

A comparison of methods for clustering longitudinal data with slowly changing trends

N. G. P. Den Teuling, S. C. Pauws & E. R. van den Heuvel

To cite this article: N. G. P. Den Teuling, S. C. Pauws & E. R. van den Heuvel (2023) A comparison of methods for clustering longitudinal data with slowly changing trends, Communications in Statistics - Simulation and Computation, 52:3, 621-648, DOI: [10.1080/03610918.2020.1861464](https://doi.org/10.1080/03610918.2020.1861464)

To link to this article: <https://doi.org/10.1080/03610918.2020.1861464>



© 2021 Koninklijke Philips N.V. Published with license by Taylor and Francis Group, LLC



[View supplementary material](#)



Published online: 19 Jan 2021.



[Submit your article to this journal](#)



Article views: 13964



[View related articles](#)





[View Crossmark data](#)



Citing articles: 5 [View citing articles](#)

A comparison of methods for clustering longitudinal data with slowly changing trends

N. G. P. Den Teuling^{a,b} , S. C. Pauws^{b,c}, and E. R. van den Heuvel^a 

^aDepartment of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands; ^bPhilips Research, Eindhoven, the Netherlands; ^cDepartment Communication and Cognition, Tilburg University, Tilburg, the Netherlands

ABSTRACT

Longitudinal clustering provides a detailed yet comprehensible description of time profiles among subjects. With several approaches that are commonly used for this purpose, it remains unclear under which conditions a method is preferred over another method. We investigated the performance of five methods using Monte Carlo simulations on synthetic datasets, representing various scenarios involving polynomial time profiles. The performance was evaluated on two aspects: The agreement of the group assignment to the simulated reference, as measured by the split-join distance, and the trend estimation error, as measured by a weighted minimum of the mean squared error (WMMSE). Growth mixture modeling (GMM) was found to achieve the best overall performance, followed closely by a two-step approach using growth curve modeling and *k*-means (GCKM). Considering the model similarities between GMM and GCKM, the latter is preferred for large datasets for its computational efficiency. Longitudinal *k*-means (KML) and group-based trajectory modeling were found to have practically identical solutions in the case that the group trajectory model of the latter method is correctly specified. Both methods performed less than GMM and GCKM in most settings.

ARTICLE HISTORY

Received 21 February 2020
Accepted 4 December 2020


KEYWORDS

Group-based trajectory modeling; Growth mixture modeling; Intensive longitudinal data; Latent-class trajectory modeling; Longitudinal clustering; Simulation study

1. Introduction

The connectivity, storage, and sensor solutions of today enable researchers to collect many data points from subjects over any period of time. The larger volume of data collected presents both new opportunities and challenges for longitudinal data analysis when it comes to understanding the data. Notably, a higher number of subjects allows for a data-driven exploration under the assumption of heterogeneity for subgroups of subjects with different trends (i.e., group trajectories). It is then important to profile subjects into different subgroups. While longitudinal cluster analyses have typically comprised only a small number of repeated measurements over time per subject (e.g., less than ten), there is a growing availability of high-frequent longitudinal datasets, referred to as intensive longitudinal data (ILD) (Walls and Schafer 2006). This type of data enables the estimation of subject-specific trajectories, especially under the presence of high within-subject or between-subject variability. Moreover, the increased number of observations allow for the estimation of more time-sensitive changes.

CONTACT N. G. P. Den Teuling  n.g.p.den.teuling@tue.nl  Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands

 Supplemental data for this article is available online at <https://doi.org/10.1080/03610918.2020.1861464>.

© 2021 Koninklijke Philips N.V. Published with license by Taylor and Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

The application of longitudinal clustering spans many domains, including criminology, sociology, medicine, and ecology. Recent examples of applications include the identification of subgroups with different cigarette smoking patterns with the aim of predicting health outcome (Lee et al. 2016), and describing adolescent substance use trajectories and its association to leisure experience (Weybright et al. 2016).

Longitudinal data often comprises trajectories with different observation times, or a different number of observations. ILD comes with additional challenges over repeated measurements data, such as the high volume of data, modeling the dynamics or volatility of trajectories, and accounting for strong correlations due to measurements being close in time.

As demonstrated by past ILD applications, traditional methods are generally applicable to ILD in spite of the increased volume of data, although they may not address all challenges. For example, Shiyko, Li, and Rindskopf (2012) applied growth mixture modeling to ILD for the flexible identification of patterns of smoking cessation behavior of up to 29 days as a way to account for the correlation between observations. Babbin et al. (2015) explored patterns of ILD comprising daily therapy usage among patients undergoing sleep apnea treatment during the first six months of therapy. They used a non-parametric trajectory representation, allowing for high flexibility in the shape of the group trajectories. Lastly, Ernst et al. (2019) analyzed ecological momentary assessments of subjects, assessing their emotional state three times per day over a period of 30 days. Subgroups with different emotion dynamics were discovered by clustering subjects based on individual vector autoregressive model coefficients.

The methods that have been introduced over the past two decades for the purpose of longitudinal clustering can be divided into three categories: First, the naive approach clusters on the observations, in which the temporal relation between the measurements is not modeled. Second, a two-step approach that first describes subject trajectories in terms of a statistical model or other metrics (which can be regarded as dimensionality reduction), and then clusters on the model parameters. Lastly, the mixture model approach describes the clusters using a mixture of statistical models (Muthén and Shedden 1999).

Considering the different methods for longitudinal clustering that are available, the question of which method is preferred for a given context arises naturally. In most published applications, the rationale for selecting a particular longitudinal cluster method is not provided. This could be because alternative methods were not considered, or because the existing body of work on comparisons between methods did not address the relevant context.

Some combinations of methods have been compared, with contradicting findings, suggesting that the optimal method depends on the scenario being considered. Martin and von Oertzen (2015) compared growth mixture modeling (GMM) against naive approaches such as longitudinal k-means (KML) on synthetic data comprising five repeated measurements, and two or three groups. They found that GMM outperforms the other methods even for small sample size. Feldman, Masyn, and Conger (2009) preferred the group-based trajectory model (GBTM) over GMM due to the complexity of the latter, and the convergence problems that arise from it (Frankfurt et al. 2016). They noted similarities in performance between longitudinal latent class analysis (LLCA) and GBTM, in contrast to Twisk and Hoekstra (2012), who found LLCA to be more similar to KML and a two-step approach involving clustering the random effects of a mixed model. Overall, there seems to be a preference for mixture-based methods, but considering the different approaches to mixture models, the results are not conclusive.

To the best of our knowledge, comparisons between methods involving ILD have not yet been done. Most of the comparison studies investigate longitudinal data involving 4-6 repeated measurements, with few studies exceeding 10 measurements. It is questionable whether these findings generalize to ILD. For example, having an increased number of observations per trajectory enables a more reliable estimation of longitudinal change under a higher degree of variability, which some methods will benefit more from than others. Moreover, with a growing number of

observations per trajectory, a faster but less data-efficient method may become preferable over a better but considerably more computationally demanding method.

We contribute to the existing body of work by evaluating the performance of five longitudinal clustering methods in an exploratory setting, applied to many scenarios comprising group trajectories that smoothly and slowly change over time. The methods are longitudinal k -means, a mixed-effects model combined with k -means, group-based trajectory modeling, growth mixture modeling, and time-varying effect mixture modeling. These methods are chosen because they take different approaches to clustering, are commonly used in applications, or are applicable in an exploratory setting without prior knowledge on the clusters. We investigate how well the methods are able to identify the underlying groups (in terms of subject assignments) and group trajectories in each of these scenarios. In addition, we assess the sensitivity of the methods to different forms of heterogeneity by evaluating the effect on performance for different distributions of the groups. The scenarios involve a large number of permutations on the different levels of within-group variability, sample sizes, number of observations, and levels of heteroskedasticity of the residual variance. Lastly, we study the effect of a proportional measurement error on the model estimation, and assess the reliability of selecting the correct number of groups per method. The simulated datasets comprise heterogeneous subgroups with varying degrees of overlap, described by quadratic group trajectories. This establishes a ground truth against which the output of the methods can be evaluated. The comparison also serves as a benchmark for the computation time with respect to the number of trajectories, number of observations, and number of groups. In view of the exploratory nature, the mixture methods are all estimated using maximum likelihood estimation instead of a Bayesian approach. In addition to the simulation study, a case study involving therapy compliance of sleep apnea patients is included to relate the findings from the simulations to a real-life setting.

The rest of the paper is organized as follows. [Section 2](#) briefly describes the selected methods. The simulation scenarios are described in [Sec. 3](#). In [Sec. 4](#), the simulation results are reported, along with the description and results of the case study. The resulting findings and recommendations are discussed in [Sec. 5](#), and conclusions are given in [Sec. 6](#).

2. Methods

We denote a trajectory of subject $i \in I$ of the available set of subjects I with T measurements by $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \dots, y_{i,T})$, where the measurement $y_{i,j}$ is recorded at time $t_{i,j}$. We denote the ILD $\psi_i(t)$ for individual $i \in I$ by $\mathbb{E}y_{i,j} = \psi_i(t_{i,j})$ with $y_{i,j}$ the measurement at time $t_{i,j}$, with $j = 1, 2, \dots, T$, and with ψ_i a continuous function $\psi_i: \mathbb{R}_+ \rightarrow \mathbb{R}$. We will focus on polynomial time profiles: $\psi_i(t) = \sum_{r=0}^p \beta_{i,r} t^r$. Alternatively, we may also study piecewise linear profiles when we apply naive clustering methods, but these profiles would be developed over groups of participants. At a subject level the piece-wise linear model is defined by

$$\psi_i(t) = \sum_{r=1}^T (\alpha_{i,r} + \beta_{i,r} t) 1_{(t_{i,r-1}, t_{i,r}]}(t) \quad (1)$$

with restrictions $\alpha_{i,r} = \alpha_{i,r-1} + \beta_{i,r-1} t_{i,r}$. The two-step approach and the mixture methods allow for the inclusion of covariates, but this is not evaluated in this work.

2.1. Longitudinal k -means

Longitudinal k -means (KML) is a commonly used naive approach (Genolini and Falissard 2010). The vectors of observations are assumed to be of equal length and aligned, i.e., $t_{a,j} = t_{b,j}$ for $a, b \in I, j = 1, \dots, T$. The vectors are passed as observations to the k -means clustering algorithm (MacQueen 1967; Genolini and Falissard 2010). The k -means algorithm aims to find the partitioning

I_1, I_2, \dots, I_G with $\cup_{g=1}^G I_g = I$ and $I_g \cap I_h = \emptyset$ when $g \neq h$, that minimizes the within-cluster variance, which in term maximizes the between-cluster variance. The objective function is given by

$$\arg \min_{I_1, I_2, \dots, I_G} \sum_{g=1}^G \sum_{i \in I_g} \|y_i - \hat{\mu}_g\|^2, \quad (2)$$

with $\hat{\mu}_g$ the mean vector of the group elements, i.e., $\hat{\mu}_g = |I_g|^{-1} \sum_{i \in I_g} y_i$, where summation is performed element-wise. The algorithm uses an iterative approach to arrive at a solution. Starting from a random partitioning, the algorithm refines the partitioning at each iteration until the solution cannot be further improved, i.e., converges. In the case of KML, the resulting cluster centers represent the group trajectory of each cluster. The resulting groups are assumed to be homogeneous, i.e., subjects belonging to a given group are assumed to follow the group trajectory $\hat{\mu}_g$.

2.2. Two-step clustering

We represent the two-step clustering approach by modeling the trajectories using a growth curve model (GCM), and clustering the subject parameter estimates (i.e., the random effects) using k -means (MacQueen 1967). We will refer to this method as GCKM. This method is also described in the comparison of Twisk and Hoekstra (2012). The GCM is estimated in a mixed model framework (Laird and Ware 1982). The model represents the longitudinal dataset in terms of a single group trajectory (i.e., the fixed effects), and for each subject, their deviation from this trajectory (the random effects). The trajectories are typically described by a polynomial of order 1 or 2 (Nagin and Odgers 2010a). A trajectory described in terms of a polynomial of order K and random effects in all terms is given by

$$y_{i,j} = \sum_{k=0}^K \beta_{k,i} t_{i,j}^k + \varepsilon_{i,j}, \quad (3)$$

$$\beta_{k,i} = \alpha_k + \zeta_{k,i}.$$

Here, α_k represents the k th order coefficient of the polynomial trajectory, $\zeta_{k,i}$ denotes the random effect of subject i for the k th coefficient (i.e., the between-subject variability), and $\varepsilon_{i,j}$ denotes the measurement error (within-subject variability). The random effects are assumed to be multivariately normally distributed with zero mean, possibly with an unstructured variance-covariance matrix, and uncorrelated with the measurement error ε . The measurement error is assumed to be independently normally distributed with zero mean and common variance, although an autoregressive correlation structure would be possible too. The model is estimated using maximum likelihood (ML) estimation (Verbeke and Molenberghs 2000). Alternatively, the model parameters can be inferred using a Bayesian approach (Gelman et al. 2013).

The random effects $\zeta_{k,i}$ of each trajectory can be predicted using the best linear unbiased predictors (BLUPs), and they are passed to the k -means algorithm as input vectors $y_i = (\hat{\zeta}_{0,i}, \hat{\zeta}_{1,i}, \dots, \hat{\zeta}_{K,i})$. The estimation of the input vectors is independent of the number of groups to be identified in the second step and therefore needs to be performed only once. Scaling or standardization may be required to ensure equal weights across the BLUPs, depending on the difference in size of the variance components of $\zeta_{k,i}$. Similarly, covariates could be included into the model as additional random effects to account for other factors, and can be clustered accordingly.

2.3. Group-based trajectory modeling

A group-based trajectory model (GBTM) describes a longitudinal dataset in terms of a mixture of group trajectories, without regard of within-group variability (Nagin and Land 1993; Nagin and Odgers 2010a). This draws similarities to k -means in the sense that the subjects in a group are

assumed to follow the group profile, but in the case of GBTM these profiles can be smooth (Nagin and Tremblay 2005). GBTM is also commonly referred to as latent class growth analysis (LCGA), and semi-parametric group-based modeling (SGBM) (Nagin 1999).

For a given trajectory y_i , its observations are described by the group trajectory of group g as follows:

$$y_{i,j}^{(g)} = \sum_{k=0}^K \alpha_k^{(g)} t_{i,j}^k + \varepsilon_{i,j}, \quad (4)$$

where $\alpha_k^{(g)}$ denotes the k th coefficient for the polynomial of group g , and $\varepsilon_{i,j}$ describes the residual at time $t_{i,j}$. In this setting the subject trajectories $\psi_i(t_{i,j})$ are all the same to $\sum_{k=0}^K \alpha_k^{(g)} t_{i,j}^k$ in (4) when subject i belongs to group I_g . The residual is assumed to be independently normally distributed with zero mean. The marginal mean of a GBTM is given by

$$\mathbb{E}(y_{i,j}) = \sum_{g=1}^G \pi^{(g)} \sum_{k=0}^K \alpha_k^{(g)} t_{i,j}^k. \quad (5)$$

Here, $\pi^{(g)}$ denotes proportion of group g , with $0 \leq \pi^{(g)} \leq 1$ and $\sum_g \pi^{(g)} = 1$. The model is fitted to the data using ML estimation. The appeal of GBTM comes from the relatively simple group model, combined with a parametric approach that enables researchers to incorporate domain knowledge. Moreover, other factors can be corrected for by the inclusion of time-variant and time-invariant covariates into the model, both in the time profile and the proportion for group g (Nagin and Odgers 2010a).

2.4. Growth mixture modeling

Growth mixture modeling (GMM) is a method for identifying heterogeneous subgroups in the data using a mixture of growth curve models (Verbeke and Lesaffre 1996; Muthén et al. 2002; Muthén and Shedden 1999). It is a generalization of GBTM by taking the coefficients $\alpha_k^{(g)}$ in (4) to be subject-specific, essentially introducing a mixed-effects model in each group g . Thus, a given trajectory y_i is described by group g by

$$\begin{aligned} y_{i,j}^{(g)} &= \sum_{k=0}^K \beta_{k,i}^{(g)} t_{i,j}^k + \varepsilon_{i,j}^{(g)}, \\ \beta_{k,i}^{(g)} &= \alpha_k^{(g)} + \zeta_{k,i}^{(g)}. \end{aligned} \quad (6)$$

The group-dependent fixed effects are denoted by $\alpha_k^{(g)}$. In any case, the model complexity of GMM significantly exceeds that of GBTM due to the additional estimation at group level of the random effects $\zeta_{k,i}^{(g)}$, residual $\varepsilon_{i,j}^{(g)}$, and the variance-covariance matrix. In practice, it is desirable to restrict or share some of the parameters across groups to reduce the challenge of finding a numerical solution. The residual is assumed to be independently and normally distributed with zero mean, and uncorrelated with $\zeta_{k,i}^{(g)}$. The random effects are assumed to be normally distributed with zero mean, but may be correlated within group g (but not with random effects across groups). The marginal mean of a GMM is computed by (4) for $\mathbb{E}(\zeta_{k,i}^{(g)}) = 0$, with group proportion $\pi^{(g)}$ defined as before.

GMM is commonly used for its flexibility, enabling researchers to specify the random effects and relations between them, in addition to the inclusion of covariates (Frankfurt et al. 2016). However, this may come at the cost of a greater difficulty in identifying the most appropriate model.

2.5. Mixed time-varying effect modeling

In a time-varying effect model (TVEM) (Tan et al. 2012), the regression coefficients that describes the relation between covariates and outcome can vary over time. The relation between time and outcome is described by a smooth function $\psi(t)$, thus longitudinal trajectories can be described by the intercept-only TVEM given by

$$y_{i,j} = \psi(t_{i,j}) + \varepsilon_{i,j}. \quad (7)$$

The ψ function is modeled using a penalized spline (referred to as a P-spline) (Song and Lu 2010). A P-spline represents a series of time intervals using low-order polynomials with smooth transitions between intervals. A smooth fit is ensured by imposing a penalty on the second derivative. The P-splines can be represented as a linear model and therefore estimated efficiently using ordinary least-squares (Ruppert 2002), although smoothing can also be obtained by introducing random effects for the time variable (Lu and Song 2012).

We evaluate the mixture model proposed by Dziak et al. (2015), named MixTVEM, which comprises a mixture of TVEMs. Thus a smooth function $\psi^{(g)}$ of splines is estimated for group g and $\pi^{(g)}$ represents a proportion for group g . Its flexible trajectory estimates make it suitable for intensive longitudinal data, and as an exploration tool for uncovering unforeseen trajectories. The method is a semi-parametric version of GBTM or GMM, depending on whether random effects are taken into account. Dziak et al. suggest the inclusion of a first-order autoregressive (AR) structure as an alternative to random effects, as this is less computationally intensive and allows for constant heteroskedasticity over time. The correlation between any two measurements $y_{i,j}$ and $y_{i,j'}$ is given by $\rho^{|t_{i,j}-t_{i,j'}|}$, where ρ denotes AR-1 component. For numerical estimation purposes, they introduce an additional variance component that is proportional to the AR-1 component, referred to as the nugget effect, given by

$$\text{cov}(y_{i,j}, y_{i,j'}) = \sigma_\rho^2 \rho^{|t_{i,j}-t_{i,j'}|} + \sigma_\varepsilon^2. \quad (8)$$

The ratio of the measurement variance σ_ε^2 to the total variance $\sigma_\rho^2 + \sigma_\varepsilon^2$ is assumed to be fixed across groups.

2.6. Number of groups

Determining the number of subgroups to describe the data is a well-known problem in the area of cluster analysis. In practice, clusters are rarely distinct, meaning that the subgroups are not well-separated and it is therefore difficult to cluster every subject without error. Nagin and Odgers (2010a) suggest to combine the use of an objective criterion for determining the number of groups with domain knowledge in order to arrive at a reasonable solution. For the mixture methods, the Bayes information criterion (BIC) appears to be generally applicable (Nagin and Odgers 2010a; Nylund, Asparouhov, and Muthén 2007; McNeish and Harring 2017; Dziak et al. 2015). For the purpose of consistency, we use the BIC for KML and GCKM too. The likelihood of a k -means solution can be computed by regarding the clusters as a mixture of spherical Gaussians, enabling the computation of the BIC (Pelleg and Moore 2000).

A recommended alternative to the BIC is the bootstrapped likelihood ratio test (BLRT), which is regarded as a more suitable criterion than the BIC (Nylund, Asparouhov, and Muthén 2007; Jung and Wickrama 2008; Tolvanen 2007). The BLRT is a computationally intensive criterion, as it requires the estimation of the model on each of the generated bootstrap samples (with a recommendation of 500 bootstrap samples). We will therefore first conduct a preliminary evaluation on GBTM and GMM before evaluating the other methods. The evaluation on GBTM and GMM compares how the BLRT performs on models assuming homogeneous subgroups and heterogeneous subgroups, respectively.

An increase in BIC score of more than 10 is considered to be of significance, meaning that the additional description it provides warrants the increased model complexity from introducing an additional group (Raftery 1995; Frankfurt et al. 2016). A robust alternative to this approach, commonly referred to as the elbow method, investigates the relative improvement in the objective function for a lower number of groups. The improvements tend to decline with an increasing number of groups, but often with a turning point (i.e., the elbow) (Hardy 1994). While this method is usually assessed visually, we approximate it by estimating a piecewise-linear change point model such that it can be evaluated on the many datasets and scenarios automatically. The model comprises a variable change point and is optimized to maximize the fit to the BIC points over the different number of groups.

2.7. Computer software

All methods are evaluated in R 3.4.2 (R Core Team 2017), running on Intel Xeon E5-2660 (2.6 GHz) processors. The analysis code used in the simulation study and case study is available in the [Supplementary material](#). The implementation of KML was based on version 2.4.1 of the `kml` package (Genolini et al. 2015). GCKM was evaluated by estimating a GCM using the `lcmm` package (version 1.7.8) (Proust-Lima, Philipps, and Liqueur 2017), and clustering the random effects using the `kml` package¹. The `lcmm` package is also used to evaluate GBTM and GMM. For the implementation of MixTVEM, we use an R script² that has been made available by Dziak et al. (2015) (version 1.1), and run it with the default settings. Preliminary tests indicated that MixTVEM generally performed better with the inclusion of the AR-1 structure, and we therefore used it in all evaluations.

The estimation of the models can be a challenging task due to the large number of parameters involved, a problem that grows with the number of groups. The iterative optimization procedures converge to a local optimum, depending on the starting position. This is an issue sometimes observed in GBTM (Skardhamar 2010; Twisk and Hoekstra 2012), and especially in GMM (Twisk and Hoekstra 2012; McNeish and Harring 2017). The accepted approach in dealing with this involves many repeated random starts (e.g., 100 random starts, although the number depends on the data and model complexity) after which the estimation proceeds with the most likely start (Jung and Wickrama 2008; McNeish and Harring 2017), which is a time-wise costly procedure indeed. In view of the many scenarios under which the mixture methods need to be evaluated, we settled for 20 random starts in order to reduce the computation time. A preliminary evaluation suggested that this does not have a practically significant effect on the performance. For KML and GCKM, the `k-means++` algorithm is used for selecting a better starting position, with 25 repeated runs (Arthur and Vassilvitskii 2007).

3. Simulation

We evaluate the methods across different scenarios using Monte Carlo simulations³. The scenarios comprise multiple settings, with each method being evaluated on each permutation of settings on 100 synthetic datasets. The generated group trajectories are consistent between scenarios and settings for the respective number of groups in the data (unless mentioned otherwise). This enables a comparison between scenarios with a higher sensitivity to the effect of changing settings (Burton et al. 2006), with the advantage of requiring fewer simulations.

¹Although the random effects are not longitudinal data, the `kml` package was used here to ensure an identical application of `k-means`.

²Version 1.1, available at <https://www.methodology.psu.edu/downloads/mixtvem/>.

³The Mersenne Twister algorithm is used for random number generation (Matsumoto and Nishimura 1998).

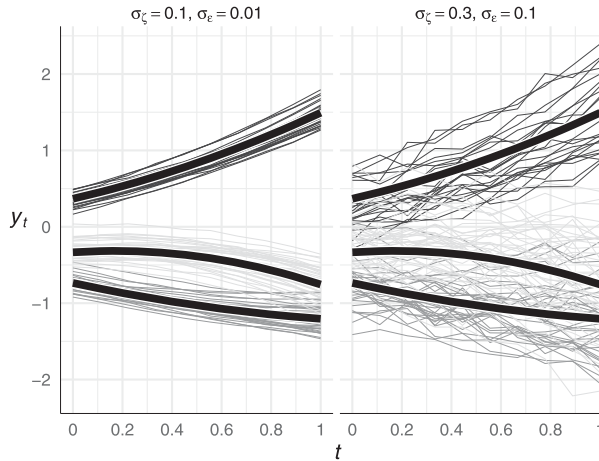


Figure 1. Example datasets for low and high within-group variability and measurement error, respectively. The datasets each contain 100 trajectories of 10 observations. The black lines denote the three group trajectories.

The datasets are generated using growth curve models describing second-order polynomial trajectories, to comprise a mixture of heterogeneous groups. The measurement times are evenly spaced between $[0, 1]$. The group trajectory $\mathbf{y}^{(g)}$ is represented by the model from Equation (3) for $K=2$. The three fixed effects (that is, the intercept $\alpha_0^{(g)}$, slope $\alpha_1^{(g)}$, and quadratic term $\alpha_2^{(g)}$) are sampled from a uniform distribution between -1 and 1 . Thus for some datasets we have well-separated group trajectories, and for other datasets there will be overlapping group trajectories. The random effects $\zeta_0^{(g)}, \zeta_1^{(g)}$ and $\zeta_2^{(g)}$ for the subjects in each group follow a normal distribution (unless mentioned otherwise) with zero mean and standard deviation of $0.1, 0.2$, or 0.3 in the simulation scenarios involving low, medium, and high variability, respectively. For the residuals, we have $\varepsilon_{i,j} \sim N(0, \sigma^2)$, with a standard deviation of 0.01 or 0.1 , representing negligible noise and considerable noise, respectively. An example of how a dataset with relatively well-separated group trajectories can vary in difficulty depending on the settings is illustrated in Figure 1.

We generate groups of different sizes (i.e., number of subjects), whilst ensuring that the smallest group is of sufficient size to be detected. This is achieved by using group proportions $\pi^{(g)} \propto \sqrt{g}$, normalized by a factor of $\sum_{g=1}^G \sqrt{g}$. In the datasets containing six groups, the smallest and largest groups comprise 9% and 55% of the subjects, respectively.

3.1. Design

In the first scenario, we investigate how the methods perform given that the number of groups is correctly specified. We compare the performance of the methods under each permutation of settings involving sample size, number of repeated observations, within-group variability, unexplained variability, and number of groups. Drawing from previous research on sample size requirements of the mixture models (Loughran and Nagin 2006; Nylund, Asparouhov, and Muthén 2007; Tolvanen 2007), we evaluate the sample size at three levels ($N = 200, 500, 1000$). The effect of the number of repeated observations on the performance is evaluated at settings ($T = 4, 10, 25$). The random effects are evaluated for three levels of within-group variability, sampled from a normal distribution with ($\sigma_{\zeta_k} = 0.1, 0.2, 0.3$ where $\sigma_{\zeta_0} = \sigma_{\zeta_1} = \sigma_{\zeta_2}$). Lastly, the within-subject variability is investigated using two levels of white noise ($\sigma_\varepsilon = 0.01, 0.1$), while keeping the variability between subjects constant.

Secondly, we assess the impact of model misspecification on the identification of subgroups. It is known that the model fit of a LME model is sensitive to the assumption on normality in the random effects (Verbeke and Lesaffre 1996; Muthén and Asparouhov 2009), but it remains to be

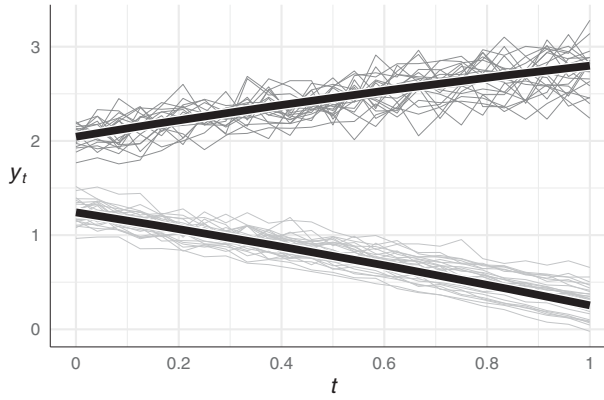


Figure 2. Example dataset for the proportional noise scenario, with $c = 0.05$, $\sigma_\zeta = 0.1$.

seen how this affects the grouping when random effects are lognormal. The datasets comprise random effects $\zeta_k^{(g)} \sim \text{lognormal}(\mu, \frac{1}{2})$ with $\mu = \log(0.15)$, $\log(0.30)$, and $\log(0.45)$ for low, medium, and high variability, respectively. The variability of these scenarios approximately corresponds to the similarly named scenarios involving normally distributed random effects. The positive range on the distribution would mean that the group trajectory is the trajectory with the lowest coefficients. We ensure the group trajectory is representative by centering the random effects around zero by subtracting the median $\exp(\mu)$ (0.15, 0.30 and 0.45 for the three settings, respectively).

Similarly, we also study the effect of a misspecified measurement error by introducing proportional within-subject variability (heteroskedastic), a feature that was observed in the case study. Here, the measurement error is specified as $\varepsilon_{i,j} \sim N(0, \max(0.01, c \cdot \hat{y}_{i,j})^2)$ with scaling factors $c = (0.01, 0.03, 0.05)$. In this scenario, we take $\alpha_0^{(g)} \sim U(1, 3)$ to ensure that the proportional deviation $c \cdot \hat{y}_{i,j}$ exceeds the minimum standard deviation of 0.01 in most cases. An example dataset is shown in Figure 2.

In the last scenario we assess the preferred fit of each method in terms of the number of groups that achieves the best result according to the model selection criteria described below. Some of the methods may produce a better fit for a different number of groups than the correct number due to differences in the group representation. We generate 100 datasets for each number of groups ranging from 3 to 5, on which the methods are evaluated for 2 to 7 groups. We investigate how well the methods perform at recovering the true number of groups using the BIC, under a low and high random effect variability with $\sigma_\zeta = \{0.1, 0.3\}$, respectively.

3.2. Evaluation

Although the large number of datasets precludes a subjective analysis on the fit of the methods, as is commonly done in applications (Nagin and Odgers 2010b), the evaluation provides a balanced assessment across different group trajectories and scenarios. We evaluate the fit of the methods to each dataset using three metrics, namely the correctness of the trajectory group membership, the group trajectory, and the number of groups. Cases in which the model could not be estimated are excluded from the evaluation, and we report how often this happens.

3.2.1. NSJ

First and foremost, the group assignments of the N trajectories are compared to their true group membership. The split-join distance introduced by Van Dongen (2000) measures the similarity between the partitions \mathcal{A} and \mathcal{B} in terms of the number of subset reassignments that are needed

to project the partition onto the other, and back. The partitions comprise sets of subjects that are in the same group g , denoted by $\mathcal{A} = \{a_1, a_2, \dots, a_G\}$ and $\mathcal{B} = \{b_1, b_2, \dots, b_G\}$. We have $a_g = I_g$ with $\bigcup_{g=1}^G I_g = I$ and $I_g \cap I_h = \emptyset$ when $g \neq h$, and the same holds for b_g . The number of groups may differ between the partitions. The number of matching assignments between \mathcal{A} and $\mathcal{A} \cap \mathcal{B}$ is denoted by $N_{\mathcal{A}}(\mathcal{B})$ and computed by

$$N_{\mathcal{A}}(\mathcal{B}) = \sum_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} |a \cap b|, \quad (9)$$

where $|a \cap b|$ denotes the number of subjects that occur in both sets. The distance between the partitions is asymmetric, with

$$\begin{aligned} d(\mathcal{A}, \mathcal{A} \cap \mathcal{B}) &= N - N_{\mathcal{A}}(\mathcal{B}), \\ d(\mathcal{B}, \mathcal{A} \cap \mathcal{B}) &= N - N_{\mathcal{B}}(\mathcal{A}). \end{aligned} \quad (10)$$

It can be seen that if \mathcal{A} contains sets which are proper subsets of a set of \mathcal{B} , then the projection from \mathcal{A} onto \mathcal{B} requires fewer adjustments with respect to these elements than vice versa. A distance of 0 implies that the partition is a subpartition (Van Dongen 2000), which makes the metric suitable for comparing partitions with a different number of groups. Combining the pairwise distances, we obtain the split-join distance

$$d(\mathcal{A}, \mathcal{B}) = d(\mathcal{A}, \mathcal{A} \cap \mathcal{B}) + d(\mathcal{B}, \mathcal{A} \cap \mathcal{B}) \quad (11)$$

The scale of the metric is dependent on the sample size. In order to evaluate the split-join distance across simulation scenarios with different sample sizes, we use the normalized metric

$$\text{NSJ}(\mathcal{A}, \mathcal{B}) = \frac{d(\mathcal{A}, \mathcal{B})}{2N}, \quad (12)$$

which expresses the distance on a scale from 0 to 1 (lower is better). We will refer to this metric as the normalized split-join (NSJ) distance.

3.2.2. WMMSE

While a low NSJ score is expected to be associated with a good fit of the group trajectories, there are cases in which this need not be the case. For example, in the case of overlapping groups it is still possible to obtain proper group trajectory fits, yet there is uncertainty on the subject group membership, impacting the NSJ score. We therefore assess the fit of the group trajectories as a secondary metric. The group trajectories are compared in terms of the mean squared error (MSE) at each observation in time. A challenging aspect of this evaluation is that there is no guaranteed one-to-one mapping between the group and reference group trajectories. We therefore associate each group trajectory $y_{\text{group}}^{(g)}$ with the nearest reference group trajectory y_{ref} , and weigh the score by the group proportion $\pi^{(g)}$. The score, which we shall call the weighted minimum MSE, is denoted by

$$\text{WMMSE} = \frac{1}{T} \sum_{g=1}^G \pi^{(g)} \min_{g' \in G_{\text{ref}}} \sum_{j=1}^T \left(y_{\text{group},j}^{(g)} - y_{\text{ref},j}^{(g')} \right)^2, \quad (13)$$

with $G_{\text{ref}} = \{1, 2, \dots, G_{\text{ref}}\}$, and G_{ref} refers to the true number of groups in the dataset. Due to the relatively small scale of the observed values and the resulting small observation error, the reported WMMSE values are multiplied by 1000.

Table 1. Percentage of convergence issues across all scenarios and datasets.

Method	Did not converge	Empty groups	Solitary groups	Total
KML	0.0%	0.0%	0.11%	0.11%
GCKM	0.0%	0.0%	0.09%	0.09%
GBTM	0.28%	< 0.01%	0.08%	0.36%
GMM	0.04%	2.0%	0.90%	3.0%
MixTVEM	26%	2.5%	4.3%	31%

4. Results

4.1. Simulations

4.1.1. Numerical convergence

We observed problems with the model estimation across the simulation scenarios. The main effects are reported in Table 1. Any convergence issues of GCKM were established by the GCM in the first step, considering that the k -means algorithm is guaranteed to converge. Due to the higher number of parameters in a GMM, it is numerically less stable than KML, GCKM and GBTM. Nevertheless, only MixTVEM exhibits significant convergence problems across scenarios, with an overall nonconvergence rate of 26%. The convergence problems appear to only occur for a large number of observations, independent of the other simulation settings. Whereas a negligible number of problems occur at $T = 4$ (0.22%), for an increasing number of T the convergence problems worsen, with 14% at $T = 10$, and already 78% at $T = 25$. Moreover, the rate of convergence problems increase with the number of groups, with 4.8% at 2 groups and 18% at 6 groups for $T = 10$. A comparison of MixTVEM configurations revealed that the convergence issues mostly occurred when the AR-1 structure was included, although it is unclear why the estimation is affected for a higher number of repeated observations.

A converged fit is not necessarily without problems. In particular, a group can be empty if the posterior class probability of all trajectories is greater for other groups. In the simulations, these problems only occurs at a considerable rate for GMM (2.0%) and MixTVEM (2.5%). GMM primarily exhibits this problem for higher number of groups, with 5.5% at 6 groups across scenarios, whereas hardly any problems occur at 2–4 groups (< 0.48%). Furthermore, the problem more often occurs with low measurement error (8.8%), compared to high measurement error (3.8%). A possible explanation for this discrepancy is that the higher measurement error provides additional possible group trajectories due to the increased overlap between trajectories. Log-normally distributed random effects result further in an increased number of solutions with empty groups. The scenario with log-normally distributed random effects at 6 groups with low measurement error exhibits empty groups in 21% of all converged cases. In MixTVEM, the problem occurs mostly in the setting with 25 observations (34% of converged cases), whereas it occurs infrequently (1% of converged cases) at 10 observations.

Another possible problem with a solution is the presence of groups consisting of a single subject trajectory, assuming that such a solution is not meaningful in practice (in the simulations it is not considered to be meaningful). We refer to this as a solitary group. In GMM, this occurs relatively often only for a higher number of groups (3.8% for 6 groups, while below 0.01% for 2 groups), under log-normally distributed random effects (5.9% at 6 groups), and especially when the number of groups does not match the true number of groups (9.0% at 6 groups). For MixTVEM, solitary groups occurs frequently on data comprising 25 observations (20%), whereas for 10 observations only 5.3% of converged cases have solitary groups. Independent of the number of observations, solitary groups occur more often on log-normal data (15% compared to 2.6% on the normal data), and with smaller sample size (7.3% for $N = 200$, compared to 2.1% for $N = 1000$).

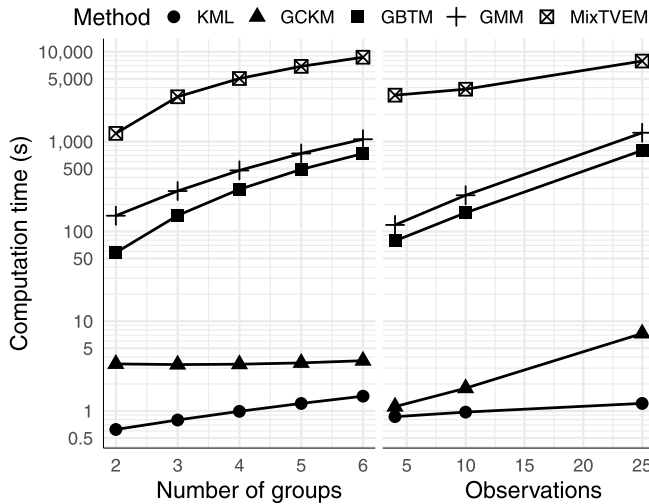


Figure 3. Computation time per model (in seconds) over the number of groups and observations.

4.1.2. Computation time

We assess how well the methods scale relative to an increasing volume of data and increasing number of parameters. In order to illustrate how the methods scale differently, the effects of the number of groups and number of observations on computation time are visualized in Figure 3. The effect of sample size is not shown because all methods scale in computation time with respect to sample size in the same way. The base computation time differs considerably between methods. Whereas *k*-means requires only 1 s on average per dataset run, GMM takes 9 min. This is largely due to GMM comprising a more complex computational problem. MixTVEM stands out from the mixture methods with computation times of over 1 h on average, likely due to its relatively unoptimized implementation compared to the mature R packages available for GBTM and GMM. However, MixTVEM scales relatively well with an increasing number of observation compared to GMM, GBTM and GCKM, suggesting that the method may be favorable for a larger number of observations, given a more optimized implementation. As a result of the independent assessment of the methods in each scenario, the first step of GCKM was recomputed each time, and therefore the computation time is higher than it would be in practice, because the results of the first step could be reused. KML is least impacted among the methods by the inclusion of additional groups and observations. The GCM computation in the first step of GCKM accounts for most of the computation time of the method, hence the near-constant time over the number of groups, and the similar scaling to the mixture methods along the number of observations.

4.1.3. Group assignment for the correct number of groups

The methods are assessed across simulation settings for all possible combinations. This results in 270 unique cases⁴ to be evaluated per method on 100 datasets generated with normally distributed random effects and normal residual. We first evaluate the correctness of the assignment of subject trajectories using the NSJ distance. The main effects of each of the independent factors (that make up the settings) are assessed by means of a linear model. Runs in which a model did not converge are excluded. The results are reported in Table 2. Due to the small standard errors below 0.01, all differences between methods can be considered statistically significant and we do not report on this further. Instead, we focus on the practical significance of the score differences.

⁴We arrived at 270 cases by evaluating all permutations of 5 different number of groups, 3 sample sizes, 3 values for the number of observations, 3 random effect deviations, and 2 measurement errors.

Table 2. Effects of the scenario settings on group assignment per model, averaged over 100 datasets, as measured by the NSJ distance (lower is better). The 'All' row reports the average performance over all cases.

	KML	GCKM	GBTM	GMM	MixTVEM
All	.21	.10	.21	.084	.22
Number of groups G					
2	.10	.039	.10	.034	.088
3	.17	.069	.17	.058	.17
4	.23	.097	.23	.083	.24
5	.27	.13	.27	.11	.28
6	.30	.15	.30	.14	.30
Sample size N					
200	.22	.099	.22	.087	.22
500	.21	.097	.21	.085	.22
1000	.21	.093	.21	.081	.21
Number of observations T					
4	.20	.11	.20	.090	.19
10	.21	.096	.21	.085	.18
25	.22	.085	.22	.078	.28
Random effects deviation σ_{ζ}					
.1	.084	.040	.085	.027	.13
.2	.22	.088	.22	.078	.21
.3	.33	.16	.33	.15	.31
Measurement error σ_{ε}					
.01	.21	.064	.21	.064	.20
.1	.21	.13	.21	.10	.23

The overall NSJ scores show that GMM and GCKM perform significantly better on average than the other models, with scores of 0.084 and 0.10, respectively. In contrast, KML, GBTM, and MixTVEM achieve an average score of approximately 0.21. By comparing the NSJ of KML and GBTM across the main effects, it becomes clear that their results are identical. Although MixTVEM obtains a similar average score, it deviates from the other two methods in some settings, in particular for the number of observations.

The number of possible assignment errors increases with the number of groups. The performance declines at a similar rate across methods for an increasing number of groups. To put the NSJ scores into perspective, the expected NSJ by random assignment, assuming correct group proportions, are 0.41, 0.58, 0.67, 0.72, and 0.76, for $G = 2, \dots, 6$, respectively. In this respect, all of the methods perform significantly better than random grouping even for a larger number of groups.

GCKM and GMM benefit from a larger sample size, and number of observations under the presence of noise, where GCKM approaches the performance of GMM with an increased number of observations. MixTVEM often fails to find a proper fit for $T = 25$, in addition to the convergence problems reported above. The size of the random effects has a considerable impact on the difficulty of the dataset, yet the differences between the methods are relatively stable. The performance of KML, GBTM, and MixTVEM are unaffected by the presence of measurement error, while GCKM and GMM are negatively affected by it. Still, GCKM and GMM outperform the other methods even in these conditions.

4.1.4. Group trajectory estimation with the correct number of groups

In addition to the group assignment accuracy, we investigate the estimation of the group trajectories. The results are reported in Table 3. Overall, GCKM and GMM achieve the best group trajectory estimates. The methods have near-identical performance on average, and the same holds for KML and GBTM. Comparing the findings with the NSJ scores of Table 2, it is evident that on average a lower NSJ is associated with a better group trajectory fit (i.e., lower WMMSE). In contrast, the group trajectory estimation of GCKM improves with an increasing number of observations, surpassing GMM at $T = 25$, even though this is not reflected in the NSJ scores. Another

Table 3. Effects of group trajectory estimation error across simulation scenarios, measured by the WMMSE multiplied by 1000.

	KML	GCKM	GBTM	GMM	MixTVEM
All	15	3.5	15	3.4	38
Number of groups G					
2	10	1.4	10	.14	27
3	13	2.4	13	1.8	32
4	15	3.5	15	3.4	38
5	17	4.5	17	5.0	44
6	19	5.5	20	6.7	49
Sample size N					
200	16	4.6	16	4.4	38
500	15	3.7	15	3.6	38
1000	14	2.1	14	2.2	38
Number of observations T					
4	15	4.4	16	3.1	15
10	15	3.3	15	3.3	15
25	15	2.7	15	3.8	83
Random effects deviation σ_{ζ}					
.1	1.4	.87	1.4	1.7	19
.2	12	2.2	12	2.6	36
.3	32	7.3	32	5.9	59
Measurement error σ_{ε}					
.01	15	2.3	15	3.4	32
.1	15	4.6	15	3.4	44

discrepancy is observed in the worsening WMMSE scores for MixTVEM with an increasing number of groups compared to KML and GBTM, whereas this pattern is not visible when assessing the NSJ scores. This indicates that MixTVEM is able to achieve a similar subject assignment despite worse group trajectory estimates. The high average WMMSE of MixTVEM arises from the poor model fit at $T=25$, whereas for fewer observations the WMMSE of MixTVEM is not significantly different from KML and GBTM.

All methods except MixTVEM benefit from an increased sample size, with GCKM and GMM showing the greatest relative improvement. The magnitude of variation of the random effects affects all methods, but KML and GBTM in particular (with a WMMSE of 32 at $\sigma_{\zeta} = 0.3$ compared to 1.4 at $\sigma_{\zeta} = 0.1$). Moreover, the associated error with an increasing number of groups differs significantly per level of σ_{ζ} . Notably, the fit of GCKM is considerably less accurate than GMM with $\sigma_{\zeta} = 0.1$ for a low number of observations ($T=4$).

4.1.5. Proportional measurement error

We assess the sensitivity of the methods on longitudinal observations with a proportional measurement error. The performance of the methods is shown in Tables 4 and 5 in terms of the NSJ and WMMSE. The methods are evaluated on datasets comprising 2-6 groups, with a standard deviation on the random effects of 0.1 and 0.3.

The results on group assignments follow the observations from the earlier experiment of Table 2 involving two levels of heteroskedasticity, with KML and GBTM being insensitive to the level, and GMM and GCKM benefiting from lower degrees of heteroskedasticity (GCKM from 0.14 to 0.072, GMM from 0.11 to 0.071). MixTVEM shows a relatively small improvement of -0.03 for lower error. Overall, the performance of the models is similar to those on the homoskedastic residual variance. In case of the group trajectory estimation error, the result appears to be unaffected by the level of heteroskedasticity.

4.1.6. Log-normal groups

The results of the scenarios involving log-normally distributed random effects are reported in Tables 6 and 7 in terms of the NSJ and WMMSE, respectively. The scores are compared to those

Table 4. Averaged effects of group assignment error (NSJ) under proportional measurement error.

	KML	GCKM	GBTM	GMM	MixTVEM
Proportional measurement error c					
.01	.21	.072	.21	.071	.17
.03	.21	.10	.21	.089	.19
.05	.21	.14	.21	.11	.20

Table 5. Averaged effects of the group trajectory estimation error (WMMSE $\times 1000$) under proportional measurement error.

	KML	GCKM	GBTM	GMM	MixTVEM
Proportional measurement error c					
.01	16	.23	16	3.2	15
.03	16	3.8	17	3.4	16
.05	17	5.3	17	3.6	18

Table 6. Averaged effects of group estimation (NSJ) under log-normally distributed random effects.

	KML	GCKM	GBTM	GMM	MixTVEM
All	.26	.14	.26	.14	.23
Log-normal random effects mean $\exp(\mu_\zeta)$					
.15	.13	.060	.13	.051	.13
.30	.28	.13	.27	.14	.24
.45	.37	.22	.36	.23	.31
Measurement error σ_ε					
.01	.26	.11	.25	.12	.22
.1	.26	.17	.26	.16	.24

under the standard scenario of Tables 2 and 3, respectively. Overall, all methods except MixTVEM achieve a worse performance compared to the standard scenario. Especially GMM is significantly impacted, although it is the best performing method regardless, indicating the importance of the correct specification of the subgroup distribution.

In terms of group trajectory estimation, KML, GCKM and GBTM achieve the best estimates, with an error of 24. GMM has a higher error of 37 on average compared to GCKM, especially for larger between-group variation. MixTVEM performs relatively poorly across all scenarios.

4.1.7. Finding the number of groups

We investigate how well the methods are able to identify the simulated number of groups using the BLRT and BIC. We generated datasets comprising 500 trajectories across 2–5 groups, and evaluated the methods with a specification of the number of groups ranging from 2 to 7 groups. We consider two scenarios involving low and high within-group heterogeneity ($\sigma_\zeta = 0.1$ and $\sigma_\zeta = 0.3$, respectively). Each of the scenarios is evaluated with 100 generated datasets, resulting in a total of 800 evaluations per method.

Table 8 reports the proportion of correct cases and cases in which the optimal solution was off by one group when using BIC_{\min} or $\text{BIC}_{\text{elbow}}$. The results are computed across the two settings for σ_ζ . The NSJ and WMMSE criteria serve as a reference for the number of groups needed to optimally match the group membership assignment and group trajectory fit, respectively. Due to model limitations, it is possible for a criterion to exceed these values. The results of the BLRT evaluation are shown in Table 9.

GMM has the highest number of correct cases across all criteria; in particular with the BLRT and BIC_{\min} . The BLRT outperforms BIC_{\min} , correctly identifying the number of groups in 86% of datasets, as opposed to 71% using BIC_{\min} . In contrast, GBTM consistently (99.9%) overestimates the number of groups with both criteria. The same pattern of overestimation is observed

Table 7. Averaged effects of group trajectory estimation error (WMMSE $\times 1000$) under log-normally distributed random effects.

	KML	GCKM	GBTM	GMM	MixTVEM
All	25	25	25	37	51
Log-normal random effects mean exp (μ_ζ)					
.15	.63	6.8	6.6	7.9	17
.30	23	.21	23	29	46
.45	45	47	45	74	89
Measurement error σ_ε					
.01	25	25	.25	36	55
.1	25	25	25	38	46

Table 8. Percentage of cases in which the solution as determined by the respective criterion corresponds to the true number of groups, computed across all cases. The NSJ and WMMSE provide a reference to how often the optimal fit in terms of group membership assignment and group trajectory fit correspond to the correct number of groups.

	Group error	KML	GCKM	GBTM	GMM	MixTVEM
Ref _{NSJ}	−1	32%	29%	32%	20%	29%
	0	43%	63%	42%	71%	37%
	+1	1.2%	0.17%	1.0%	1.8%	13%
Ref _{WMMSE}	−1	17%	9.8%	16%	6.5%	20%
	0	55%	61%	53%	66%	41%
	+1	10%	19%	11%	20%	19%
BIC _{min}	−1	0%	15%	0%	22%	17%
	0	0%	30%	0%	71%	28%
	+1	33%	17%	34%	0%	25%
BIC _{elbow}	−1	32%	35%	32%	32%	30%
	0	42%	46%	41%	55%	36%
	+1	7.2%	2.7%	7.5%	1.1%	13%

Table 9. Percentage of cases in which the solution determined by the BLRT corresponds to the correct number of groups.

Scenario	Method	Group error				
		< −1	−1	0	+ 1	> 1
All	GBTM	0%	0%	< 1%	< 1%	99%
	GMM	< 1%	7.3%	86%	5.3%	< 1%
$\sigma_\zeta = .1$	GBTM	0%	0%	< 1%	1.0%	99%
	GMM	< 1%	< 1%	93%	5.3%	0%
$\sigma_\zeta = .3$	GBTM	0%	0%	0%	0%	100%
	GMM	1.0%	14%	79%	5.3%	< 1%

when applying BIC_{min} for KML, GCKM, and MixTVEM. In view of these findings and the computationally intensive aspect of BLRT, we therefore do not evaluate the BLRT for the other methods.

KML, GBTM and MixTVEM achieve far better results using BIC_{elbow}, and come close to the performance with the NSJ as a reference. The solutions of KML and GBTM for minimum WMMSE tend to be closer to the correct number of groups than for the NSJ, indicating that the closest approximation of the group trajectories does not always correspond to a correct group assignment.

The magnitude of σ_ζ has a significant effect on the proportion of correct cases, as can be seen from Tables 10 and 9. KML and GBTM are most affected by the large within-group variability, with respect to both the group assignment (from 70% to 15%) and group trajectory fit (from 82% to 26%). GMM achieves a correctness of 93% under low variability for both BLRT and BIC_{min}. Under high variability, the performance degrades to 49% when using BIC_{min} but only to 79% when using BLRT.

Table 10. Percentage of cases with the correct number of groups according to the respective criterion.

	σ_ζ	KML	GCKM	GBTM	GMM	MixTVEM
Ref _{NSJ}	.1	71%	75%	70%	90%	49%
	.3	15%	51%	14%	52%	24%
Ref _{WMMSE}	.1	83%	58%	80%	75%	50%
	.3	27%	64%	26%	56%	33%
BIC _{min}	.1	0%	7.7%	0%	93%	40%
	.3	0%	53%	0%	49%	16%
BIC _{elbow}	.1	50%	52%	49%	44%	37%
	.3	34%	40%	33%	37%	34%

Overall, the performance of the methods is consistent between the NSJ and WMMSE criteria except for GCKM, in which the group trajectory estimation does not improve under lower variability. The decrease in performance over σ_ζ is less prominent in the BIC results, demonstrating that the approach is relatively robust. For high variability, the correct cases of KML and GBTM for BIC_{elbow} even exceed those of the reference criteria.

4.2. Case study

We investigate the usage data of sleep apnea patients undergoing positive airway pressure (PAP) treatment. Weaver et al. (2007) have demonstrated the importance of sufficient usage of CPAP therapy, where increasing daily usage was associated with a better outcome. Four hours of usage per day is considered to be the minimum for adequate treatment. Many patients struggle to accommodate to the therapy, resulting in much lower hours of use, or an abandonment of the therapy, whereas other patients may improve over time. Other patients establish a preferred number of hours of usage early on, and remain constant over time. In identifying the common longitudinal patterns of usage, we can describe patients in greater detail and quantify the occurrence of certain patterns. Previous studies have investigated these patterns using a two-step approach (Aloia et al. 2008; Babbin et al. 2015).

The daily usage data was collected from a retrospective observational study in the US over the past six years, comprising 2,686 patients diagnosed with sleep apnea and being on therapy for the first time. The average age of the patients is 60 years ($\sigma = 15$ years). We focus on the first 6 months of therapy, including only patients with at least 6 months of data (1,745 patients). Days on which the device data is missing indicate that the therapy was not used, and are therefore represented by usage of zero hours (accounting for 30% of all days). Given that patients used the therapy, the mean daily usage is 6.3 hours ($\sigma = 2.7$ h). Due to the computational requirements of the mixture methods, we downsample the usage data into weekly averages, resulting in 25 observations per patient.

In this exploratory analysis, we are primarily interested in identifying groups of patients that exhibited a change in usage over time, and the patterns of change associated with these groups. We consider the mean level of usage to be of less relevance when it is above 4 hours as this is generally regarded as the minimum for adherence. The balance in relevance of the mean level of usage is challenging to capture in a single metric. We therefore apply a hybrid approach, basing our decision of the preferred number of groups on a combination of the model information criterion and model results (Van de Schoot et al. 2017). We use the BIC as a starting point toward determining the ideal solution. The solutions close to the preferred number of groups indicated by the BIC are then evaluated regarding the clinical interpretation of the group trajectories (Feldman, Masyn, and Conger 2009). A solution involving more groups is preferred only if it contains a group trajectory exhibiting change or a low mean level of usage not present in solutions with a fewer number of groups. Moreover, the solution needs to have groups of considerable size (greater than 5%). Lastly, we assess

the confidence in trajectory assignments to the groups, which we measure using relative entropy (Van de Schoot et al. 2017; McNeish and Harring 2017).

4.2.1. KML

We apply KML as described in Sec. 2.1. Due to the low running time we can evaluate the method on 1 to 8 groups in a short amount of time, with even the eight-groups solution requiring only 11 seconds to compute. The sequence of BIC values of Figure 4a show a consistent but diminishing improvement of the model fit over an increasing number of groups.

We select the solution comprising 6 groups because it provides a balance between the level of detail by which the patients are described, and the number of group trajectories involved. The group trajectories are depicted in Figure 4b. The group trajectory coefficients are shown in the Appendix, in Table 11. The confidence in the classification of the trajectories is strong, with a relative entropy of 0.96. The majority of patients appear to follow a near-constant trajectory, with the group trajectories A (6.7 h), B (4.8 h), C (0.6 h), and D (8.6 h) comprising 80% of all patients. With respect to adherence, group B is of particular interest due to variability around the compliance threshold of 4 hours. The proportion of patients with near-zero usage (group C) accounts for 19% of all patients. This group, together with group E, describes patients that either stop using the therapy or use it infrequently. Group F describes patients that start out as non-compliant but improve their usage throughout the therapy.

4.2.2. GCKM

The trajectory coefficients underlying the GCKM method are obtained using a GCM with 3rd-order orthogonal polynomial random effects. We selected this model by fitting growth curve models of increasing complexity to the data and selecting the model that minimizes the BIC. The model choice is further supported by the group trajectories found by KML in Figure 4b. The *k*-means algorithm is applied in the same way as was done with KML. The first step of the approach, involving the estimation of the GCM, only needs to be done once. This took only 89 seconds, followed by *k*-means clustering, of which we evaluated the solutions from 1 to 8 number of groups.

We choose the solution involving eight groups, which is shown in Figure 5b. The group trajectory coefficients are depicted in the Appendix in Table 12. Despite the large number of groups, some groups are of considerable size. Group A, B, and C already comprise 68% of patients, meaning that the remaining groups capture trajectories that occur less frequently. The relative entropy is 0.86, indicating a good separation of groups. The group trajectories of each of the three major groups is near-constant, with an average usage of 7.7 hours, 5.1 hours, and 0.87 hours, respectively. The latter trend represents patients who were mostly non-compliant throughout the 6 months of therapy, accounting for 19% of patients. Group D describes patients (8%) that were non-compliant at the start of therapy but improved later on. The remaining groups describe group trajectories with periods of noncompliance at specific moments in time.

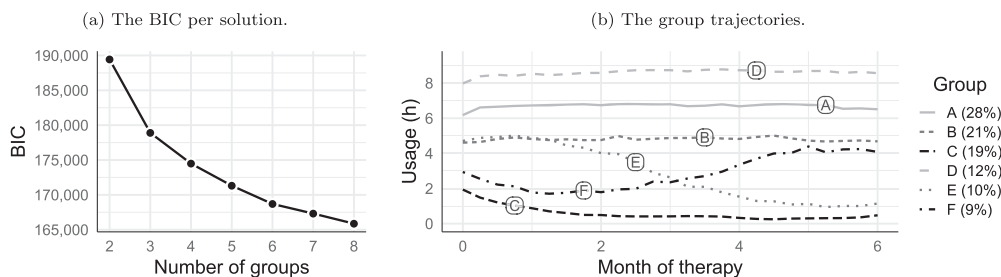


Figure 4. Case study analysis using KML.

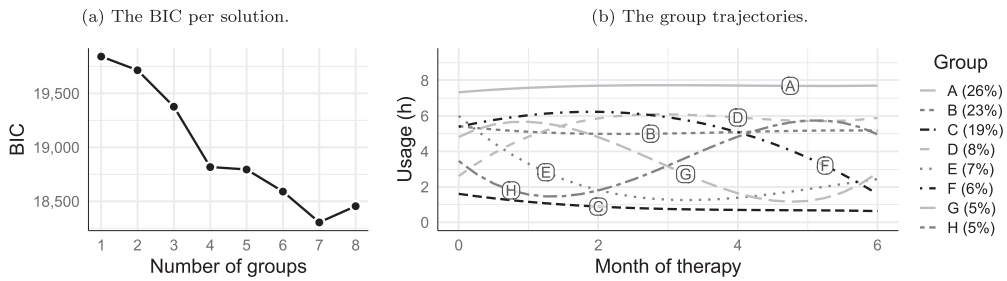


Figure 5. Case study analysis using GCKM.

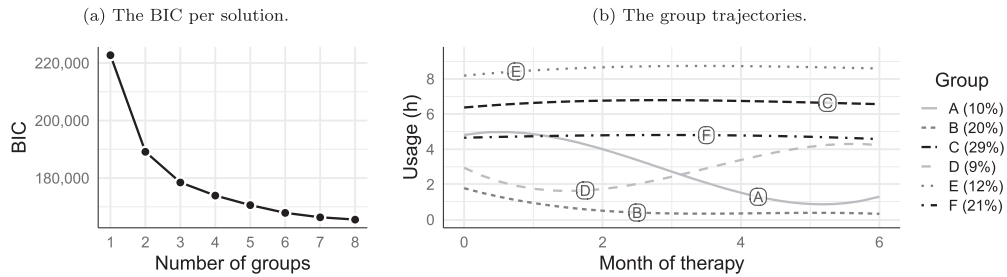


Figure 6. Case study analysis using GBTM.

4.2.3. GBTM

The GBTM model is initialized using 50 random starts, and the best candidate model is estimated until either convergence or 500 iterations are reached. We fit GBTMs with 3rd-order orthogonal polynomial group trajectories, based on the reasoning provided in the GCKM analysis described above. In total, the estimation of two groups takes about 6 min, whereas the eight-groups model takes almost 2 h of computation time. The 2 h of computation time consist of 80 minutes for estimating the random starts, and 39 min for optimizing the final model. A converged solution was obtained for all evaluated number of groups.

We arrive at the solution in Figure 6, representing 6 group trajectories. The group trajectory coefficients are shown in the Appendix, in Table 13. With a relative entropy of 0.96, the solution exhibits a strong separation of groups. About 82% of patients are assigned to one of the four near-constant group trajectories (B, C, E, F). Group B represents patients with near-zero usage (0.60 h on average), comprising 20% of all patients. Group A and D describe patients who are mostly compliant at the start and drop off later on, and vice versa.

4.2.4. GMM

We apply GMMs with polynomial group trajectories of degree 3, based on the reasoning that was provided in the GCKM analysis. Furthermore, we specify a random intercept, and a shared diagonal variance-covariance matrix across the latent classes. This is done in order to limit the added complexity of the model, in view of the large sample size. The model is initialized using 50 random starts, and the best candidate model is iterated until either convergence or 500 iterations are reached. Despite the restrictions imposed on the model, the model estimation remains numerically challenging. This is evident from the time needed for the model to converge. Whereas the two-class model requires 8 minutes of computation time, the eight-class model takes 4.5 h to compute. Although GMM converged for all evaluated number of groups, the solution for five and six groups was rather poor, having 7 empty groups.

We choose the solution with 7 groups, primarily because the solutions involving a lower number of groups have a group comprising 70% of all patients, with a constant trajectory and large

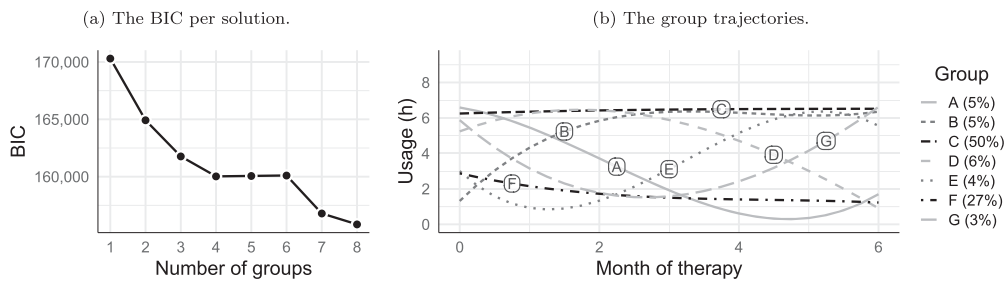


Figure 7. Case study analysis using GMM.

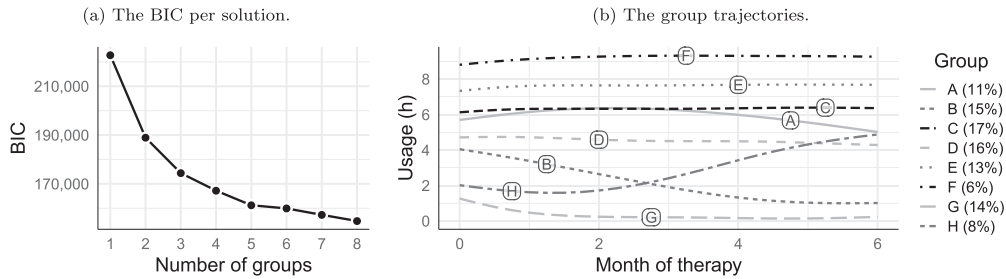


Figure 8. Case study analysis using MixTVEM.

between-patient variability. The preferred solution is shown in Figure 7, and consists of two large groups (C and F) describing patients with near-constant usage. Group C is the largest group (50% of all patients), and describes a constant usage around 6 h. Despite the large number of groups, the relative entropy is high, with a value of 0.93. The group trajectory coefficients are shown in Table 14, in the Appendix.

4.2.5. MixTVEM

We initialize the model with 50 random starts to ensure a good fit. In contrast to the simulations, we do not include the autocorrelation term here. This was done to reduce the running time, which even with this simplification requires 7 h to arrive at the 2-groups solution, and 36 hours with 8 groups. Across the different solutions, the group trajectories discovered by MixTVEM are mostly constant, likely due to the penalization factor and the variability between patients.

It is not until 8 groups that the solution contains multiple curved group trajectories. The group trajectories are shown in Figure 8. The group trajectory coefficients are shown in Table 15 in the Appendix. The relative entropy of 0.97 indicates a strong confidence in the classification of the trajectories. Group C, D, E, F and G (66% in total) represent constant usage over time. Out of these, groups G (14%) describes patients with near zero-hour usage. Group A (11%) contains a wide range of trajectories, considering its large within-group variability, and it is questionable whether the group trajectory is representative. Group B and H (23% in total) represent patients who decline or increase over time, respectively.

4.2.6. Evaluation

In general, it is difficult, if not impossible, to establish with certainty whether the data truly comprises heterogeneous subgroups, as the observed subgroups could be an artifact from model misspecification (Bauer 2007). In a more practical approach, referred to as an indirect application of clustering, the data is not assumed to comprise distinct subgroups, but instead comprises a complex spectrum which can be discretized into subgroups for ease of analysis and reporting (using e.g., KML or GBTM) (Nagin and Odgers 2010a). None of the methods found a solution involving distinct groups, so it is

difficult to establish which method achieved the best result. If one regards it as a segmentation problem, where the within-group error should be minimized, KML or GBTM would be preferable. On the other hand, GMM and GCKM focus on grouping similar trajectory shapes (i.e., the coefficients), resulting in larger groups of subjects with similar trajectories, and consequently more groups with varying shapes. Nevertheless, the confidence in the classification of the trajectories, as indicated by the relative entropy, was found to be high for all methods, despite the many number of groups. Care should be taken in interpreting these varying shapes, as the edges of the polynomial group trajectories may not be representative of the underlying data, but instead could be an artifact of the limited representation (Sher, Jackson, and Steinley 2011). Using a spline representation, such as the one used in MixTVEM, results in more reliable trajectory estimates.

5. Discussion

We evaluated the ability of longitudinal clustering methods to identify group trajectories in synthetic datasets with known groups, displaying different slowly changing longitudinal patterns. Each combination of conditions involving sample size, number of repeated observations, within-group variability, unexplained variability, and number of groups was explored. This approach allows for an objective assessment of each method, and of the degree to which performance is affected by the conditions. Although it may seem unreasonable to represent real-life datasets by relatively simple synthetic datasets comprising heterogeneous subgroups around a polynomial group trajectory (Raudenbush 2005; Bauer 2007), it embodies the assumption behind GMM; which is one of the more commonly used methods for longitudinal clustering. A similar simulation approach has been undertaken, for example, by Martin and von Oertzen (2015), and McNeish and Harring (2017). While we only evaluated each permutation of settings on 100 synthetic datasets, the actual number of evaluated datasets is much greater, primarily due to the different number of groups across settings. In the scenario involving group assignment with the correct number of groups, 27,000 cases were evaluated.

In the simulations, GMM and GCKM significantly outperform the other methods across all scenarios, both in terms of group assignment and estimation of the group trajectories. The NSJ distances of the other methods were approximately twice as high, and the WMMSE was about 4 times higher. The good performance of GMM is not unexpected, after all, the heterogeneous subgroups are its best-case scenario. Moreover, similar findings have been reported in the comparison by Martin and von Oertzen (2015), although they did not evaluate GCKM. At the same time, its performance is closely matched with the two-step approach of GCKM, with the benefit that the computation time of the latter is two orders of magnitude lower. The discrepancy in performance of KML and GCKM demonstrates the benefit of dimensionality reduction in the first step, describing the characteristics of the trajectory more concisely. These results are in contrast to the findings by Twisk and Hoekstra (2012), who concluded that KML and GCKM gave similar results. KML and GBTM were found to have near-identical solutions under all scenarios. This relates to the conclusion of Feldman, Masyn, and Conger (2009), who found that longitudinal latent class analysis (LLCA), a method that could be regarded as a naive clustering approach such as KML, obtains similar results to GBTM. Our findings suggest that, in general, KML is the preferred choice over GBTM because of its considerable flexibility in describing the trajectories, lower computation time, and better scaling. However, GBTM is preferred when the data contains missing or nonaligned observations, when there is prior knowledge on the shape of the group trajectories, or when covariates are to be included into the trajectories. MixTVEM performed marginally better than KML and GBTM (a difference in NSJ distance of -0.02) for 4 and 10 observations, despite its group trajectory estimation error being approximately 2.5 times higher. Its poor performance for 25 observations raised the overall average NSJ distance above the other methods, however. Due to its performance in our simulation and its computational burden, we

consider the other methods to be preferable. In contrast to our findings regarding MixTVEM, Yang, Shao, and Cai (2019) found that MixTVEM works well in most cases in identifying the number of groups and the group trajectories, even for a larger number of observations. The difference in findings could be due to the higher level of subgroup heterogeneity in our simulations.

Having evaluated the effect of sample size under various settings, all methods except MixTVEM marginally benefited from an increased sample size. This is in agreement with previous studies (Martin and von Oertzen 2015; Peugh and Fan 2012). Interestingly, the sample size requirements did not go up with an increasing number of groups (up to 6). This indicates that for datasets with sufficiently distinct group trajectories, a small sample size of 200 trajectories may suffice. Although GMM and GCKM produced similar results, a shortcoming of GCKM becomes evident when comparing the performance across different number of observations. GMM performs relatively well under a low number of observations, whereas GCKM benefits from having more observations to obtain reliable estimates of the random effects. Due to the similar performance to GMM for a higher number of observations, and a better run time scaling with model complexity, GCKM is the favorable option for ILD, due to computational speed.

The assessment of the optimal solution based on the NSJ distance showed that KML, GCKM, GBTM and especially GMM were able to represent the underlying groups well for datasets with low within-group variability. Under larger within-group variability however, only GCKM and GMM were able to do this to a satisfactory degree. Optimizing for the WMMSE, it was found that especially KML and GBTM excel in representing the true underlying group trajectories, although they are surpassed by GCKM and GMM on datasets involving large within-group variability.

Applying GMM in combination with either the BIC or BLRT resulted in the correct selection of the number of groups at a high rate. The BLRT outperformed the BIC in the scenario with high within-group heterogeneity (79% against 49%), whereas the recovery rate was identical under low within-group heterogeneity (93%). The selection of the number of groups turned out to be more difficult for the other methods (around 50% for low group variability) when applying the same metrics. In particular, KML and GBTM tended to consistently overestimate the number of groups when minimizing the BIC (and BLRT in case of GBTM), which is an observation that has also been noted by others (Twisk and Hoekstra 2012; Feldman, Masyn, and Conger 2009). While minimizing the BIC is recommended for GBTM by some (Nylund, Asparouhov, and Muthén 2007; Frankfurt et al. 2016), we found that better results were obtained across scenarios by applying the elbow method on the BIC.

The scenario involving a proportional measurement error confirms the insensitivity to measurement error of KML and GBTM that was observed in the standard scenario. The performance of GCKM and GMM degrades with increasing heteroskedasticity, but even under high measurement error the methods perform better than KML, GBTM and MixTVEM.

The evaluation of the log-normally distributed groups demonstrated that the group assignment accuracy of all methods except MixTVEM degraded, although only slightly. GMM showed the highest relative degradation in performance, indicating sensitivity of the performance to the correct specification of the model, though not substantively (Kreuter and Muthén 2008). The group trajectory estimation error is elevated for all methods. Most notably, GMM exhibited large group trajectory estimation errors compared to the standard scenario, significantly exceeding the errors of GCKM, KML, and GBTM.

Regarding the evaluation of the case study, there is a compelling agreement between KML and GBTM. The similarity of the group trajectories is apparent from Figures 4b and 6b, and further confirmed by a WMMSE of 0.91). Consequently, the group assignment agreement is high as well, with an NSJ of only 0.023. In contrast, the NSJ between KML and GCKM is 0.38. The results of GCKM are closest to those of GMM (NSJ = 0.19, WMMSE = 25), although not as close as the simulation results would suggest. MixTVEM appeared to be conservative in its estimation of the group trajectories, resulting mostly in constant trajectories whereas the other methods showed a more varied range of curves. This is likely a consequence of the regularization term.

Although mixture methods are numerically challenging to estimate, GBTM and GMM experienced few convergence problems on the synthetic datasets. This convergence rate of GMM is in contrast to our case study evaluation, as well as from experiences described by others (Tolvanen 2007; Feldman, Masyn, and Conger 2009; Twisk and Hoekstra 2012; Frankfurt et al. 2016; McNeish and Harring 2017), where GMM was found to exhibit convergence problems, especially for more complex specifications. It is for this reason that GBTM is the recommended method by Frankfurt et al. (2016). The discrepancy in convergence rate could be due to the satisfaction of all assumptions of the model for the synthetic data, whereas on real-life datasets the groups, if they exist, the subgroups are more heterogeneous. Moreover, we applied GMM with a shared variance-covariance matrix, resulting in a less complex model. The solutions of GMM comprising one or more empty groups could be partly due to a few synthetic datasets comprising duplicate groups, i.e., groups with nearly the same group trajectory. Therefore, these datasets effectively have a lower true number of groups. MixTVEM exhibited convergence problems, especially for a larger number of observations, independent of the other simulation settings. The cause of the frequent convergence problems (78% at 25 observations) is unclear. On the same topic, the numerical complexity increases significantly for the mixture models with an increasing number of observations, which results in poor scalability of these methods on ILD. In fitting a GMM or GBTM with four or 25 observations, we observed a ten-fold increase in computation time. On this aspect, GCKM and KML have a clear advantage.

6. Conclusion

The simulations showed that GMM and GCKM outperform KML, GBTM, and MixTVEM on datasets comprising heterogeneous subgroups. In view of the strong assumption of heterogeneous subgroups, the other methods cannot be ruled out in real-life situations for explaining heterogeneity. KML and GBTM were found to have nearly identical results when the group trajectory of GBTM were properly specified, suggesting that KML could provide a good starting point for a GBTM analysis. MixTVEM suffered from significant convergence problems at 25 observations, so under the evaluated specification, GBTM would be preferred. Overall, GMM was found to perform best. Considering the close results between GMM and GCKM however, we recommend GCKM in ILD applications due to its computational efficiency and scaling.

Acknowledgments

The authors are grateful for the thoughtful comments provided by the anonymous reviewer, which have helped to improve the content of this paper.

Funding

This work was supported by Philips Research, Eindhoven, the Netherlands. Niek Den Teuling and Steffen Pauws are employees of Philips.

ORCID

N. G. P. Den Teuling  <http://orcid.org/0000-0003-1026-5080>
E. R. van den Heuvel  <http://orcid.org/0000-0001-9157-7224>

A case study models

A.1 KML

Table 11. KML group trajectory parameters for the selected solution.

	Group g					
	A	B	C	D	E	F
$\pi^{(g)}$	28%	21%	19%	12%	10%	9%
$\hat{\mu}_{g,1}$	6.2	4.6	1.9	8.0	4.7	2.9
$\hat{\mu}_{g,2}$	6.6	4.6	1.5	8.4	4.9	2.6
$\hat{\mu}_{g,3}$	6.7	4.8	1.2	8.5	4.9	2.2
$\hat{\mu}_{g,4}$	6.7	4.9	1.0	8.4	5.0	2.1
$\hat{\mu}_{g,5}$	6.7	4.8	.88	8.5	4.9	1.8
$\hat{\mu}_{g,6}$	6.7	4.8	.70	8.5	4.8	1.7
$\hat{\mu}_{g,7}$	6.8	4.8	.60	8.5	4.5	1.8
$\hat{\mu}_{g,8}$	6.8	4.8	.50	8.6	4.3	1.9
$\hat{\mu}_{g,9}$	6.7	4.8	.49	8.6	4.0	1.8
$\hat{\mu}_{g,10}$	6.8	5.0	.43	8.7	4.0	1.9
$\hat{\mu}_{g,11}$	6.8	4.8	.41	8.7	3.5	2.0
$\hat{\mu}_{g,12}$	6.8	4.8	.41	8.7	2.9	2.4
$\hat{\mu}_{g,13}$	6.8	4.9	.42	8.7	2.6	2.4
$\hat{\mu}_{g,14}$	6.7	4.9	.43	8.7	2.1	2.6
$\hat{\mu}_{g,15}$	6.7	4.9	.42	8.8	2.1	2.7
$\hat{\mu}_{g,16}$	6.8	4.8	.40	8.8	1.8	3.0
$\hat{\mu}_{g,17}$	6.7	4.8	.33	8.7	1.5	3.4
$\hat{\mu}_{g,18}$	6.7	4.9	.27	8.7	1.3	3.7
$\hat{\mu}_{g,19}$	6.8	5.0	.25	8.6	1.3	4.0
$\hat{\mu}_{g,20}$	6.8	4.9	.29	8.6	1.1	4.0
$\hat{\mu}_{g,21}$	6.8	4.7	.30	8.7	1.1	4.4
$\hat{\mu}_{g,22}$	6.8	4.7	.31	8.7	.95	4.1
$\hat{\mu}_{g,23}$	6.5	4.7	.32	8.6	1.0	4.2
$\hat{\mu}_{g,24}$	6.6	4.7	.36	8.6	1.0	4.2
$\hat{\mu}_{g,25}$	6.5	4.7	.48	8.6	1.1	4.1

A.2 GCKM

Table 12. GCKM group trajectory parameters for the selected solution.

	Group g						
	A	B	C	D	E	F	G
$\pi^{(g)}$	44%	23%	9%	8%	5%	5%	5%
$\zeta_0^{(g)}$	5.7	3.0	5.1	3.6	4.1	4.1	4.9
$\zeta_1^{(g)}$	7.0	−44	130	−160	−320	270	−240
$\zeta_2^{(g)}$	−.81	23	−120	190	40	68	−170
$\zeta_3^{(g)}$.31	−10	56	−37	150	−150	−4.7

A.3 GBTM

Table 13. GBTM group trajectory parameters for the selected solution. The standard error is reported in brackets.

	Group g					
	A	B	C	D	E	F
$\pi^{(g)}$	10%	20%	29%	9%	12%	21%
$\beta_0^{(g)}$	2.8 (.044)	.60 (.021)	6.7 (.025)	2.9 (.035)	8.6 (.033)	4.7 (.028)
$\beta_1^{(g)}$	−320 (7.6)	−67 (3.8)	6.0 (3.3)	170 (7.8)	21 (4.6)	−3.7 (5.1)
$\beta_2^{(g)}$	22 (5.8)	48 (3.6)	−22 (3.0)	78 (6.3)	−23 (4.5)	−13 (3.8)
$\beta_3^{(g)}$	78 (5.4)	−20 (3.6)	4.8 (3.0)	−73 (5.8)	3.9 (4.5)	−.94 (3.8)

A.4 GMM

Table 14. GMM group trajectory parameters for the selected solution. The standard error is reported in brackets.

	Group g						
	A	B	C	D	E	F	G
$\pi^{(g)}$	5%	5%	50%	6%	4%	27%	3%
$\beta_0^{(g)}$	2.7 (.19)	5.4 (.19)	6.4 (.071)	4.9 (.18)	3.5 (.22)	1.7 (.11)	3.3 (.26)
$\beta_1^{(g)}$	−420 (8.9)	220 (7.2)	16 (2.3)	−290 (7.1)	380 (7.6)	−86 (3.5)	77 (9.0)
$\beta_2^{(g)}$	150 (7.0)	−170 (7.3)	−5.9 (2.1)	−190 (6.4)	79 (7.4)	37 (3.2)	310 (11)
$\beta_3^{(g)}$	88 (6.9)	73 (6.6)	.93 (1.9)	17 (6.8)	−180 (7.2)	−14 (3.1)	−24 (9.0)

A.5 MixTVEM

Table 15. MixTVEM group trajectory parameters for the selected solution. Here, $B_c^{(g)}(t)$ denote the group-specific spline basis function coefficients (Dziak et al. 2015).

	Group g						
	A	B	C	D	E	F	G
$\pi^{(g)}$	11%	15%	17%	16%	14%	6%	14%
$B_1^{(g)}(t)$	5.3	4.6	5.8	4.6	6.9	8.4	2.2
$B_2^{(g)}(t)$	5.7	4.1	6.2	4.7	7.4	8.8	1.2
$B_3^{(g)}(t)$	6.1	3.5	6.3	4.8	7.6	9.1	.45
$B_4^{(g)}(t)$	6.4	2.9	6.3	4.6	7.7	9.2	.24
$B_5^{(g)}(t)$	6.4	2.2	6.3	4.5	7.6	9.3	.21
$B_6^{(g)}(t)$	6.2	1.6	6.3	4.5	7.7	9.3	.21
$B_7^{(g)}(t)$	5.9	1.2	6.4	4.5	7.7	9.3	.14
$B_8^{(g)}(t)$	5.5	.98	6.4	4.4	7.7	9.3	.13
$B_9^{(g)}(t)$	5.0	1.0	6.4	4.3	7.7	9.3	.23
$B_{10}^{(g)}(t)$	4.5	1.1	6.3	4.2	7.7	9.2	.34

References

- Aloia, M. S., M. S. Goodwin, W. F. Velicer, J. T. Arnedt, M. Zimmerman, J. Skrekas, S. Harris, and R. P. Millman. 2008. Time series analysis of treatment adherence patterns in individuals with obstructive sleep apnea. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 36 (1):44–53. doi:[10.1007/s12160-008-9052-9](https://doi.org/10.1007/s12160-008-9052-9).
- Arthur, D., and S. Vassilvitskii. 2007. *k*-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, Philadelphia, PA, USA, pp. 1027–35. ACM, New York.
- Babbitt, S. F., W. F. Velicer, M. S. Aloia, and C. A. Kushida. 2015. Identifying longitudinal patterns for individuals and subgroups: An example with adherence to treatment for obstructive sleep apnea. *Multivariate Behavioral Research* 50 (1):91–108. doi:[10.1080/00273171.2014.958211](https://doi.org/10.1080/00273171.2014.958211).
- Bauer, D. J. 2007. Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research* 42 (4):757–86. doi:[10.1080/00273170701710338](https://doi.org/10.1080/00273170701710338).
- Burton, A., D. G. Altman, P. Royston, and R. L. Holder. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine* 25 (24):4279–92. doi:[10.1002/sim.2673](https://doi.org/10.1002/sim.2673).
- Dziak, J. J., R. Li, X. Tan, S. Shiffman, and M. P. Shiyko. 2015. Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological Methods* 20 (4):444–69. doi:[10.1037/met0000048](https://doi.org/10.1037/met0000048).
- Ernst, A. F., M. E. Timmerman, B. F. Jeronimus, and C. J. Albers. 2019. Insight into individual differences in emotion dynamics with clustering. *Assessment*. Advance online publication. doi:[10.1177/1073191119873714](https://doi.org/10.1177/1073191119873714)
- Feldman, B. J., K. E. Masyn, and R. D. Conger. 2009. New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology* 45 (3):652–76. doi:[10.1037/a0014851](https://doi.org/10.1037/a0014851).
- Frankfurt, S., P. Frazier, M. Syed, and K. R. Jung. 2016. Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. *The Counseling Psychologist* 44 (5):622–60. doi:[10.1177/0011000016658097](https://doi.org/10.1177/0011000016658097).
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian data analysis*. New York: Chapman and Hall/CRC.
- Genolini, C., X. Alacoque, M. Sentenac, and C. Arnaud. 2015. kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software* 65 (4):1–34. doi:[10.18637/jss.v065.i04](https://doi.org/10.18637/jss.v065.i04).
- Genolini, C., and B. Falissard. 2010. Kml: *K*-means for longitudinal data. *Computational Statistics* 25 (2):317–28. doi:[10.1007/s00180-009-0178-4](https://doi.org/10.1007/s00180-009-0178-4).
- Hardy, A. 1994. An examination of procedures for determining the number of clusters in a data set. In *New approaches in classification and data analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, 178–85. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Jung, T., and K. A. S. Wickrama. 2008. An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass* 2 (1):302–17. doi:[10.1111/j.1751-9004.2007.00054.x](https://doi.org/10.1111/j.1751-9004.2007.00054.x).
- Kreuter, F., and B. Muthén. 2008. Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology* 24 (1):1–31. doi:[10.1007/s10940-007-9036-0](https://doi.org/10.1007/s10940-007-9036-0).
- Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* 38 (4):963–74. doi:[10.2307/2529876](https://doi.org/10.2307/2529876).
- Lee, J. Y., J. S. Brook, S. J. Finch, and D. W. Brook. 2016. Trajectories of cigarette smoking beginning in adolescence predict insomnia in the mid thirties. *Substance Use & Misuse* 51 (5):616–24. doi:[10.3109/10826084.2015.1126747](https://doi.org/10.3109/10826084.2015.1126747).
- Loughran, T., and D. S. Nagin. 2006. Finite sample effects in group-based trajectory models. *Sociological Methods & Research* 35 (2):250–78. doi:[10.1177/0049124106292292](https://doi.org/10.1177/0049124106292292).
- Lu, Z., and X. Song. 2012. Finite mixture varying coefficient models for analyzing longitudinal heterogeneous data. *Statistics in Medicine* 31 (6):544–60. doi:[10.1002/sim.4420](https://doi.org/10.1002/sim.4420).
- MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Volume 1, pp. 281–97. Oakland, CA, USA: Univ. California Press, Berkeley, Calif.
- Martin, D. P., and T. von Oertzen. 2015. Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal* 22 (2):264–75. doi:[10.1080/10705511.2014.936340](https://doi.org/10.1080/10705511.2014.936340).
- Matsumoto, M., and T. Nishimura. 1998. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* 8 (1):3–30. doi:[10.1145/272991.272995](https://doi.org/10.1145/272991.272995).

- McNeish, D., and J. R. Harring. 2017. The effect of model misspecification on growth mixture model class enumeration. *Journal of Classification* 34 (2):223–48. doi:[10.1007/s00357-017-9233-y](https://doi.org/10.1007/s00357-017-9233-y).
- Muthén, B., and T. Asparouhov. 2009. Growth mixture modeling: Analysis with non-Gaussian random effects. In *Longitudinal data analysis*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, 143–65. Boca Raton, FL: CRC Press.
- Muthén, B., C. H. Brown, K. Masyn, B. Jo, S.-T. Khoo, C.-C. Yang, C.-P. Wang, S. G. Kellam, J. B. Carlin, and J. Liao. 2002. General growth mixture modeling for randomized preventive interventions. *Biostatistics (Oxford, England)* 3 (4):459–75. doi:[10.1093/biostatistics/3.4.459](https://doi.org/10.1093/biostatistics/3.4.459).
- Muthén, B., and K. Shedden. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55 (2):463–69. doi:[10.1111/j.0006-341x.1999.00463.x](https://doi.org/10.1111/j.0006-341x.1999.00463.x).
- Nagin, D. S. 1999. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods* 4 (2):139–57. Early comprehensive paper-length review of GBTM. doi:[10.1037/1082-989X.4.2.139](https://doi.org/10.1037/1082-989X.4.2.139).
- Nagin, D. S., and K. C. Land. 1993. Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology* 31 (3):327–62. doi:[10.1111/j.1745-9125.1993.tb01133.x](https://doi.org/10.1111/j.1745-9125.1993.tb01133.x).
- Nagin, D. S., and C. L. Odgers. 2010a. Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology* 6 (1):109–38. doi:[10.1146/annurev.clinpsy.121208.131413](https://doi.org/10.1146/annurev.clinpsy.121208.131413).
- Nagin, D. S., and C. L. Odgers. 2010b. Group-based trajectory modeling (nearly) two decades later. *Journal of Quantitative Criminology* 26 (4):445–53. doi:[10.1007/s10940-010-9113-7](https://doi.org/10.1007/s10940-010-9113-7).
- Nagin, D. S., and R. E. Tremblay. 2005. Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology* 43 (4):873–904. doi:[10.1111/j.1745-9125.2005.00026.x](https://doi.org/10.1111/j.1745-9125.2005.00026.x).
- Nylund, K. L., T. Asparouhov, and B. O. Muthén. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal* 14 (4):535–69. doi:[10.1080/10705510701575396](https://doi.org/10.1080/10705510701575396).
- Pelleg, D., and A. W. Moore. 2000. X-means: Extending *k*-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Volume 1 of *ICML 2000*, San Francisco, CA, USA, pp. 727–34. Morgan Kaufmann Publishers Inc.
- Peugh, J., and X. Fan. 2012. How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining gmm's performance characteristics. *Structural Equation Modeling: A Multidisciplinary Journal* 19 (2):204–26. doi:[10.1080/10705511.2012.659618](https://doi.org/10.1080/10705511.2012.659618).
- Proust-Lima, C., V. Philipps, and B. Lique. 2017. Estimation of extended mixed models using latent classes and latent processes: the *r* package lcmm. *Journal of Statistical Software* 78 (2):1–56. doi:[10.18637/jss.v078.i02](https://doi.org/10.18637/jss.v078.i02).
- R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–63. doi:[10.2307/271063](https://doi.org/10.2307/271063).
- Raudenbush, S. W. 2005. How do we study “what happens next”? *The ANNALS of the American Academy of Political and Social Science* 602 (1):131–44. doi:[10.1177/0002716205280900](https://doi.org/10.1177/0002716205280900).
- Ruppert, D. 2002. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 11 (4):735–57. doi:[10.1198/106186002853](https://doi.org/10.1198/106186002853).
- Sher, K. J., K. M. Jackson, and D. Steinley. 2011. Alcohol use trajectories and the ubiquitous cat's cradle: cause for concern? *Journal of Abnormal Psychology* 120 (2):322–35. doi:[10.1037/a0021813](https://doi.org/10.1037/a0021813).
- Shiyko, M. P., Y. Li, and D. Rindskopf. 2012. Poisson growth mixture modeling of intensive longitudinal data: An application to smoking cessation behavior. *Structural Equation Modeling: A Multidisciplinary Journal* 19 (1):65–85. doi:[10.1080/10705511.2012.634722](https://doi.org/10.1080/10705511.2012.634722).
- Skardhamar, T. 2010. Distinguishing facts and artifacts in group-based modeling. *Criminology* 48 (1):295–320. doi:[10.1111/j.1745-9125.2010.00185.x](https://doi.org/10.1111/j.1745-9125.2010.00185.x).
- Song, X.-Y., and Z.-H. Lu. 2010. Semiparametric latent variable models with Bayesian P-splines. *Journal of Computational and Graphical Statistics* 19 (3):590–608. With supplementary material available online. doi:[10.1198/jcgs.2010.09094](https://doi.org/10.1198/jcgs.2010.09094).
- Tan, X., M. P. Shiyko, R. Li, Y. Li, and L. Dierker. 2012. A time-varying effect model for intensive longitudinal data. *Psychological Methods* 17 (1):61–77. doi:[10.1037/a0025814](https://doi.org/10.1037/a0025814).
- Tolvanen, A. 2007. *Latent growth mixture modeling: A simulation study*. Jyväskylä, Finland: University of Jyväskylä.
- Twisk, J., and T. Hoekstra. 2012. Classifying developmental trajectories over time should be done with great caution: A comparison between methods. *Journal of Clinical Epidemiology* 65 (10):1078–87. doi:[10.1016/j.jclinepi.2012.04.010](https://doi.org/10.1016/j.jclinepi.2012.04.010).
- Van de Schoot, R., M. Sijbrandij, S. D. Winter, S. Depaoli, and J. K. Vermunt. 2017. The grolts-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling: A Multidisciplinary Journal* 24 (3):451–67. doi:[10.1080/10705511.2016.1247646](https://doi.org/10.1080/10705511.2016.1247646).

- Van Dongen, S. 2000. Performance criteria for graph clustering and Markov cluster experiments. techreport INS-R0012, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands.
- Verbeke, G., and E. Lesaffre. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 91 (433):217–21. doi:[10.1080/01621459.1996.10476679](https://doi.org/10.1080/01621459.1996.10476679).
- Verbeke, G., and G. Molenberghs. 2000. *Linear mixed models for longitudinal analysis*. New York: Springer-Verlag.
- Walls, T. A., and J. L. Schafer. 2006. *Models for intensive longitudinal data*. New York, NY: Oxford University Press, Oxford.
- Weaver, T. E., G. Maislin, D. F. Dinges, T. Bloxham, C. F. P. George, H. Greenberg, G. Kader, M. Mahowald, J. Younger, and A. I. Pack. (2007, June). Relationship between hours of CPAP use and achieving normal levels of sleepiness and daily functioning. *Sleep* 30 (6):711–9. doi:[10.1093/sleep/30.6.711](https://doi.org/10.1093/sleep/30.6.711).
- Weybright, E. H., L. L. Caldwell, N. Ram, E. A. Smith, and L. Wegner. 2016. Trajectories of adolescent substance use development and the influence of healthy leisure: A growth mixture modeling approach. *Journal of Adolescence* 49:158–69. doi:[10.1016/j.adolescence.2016.03.012](https://doi.org/10.1016/j.adolescence.2016.03.012).
- Yang, J., M. Shao, and G. Cai. 2019. On the performance of MixTVEM: a simulation study. *Communications in Statistics - Simulation and Computation* 48 (9):2830–44. doi:[10.1080/03610918.2018.1468458](https://doi.org/10.1080/03610918.2018.1468458).