Comparison of methods for clustering in longitudinal categorical data

JeongSook Kim

The Graduate School
Yonsei University
Department of Biostatistics and Computing

Comparison of methods for clustering in longitudinal categorical data

A Master's Thesis
Submitted to the Department of Biostatistics and Computing
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Master of Science

JeongSook Kim

June 2024

This certifies that the master's thesis of ***JeongSook Kim*** is approved.

Inkyung Jung: Thesis Supervisor

ChungMo Nam: Thesis Committee Member #1

Jisu Moon: Thesis Committee Member #2

The Graduate School

Yonsei University

June 2024

Contents

List of Tables

List of Figures

Abstract

Comparison of method for clustering in longitudinal categorical data

The longitudinal data analysis with a categorical dependent variable frequently occurs in research, offering insights into individual patterns and facilitating tailored interventions. However, compared to longitudinal continuous data, there has been limited exploration of methodologies for analyzing longitudinal categorical data.

The study explores methodologies for identifying similar patterns in categorical dependent variables across diverse contexts. Simulations were conducted to generate longitudinal binary data, employing models with random intercepts, multivariate binary models, and Grouped Generalized Estimating Equation models. Results indicate the group-based trajectory model consistently outperformed others in accurately estimating cluster numbers. However, limitations were identified in representing binary data, particularly in trajectories with four clusters. Performance metrics such as the adjusted Rand index were used but raised doubts about adequacy, urging the need for more comprehensive evaluation metrics.

Keywords: Longitudinal data, Group GEE, Group-based Trajectory model, Growth mixture model, Trajectory clustering

1. Introduction

In various health studies, measured outcomes are typically aggregated and analyzed across the entire study population or predefined subgroups. However, in most cases, unknown or unexpected subgroups exhibit similar patterns of clinical symptoms, behaviors, or healthcare utilization. Therefore, relying solely on mean estimates to simplify the complex intra- and inter-individual variability may underestimate the intricacies of the real-life clinical context.

Additionally, analyzing longitudinal data with categorical dependent variables is a common research practice, providing valuable insights into individual patterns and facilitating customized applications. However, compared to continuous data, methodologies for analyzing longitudinal categorical data have been relatively limited in exploration.

This paper aims to compare and explore methodologies for analyzing longitudinal categorical data by applying growth mixture models, group-based trajectory models, and Group GEE models.

Grouped Generalized Estimating Equations (GGEE) represent an extension of the standard GEE analysis tailored to address potential heterogeneity within longitudinal data. This approach adopts a grouping mechanism commonly employed in panel data analysis literature (Ito, 2023). Specifically, GGEE models operate under the assumption that individuals within longitudinal datasets can be categorized into a finite number of groups. Within each group, individuals share identical regression coefficients, implying homogeneity in regression coefficients among individuals belonging to the same group. By implementing this grouping strategy, GGEE facilitates the exploration of nuanced variations within longitudinal data, accounting for potential differences across distinct groups while maintaining computational feasibility and interpretability (Ito, 2023).

Group-based trajectory modeling assumes that the entire population is composed of several groups experiencing different changes over time. It estimates the probability density function of individuals being assigned to specific trajectory groups at each time point. The probability density function is estimated based on the probability of subject i belonging to a specific group multiplied by the probability density function of the states of members within that group. This allows for the derivation of the probability of individual samples belonging to a specific group over time.

Growth mixture model (GMM) extends GBTM with the inclusion of parametric random effects, enabling a better fit to the data under the assumption of within-cluster variability (Grimm, 2009). The method is also described as a longitudinal latent-class mixed model, a multilevel mixture model, or a finite mixture of mixed models.

Our study endeavors to compare and contrast the efficacy of three methodologies in delineating the number of genuine clusters for a pre-established trajectory in data generated through distinct mechanisms: models incorporating random intercepts, multivariate binary distribution models integrating conditional expectations, and Grouped Generalized Estimating Equation (GEE) models. By scrutinizing the capacity of these methodologies to discern the true number of clusters and evaluating their respective performances, we aim to ascertain their effectiveness in practical settings.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of the methods used. Section 3 discusses the methods for determining the number of groups, while Section 4 describes the methods for evaluating clustering. Section 5 covers simulation, Section 6 focuses on application, and Section 7 presents the discussion.

2. Method

2.1 Longitudinal studies

Let Y_{ij} denote a response variable for the i^{th} individual $(i = 1,2, \dots, N)$ at j^{th} observation $(i = 1,2, \dots, n_i)$, where N indicates the total number of individuals and n_i indicates the number of observed responses on the i^{th} individual (Fitzmaurice, 2012). The model for changes in the mean response over time and for relating the changes to the covariates can be expressed as:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \varepsilon_{ij} \ , j = 1,2, \dots, n_i$$

where $\beta_1, \beta_2, \dots, \beta_p$ are unknown regression coefficients relating the mean of Y_{ij} to its corresponding covariates. Given we have n_i repeated measurements of the response variable on the same individual I, $n_i \times 1$ response vector Y_i is denoted as (Fitzmaurice, 2012):

$$Y_i = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \cdots \\ X_{ijp} \end{pmatrix} \ , i = 1,2, \dots, N; j = 1,2, \dots, n_i$$

Every row of X_{ij} corresponds to different covariates, and the covariates could be time-dependent or independent. The covariates' vectors could also be grouped into a $n_i \times p$ matrix of covariates (Fitzmaurice, 2012).

$$X_{ij} = (X_{ij1} \quad \cdots \quad X_{ijp}), \qquad i = 1,2,\dots,N; j = 1,2,\dots,n_i$$

Every row of X_{ij} corresponds to different covariates, and the covariates could be time-dependent or independent (Fitzmaurice, 2012).

$$X_i = \begin{pmatrix} X_{i1}' \\ X_{i2}' \\ \dots \\ X_{in_i}' \end{pmatrix} = \begin{pmatrix} X_{i11} & \dots & X_{i1p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{pmatrix}, \; i = 1,2,\dots,N$$

These covariates' vectors could also be grouped into a $n_i \times p$ matrix of covariates for the i^{th} individual at the j^{th} observation (Fitzmaurice, 2012). Lastly, $the \; n_i \times 1$ vector of random errors would be:

$$\varepsilon_{ij} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdots \\ \varepsilon_{in_i} \end{pmatrix}, i = 1,2,\dots,N$$

2.2 Trajectory model

Trajectory analysis refers to examining changes in variables over time or concerning age, and studying how covariates influence these trajectories. This analytical approach encompasses not only the trajectory of variables over time but also investigates how covariates shape these trajectories. In this paper, we present trajectory analysis as a method for studying longitudinal data, emphasizing its utility in uncovering patterns of change and understanding the effects of covariates on these patterns.

Trajectory modeling is currently a subject of considerable importance. It involves identifying patterns in variables that impact diseases, such as BMI, cholesterol, hypertension, and exercise habits. By discerning these patterns, it becomes possible to identify high-risk population groups and tailor interventions accordingly for individuals. Furthermore, leveraging these patterns enables effective disease prediction and preventive measures to be implemented.

In this paper, we will explore various methods for trajectory analysis, including group-based trajectory modeling, growth mixture modeling, and grouped GEE modeling. Through these methods, we aim to examine and compare the effectiveness of different approaches in capturing and interpreting longitudinal patterns in the data.

2.3 Finite mixture model

2.3.1 Introduction

Mixture models offer a framework for describing a distribution by positing a combination of underlying distributions, operating under the premise that the observed distribution is composed of multiple data-generating processes and random variables. Within this framework, the sub-models assume a common parametric distribution but with varying coefficients (N. G. P. Den Teuling, 2023).

In the context of longitudinal data analysis, a longitudinal mixture model aims to characterize the distribution of longitudinal observations Y_i. In this model, the mixture density $f(Y_i|\theta)$, where $\theta = (\pi, \theta_1, \dots, \theta_G)$ represents the model parameters, is defined as:

$$f(Y_i|\theta) = \sum_{g=1}^{G} \pi_g f(Y_i|\theta_g).$$

Here, $f(Y_i|\theta_g)$ denotes the conditional density of Y_i given that i belongs to cluster g (N. G. P. Den Teuling, 2023).

The probability of observing Y_i given that i belongs to cluster g and under model parameters θ is expressed as:

$$\Pr(Y_i | i \in I_g, \theta) = \frac{\pi_g f(Y_i|\theta_g)}{\sum_{g'=1}^{G} \pi_{g'} f(Y_i|\theta_{g'})}$$

This formulation reflects the probability of Y_i conditioned on membership in cluster g and the parameters θ, where π_g represents the mixing proportion associated with cluster g, and $f(Y_i|\theta_g)$ denotes the density function corresponding to cluster g (N. G. P. Den Teuling, 2023).

2.3.2 Model Estimation

For given membership in group g, the measurements of outcome Y_{ij} of subject i at time j is,

$$Logit[P(Y_{ij} = 1|C = g)] = \beta_{0g} + \beta_{1g} * Time_{ij} + \beta_{2g} * Time_{ij}^{2}$$

The link function varies depending on the form of Y. The probability of observing an individual $i's$ longitudinal sequence of behavioral measurements Y_i is

$P(Y_i) =$

$\sum_g \pi_g P^g(Y_g)$ *the probability of* Y_i *given membership in group* g *is* $P^g(Y_g)$ and the probability of membership in group g is π_g.

The probability $P^g(Y_g)$ can be obtained by multiplying the observed values of Y for the subject I at each repeated measurement occasion j within each group g. The values at the repeated measurement occasions are all assumed to be independent.

$$P^g(Y_g) = \prod_{j=1}^{J} P^g(Y_{ij}) = P^g(Y_{i1}) * P^g(Y_{i2}) * \ldots * P^g(Y_{ij})$$

Model estimation proceeds with the following equation, and maximum likelihood estimates are obtained using the EM-Quasi Newton method.

$$L = \prod_{0}^{N} P(Y_i) = \prod_{0}^{N} \sum_{g} \pi_g P^g(Y_g) = \prod_{0}^{N} \sum_{g} \pi_g \prod_{j=1}^{J} P^g(Y_{ij})$$

2.3.3 Group-based trajectory model (GBTM)

Group-based trajectory modeling assumes that the entire population is composed of several groups experiencing different changes over time. It estimates the probability density function of individuals being assigned to specific trajectory groups at each time point. The probability density function is estimated based on the probability of subject i belonging to a specific group multiplied by the probability density function of the states of members within that group. This allows for the derivation of the probability of individual samples belonging to a specific group over time. Depending on the characteristics of the dependent variable, the probability of being included in each type is estimated differently.

The dependent variables encompass a range of models including the censored normal model, Poisson-based model, and logit-based model. In this paper, our focus lies on discussing categorical data, particularly emphasizing the logit-based model. Additionally, the group-based trajectory model adheres to the following formula, as outlined in the paper:

$$Logit[P(Y_{ij} = 1|C = g)] = \beta_{0g} + \beta_{1g} * Time_{ij} + \beta_{2g} * Time_{ij}{}^2, \qquad i \in I_g$$

Here, β_{0g}, β_{1g}, and β_{2g} denote the cluster-specific regression coefficients.

2.3.4 Growth mixture model (GMM)

The growth mixture model extends the group-based trajectory model by incorporating random effects, allowing for within-group variability. It represents cases where distinct subgroups are delineated in previous theories.

Growth mixture model follows the following formula:

$$Logit[P(Y_{ij} = 1|C = g)] = \beta_{0g} + \beta_{1g} * Time_{ij} + \beta_{2g} * {Time_{ij}}^2 + Z_{ij}u_{ig},\ i \in I_g$$

$$u_{ig} \sim MVN(0, \Sigma_g)$$

Due to its flexibility, GMM is widely used, allowing researchers to specify random effects and their relationships, as well as include covariates (N. G. P. Den Teuling, 2023). However, this may lead to difficulties in identifying the most appropriate model.

2.4 Grouped Generalized Estimating Equations

For longitudinal data analysis, let Y_{ij} denote the response variable of interest and X_{ij} represent a p-dimensional vector containing covariate information of subject i at time j, where $i = 1, \dots, n$ and $j = 1, \dots, n_i$. For ease of notation, we set $J_i = J$ for all i, representing a balanced data case, but the extension to an unbalanced case is straightforward (Ito, 2023).

We adopt a generalized linear model for Y_{ij}, which is given by:

$$f(Y_{ij}|X_{ij;\beta,\phi}) = \exp[\{Y_{ij}\theta_{ij} - a(\theta_{ij}) + b(Y_{ij})\}/\phi]$$

where $a(\cdot) and\ b(\cdot)$ are known functions, and $\theta_{ij} = u(X_{ij}^T\beta_i)$ for a known monotone function $u(\cdot)$. By the model, the canonical link function $u(x) = x$ is commonly utilized (Ito, 2023).

Here, β_i denotes the regression parameter of interest, which may exhibit heterogeneity across subjects, while ϕ represents a known scale parameter shared among all subjects (Ito, 2023).

Under this model, the first two moments of Y_{ij} are expressed as $m(X_{ij}^T\beta_{ij}) = a'(\theta_{ij})$ and $\sigma^2(X_{ij}^T\beta_{ij}) = a''(\theta_{ij})\phi$, respectively. For instance, in the scenario of binary response, the function $a(x) = \log\{1 + \exp(x)\}$ is applied, leading to the logistic model formulated as $m(X_{ij}^T\beta_{ij}) = \{1 + \exp(-X_{ij}^T\beta_i)\}^{-1}$ (Ito, 2023).

In the standard Generalized Estimating Equations (GEE) analysis, the regression parameters are assumed to be homogeneous, meaning $\beta_i = \beta$, while allowing for potential heterogeneity among subjects. However, estimating β_i accurately becomes challenging as the number of subjects increases, especially when J is not sufficiently large, which is a common scenario in longitudinal data analysis (Ito, 2023).

To address this issue, we propose a grouped structure for the subjects, where the n subjects are divided into G groups Subjects within the same group share the same regression coefficients. Specifically, we introduce an unknown grouping variable g_i belonging to {1, …, G}, which determines the group to which the i^{th} subject belongs. We define $\beta_i = \beta_{g_i}$, where the unknown regression parameters are $\beta_1, \dots, \beta_G$. Therefore, if G is not considerably large compared to n and T, then $\beta_1, \dots, \beta_G$ can be estimated accurately. Additionally, due to the grouped nature, the estimation results of g_i provide a grouping of subjects in terms of regression coefficients, making the results easily interpretable for users. We also consider G as an unknown parameter, although we assume G to be known for the time being (Ito, 2023).

3. Number of groups

Determining the number of groups when clustering is a crucial issue. In practice, clusters are often indistinct, making it challenging to differentiate between subgroups and thereby cluster all subjects accurately.

In this paper, we aim to determine the number of groups using the cross-validation averaging method proposed in the group GEE model, as well as the widely used Bayes Information Criterion (BIC) in mixture models.

3.1 Cross-validation with averaging (CVA)

The cross-validation averaging method involves dividing the N subjects into three subsets. Two training sets of size M and one testing set of size N-2M are then created. Through the group GEE method, regression coefficients and working correlation matrices are estimated using the two training sets. The estimated regression coefficients and working correlation are then utilized to determine the optimal number of groups based on performance on the testing set.

The following formula sets the number of groups as the one that minimizes the equation, which involves substituting the regression coefficients and working correlation obtained from the training set into the test data. The formal representation of the equation is as follows:

$$\widehat{g_i}^{(h)} = argmin\left\{y_i - m\left(X_i\ \widehat{\beta_g}^{(h)}\right)\right\}^T \left\{\hat{R}^{(h)}\right\}^{-1}\left\{y_i - m\left(X_i\ \widehat{\beta_g}^{(h)}\right)\right\},$$

$$h = 1,2(training\ \ set)$$

where $\widehat{\beta_g}^{(h)}$ and $\hat{R}^{(h)}$ are estimates of regression coefficients and working correlation based on h^{th} training data for $h = 1,2$.

$$\widehat{S^c} = \sum_{i,j \in test\ \ data} 1\left\{1\left(\widehat{g_i}^{(1)} = \widehat{g_j}^{(1)}\right) + 1\left(\widehat{g_i}^{(2)} = \widehat{g_j}^{(2)}\right) = 1\right\}$$

By averaging, we select g as the minimizer of the criterion among some candidates of g.

3.2 Bayesian information criterion (BIC)

In mixture models, the Bayes Information Criterion (BIC) is widely applied, and the number of groups is determined by selecting the smallest BIC value. The BIC is defined by Nagin as follows

$$BIC = logL(\hat{\pi}, \hat{\theta}) - 0.5plog(n),$$

where p is the number of parameters of the model, $n\ is$ the number of patients, and $L(\hat{\pi}, \hat{\theta})$ the likelihood of the model, evaluated at the maximum likelihood estimates. In model selection, BIC tends to favor models with a greater number of groups, prompting the proposal of the bootstrapped likelihood ratio test as an alternative approach.

4. Evaluation

When conducting evaluations, we primarily employ the Adjusted Rand Index for scenarios where the actual number of clusters is known, while utilizing methodologies relying on distance metrics like the Calinski-Harabasz Index and Davies-Bouldin Index for cases where the actual number of clusters is unknown.

Firstly, we determine the number of groups using the Bayesian Information Criterion (BIC) for the Group-Based Trajectory Model and Growth Mixture Model, and through cross-validation with averaging for the Grouped Generalized Estimating Equations. Subsequently, we evaluate the performance using these three evaluation metrics. The higher the values of the Adjusted Rand Index and the Calinski-Harabasz Index, the better the performance. Conversely, a lower value of the Davies-Bouldin Index indicates better performance.

4.1 Adjusted Rand index

The Adjusted Rand Index (ARI) serves as a widely adopted metric to evaluate the resemblance between two clustering outcomes. Given a set of n objects $S = \{O_1, \dots, O_n\}, suppose\ U = \{u_1, \dots, u_R\}$ and $V = \{v_1, \dots, v_C\}$ show two distinct partitions of the objects in S such that $\bigcup_{i=1}^{R} u_i = S = \bigcup_{j=1}^{C} v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \le i \ne i' \le R$ and $1 \le i \ne i' \le C$. Assuming U as the external criterion and V as a clustering result, let a denote the count of object pairs placed in the same class in U and the same cluster in V, b be the count of object pairs in the same class in U but not in the same cluster V, c be the count of object pairs in the same cluster in V but not in the same class in U, and d be the count of object pairs in different classes and different clusters in both partitions. The quantities a and d can be interpreted as agreements, and b and c as disagreements. The Rand index is simply $\frac{a+d}{a+b+c+d}$. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1 (Gao, 2023).

One limitation of the Rand index is that the expected value of the Rand index of two random partitions does not remain constant. The adjusted Rand index proposed by (Hubert and Arabie, 1985) assumes the generalized hypergeometric distribution as the model of randomness, the U and V partitions are randomly such that the number of objects in the classes and clusters is fixed (Gao, 2023).

Let n_i and n_j denote the number of objects in class u_i, and cluster v_j respectively. The general form of an index with a constant expected value is

$\frac{index - expected\ index}{maximum\ index - expected\ index}$, which is bounded above by 1, and takes the value 0 when the index equals its expected values (Gao, 2023).

4.2 Calinski-Harabasz index

The Calinski-Harabasz index consists of a numerator representing between-group variation and a denominator representing within-group variation. If clustering is done well, each cluster should be as far apart as possible, meaning that between-group variation should increase. Additionally, data within each cluster should be as close as possible. Hence, within-group variation should be minimized. Therefore, as this index increases, it can be considered that clustering has been done well.

The Calinski-Harabasz index enables comparison of results between clustering algorithms, with the algorithm producing the highest value considered the best for clustering. Moreover, when the number of clusters is unknown, one can increment the number and calculate the index's value, selecting the value of k that maximizes the index as the final number of clusters.

Let K denote the total number of clusters, C_k represent the centroid vector of a cluster, n_k denote the number of data points belonging to a cluster, and $c = \sum_{i=1}^{n} \frac{x_i}{n}$ signify the centroid vector of all data points. When x_i^k refers to the data points belonging to the kth cluster, and $\|a\|_2$ denotes the P-dimensional Euclidean distance.

$$CH = \left[\frac{\sum_{k=1}^{K} n_{k\|C_k - C\|_2^2}}{K-1}\right] / \left[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\|x_i^k - C_k\right\|_2^2}{n-K}\right]$$

4.3 Davies-Bouldin index

Let's consider that there are n data points. Here, $x_i \in R^m$, x_i is an m-dimensional vector. Assuming that clusters have been assigned to the data points using a clustering algorithm, let the total number of clusters be K, c_k, k=1, ..., K be the centroid vector of a cluster, and n_k denote the number of data points belonging to a cluster. In this context, we define the following.

$$S_k = (\frac{1}{n_k} \sum_{i=1}^{n_k} \|x_i^k - c_k\|_p^q)^{1/q}$$

$$M_{k,l} = \|c_k - c_l\|_p = \left(\sum_{j=1}^{m} \|c_{k,j} - c_{l,j}\|^p \right)^{\frac{1}{p}}$$

where $\|a\|_p$ is $L_p - Norm$. In this case, $q = 1$ and $p = 2$ are $commonly$ $used$.

S_k represents within-cluster variation, hence smaller values indicate higher similarity among data points within the cluster. $M_{k,l}$ indicate better performance, serving as a measure of how well two clusters are separated:

$$R_{k,l} = \frac{S_k + S_l}{M_{k,l}}$$

5. Simulation

5.1 Data generation

Trajectory data is generated using binary distributions. the trajectory data is as follows:

$$Logit[P(Y_{ij} = 1)] = \beta_{0g} + \beta_{1g} * Time_{ij} + \beta_{2g} * Time_{ij}^{2}$$

To generate trajectory data, the distribution of each trajectory was taken into account, and data generation was conducted accordingly for each trajectory. the process is detailed as follows.

The first data-generating method considers a model with random intercepts. Random intercepts were assigned to each subject by generating random numbers from a normal distribution with a mean of 0 and a standard deviation of 0.5. These values, along with those specific to each trajectory, were added to create probabilities, which were then used to generate binomial distributions.

The second data-generating method involves generating covariates, followed by creating correlated binary response variables based on a multivariate binary distribution, as described (Jung, 2013). This method requires assumptions about the $n_i \times 1$ mean vector π_i, $n_i \times n_i$ covariance matrix V_i, and $n_i \times n_i$ correlation matrix C_i.

First, by applying the parameters in each trajectory, we can obtain the mean vectors through a logit model. The covariance matrix is $V_i = A_i C_i A_i$, where A_i is diag $\{v_{it}^{1/2}\}$

and v_{it} here is $\pi_{it}(1-\pi_{it})$. The correlation matrix C_i is an exchangeable matrix, with a correlation parameter assumed to be 0.5. Given the assumed mean vector, covariance matrix, and correlation matrix as described above, the conditional mean v_{it} is defined by the following equation, where $Z_t = (Y_1, \ldots, Y_{t-1})^{\intercal}$, $\mu_t = E(Z_t)$, $G_t = cov(Z_t)$, $s_t = cov(Z_t, Y_t)$, $b_t = G_t^{-1}s_t$.

$$v_t = v_t(z_t;\pi,V) := P(Y_{it}=1|Z_t=z_t) = \pi_t + b_t^{\intercal}(z_t-\mu_t)$$
$$= \pi_t + \sum_{j=1}^{t-1} b_{tj}(y_j-\pi_j) \quad (t=2,\ldots,T).$$

The binary response variable Y is generated as follows: Y_1 follows a Bernoulli distribution with mean π_1 and is generated using random numbers. Y_t(t=2, ...,6) follows a Bernoulli distribution with conditional mean v_t, and is generated using random numbers. In this way, the response variables at the initial time point are generated using a mean vector, while the response variables at subsequent time points are generated using conditional means, based on a multivariate binomial distribution with a conditional linear property. Thus, the generated binary response variables exhibit correlation.

The third-generation method utilized the grouped Generalized Estimating Equations (GEE) approach. Based on the probability π_{it}, we generated $(Y_{i1}, \ldots, Y_{it})$ from a correlated binary vector using the R package “bindata” with an exchangeable correlation matrix with a 0.5 correlation parameter.

5.2 Simulation Setting

We aim to assess whether each clustering method accurately estimates the true number of clusters under various conditions and determine which clustering method is appropriate under these conditions. we consider three data generation methods.

All three generating methods apply to a cohort of 300 subjects per trajectory. Covariates include time and its square, generated through random number generation following a uniform distribution. six-time points per subject were generated. additionally, When the number of trajectories is 2, each has a probability of 0.5. For 3 trajectories, the probabilities are 0.34, 0.33, and 0.33 respectively. When there are 4 trajectories, each group's entry probability is set at 0.25.

Furthermore, when there are 2 trajectories, the regression coefficients are specified as $\{\beta_{01}, \beta_{11}, \beta_{21}\} = \{-2,2,-0.2\}$ and $\{\beta_{02}, \beta_{12}, \beta_{22}\} = \{2,-2,0.2\}$. For 3 trajectories, they are set a $\{\beta_{01}, \beta_{11}, \beta_{21}\} = \{6,-2,0.01\}$ and $\{\beta_{02}, \beta_{12}, \beta_{22}\} = \{-6,2,0.01\}$ and $\{\beta_{03}, \beta_{13}, \beta_{23}\} = \{2,-0.01,0.01\}$. When there are 4 trajectories, they are $\{\beta_{01}, \beta_{11}, \beta_{21}\} = \{6,-2,0.01\}$ and $\{\beta_{02}, \beta_{12}, \beta_{22}\} = \{-6,2,0.01\}$ and $\{\beta_{03}, \beta_{13}, \beta_{23}\} = \{-0.5,0.01,0.01\}$ and $\{\beta_{04}, \beta_{14}, \beta_{24}\} = \{2,-2,0.2\}$.

Table 1. Parameter settings according to group.

Groups	N	β
Group=2	600	$\beta_{01} = -2, \beta_{11} = 2, \beta_{21} = -0.2$ $\beta_{02} = 2, \beta_{12} = -2, \beta_{22} = 0.2$
Group=3	900	$\beta_{01} = 6, \beta_{11} = -2, \beta_{21} = 0.01$ $\beta_{02} = -6, \beta_{12} = 2, \beta_{22} = 0.01$ $\beta_{03} = 2, \beta_{13} = -0.01, \beta_{23} = 0.01$
Group=4	1200	$\beta_{01} = 6, \beta_{11} = -2, \beta_{21} = 0.01$ $\beta_{02} = -6, \beta_{12} = 2, \beta_{22} = 0.01$ $\beta_{03} = -0.5, \beta_{13} = 0.01, \beta_{23} = 0.01$ $\beta_{04} = 2, \beta_{14} = -2, \beta_{24} = 0.2$

5.3 Simulation Result

We confirmed trajectories with 2, 3, and 4 instances over time. These can be observed in Figure 1.

We will explain the results of the model considering the first random intercept. In the case of two trajectories, it can be observed from Table 2 that the group-based trajectory model correctly identifies the true clusters, whereas the growth mixture model and grouped generalized estimating equations model are not performing as well. Indeed, when examining the evaluation metrics, it is evident that the group-based trajectory model demonstrates superior performance.

In the case of three trajectories, the true number of clusters is accurately estimated in the following order: growth mixture model, group-based trajectory model, and grouped generalized estimating equations, as evident in Table 4. Furthermore, as seen in Table 5, the group-based trajectory model exhibits the best performance, and similar results are observed when there are four trajectories.

When generating data using a multivariate binary distribution model with conditional expectation, it is observed in Table 8 that all models accurately identify the true number of clusters when there are two trajectories. Additionally, Table 9 demonstrates that the Group-based Trajectory Model exhibits the best performance. For the case of three trajectories, Table 10 shows that the Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations model, in that order, accurately represent the true number of clusters. Furthermore, Table 11 confirms that the Group-based Trajectory Model performs the best across all evaluation metrics

As indicated in Table 12, when there are four trajectories, it's apparent that none of the models successfully identify the true number of clusters. Despite the evaluation metrics showing the Growth Mixture Model to perform the best, scrutiny of the Adjusted Rand Index demonstrates that none of the models exhibit strong performance.

Considering the simulation in the grouped generalized estimating equations paper, data was generated using time and its square as variables. When there were two trajectories in the trajectory model, both the Calinski Harabasz index and Davies Bouldin index showed that grouped generalized estimating equations performed the best, as indicated in Table 15. Additionally, as seen in Table 14, both the group-based trajectory model and grouped generalized estimating equations accurately estimated the number of clusters.

When the trajectory consisted of three groups, similarly, grouped generalized estimating equations exhibited the best performance. However, as shown in Table 16, the number of clusters was most accurately estimated by the group-based trajectory model.

In the case of four trajectories, all performance metrics suggested that each model performed better in different aspects. This discrepancy arises from the limitation of describing four trajectories solely using binary classification, as noted in the paper.

Figure 2. Trajectory figures according to the number of groups.

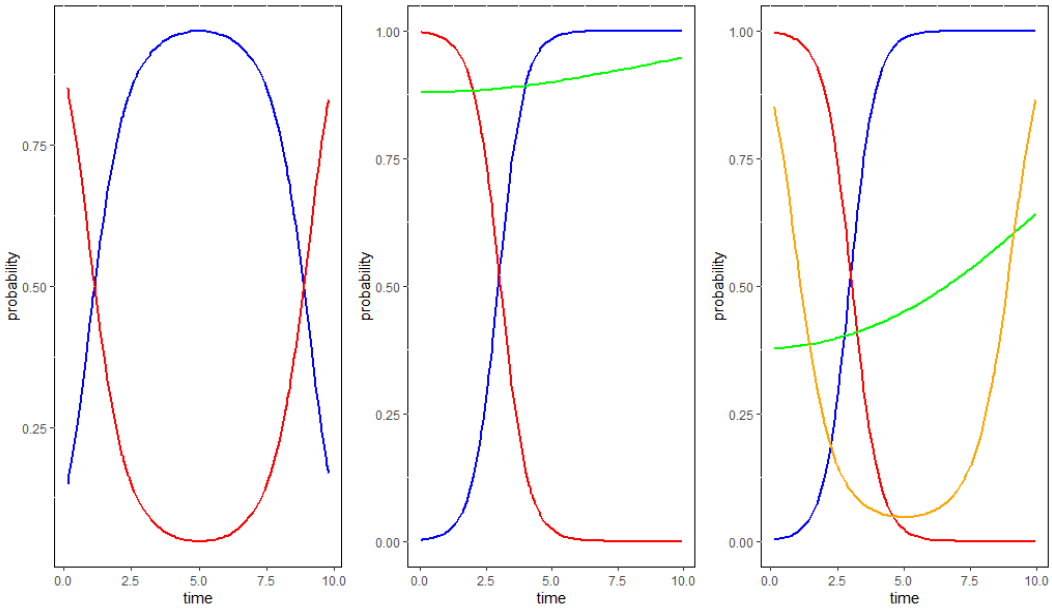


Table 2. The results of group numbers when the group is 2 in the random effect model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	94	0	0
number of groups = 3	6	0	0
number of groups = 4	0	100	100

Table 3. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 2 in the random effect model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.944	0.512	0.241
Calinski-Harabasz index	10.907	5.673	5.366
Davies-Bouldin index	27.863	38.673	57.511

Table 4. The results of several clusters for trajectories of 3 in the random effect model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	0	0	57
number of groups = 3	95	100	4
number of groups = 4	5	0	39

Table 5. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 3 in the random effect model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.846	0.824	0.447
Calinski-Harabasz index	8.819	6.656	4.327
Davies-Bouldin index	59.222	68.518	67.076

Table 6. The results of several clusters for trajectories of 4 in the random effect model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 3	0	0	92
number of groups = 4	98	99	3
number of groups = 5	2	1	5

Table 7. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 4 in the random effect model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.695	0.567	0.243
Calinski-Harabasz index	6.185	4.237	5.656
Davies-Bouldin index	88.052	112.122	84.112

Table 8. The results of several clusters for trajectories of 2 in the multivariate binary distribution model with conditional expectation.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	95	100	100
number of groups = 3	5	0	0
number of groups = 4	0	0	0

Table 9. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 2 in the multivariate binary distribution model with conditional expectation.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.458	0.392	0.267
Calinski-Harabasz index	1926.069	1554.047	17.170
Davies-Bouldin index	1.457	1.519	12.314

Table 10. The results of several clusters for trajectories of 3 in the multivariate binary distribution model with conditional expectation.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	1	0	77
number of groups = 3	99	97	22
number of groups = 4	0	2	1

Table 11. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 3 in the multivariate binary distribution model with conditional expectation.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.438	0.420	0.134
Calinski-Harabasz index	1071.829	813.067	19.447
Davies-Bouldin index	2.685	2.787	17.788

Table 12. The results of several clusters for trajectories of 4 in the multivariate binary distribution model with conditional expectation.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 3	100	100	84
number of groups = 4	0	0	0
number of groups = 5	0	0	16

Table 13. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 4 in the multivariate binary distribution model with conditional expectation.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.130	0.162	0.032
Calinski-Harabasz index	456.066	382.314	16.441
Davies-Bouldin index	12.938	6.964	48.415

Table 14. The results of several clusters for trajectories of 2 in the grouped generalized estimating equation model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	96	3	100
number of groups = 3	4	4	0
number of groups = 4	0	93	0

Table 15. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 2 in the grouped generalized estimating equation model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.974	0.441	0.980
Calinski-Harabasz index	76.590	344.270	438.771
Davies-Bouldin index	6.961	10.296	3.457

Table 16. The results of several clusters for trajectories of 3 in the grouped generalized estimating equation model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 2	0	0	34
number of groups = 3	84	56	66
number of groups = 4	16	44	0

Table 17. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 3 in the grouped generalized estimating equation model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.679	0.543	0.803
Calinski-Harabasz index	110.264	460.170	1759.573
Davies-Bouldin index	17.170	4.095	0.869

Table 18. The results of several clusters for trajectories of 4 in the grouped generalized estimating equation model.

Number of groups	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
number of groups = 3	6	0	98
number of groups = 4	68	4	0
number of groups = 5	26	96	2

Table 19. The performance metrics of Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations when the group is 4 in the grouped generalized estimating equation model.

Evaluation	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Adjusted rand index	0.794	0.679	0.377
Calinski-Harabasz index	30.960	36.607	16.601
Davies-Bouldin index	38.311	28.690	21.149

6. Application

We utilize the proposed technique on the HRS dataset, sourced from research carried out by the University of Michigan. This longitudinal panel investigation conducts thorough interviews with American adults aged 50 and above once every two years, providing insights into their health and financial situations. The primary objective of this study is to examine how participants' health statuses evolve within the HRS study and identify the factors linked to these changes.

We utilized the dataset from the HRS study, available through the R package "LMest". The sample comprises 7074 individuals tracked over approximately 8 equally spaced intervals, with no missing responses or dropouts. The response variable is self-reported health status, categorized into five levels: "poor", "fair", "good", "very good", and "excellent", ranked from 5 to 1.

We classified "good", "very good", and "excellent" as "well" (1) and the remaining values as "unwell" (0). Moreover, within the covariates, we incorporated indicator variables denoting gender (1 for male, 0 for female), indicators for race (black and other and white), indicators for educational attainment (SC: some college, CAA: college and above, Others), and age recorded at each time instance.

We assume that individuals can be categorized into groups: those who are consistently healthy, those who are consistently unhealthy, and those whose health status may change over time. Therefore, we aim to compare and evaluate the performance of group-based trajectory models, and growth mixture models and grouped generalized estimating equations as methods for clustering the variations over time.

Let y_{ij} be the binary response variable, and x_{ij} be the vector of five covariates and an intercept, for $i = 1, \dots, n(= 7074)$ and $t = 1, \dots, T(= 8)$. We consider the mean structure $E[y_{ij}|x_{ij}] = m(x_{ij}^{\intercal}\beta_{gi})$ with $m(x) = \frac{\exp(x)}{\{1+\exp(x)\}}$ and $g_i \in \{2, \dots, 7\}$.

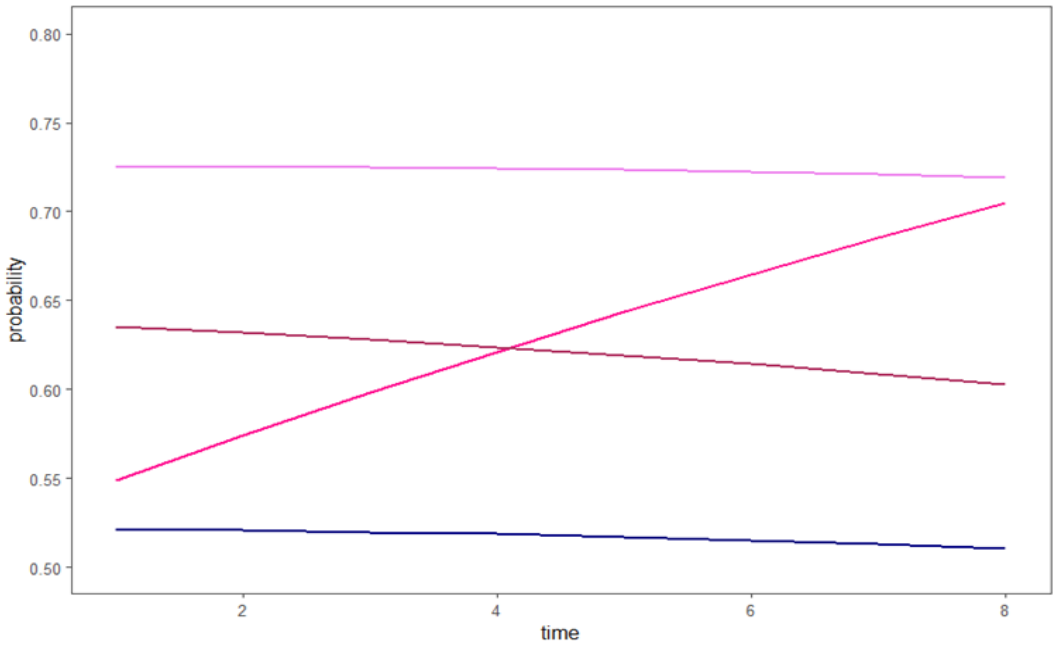
Firstly, to determine the number of groups, we considered the Bayesian Information Criterion (BIC) in both the Group-based Trajectory Model and the Growth Mixture Model, selecting 7 and 4 groups based on the lowest BIC values. In contrast, the Grouped Generalized Estimating Equations model utilized the cross-validation with averaging method to specify the number of groups as 5. Subsequently, applying each specified number of groups, we evaluated and compared the Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations using the Calinski-Harabasz Index, and Davies-Bouldin Index.

In Table 20, we observe that the Growth Mixture Model has the highest value for the Calinski-Harabasz Index and the lowest value for the Davies-Bouldin Index. This indicates that the Growth Mixture Model exhibits the best performance. Trajectory exhibits distinct characteristics.

Table 20. The performance results of the Calinski-Harabasz Index and Davies-Bouldin Index for different numbers of groups in the Group-based Trajectory Model, Growth Mixture Model, and Grouped Generalized Estimating Equations

	Group based trajectory model	Growth mixture model	Grouped generalized estimating equations
Number of groups	7	4	5
Calinski-Harabasz index	29.44	83.46	54.55
Davies-Bouldin index	219.683	126.72	149.791

Figure 2. Trajectory plots in the Growth Mixture Model



7. Discussion

In various health studies, the measured outcomes are typically aggregated and analyzed across the entire study population or predefined subgroups. However, in most cases, unknown or unexpected subgroups may exhibit similar patterns. Therefore, relying solely on mean estimates may underestimate the complexity of real-life clinical contexts. Additionally, categorical dependent variables are commonly used in research, and they offer the advantage of ease in customization when identifying similar patterns, similar to continuous dependent variables. Hence, we have embarked on a study to compare methods for identifying similar patterns in categorical dependent variables across various contexts.

We conducted simulations to generate longitudinal binary data, considering models incorporating random intercepts, multivariate binary models using conditional expectations, and Grouped Generalized Estimating Equation models. We examined trajectories with 2, 3, and 4 clusters.

In the first model considering random intercepts, regardless of the number of trajectories, we found that the group-based trajectory model performed the best and accurately estimated the number of clusters.

In the second model, the multivariate binary model using conditional expectations, we observed that for trajectories 2 and 3, the group-based trajectory model exhibited the best performance, accurately estimating the number of groups alongside the growth mixture model. However, for trajectory 4, none of the models accurately estimated the number of clusters. The Adjusted Rand Index values also indicate a random assignment pattern, similar to chance, across all models. The issue arises because when generating trajectories with four binary outcomes, there tends to be a significant overlap.

In the third simulation, following the format of the Grouped Generalized Estimating Equations paper and considering the time variable, we observed that grouped generalized estimating equations consistently performed the best across trajectories with 2 and 3 clusters. However, the group-based trajectory model exhibited the most accurate estimation of the number of clusters.

Moreover, for trajectories with 4 clusters, both the estimation of the number of clusters and overall performance favored the group-based trajectory model. However, since performance metrics indicate different models as superior, it's challenging to determine which model is best. The issue seems to stem from limitations in the metrics used to evaluate binary data.

During this study, the group-based trajectory model consistently demonstrated robust performance across various simulation scenarios, effectively estimating the number of clusters. However, a limitation was identified regarding the representation of binary data, where trajectories were encoded solely as 0s and 1s. Evidence of this limitation is observed in cases where trajectories with four clusters, except those generated by the random effect model, showed suboptimal performance in terms of both the true number of clusters and overall performance. Additionally, performance evaluation was conducted using metrics such as the adjusted Rand index, Calinski-Harabasz index, and Davies-Bouldin index. While these metrics can be applied to binary data, doubts arose regarding whether the performance evaluation was adequate, prompting the need for more detailed and diverse performance metrics.

References

Diop, A., Gupta, A., Mueller, S., Dron, L., Harari, O., Berringer, H., Kalatharan, V., Park, J. J. H., Mesidor, M., & Talbot, D. (2024). Assessing the performance of group-based trajectory modeling method to discover different patterns of medication adherence. *Pharm Stat*.

Fitzmaurice. (2012). *Applied Longitudinal Analysis*.

Gao, C. X., Dwyer, D., Zhu, Y., Smith, C. L., Du, L., Filia, K. M., Bayer, J., Menssink, J. M., Wang, T., Bergmeir, C., Wood, S., & Cotton, S. M. (2023). An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Res*, *327*, 115265.

Grimm, N. R. a. K. J. (2009). Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups. *International Journal of Behavioral Development*.

Hickson, R. P., Annis, I. E., Killeya-Jones, L. A., & Fang, G. (2021). Comparing Continuous and Binary Group-based Trajectory Modeling Using Statin Medication Adherence Data. *Med Care*, *59*(11), 997-1005.

Ito, T., & Sugasawa, S. (2023). Grouped generalized estimating equations for longitudinal data analysis. *Biometrics*, *79*(3), 1868-1879.

Jinwon Sohn, S. J., Young Min Cho, Taeyoung Park. (2023). Functional clustering methods for binary longitudinal data with temporal heterogeneity. *Computational Statistics and Data Analysis*.

Jung, B. P. a. I. (2013). Comparison of GEE Estimation Methods for Repeated Binary Data with Time-Varying Covariates on Different Missing Mechanisms. *The Korean Journal of Applied Statistics*.

Leiby, B. E. (2012). Growth curve mixture models. *Shanghai Arch Psychiatry*, *24*(6), 355-358.

LEIBY, B. E. (2012). On Growth Curves and Mixture Models. *Shanghai Archives of Psychiatry*.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern*, *43*(3), 982-994.

N. G. P. Den Teuling, S. C. P. E. R. v. d. H. (2023). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics - Simulation and Computation*.

Wu, R., Ma, C. X., Littell, R. C., Wu, S. S., Yin, T., Huang, M., Wang, M., & Casella, G. (2002). A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories. *Genet Res*, *79*(3), 235-245.

국 문 요 약

종단 범주형 데이터의 클러스터링 방법 비교

이 연구는 다양한 맥락에서 범주형 종속 변수의 유사한 패턴을 식별하기 위한 방법론을 탐구합니다. 시뮬레이션을 통해 종단적인 바이너리 데이터를 생성하며, 무작위 절편, 다변량 이항 모델 및 그룹화된 일반화 추정 방정식 모델을 사용합니다.

결과는 그룹 기반 궤적 모델이 클러스터 수를 정확하게 추정하는 데 다른 모델보다 일관되게 우수한 성능을 보였다는 것을 나타냅니다. 그러나 이진 데이터를 표현하는 데 한계가 있음을 확인했는데, 특히 네 개 클러스터를 갖는 궤적에서 뚜렷하게 드러났습니다. 조정된 랜드 지수와 같은 성능 지표를 사용했지만, 이러한 평가가 적절한지에 대한 의문이 제기되어 더 포괄적인 평가 지표가 필요합니다.

핵심되는 말: 종단 자료, Group GEE, Mixture model, 궤적 클러스터링