods.

Keywords: Classification; irregular spaced time; trajectories; linear mixed model; heterogeneous; mixture distribution of random effects

## 1. Background and Motivation:

Recognizing presence of subgroups with heterogeneous trajectories in longitudinal data, identifying potential contributors to varying trends, and utilizing this data-driven knowledge to make decisions are areas of growing research interest in this data science era (Jung et al. 2008, Arminger et al. 2013). In learning healthcare systems and personalized as well as precision medical decisions, massive amounts of naturally occurring longitudinal data are utilized to derive evidence-based knowledge by accounting for individual-level variations and recognizing distinct patterns of health and disease trajectories with the goal of facilitating precise diagnosis, appropriate strategies for prevention, and tailored treatment of health conditions (Jha et al. 2009, Wei et al. 2015).

While naturally occurring data such as electronic medical records have shown great utility in learning healthcare systems and personalized medical decisions, observation times in these datasets are often irregularly spaced as patients make clinic visits idiosyncratically, thereby producing unique sequences of measurements across individuals. The repeated measures of this type are often expressed as parametric or semiparametric functions of time and are described by levels and shapes of the curves. Linear mixed effects regressions model these data as a combination of population temporal trend that is shared by all individuals and subject-specific effects that describe how the trend over time of each individual differs from the population mean trend (Laird et al. 1982). The former are the fixed effects and the later are the random effects in the model. The random effects vary from one individual to another, thereby accounting for sources of heterogeneity in trajectories across individuals (Fitzmaurice et al. 2012, Garrett et al. 2000). The inclusion of time as random effects in the model allows us to express the covariance of repeated measures as a function of time. In the general form of the mixed effect model, random effects are assumed to be distributed as multivariate normal with mean 0 and constant variance-covariance matrix. This distributional assumption of the random effects implies that individual-level trajectories are homogeneous in shapes, thus the model under this assumption can be termed as a homogeneous linear mixed effects model. If the dataset contains heterogeneous individual-level trajectories, the above homogeneity assumption could be violated and the distribution of random effects could be expressed as the joint distribution of heterogeneous subgroups each with multivariate distribution with different parameters. With this violation, averaging over curves of heterogeneous shapes may result in missing important features of trajectories for individuals across heterogeneous subgroups causing misrepresentation of model fit. Therefore, it is important to identify potential hidden subgroups of individuals with distinct trajectories that may exist in a dataset. This requires appropriate capturing of the heterogeneity in levels and shapes of trajectories across individuals.

Verbeke et al (1996) introduced a heterogeneous linear mixed effects model where random coefficients were assumed to have a mixture of normals distribution. The method was implemented in a SAS macro (Komárek et al. 2002), but the EM algorithm used was found to be computationally expensive and failed to provide good convergence criteria and direct estimates of the variance of the parameters. Proust-Lima et al (2017) recently extended the model of Verbeke et al by expressing mixture component-specific fixed and random effects. This extended model was implemented in R package 'lcmm'. The package used Marquardt algorithm in order to minimize the EM-algorithm related limitations found in Verbeke's method. Immediately following Verbeke et al, Muthen et al (1999) introduced a similar method that combines multilevel mixed effects and mixture models and implemented the software Mplus. Muthen's method gained wide popularity as the growth

mixture model (GMM) and Mplus has been widely used, especially in the psychometric and socio-behavioral studies, for over a decade in classification of longitudinal data. Both of these methods are built on the same concept: specifically, the combination of linear mixed effects and mixture models and both methods have some inherent limitations. The method becomes mathematically and computationally complex fairly quickly. Computation time increases with sample size, degree of unbalancedness, complexity of the parametric curve, and number of random effects. The method requires pre-specification of the number of subgroups and typically uses information criteria (e.g., BIC) for the selection of number of classes, but definitive determination is exploratory. Log-likelihood functions may have local maxima; therefore, a careful choice of the initial values is crucial for ensuring convergence toward the global maxima. In Mplus, the programming structure lends itself to the situation of limited number of repeated observations per subject at a common set of measurement occasions across individuals. For computational feasibility, observations may need to be thinned and aligned to a common set of time points. In the R `lmm` package, the algorithm may reach the highest number of iterations without convergence. Another serious limitation is that the methods are unavailable in mainstream statistical software packages other than R and Mplus. Taken together, it seems that although theoretically sound, methods implemented in R and Mplus do not fully resolve the problem of clustering unbalanced longitudinal data because of the computational complexities. In a recent application, we have determined that application of post-hoc finite mixture models to the empirical best linear unbiased predictor (eBLUP) from linear or piecewise linear mixed effects model can reasonably classify heterogeneous trajectories of hidden components in distinct subgroups. Theoretical ground of this functionality is obvious as vectors of BLUPs account for the heterogeneities in shapes across individual-level trajectories, and mixture models classify individuals based on these heterogeneities (Hossain 2016). In an application on the real data, the method produced more convincing results in classifying distinct trajectories of early childhood growth patterns in 5 datasets of consisting of 2-3 components of linear, quadratic and cubic trends with varying level of separability than that of using HLME and GMM (Hossain et al, 2019). Following the internal validation of cluster evaluation, in this study, we used classifications of each method and component-based fitted piecewise linear mixed effects model to simulate 200 copies for each of the above 5 datasets of early childhood growth patterns to assess the agreement between classification of real and simulated data for each method.

**2. Overview of heterogeneous linear mixed effects model:** The general form of the linear mixed effects model is,  $Y_i = X_i\beta + Z_i b_i + \epsilon_i$ ,  $i = 1, 2, \dots, N$ ; where,  $\beta$  is the vector of fixed effects that describes the shapes of average trajectories over all individuals under study;  $b_i \sim MN(0, D)$  is the vector of random coefficients related to the  $i$ th subject;  $\epsilon_i \sim MN(0, R_i)$  is the vector of measurement or sampling errors associated with the responses of the  $i$ th subject. Vectors  $b_i$  and  $\epsilon_i$  are assumed to be independent. The distributional assumption of  $b_i$  implies that individual-level trajectories are homogeneous in shapes, thus the model under this assumption is termed as homogeneous linear mixed effects model. Under this form of the model,  $Y_i \sim MN(X_i\beta, Z_i' D Z_i + \sigma^2 I = V_i)$  and the empirical best linear unbiased predictor (BLUP) of random coefficients for given data ( $Y_i$ ) is,  $\hat{b}_i = \hat{E}(b_i | Y_i) = \hat{D} Z_i' \hat{V}_i^{-1} (Y_i - X_i \hat{\beta})$ . This normality assumption of  $b_i$ ,  $i = 1, 2, \dots, n$ ; could be violated in presence of the hidden subgroups or heterogeneous shapes of individual trajectories. In case of the presence of subgroups in shapes of trajectories, Verbeke et al (1996) suggested a Gaussian mixture distribution of  $b_i$  as  $b_i \sim \sum_{g=1}^G \pi_g MN(\mu_g, D)$  with  $\pi_g$  is the component probability,  $g = 1, 2, \dots, G$ ,  $\mu_g$  is the component specific mean and  $D$  is the common variance-covariance matrix. Then the

covariance matrix of  $b_i$ ,  $D^* = D + \sum_{g=1}^G \pi_g \mu_g \mu_g' - \sum_{l=1}^G \sum_{g=1}^G \pi_l \pi_g \mu_l \mu_g'$ , and the distribution of  $Y_i \sim \sum_{g=1}^G \pi_g MN(X_i \beta + Z_i \mu_g, V_i)$ . It is obvious that a restriction of  $\sum_{g=1}^G \pi_g \mu_g = 0$ , makes  $E(Y_i) = X_i \beta$  which is the mean trajectories under homogeneous model. Under the mixture model, the expression of the empirical BLUP becomes  $\hat{b}_i = E(b_i | Y_i, \varphi) = \hat{D} Z_i' \hat{V}_i^{-1} (Y_i - X_i \hat{\beta}) + A_i \sum_{g=1}^G \pi_{ig}(\varphi) \mu_g$ ; where  $A_i = I - D Z_i' V_i^{-1} Z_i$ ;  $\theta$  is the vector of parameters of  $\beta, \sigma, D$  and  $V_i$ , and  $\pi_{ig} = \pi_{ig}(\varphi) = \frac{\pi_g f(Y_i | \theta)}{\sum_{g=1}^G \pi_g f(Y_i | \theta)}$ ;  $\varphi' = (\pi', \theta')$  is the posterior probability for the  $i$ th individual to belong to the  $g$ th component of the mixture. The second part of the expression of  $\hat{b}_i$  is the correction term toward the component means, proportional to the posterior probability of belonging to each component. The model described above is referred to as the heterogeneous mixed effects model and has been used for the classification of longitudinal data (Verbeke, 1996). Proust-Lima et al. (2017) defined an extension of this model in which both fixed and random effects can be mixture component specific with  $b_i \sim \sum_{g=1}^G \pi_g MNV(\mu_g, D_g)$ ;  $D_g = w_g D$ , where  $w_g$  is the class-specific intensity of individual variability. They also replaced EM algorithm by Marquardt algorithm to improve computational efficiency and implemented to minimize the EM algorithm-related limitations; and implemented in the R package 'lcm'.

Immediate following Verbeke et al introduced to HLME model, Muthen et al (1999) expanded the concept through latent variable mixed effects models from the structural equation modeling approach and accommodated many linear and nonlinear mean functions over time. The method was implemented in Mplus and received popularity as growth mixture model (GMM).

For a simple linear growth curve implemented in Mplus, Muthen et al described the responses of  $i$ th individual who belongs to latent class  $c_i = g$  as

$$Y_{ij} | c_i = g = \eta_{0i} + \eta_{1i} t_{ij} + \epsilon_{ij}$$

$$\eta_{ki} | c_i = g = \alpha_{kg} + \gamma_{kg} x_i + \xi_{ki}, k = 0, 1$$

where  $\eta_{0i}$  and  $\eta_{1i}$  are random intercepts and slopes,  $\alpha_{0g}$  and  $\alpha_{1g}$  are the average intercept and slope of time varying variables,  $\gamma_{0g}$  and  $\gamma_{1g}$  intercepts and slopes of time invariant variables associated with latent class  $g$ , and  $c_i$  is the latent categorical random variable with probability of the unobserved class membership of the  $i$ th subject,  $\Pr(c_i = g) = \pi_{ig}$ . This probability follows the multinomial logistic regression with respect to time invariant covariates  $X_{ci}$  associated with  $i$ th subject as

$$\pi_{ig} = P(c_i = g | X_{ci}) = \frac{e^{\gamma_{0g} + X_{ci}^T \gamma_{1g}}}{\sum_{l=1}^G e^{\gamma_{0l} + X_{ci}^T \gamma_{1l}}}, \pi_{ig} \geq 0.$$

To obtain the desired clustering given a pre-specified number of classes, subjects can be assigned to their most likely class based on the posterior probabilities of class membership as,

$$P(c_i = g | X_{ci}, Y_i) = \frac{\pi_{ig} P(Y_i | c_i = g)}{\sum_{l=1}^G \pi_{il} P(Y_i | c_i = l)}$$

Model parameters are estimated using an EM algorithm from the likelihood function of mixture distributions for a given pre-specified number of classes. An overview of the model framework and estimation procedure is available in a book of Taylor & Francis Groups book's "Longitudinal Data Analysis" (edited by G Fitzmaurice, M Davidian, G Verbeke, G Molenberghs).

**3. Post-hoc Mixture Modeling of BLUPs:** This method is built on capturing the heterogeneity in individual trajectories through the empirical BLUP (eBLUP) from the fit of a suitable linear mixed effects model and then applying the conventional mixture model on the fitted eBLUP as a post-hoc analysis. Fixed effects in the linear mixed effects model describe the shape of the average trajectories over all individuals under study and the random coefficients explain the heterogeneity between trends in average and individual trajectories. Thus, the random coefficients convey heterogeneities across trajectories of all individuals respective to the shape of average trajectory. For individuals with trajectories of similar shapes would be expected to have similar pattern in heterogeneities, and thereby could safely be used as a primary sources of information for classification of trajectories. Without loss of generality fixed effects can be ignored for this purpose. Under the general form of linear mixed effects model, the random coefficients  $b_i, i = 1, 2, \dots, n$ , are normally distributed,  $b_i \sim MN(0, D)$ . This distributional assumption is not appropriate when there exist distinct subgroups of individual-level trajectories. Mixture model, a model-based cluster analysis technique, that assumes the distribution of the dataset under consideration of analysis as the mixture of several distributions could appropriately be used to identify subpopulations or distinct components of individual-level trajectories. Under this technique, the components are modeled separately, typically using the same parametric density family, and the overall population is modeled as a mixture or weighted sum of these subpopulations, using finite mixture models (Raftery & Dean, 2006). Gaussian mixture model is a particular case of mixture models when components are multivariate normal with different sets of parameters. Assuming a fixed number of  $G$  mixture components,  $b_i | k_i = k \sim MN(\gamma_k, D_k)$ , and the marginal distribution of  $b_i$  can be given as  $f(b_i) = \sum_{k=1}^G f(b_i, k_i = k) = \sum_{k=1}^G f(b_i | k_i = k) P(k_i = k) = \sum_{k=1}^G \pi_k f(b_i | k_i = k) = \sum_{k=1}^G \pi_k MN(\gamma_k, D_k)$ , where,  $\pi_k$  is the mixing weight or proportion of the population of the  $k$ th group with  $\sum_{k=1}^G \pi_k = 1$ , The likelihood function of  $\theta = (\gamma_1, \dots, \gamma_G, D_1, \dots, D_G, \pi_1, \dots, \pi_{G-1})$  for given  $b_i$  is  $L(\theta | b_1, b_2, \dots, b_n) = \prod_{i=1}^n \sum_{k=1}^G \pi_k MN(b_i; \gamma_k, D_k)$ . In reality, we observe data  $Y_i$ , not  $b_i$ . However, we use the empirical BLUP of  $b_i$  for given  $Y_i$ . That is, we use  $\hat{b}_i = \hat{E}(b_i | Y_i) = \hat{D} Z'_i \hat{V}_i^{-1} (Y_i - X_i \hat{\beta})$  as data for classification. Then the task of partitioning or clustering become that of estimating the parameters of the vector  $\theta$  and the posterior probability of cluster membership (Banfield et al. 1993; Everitt et al., 2001) by maximizing the log transformed above likelihood function (Dempster 1977). As the multivariate normal distribution are characterized by mean vector and covariance matrix, clusters or components are ellipsoidal, centered at the mean vector  $\mu_k$ , and with other geometric features, such as volume, shape and orientation, determined by eigen decomposition of the covariance matrix  $\Sigma_k$ . The covariance matrix of  $b_i$  can be decomposed as an eigen-decomposition through following parsimonious parameterisations,  $\Sigma_k = \lambda_k A_k D_k A_k$ , where,  $\lambda_k$  is a scalar, the largest eigenvalue of  $\Sigma_k$ , that controls the volume of the ellipsoid of the corresponding cluster,  $D_k$  is a diagonal matrix that specifies the shape of the density contours with  $\det(D_k) = 1$ , and  $A_k$  is an orthogonal matrix which determines the orientation of the corresponding ellipsoid (Banfield et al. 1993; Celeux et al, 1995). The volume, shape, and orientation can be constrained to be

equal or variable across groups, and thereby, 18 possible models with different geometric characteristics can be specified (Fraley et al. 2016). We have used R package mclust version 5 for Gaussian mixture analysis. (Scrucca et al. 2016). A total of eighteen models are analyzed simultaneously by the mclust software for one through nine clusters (this default can be increased or decreased), and each model is compared against others using the Bayesian Information Criterion (BIC) as internal validation of this cluster analysis. If the best fitting model indicates one cluster, then the dataset is multivariate normal and does not contain a mixture of heterogeneous subpopulations. Besides internal validation of classification, we also used standard methods of external validation of classification problem. Specifically, we assessed the level of agreement between the classification of original data and classifications of 200 simulated datasets using distribution of mixing components and mixing properties.

#### 4. Classification Indices and External Validation of Classifications:

This study assessed the agreement between classifications of real and 200 simulated datasets for each method for each of the five datasets. Six standard external validation methods have been used in this study to evaluate the agreement between real and simulated data.

- i. Agreement Proportion between Classification Vectors: Calculates the agreement proportion between two classification vectors. The value of this index ranges between 0 to 1, with 1 indicating perfect agreement.
- ii. Variation of Information (VarInf) in Two Partitioning Vectors: This is also known as shared information index, and is a measure of the distance between two clusterings. Suppose,  $c_1$  and  $c_2$  are two classifications, then  $\text{VarInf}, VI(c_1, c_2) = H(c_1) + H(c_2) - 2I(c_1, c_2)$ , where,  $H(c_i)$  is the entropy associated with  $c_i$  defined as  $H(c_i) = -\sum_{k=1}^{K(i)} \frac{n_k^{(i)}}{n} \log \frac{n_k^{(i)}}{n}$  and  $I(c_1, c_2) = \sum_{k=1}^{K(1)} \sum_{r=1}^{K(2)} \frac{n_{kr}}{n} \log \frac{n_{kr}n}{n_k^{(1)}n_r^{(2)}}$  represents the mutual info between two classes  $c_1$  and  $c_2$ .  $VI(c_1, c_2) = 0$ , for  $c_1 = c_2$ , and  $\log n$  is the upper bound of this index (Meilă 2003, 2007).
- iii. Rand Index (RI) and iv. Adjusted Rand Index (ARI): RI is a measure of the similarity between two classifications or clusters, and is derived as the proportion of number of pairs agree in two classifications divided by the total possible number of pairs. Specifically, if  $N_{11}$  be the number of pairs that are in the same group in both classifications,  $N_{00}$  be the number of pairs that are in different clusters in both classifications, and  $N_{10}$  and  $N_{01}$  be the number of pairs in the same group in one classification and in different groups in the other classification, then  $RI = \frac{N_{11} + N_{00}}{N_{11} + N_{00} + N_{10} + N_{01}} = \frac{N_{11} + N_{00}}{\binom{N}{2}}$ , where  $N$  are the number elements in a classification. ARI is the corrected version for chance of RI as the following expression,  $(RI - E(RI)) / (1 - E(RI))$ , where  $E(RI)$  is the expected value of RI. In the above notations,  $ARI = \frac{2(N_{11}N_{00} - N_{10}N_{01})}{(N_{11} + N_{01})(N_{00} + N_{01}) + (N_{11} + N_{10})(N_{00} + N_{10})}$ . The ARI have an upper bound equal to 1 when two comparing class variables perfectly match (Rand 1971, Vinh et al 2010), and equals 0 when the RI equals it's expected value. The ARI is bounded above by 1 when two classifications agree perfectly, and equals 0 when the RI equals its expected value (under the generalized hypergeometric distribution assumption for randomness) indicating only chance agreement between two classification vectors.

- v. Mirkin Index: This is a variation of the Rand Index and can be expressed as  $MC(c_1, c_2) = n(n-1)(1 - R(c_1, c_2))$ . MC approaches to 0 in this situation when  $c_1 = c_2$ , otherwise it takes a positive integer value.
- vi. Fowlkes and Mallows Index: This is an external evaluation method that is used to determine the similarity between two clusterings, and also a metric to measure confusion matrices. It can be expressed as  $F = \frac{\sqrt{W_1(c_1, c_2)W_2(c_1, c_2)}}{\sum_{k=1}^{K^{(i)}} n_k^{(i)}(n_k^{(i)}-1)}$ ;  $n_k^{(i)}$  size of the  $k$ th cluster in  $c_i$ ,  $K^{(i)}$  is number cluster in  $c_i$ ; In simple terms,  $F = \sqrt{PPV \cdot TPR}$ , where PPV is the positive predictive value and TPR is the true positive rate. F has upper bound equal to 1 when,  $c_1 = c_2$  (Fowlkes et al. 1983).

**4. Real Data Applications:** We have used five datasets consisting of trajectories of 2-3 distinct components with varying level of separability to compare classification performance of three methods. The datasets of mixture distributions are created by combining components of plausible homogeneous patterns of linear, quadratic and cubic trends of early childhood growth trajectories identified from a large dataset of 3,365 children. The dataset consists of the standardized scores of weight-for-length (at ages  $< 2$  years) and body mass index (BMI) (at ages  $\geq 2$  years), collected on clinic visits during their first 5 years of life. In the United States, weight-for-length and BMI are common measures of the somatic growth of children aged  $< 2$  years and  $\geq 2$  years, respectively. Because the same quantile cutoffs of the two variables are used to classify the weight status of children, the standardized score of this variable, denoted BMIz, has been used as an early childhood growth indicator. The data were retrospectively retrieved from electronic health records. Using the identified plausibly homogeneous subsets, we generated 5 datasets of two or three components with varying extent of separability. We generated classifications of 2-4 groups using three methods: post-hoc mixture model of BLUPs using R-package mclust (BLUP), HLME, and GMM in Mplus. In real data applications, post-hoc mixture model of BLUPs emerged as the best in classifying longitudinal data with unequal intervals of measurement time.

**5. Simulation and Classification:** As discussed, we recently used 5 datasets of early childhood growth patterns of varying separability as application of real datasets to compare the classification performance of the three methods, namely HLME, GMM in Mplus and post-hoc mixture modelling of BLUPs (Hossain et al. 2019). The piecewise linear mixed effects model identified as the best fit to a dataset was used for classification of the corresponding dataset using all three methods. The mixed effects model originally used for classification is applied to the dataset stratified by groups generated by a classification method to produce 200 sets of simulated data for each method and subsequently the corresponding method is applied to classify all of the 200 simulated datasets. Each method of classification was used in simulation and subsequent classification of simulated datasets for each real dataset to assess the agreement between classification of the real and simulated data sets to assess the agreement between classification of the real and simulated data sets. As mentioned, six methods of classification evaluation indices were used to compare the agreement between the classifications of original and simulated datasets.

Dataset 1 consists of three quite distinct components (Figure 1) of trajectories with plausibly linear, quadratic, and cubic trends of BMIz. Figure 1 shows the average agreement between classification of real and simulated datasets using post-hoc mixture modeling of BLUPs, HLME, and GMM in Mplus. Mean agreement proportion between real and simulated data classifications lies between 0.948 (HLME) and 0.9759 (post-hoc

mixture modeling of BLUPs) in three methods. Similarly, mean variation of information index lies between 0.2114 (post-hoc mixture modeling of BLUPs) and 0.307 (GMM in Mplus), mean Rand index and adjusted Rand index ranges 0.9361 (GMM in Mplus) to 0.9638 (post-hoc mixture modeling of BLUPs), and 0.8721 (GMM in Mplus) to 0.9264 (post-hoc mixture modeling of BLUPs), respectively. Also, mean Fowlkes and Mallows index ranges between 0.9319 (HLME) and 0.9586 (post-hoc mixture modeling of BLUPs), and mean Mirkin index ranges between 6406 (post-hoc mixture modeling of BLUPs) and 11303 (GMM in Mplus), respectively. The six indices showed reasonably nice agreement between real and simulated data classifications using all methods for this dataset and the classification using post-hoc mixture model was identified as the best. The performance of HLME and GMM was quite similar with negligible differences in indices with no consistent superiority of any of these two methods.

Dataset 2 also consists of three well separable components (Figure 2) of trajectories with linear and two opposite quadratic trends of BMIz. Figure 2 also exhibits the classification indices to compare the similarities between classifications of real and simulated datasets. Post-hoc mixture modeling and HLME showed an excellent level of agreements, and GMM in Mplus performed acceptable level of agreements between classes of original and simulated data. Once again, post-hoc mixture modeling performs the best in terms of the similarity measures between real and simulated datasets, followed by HLME and then GMM in Mplus.

Dataset 3 comprises three relatively less separable components (Figure 3). Component 1 consists of a weak cubic mean trend in BMIz, components 2 and 3 contains with mean trends of opposite patterns at the beginning of the life. Once again, the post-hoc mixture model of BLUPs showed the highest similarities between classifications of original and simulated data with mean indices of agreement proportion, Rand index, adjusted Rand index and Fowlkes and Mallows index ranging 0.9164 to 0.9719, followed by HLME that ranges 0.8462 to 0.9238. GMM in Mplus showed a poor agreement between real and simulated datasets with indices ranging 0.4259 to 0.7194.

Dataset 4 is composed of two clearly diverged trends of BMIz with opposite quadratic curves (Figure 4). Both post-hoc mixture modeling of BLUPs and GMM in Mplus showed excellent agreement between classifications of original and simulated data with mean indices of agreement proportion, Rand, adjusted Rand and Fowlkes and Mallows to be closed to 1. HLME showed poor agreement between classifications of original and simulated data with above agreement indices ranging from 0.158 (adjusted Rand index) to 0.764 (Fowlkes and Mallows index).

Dataset 5 also consists of two reasonably separable components of BMIz trajectories (Figure 5). The post-hoc mixture model of BLUPs once again showed strong agreement between classifications of real and simulated data with 4 indices ranging between 0.9873 (adjusted Rand index) to 0.9968 (agreement proportion). Both HLME and GMM in Mplus showed poor agreement between classifications of real and simulated datasets with indices ranging 0.5094 to 0.8348 and 0.2076 to 0.7579, respectively.

## **5. Discussions and conclusions:**

In real data application to 5 datasets of early childhood growth patterns, the post-hoc mixture model of BLUPs showed superior performance in identifying distinct subgroups of trajectories in longitudinal data with unequal-spaced intervals compared to existing

methods of HLME and GMM in Mplus although the same piecewise linear mixed effects model is used in all three methods of classification. As discussed, this study conducted classification-specific simulations of real dataset and used the corresponding method for classification of simulated datasets, and subsequently examined the level of agreement between classifications of real and simulated datasets as the external validation of clustering performance of three methods. In other words, this study compared the ability of classifying similar structures in real and simulated datasets in the same group.

Likewise in real data applications, post-hoc mixture model of BLUPs showed superior performance in agreement between classifications of real and simulated datasets compared to that in two existing methods of HLME and GMM in Mplus. The method demonstrated excellent performance in identifying similar components of distinct trajectories in real and corresponding simulated datasets. As mentioned, a total of six indices used to examine the similarity in classifications of 5 datasets of early childhood growth patterns and the corresponding simulated datasets. Comparison of classification of each simulated dataset with that of the corresponding real dataset provided a score for each index and the mean score of an index reflect the average of the 200 scores produced through the comparisons of classifications of 200 simulated datasets with that of the real dataset. Four of the six indices such as agreement proportion between classification vectors, Rand index, adjusted Rand index, Fowlkes and Mallows index would have an upper bound of 1 when classification vector of a simulated dataset is identical to that of the corresponding real dataset. The mean score of all these 4 indices is close to 1 for post-hoc mixture modelling for all 5 datasets and is also much higher than that for two other methods. Specifically, the mean agreement proportion between post-hoc mixture modelling classifications of real and corresponding simulated datasets are 0.9759, 0.9937, 0.9719, 0.99996, and 0.9968 for datasets 1, 2, 3, 4 and 5, respectively. The corresponding mean agreement proportion using HLME classification is 0.948, 0.9238, 0.9633, 0.7173, and 0.8348, and using GMM in Mplus classification is 0.9551, 0.5854, 0.888, 0.9997, and 0.7122, respectively. Similarly, the mean Rand index to measure the similarity between classifications of real and simulated datasets are 0.9638, 0.9623, 0.9896, 0.9999, and 0.9937 using post-hoc mixture modelling of BLUPs, 0.9401, 0.923, 0.9598, 0.6409, and 0.7674 using HLME and 0.9361, 0.7194, 0.855, 0.9993, and 0.9993 for datasets 1, 2, 3, 4, and respectively 5. After adjustment for randomness, the adjusted Rand index for 5 datasets becomes 0.9264, 0.9164, 0.9791, 0.9998, and 0.9873 for post-hoc mixture modelling of BLUPs, 0.8784, 0.8462, 0.9195, 0.158, and 0.5094 for HLME, and 0.8721, 0.4259, 0.7115, 0.9986, and 0.2076 for GMM in Mplus. These three indices indicate that post-hoc mixture modelling of BLUPs performed outstanding performance in classifying similar structures of real and simulated data in the same group for all datasets. HLME showed excellent performance in one dataset, acceptable performance in two datasets and poor performance in two datasets to classify similar structures of real and simulated datasets. GMM also demonstrated excellent performance in one dataset, acceptable/ marginally acceptable performance in two datasets and poor performance in two datasets to identify classes of the similar structures in real and simulated datasets. The findings of the three indices are well supported by the mean Fowlkes and Mallows index, mean variation of information index, and mean Mirkin index.

In summary, classifications of real and simulated datasets demonstrated that the post-hoc mixture modelling of BLUPs from linear/ piecewise linear mixed effects model performs better in clustering longitudinal data with unequal intervals than two existing methods. The results of this study confirmed the findings of the real data application to compare the performance of three methods in classifying longitudinal unbalanced data. More application of three methods to real and simulated datasets could provide further insight of

strength and weakness of these methods in classifying longitudinal data with irregular spaces between measurement times.

**Acknowledgments**

The first author received partial salary support from the National Institutes of Health (NIH) COBRE Grant 8P20GM103464-9 (PI: Shaffer) and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the NIH under Grant no. U54-GM104941 (PI: Binder-Macleod).

**References:**

- Arminger, G., Clogg, C.C. and Sobel, M.E. eds., 2013. Handbook of statistical modeling for the social and behavioral sciences. Springer Science & Business Media.
- Banfield, J.D. and Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pp.803-821.
- Fraley, C., Raftery, A. E. and Scrucca, L, 2016. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation, 2016. URL <https://CRAN.R-project.org/package=mclust>. R package version 5.2. [p289]
- Celeux, G. and Govaert, G., 1995. Gaussian parsimonious clustering models. *Pattern recognition*, 28(5), pp.781-793.
- Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), pp.1-22.
- Fitzmaurice, G.M., Laird, N.M. and Ware, J.H., 2012. Applied longitudinal analysis (Vol. 998). John Wiley & Sons.
- Fowlkes, E.B. and Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383), pp.553-569.
- Garrett, E.S. and Zeger, S.L., 2000. Latent class model diagnosis. *Biometrics*, 56(4), pp.1055-1067.
- Hossain, M. J. 2016. Clustering Longitudinal Unbalanced Data: An Application to the Early Childhood Growth Pattern, Section on Statistics in Epidemiology. Alexandria, VA: American Statistical Association, pp.
- Hossain, M.J. and Leiby, B. E. 2019. Classification of Distinct Trajectories in Longitudinal Data With Irregular Spaced Intervals: Heterogeneous Linear Mixed Model vs. Mixture Modeling of BLUPs From Linear Mixed Model. In JSM Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association. 2054-2071.
- Jha, A.K., DesRoches, C.M., Campbell, E.G., Donelan, K., Rao, S.R., Ferris, T.G., Shields, A., Rosenbaum, S. and Blumenthal, D., 2009. Use of electronic health records in US hospitals. *New England Journal of Medicine*, 360(16), pp.1628-1638.
- Jung, T. and Wickrama, K.A.S., 2008. An introduction to latent class growth analysis and growth mixture modeling. *Social and personality psychology compass*, 2(1), pp.302-317.
- Komárek, A., Verbeke, G. and Molenberghs, G., 2002. A SAS-Macro for Linear Mixed Models with Finite Normal Mixtures as Random-Effects Distribution, version 1.1. *Katholieke Universiteit Leuven Biostatistisch Centrum*. <http://www.med.kuleuven.ac.be/biostat/research/software.htm>, 2001.
- Laird, N.M. and Ware, J.H., 1982. Random-effects models for longitudinal data. *Biometrics*, 38(4), pp.963-974.
- Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2), pp.431-441.
- Meilă, M., 2003. Comparing clusterings by the variation of information. In *Learning theory and kernel machines* (pp. 173-187). Springer, Berlin, Heidelberg.
- Muthén, B. and Shedden, K., 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2), pp.463-469.

- Muthén, B. and Asparouhov, T., 2008. Growth mixture modeling: Analysis with non-Gaussian random effects.
- P. Browne, A. ElSherbiny, and P. D. McNicholas. mixture: Mixture Models for Clustering and Classification, 2015. URL <https://CRAN.R-project.org/package=mixture>. R package version 1.4.
- Proust-Lima, C., Philipps, V. and Liquet, B., 2015. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. arXiv preprint arXiv:1503.00890.
- Proust, C. and Jacqmin-Gadda, H., 2005. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer methods and programs in biomedicine*, 78(2), pp.165-173.
- Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), pp.846-850.
- Scrucca, L., Fop, M., Murphy, T.B. and Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, 8(1), p.289.
- Spießens, B., Verbeke, G. and Komárek, A., 2002. A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed models. <http://www.med.kuleuven.ac.be/biostat/research/software.htm>, 2002.
- Verbeke, G. and Lesaffre, E., 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433), pp.217-221.
- Vinh, N.X., Epps, J. and Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11, pp.2837-2854.
- Wang, M. and Bodner, T.E., 2007. Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, 10(4), pp.635-656.
- Wei, W.Q. and Denny, J.C., 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine*, 7(1), p.41.

Figure 1: Classification and Performance -Dataset 1

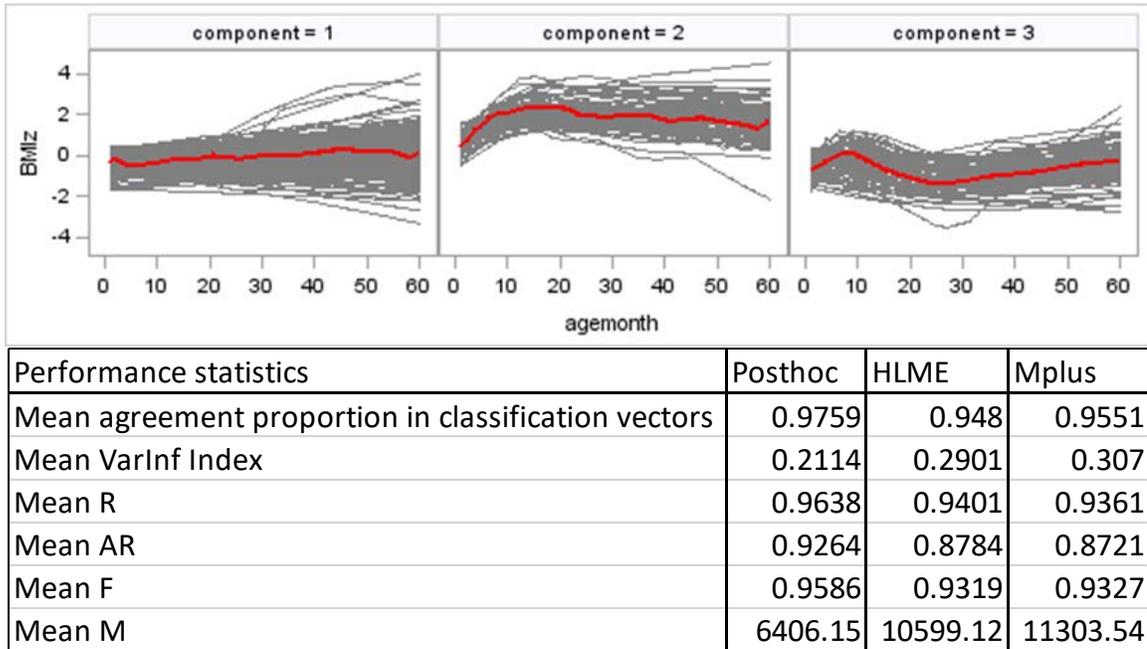


Figure 2: Classification and Performance- Dataset 2

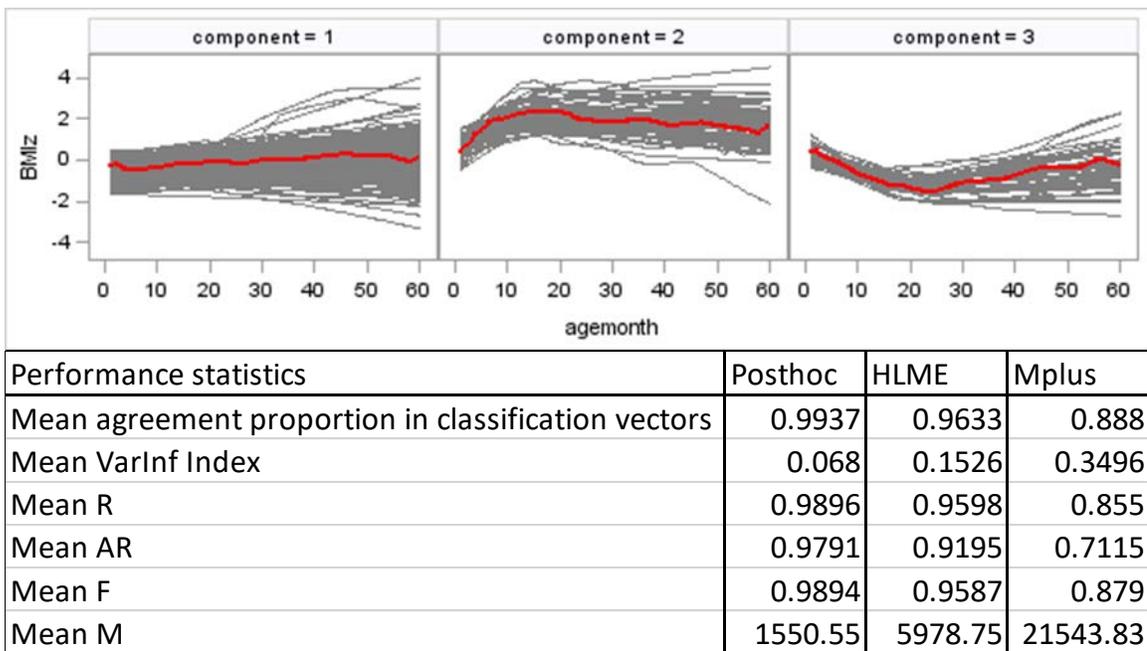


Figure 3: Classification and Performance- Dataset 3

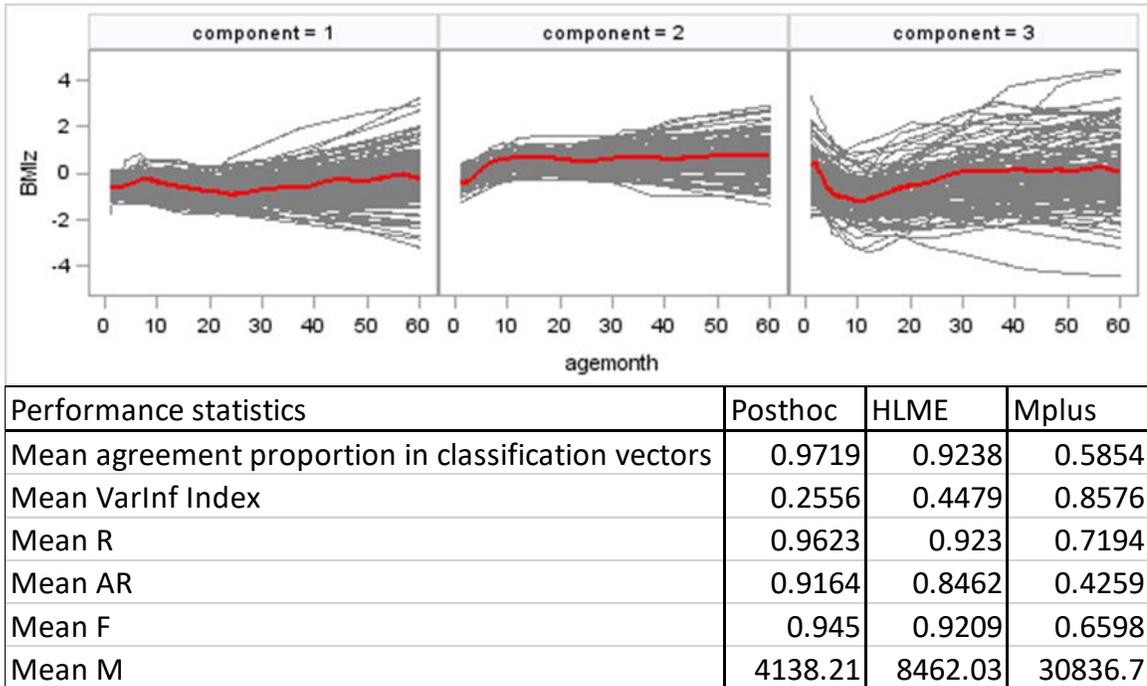


Figure 4: Classification and Performance- Dataset 4

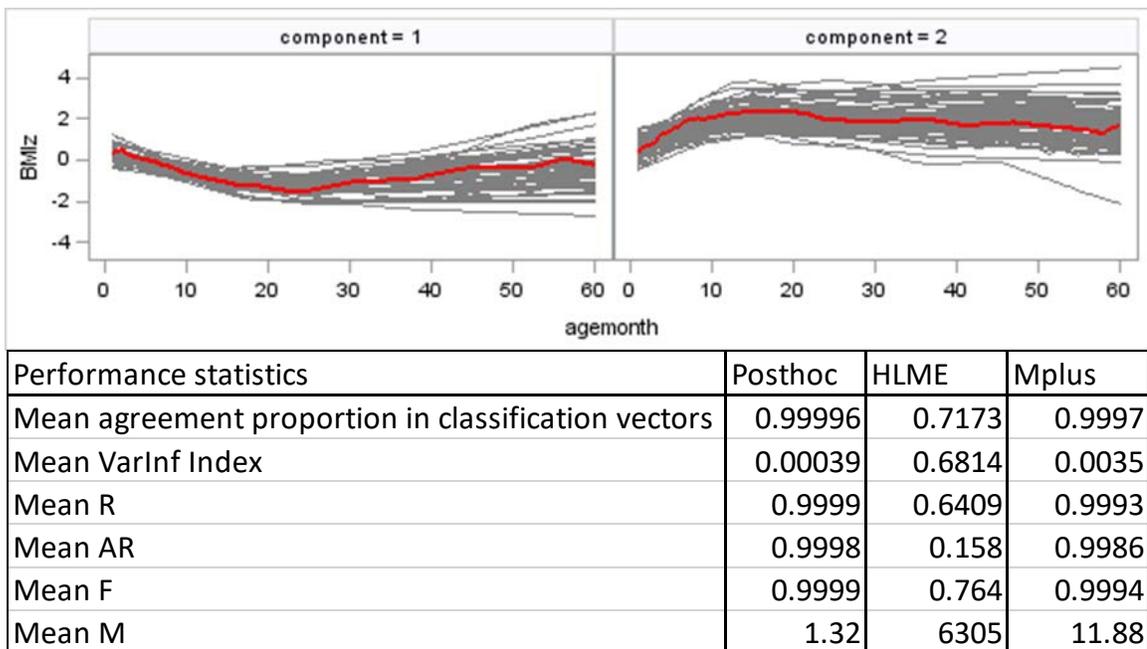


Figure 5: Classification and Performance- Dataset 5

