Advances in Life Course Research 43 (2020) 100323

Contents lists available at ScienceDirect

Advances in Life Course Research

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

An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software

Gavin van der Nest[a,*], Valéria Lima Passos[a], Math J.J.M. Candel[a], Gerard J.P. van Breukelen[a,b]

[a] Department of Methodology and Statistics, and Care and Public Health Research Institute (CAPHRI), Maastricht University, the Netherlands
[b] Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, the Netherlands

ARTICLE INFO

Keywords:
Growth mixture model
Classification
Trajectory
Hidden heterogeneity
Latent class growth analysis
Repeated measures

ABSTRACT

The use of finite mixture modelling (FMM) is becoming increasingly popular for the analysis of longitudinal repeated measures data. FMMs assist in identifying latent classes following similar paths of temporal development. This paper aims to address the confusion experienced by practitioners new to these methods by introducing the various available techniques, which includes an overview of their interrelatedness and applicability. Our focus will be on the commonly used model-based approaches which comprise latent class growth analysis (LCGA), group-based trajectory models (GBTM), and growth mixture modelling (GMM). We discuss criteria for model selection, highlight often encountered challenges and unresolved issues in model fitting, showcase model availability in software, and illustrate a model selection strategy using an applied example.

1. Introduction

This paper compares statistical model-based approaches for uncovering latent (unobserved) evolutions in longitudinal data of the repeated measures type, i.e. multiple time points of measurements per subject (Burton-Jeangros et al., 2015). These methods provide the means to evaluate individual variation in responses to interventions (e.g. in randomized controlled trials) as well as to test hypotheses of subgroups within the population (known as latent classes) following distinct developmental paths over time (trajectories) (Nagin & Odgers, 2010a) without *a priori* knowledge of grouping variables.

Such approaches have a direct application in life course research, in particular when addressing questions of whether groups of individuals exhibit different responses or development in a variety of behaviours, physical health, life satisfaction, and disorders (Nagin & Odgers, 2010b) over their life course. Some recent applications include uncovering distinct trajectories of treatment response for adults with obsessive-compulsive disorder (Falkenstein et al., 2019), disparate patterns of change over time in terms of criminogenic risks of juvenile offenders (Hilterman, Bongers, Nicholls, & van Nieuwenhuizen, 2018), divergent general psychopathology trajectories and their link to social outcomes (Lee, Wickrama, O'Neal, & Lorenz, 2017), examining group differences in the link between alcohol consumption evolution and cardiovascular events (Lima Passos, Klijn, van Zandvoort, Abidi, & Lemmens, 2017), and relating distinctive cannabis use patterns among adolescents to life satisfaction, academic achievement and other psychoactive substance usage (Grevenstein & Kröninger-Jungaberle, 2015).

As latent evolution models, broadly referred to as longitudinal latent growth models (LGM), they are flexible in estimating temporal changes in one (univariate) outcome as well as measuring the degree of temporal interrelationships between several outcomes (multivariate models). These properties make these techniques useful statistical tools in addressing the complexity underlying the abundance of information contained in longitudinal studies.

This paper will introduce the most popular longitudinal model-based approaches for latent evolution and will show how they are interrelated. Model fit and selection criteria for selecting the best model will be discussed. The paper will further cover software available for the estimation of these models by delineating their various capabilities. Finally, an empirical example will be provided to illustrate a detailed

Abbreviations: AIC, Akaike Information Criterion; aLMR, Adjusted Lo-Mendell-Rubin test; APPA, Average posterior probability of assignment; BF, Bayes Factor; BIC, Bayesian Information Criterion; BLRT, Bootstrap likelihood ratio test; FMM, Finite mixture model; GBTM, Group-based trajectory model; GCM, Growth curve model; GMM, Growth mixture model; LCGA, Latent Class Growth Analysis; LGM, Latent growth model; LLIC, log-likelihood information criteria; OLS, Ordinary least squares; LRT, Likelihood ratio test; VLMR, Vuong-Lo-Mendell-Rubin test

* Corresponding author at: Department of Methodology and Statistics, and Care and Public Health Research Institute (CAPHRI), Maastricht University, P.O. Box 616, 6200MD, Maastricht, the Netherlands.

E-mail addresses: g.vandernest@maastrichtuniversity.nl (G. van der Nest), valeria.limapassos@maastrichtuniversity.nl (V. Lima Passos), math.candel@maastrichtuniversity.nl (M.J.J.M. Candel), gerard.vbreukelen@maastrichtuniversity.nl (G.J.P. van Breukelen).

https://doi.org/10.1016/j.alcr.2019.100323
Received 19 September 2019; Received in revised form 28 November 2019; Accepted 20 December 2019
Available online 25 January 2020

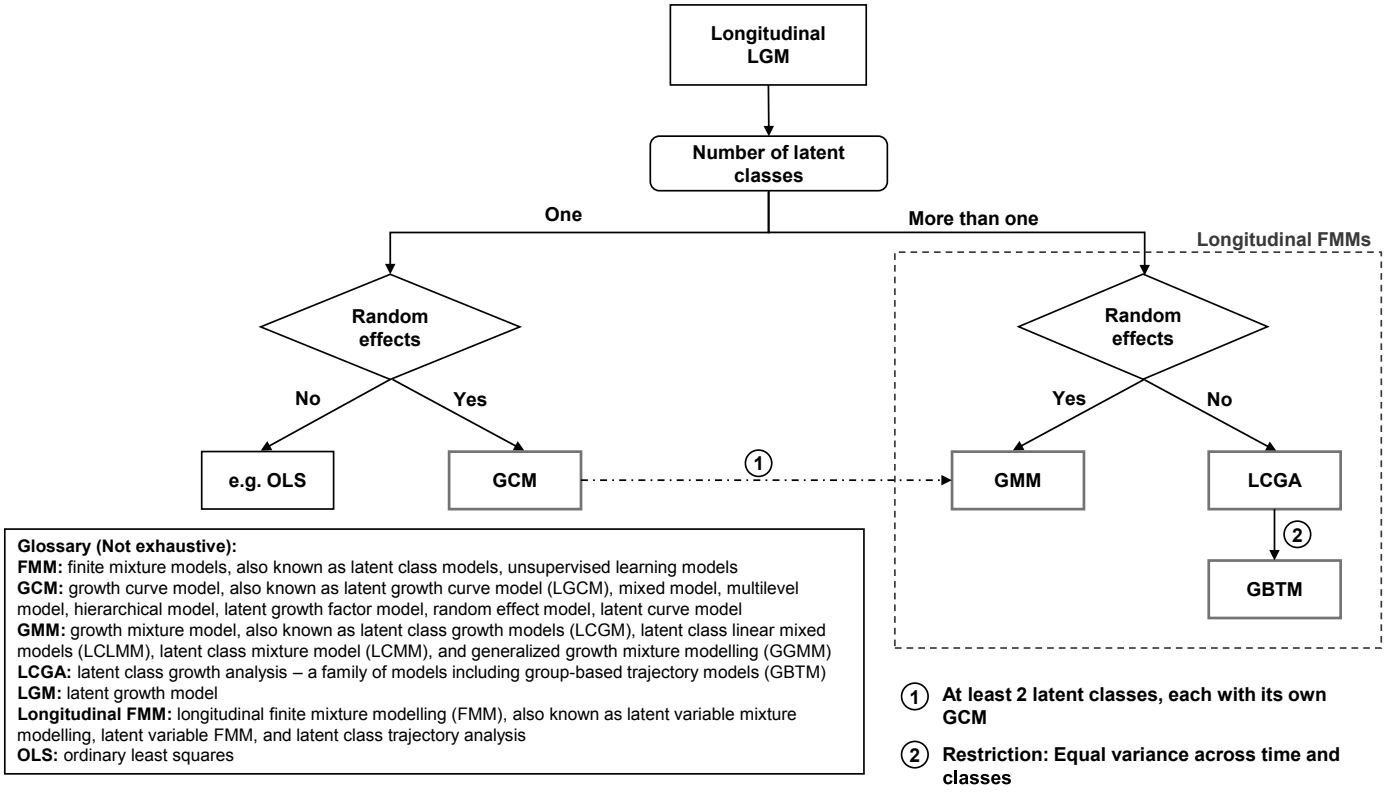

Fig. 1. Interrelatedness of longitudinal LGM models.

strategy for fitting these models.

2. Types of longitudinal growth models and their interrelatedness

There are several model-based techniques for analysing outcome development over time (Verbeke, Fieuws, Molenberghs, & Davidian, 2014), in particular for longitudinal repeated measures data. They fall under the shorthand term of longitudinal LGM. These approaches accommodate inter-individual variability (between-subjects) and intra-individual (within-subjects) patterns of change over time (Curran, Obeidat, & Losardo, 2010; Muthén, 2008), which are typically represented as time trends, time paths, growth curves or latent trajectories (Curran et al., 2010).

Within the family of longitudinal LGM models, the same model may be termed differently. This often creates confusion in literature and in practice, specifically concerning model commonalities and applications. For clarification, a partial glossary of these terms and their aliases are contained in Fig. 1. As Fig. 1 shows, longitudinal LGMs are divided into models comprising the estimation of one latent class (characterised as one population mean trajectory) such as growth curve models (GCM), or more than one latent class (represented as one mean trajectory per class) such as growth mixture models (Muthén & Shedden, 1999) (GMM), latent class growth analysis (Berlin, Parra, & Williams, 2014) (LCGA), and group-based trajectory models (Nagin, 2005) (GBTM). Their trajectories are modelled as functions of time and represent the mean development of an outcome over time within the latent class. Each latent class may be thought of as a group of subjects sharing similar development patterns which are not immediately evident from the data. Single-class ordinary least squares (OLS) models are excluded since they cannot handle repeated measures data.

The models discussed in this paper are assumed to include only time as a within-subject predictor and are considered in a univariate setting (one outcome). They may all be extended to include between-subject predictors (such as sex, age, treatment group), other within-subject predictors besides time (e.g. to model behavioural change as a function of major life events occurring during the follow-up time interval), and multiple outcomes.

2.1. Growth curve models

The single-class GCM models are not concerned with categorizing subjects, but rather with modelling the relationship between explanatory variables and the development of a repeatedly measured outcome (Laursen & Hoff, 2006). Therefore, they are well suited to studies concerning the relative contributions predictors make to explain the variability of an outcome. Conventional applications of a GCM assume that the sample under study is drawn from a single population described by a single set of parameters (e.g. means, variances and covariances) (Muthén & Muthén, 2000; Ram & Grimm, 2009).

The equation for the single trajectory of a GCM in scalar form is presented in Eq. (1) in Table 1. A matrix formulation is provided in Table 6 in the appendix. In the scalar formulation, y_{it} is the measured outcome for subject i at time $t = 1, ..., T$, and X_{it} denotes the value of a predictor X for subject i at time t. Consider for example that y_{it} is observed alcohol consumption, and X_{it} is the subject's age at which alcohol consumption is measured. For our example, alcohol consumption is measured at the same age for each time point across subjects. Then, for equidistant values of X, X_{it} may be coded by t itself i.e. $X_{it} = t$ (x-axis of Fig. 2(a)) (Lima Passos et al., 2017). For simplicity, we assume that the outcome trend across time, as represented by the effect of X on y, follows a second-order polynomial in time, but this can either be extended to higher orders or constrained to a linear trend. β_0, β_1 and β_2 are fixed effects, which quantify the population average growth curve i.e. alcohol consumption averaged over all individuals across time. This is represented by the single thick line in Fig. 2(a). b_{0i}, b_{1i} and b_{2i} are random effects, which allow for individual differences in alcohol consumption from the average time trend (inter-individual variability). ε_{it} represents the errors (intra-individual random variability). Total individual differences (the sum of random effects and error) are represented by the subject-specific lines' (thin lines) deviation from the

Table 1
Model trajectory specification (Davies et al., 2017; Demidenko, 2013).

General:

$k = 1, \ldots, K$ is the class
$t = 1, \ldots, T$ is the time point
$i = 1, \ldots, n$ is the subject
X_{it} = predictor value of subject i at time point t

Model	Trajectory Specification	Assumptions
GCM	$y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})X_{it} + (\beta_2 + b_{2i})X_{it}^2 + \varepsilon_{it}$ (1)	$b_{ji} \sim N(0, \sigma_{b_j}^2)$, $j = 0,1,2$ $\varepsilon_{it} \sim N(0, \sigma_{\varepsilon_t}^2)$ $cov(b_{ji}, b_{hi}) \neq 0$, $j \neq h$, $h = 0,1,2$
LCGA	$y_{it}^k = \beta_0^k + \beta_1^k X_{it} + \beta_2^k X_{it}^2 + \varepsilon_{it}^k$ (2)	$\varepsilon_{it}^k \sim N(0, \sigma_{\varepsilon_{kt}}^2)$
GMM	$y_{it}^k = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2 + \varepsilon_{it}^k$ (3)	$b_{ji}^k \sim N(0, \sigma_{b_j^k}^2)$, $j = 0,1,2$ $\varepsilon_{it}^k \sim N(0, \sigma_{\varepsilon_{kt}}^2)$ $cov(b_{ji}^k, b_{hi}^k) \neq 0$, $j \neq h$, $h = 0,1,2$

average trend in Fig. 2(a).

The random effects and errors are assumed to be normally distributed with zero mean and have their own covariance structure. Specifically, each of the three random effects has its own variance, and so, for instance, individuals may differ in intercept and linear change, but much less so in the quadratic deviation from linearity (i.e. $\sigma_{b_2}^2$ may be small compared to $\sigma_{b_0}^2$ and $\sigma_{b_1}^2$). Also, each pair of random effects can have its own covariance. The error variance can depend on time (e.g. increase over time), and successive errors can be correlated, for instance by a first-order autoregressive AR(1) structure in which each error is a function of the preceding error. We refer the reader to Verbeek (2012) for more details on the various types of autocorrelation.

A GCM may be extended to examine differences in outcome development between known subgroups, for instance between males and females. As an example, a GCM may differentiate linear trends between sexes by adding sex and a sex by time interaction term to the model,

$$y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 sex_i + \beta_3 sex_i X_{it} + b_{0i} + b_{1i} X_{it} + \varepsilon_{it} \quad (4)$$

There are then separate growth trajectories for each level of sex. For example, if $sex_i = 0$ for males and $sex_i = 1$ for females, then $\beta_0 + \beta_1 X_{it}$ is the average growth curve for males, and is the average growth curve for females. Individual deviations from the sex-specific average trend are again captured through the random effects (b_{0i} and b_{1i}). Applications of these extended GCMs require *a priori* knowledge of the number of subgroups and of the subgroup membership of each study participant (Ram & Grimm, 2009). Since GCM is not designed to uncover latent classes it will not be considered further in the remainder of this study.

2.2. Longitudinal FMMs

The identification of multiple latent classes of outcome development is possible using multi-class longitudinal models which are collectively known as longitudinal finite mixture models (FMMs) (McLachlan &

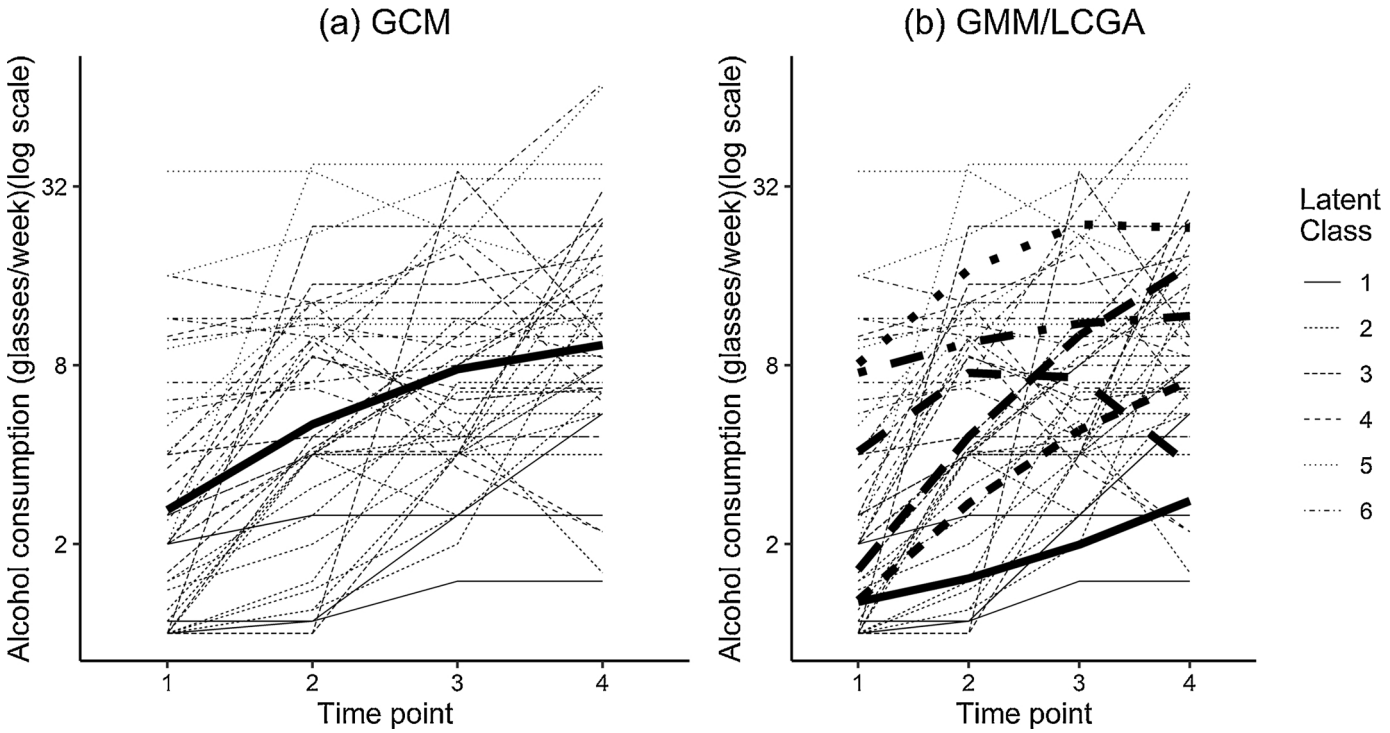

Fig. 2. An illustration of GCM and GMM/LCGA approaches. Thin lines correspond to subject trajectories, thick lines correspond to the average trend (GCM) or the average trend in each class (GMM/LCGA).

Peel, 2000). These models assume that the population under study is composed of distinct, latent subgroups or classes (Berlin, Williams, & Parra, 2014). These classes represent a heterogeneous population in the sense that predictors (e.g. time) may act differently on the outcome per class, where classes need not be defined *a priori* in terms of some observed variable such as sex. However, FMM assumes that the number of classes is known but this is often difficult to deduce from the data and various methods exist to estimate the appropriate number of classes (See Section 3).

Longitudinal FMMs have the distinct feature of being able to capture the concealed variation in development patterns between groups (hidden heterogeneity) without the explicit need of additional predictors besides time. This is done through the inclusion of K latent classes (represented as latent categorical variables), each with its own mathematical model for the trajectory. The assignment of individuals to classes is then based on the degree of similarity of developmental courses between individuals (Diallo, Morin, & Lu, 2017). For this reason, FMMs have been frequently used in exploratory contexts, in which researchers are unaware of the underlying drivers of distinct developmental trajectories or in cases where a defining characteristic separating groups could not be measured (e.g. undiscovered genotype, or drug use). In contrast to GCM and the methods introduced in Mund and Nestler (2019), FMMs provide for the *post hoc* identification and description of class differences in change (Ram & Grimm, 2009). Furthermore, FMMs extend these methods by combining the use of latent classes with random effects to account for both individual and class differences in development across a heterogeneous population (Curran et al., 2010; Muthén, 2008).

Typical longitudinal FMM models include; growth mixture modelling (Muthén & Shedden, 1999) (GMM), latent class growth analysis (Nagin, 2005) (LCGA), mixture latent transition analysis (LTA) (Collins & Lanza, 2010) (also known as mixture hidden Markov models), and survival mixture analysis (SMA) (Muthén & Masyn, 2005) amongst others (Berlin, Parra et al., 2014; Muthén, 2008). They all differ according to their underlying assumptions.

Only GMM and LCGA will be discussed in more depth in the next section due to space limitations and since these appear to be more popular longitudinal mixture approaches according to a recent review (Killian, Cimino, Weller, & Hyun Seo, 2019). Mixture LTA is also excluded as it introduces an additional layer of complexity in the form of discrete time-invariant latent states. The primary objective of mixture LTA is to study the probability of transitioning from one state to another at different time points and to uncover heterogeneous latent classes characterised by different transition probabilities for these latent states (Magidson, Vermunt, & Tran, 2009; Muthén & Muthén, 2000; Piccarreta & Studer, 2019). An example might include studying the probability of transitioning from a healthy to an unhealthy state (of some health outcome e.g. stroke) for different latent classes distinguished by individuals showing different alcohol consumption patterns over time. Mixture LTA may be estimated in software including Mplus and LatentGOLD. SMA is excluded since it models the waiting time until an event (e.g. death) occurs, whereas this review focuses on models for repeated outcome measures at fixed time points.

2.2.1. Latent class growth analysis (LCGA)

Eq. (2) shows the class-specific equation for the trajectory in an LCGA. The superscript k shows that the various parameters are class-specific. They are the same for all subjects within a class, who are assumed to follow the estimated mean trajectory per class, but are different between classes. For instance, one class may have a linear (or increasing) mean growth curve, whereas another class has a quadratic (or decreasing) growth curve.

The LCGA has no random effects to capture individual differences in a continuous way. Instead, it allows for discrete individual differences by letting fixed effects (given by the trend) differ between classes (Pennoni & Romeo, 2017). This is represented by the bold lines in Fig. 2(b). Individual deviations from the class-specific trend are treated as residual error and corresponds to the distance from the class-specific bold lines to the individual thin lines of subjects assigned to that class. Furthermore, the error variance may vary between time points as well as between classes. The group-based trajectory model (GBTM) is a popular special case of the LCGA in which the error variance is assumed to be the same for all classes and all time points (Nagin, 2005, p. 337; Nagin & Land, 1993).

As LCGA exhibits no between-subject variability within a class, far fewer parameters need to be estimated. Therefore, it may be useful in cases of smaller sample sizes or in the presence of more complex models that fail to converge, produce out of range estimates, or it may be used as an initial modelling step before specifying a GMM (Jung & Wickrama, 2008).

2.2.2. Growth mixture models (GMM)

The class-specific trajectory for a GMM is represented in Eq. (3), which is an amalgamation of Eqs. (1) and (2). This allows for multiple latent classes with each class having its own GCM. The average class-specific time trend is again given by the fixed effects as is represented by the bold lines in Fig. 2(b). Random effects are used to capture individual differences in trajectories within a class (Muthén & Muthén, 2000), since the outcome at the start (the intercept) and the rate of change (the slope) may vary between individuals within a class. The latter distinguishes it from the LCGA. The distance between the class-specific average trend in Fig. 2(b) and the thin individual lines for individuals belonging to that class, is now modelled as the sum of the random effect and random error instead of just random error as in the LCGA. Furthermore, the random effects and errors follow the same assumptions as in GCM, but now per latent class.

2.2.3. More general formulation of LCGA and GMM

As longitudinal finite mixture models (FMM), GMM and LCGA comprise a combination of two or more probability functions. A longitudinal FMM for latent evolutions states that for K latent classes, the marginal probability distribution of a randomly chosen trajectory is modelled as (Nagin, 2005),

$$P(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k P^k(\boldsymbol{y}_i), \tag{5}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$ is the vector of measured outcomes (e.g. alcohol consumption) for subject i, $i = 1, \ldots, n$, at time $t = 1, \ldots, T$, and $P^k(\boldsymbol{y}_i)$ is the conditional distribution of the longitudinal sequence, $\boldsymbol{y}_i$, given that individual i is in latent class k. In our paper, this is uniquely defined by the trajectory specification for each class. π_k is the class membership probability (also referred to as mixing weight, class size or mixing proportion in the literature) such that $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, and $K > 1$. Eq. (5) shows that $P(\boldsymbol{y}_i)$ is the sum over a finite number of discrete classes, each with its own trajectory and class size.

The combination of the properties of the K individual conditional distribution functions (i.e. the $P^k(\boldsymbol{y}_i)$ on the right side of Eq. (5)) with the class membership probabilities (the π_k on the right side of (5)) allows the mixture model to approximate any arbitrary marginal distribution (the $P(\boldsymbol{y}_i)$ on the left side of (5)). It is this property which makes FMMs a powerful and flexible tool for the modelling of complex data (McLachlan & Peel, 2000), such as highly asymmetrical and multimodal data.

In the longitudinal context, with a repeatedly measured continuous outcome, $P^k(\boldsymbol{y}_i)$ could be the multivariate normal (MVN) density function (in line with the model assumptions in Eqs. (1) and (3) in Table 1). So, for subject i in class k,

$$\boldsymbol{y}_i^k \sim MVN(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k) \tag{6}$$

with $\boldsymbol{y}_i^k$ the vector of T succesive repeated measures (alcohol consumption at each time point) of subject i in latent class k, and $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ the mean vector (the average alcohol consumption at each time point)

and covariance matrix (the variances and covariance of alcohol consumption across time points) for class k respectively (Davies, Giles, & Glonek, 2018). For this model, the GCM of Eq. (1) is assumed to hold per class.

In GMM models, the $\mathbf{\Sigma}^k$ consists of two sources of variation: inter-individual variation (given by random intercept and slope, the covariance matrix $\mathbf{D}^k$) and intra-individual variation (given by the errors, which may be independent or autocorrelated, the $\mathbf{R}^k$ matrix). In a GMM, the $\mathbf{D}^k$ matrix may be set equal or allowed to vary freely between groups (Davies, Glonek, & Giles, 2017). In GBTM and LCGA, the $\mathbf{D}^k$ matrix is zero (so that $\mathbf{\Sigma}^k = \mathbf{R}^k$) and the $\mathbf{R}^k$ matrix is diagonal. This assumption about $\mathbf{D}^k$ implies the absence of random effects. The assumption about $\mathbf{R}^k$ implies absence of autocorrelation. Imposing the further restriction on the diagonal $\mathbf{R}^k$ that the residual variance is the same for all time points and all classes reduces the LCGA to a GBTM (Nagin, 2005, p. 337; Nagin & Land, 1993). Possible specifications for the $\mathbf{\Sigma}^k$ and $\mathbf{R}^k$ matrices are presented in Table 7 in the appendix.

2.2.4. Extension beyond continuous outcomes and a polynomial trend

The MVN assumption on $\boldsymbol{y}_i$ in Eq (6) may be relaxed to accommodate outcomes that are not continuous. $P^k(\boldsymbol{y}_i)$ may then take on various distributional forms, such as Poisson (for count data) and Binary Logit (for binary data) as has been applied in GMM (Reinecke & Seddig, 2011) and LCGA (Nagin, 2005) studies.

For count data, the conditional distribution of the realization y_{it} (where $y_{it} = 0,1, 2, ...$) in class k follows the Poisson distribution,

$$p^k(y_{it}) = \frac{\lambda_{kt}^{y_{it}} e^{-\lambda_{kt}}}{y_{it}!}. \tag{7}$$

In a GMM, the trajectory for a quadratic time effect is then defined by $\ln(\lambda_{kt}) = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2$, where λ_{kt} is the mean rate of occurrence of the event for all individuals in class k at time t.

For binary data, $p^k(y_{it} = 1)$ may be described by the logit model:

$$p^k(y_{it} = 1) = \frac{\exp((\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2)}{1 + \exp((\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2)}. \tag{8}$$

An alternative for binary data is the probit model, which is almost empirically indistinguishable from the logit model (Chen & Tsurumi, 2010). However, the logit is often chosen due to having a closed-form equation (Nagin, 2005).

An alternative approach to modelling the outcome trend as a polynomial function of time is piecewise regression (see Chamroukhi (2016)) which we briefly address here for a continuous outcome. An example is a linear piecewise regression model. In the case of two nodes at $X_{it} = c_1$ and $X_{it} = c_2$, with $c_1 < c_2$, the trajectory is modelled as:

$$y_{it}^k = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)(X_{it} - c_1)D_1 + (\beta_3^k + b_{3i}^k)(X_{it} - c_2)D_2 + \varepsilon_{it}^k. \tag{9}$$

where D_1 is a dummy that is 0 for $X_{it} < c_1$ and 1 for $X_{it} > c_1$, and D_2 is a dummy that is 0 if $X_{it} < c_2$ and 1 if $X_{it} > c_2$. Such an approach is useful to test critical points along the trajectory (in Eq. (9) where $X_{it} = c_1$ and c_2) in which the relationship between the predictor (time) and the measured outcome (alcohol consumption) abruptly changes.

3. Criteria for model selection

In this section, we will focus on the criteria for longitudinal FMMs' model selection. Although no automated process exists for the often lengthy and iterative model-fitting procedure, a two-step procedure has been recommended (Nagin, 2005). The first step entails selecting the number of latent classes, K, for a fixed trajectory specification. This step is often referred to as class enumeration. The second step involves refining the polynomial order of the time effect (or other smoothing functions e.g. B-splines) that best describes the shape of the latent trajectories for a fixed K as determined in step one.

An innate problem of class enumeration for mixture models is that models comprising different numbers of classes are, in general, not nested. Consequently, standard likelihood ratio tests (LRTs) to test models against each other cannot be conducted. Nonetheless, a plethora of fit indices, including modified LRTs, exist to assist in the choice of K, which are discussed shortly. However, all current indices suffer from inherent weaknesses since their accuracy in determining the true number of latent classes largely depends on the underlying data features (e.g. such as level of class separation i.e. how distinct classes are from each other, sample size, class size). For this reason, the question of which one is the most valid remains largely unresolved (McLachlan & Peel, 2000, p. 175; Nylund, Asparouhov, & Muthén, 2007). Finally, these model fit statistics may be used in conjunction with Wald tests and LRTs to determine the final polynomial order.

3.1. Statistical fit indices for determining K

During class enumeration, it is recommended to determine the best-fitting K for which all the classes are still distinct in terms of their trajectories as given by $P^k(\boldsymbol{y}_i)$ in (Eq. (5)), and all their associated class probabilities (mixing weights), π_k, are non-zero (McLachlan & Peel, 2000, p. 177).

Finding the best K is aided by using statistical fit indices. These indices generally fall into three broad categories: (a) log-likelihood based statistics, (b) statistics based on the classification of individuals, and (c) statistics based on distributional properties of the data. Table 2 (Blaze, 2013; Henson, Reise, & Kim, 2007; Mardia, 1970) presents an overview of the most frequently cited of these fit indices which will be discussed in detail.

3.1.1. Log-likelihood criteria

The log-likelihood information criteria (LLIC) statistics have the general form (Sclove, 1987),

$$-2\log[L(K)] + a(n)m(K) \tag{10}$$

where $L(K)$ is the maximum likelihood of the data for a model with K classes, n is the sample size, $a(n)$ is a function of the sample size, and $m(K)$ is the number of independent parameters in the model with K classes. Smaller values of Eq. (10) correspond with better models, and $a(n)m(K)$ is a penalty for lack of model parsimony. A better fitting model is one for which the increase in model fit, as expressed by the decrease in $-2\log[L(K)]$, outweighs the penalty of increased model complexity, as expressed by the number of unknown parameters.

All LLIC statistics have a common form but differ in the calculation of the penalty statistic. The Bayesian Information Criterion (BIC) favours more parsimonious models relative to the Akaike Information Criterion (AIC). The AIC is not asymptotically optimal since the probability of choosing the correct number of classes does not approach 1 as n approaches infinity (Blaze, 2013). To address this drawback, the Consistent Akaike Information Criterion (CAIC) was proposed, which favours parsimonious models slightly more than the BIC given the addition of 1 to the penalty term. The sample-size adjusted BIC's (ssBIC) penalty term is not as harsh as the BIC's and may be beneficial in the case of small sample sizes or many parameters. It is useful to note that the ordering of the severity of the penalty term of the LLIC for $n < 176$ is ssBIC<AIC < BIC<CAIC, and for $n \geq 176$ is AIC<ssBIC < BIC<CAIC. Simulation results show that the AIC has a tendency to overestimate the true number of components in a mixture relative to the other three information criteria (BIC, ssBIC, CAIC) with the BIC and CAIC tending to underestimate the number of components (Henson et al., 2007).

The Bayes Factor (Kass & Raftery, 1995) (BF) is a criterion which may be used to compare the magnitude of change in the BIC between any two models. It is the ratio of the likelihood of the data under the two models (McLachlan & Peel, 2000, p. 210; Wagenmakers, 2007),

Table 2
Typical statistical criteria used for class enumeration.

Type	Measure	Equation	Model selection (*)
Log-Likelihood Statistics	AIC	$-2\log[L(K)] + 2[m(K)]$	Smallest value
	BIC	$-2\log[L(K)] + \log(n)[m(K)]$	Smallest value
	CAIC	$-2\log[L(K)] + (\log(n) + 1)[m(K)]$	Smallest value
	ssBIC	$-2\log[L(K)] + \log\left(\frac{n+2}{24}\right)[m(K)]$	Smallest value
	VLMR	$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\hat{P}(\mathbf{y}_i \mid m(K_1))}{\hat{P}(\mathbf{y}_i \mid m(K_1-1))}\right)^2$	H_0: $K = K_1 - 1$ H_1: $K = K_1$ If likelihood ratio $p \leq 0.05$, then choose K_1, else choose $K_1 - 1$
	aLMR	$\frac{VLMR}{1 + ([m(K_1-1) - m(K_1)]\log(n))^{-1}}$	H_0: $K = K_1 - 1$ H_1: $K = K_1$ If likelihood ratio $p \leq 0.05$, then choose K_1, else choose $K_1 - 1$
	BLRT	Bootstrapped: $LR = -2(\log[L(K_1 - 1)] - \log[L(K_1)])$	H_0: $K = K_1 - 1$ H_1: $K = K_1$ If bootstrapped $p \leq 0.05$, then choose K_1, else choose $K_1 - 1$
Classification statistics	sE	$Scaled\ Entropy(K) = 1 - \frac{E(K)}{n\log(K)}$	Largest
	NEC	$NEC(K) = \frac{E(K)}{LL(K) - LL(1)}$	Smallest
	APPA	Defined per class: $APPA_k = \frac{1}{n_k}\sum_{i=1}^{n_k} pp_{ik}$ where n_k= number of individuals assigned to class k, and sum only the respective pp_{ik} of subjects assigned to class k. Individual $i$is assigned to class k if pp_{ij} is larger than that person's pp_{ij} for any class j other than k.	Values closer to 1 indicate a good fit. Usual acceptable threshold > 0.7 for all classes
	OCC	Defined per class: $OCC_k = \frac{APPA_k/(1 - APPA_{kk})}{\hat{\pi}k/(1 - \hat{\pi}_k)}$	Higher values (preferably > 5) for all classes
	CLC	$CLC = -2\log L(K) + 2E(K)$	Smallest
	ICL-BIC	ICL-BIC $= -2\log L(K) + \log(n)m(K) + 2E(K)$	Smallest
Distributional statistics	MVS MVK		H_0: Kclass model H_1: Not$K$class model

*Notes**: Not all software defines fit statistics in the same way, which may lead to a different value for model selection e.g. in Proc traj select largest BIC.
K: number of classes.
$L(K)$, $L(K_1 - 1)$, $L(K_1)$: Maximum likelihood of K-class, null and alternative model respectively.
$m(K)$, $m(K_1 - 1)$, $m(K_1)$: Number of parameters of K-class, null and alternative model respectively.
$LL(K)$, $LL(1)$: log-likelihood of K and one-class model.
pp_{ik}: posterior probability of subject i for class k.
$\hat{\pi}_k$: estimated proportion of population in class k.
$E(K)$: entropy of K class model.
$\log(x)$: the natural logarithm of x.
n: sample size.

$$BF_{10} = \frac{P(\mathbf{y}|K_1)}{P(\mathbf{y}|K_0)} \tag{11}$$

where K_0 and K_1 are the null and alternative model, respectively. A value greater than one would suggest that the data is more likely given the alternative model. It has been shown that the BF is asymptotically equal to $BF_{10} = \exp(\Delta BIC_{01}/2)$ (Faulkenberry, 2018; Kass & Wasserman, 1995; Raftery, 1995; Wagenmakers, 2007, p. 796, 804), where $\Delta BIC_{01} = BIC(K_0) - BIC(K_1)$. A value of BF_{10} greater than 10 is cited as a reasonable standard for strong evidence in favour of the alternative model (Wasserman, 2000).

The Vuong-Lo-Mendell-Rubin test (VLMR) is a modified LRT (Lo, Mendell, & Rubin, 2001). It seeks to address distributional assumption violations of conventional LRTs in cases where the difference statistic is not chi-square distributed when comparing non-nested K_1 class to $K_1 - 1$ class mixture models (McNeish & Harring, 2017). The VLMR test seeks to circumvent these violations by analytically deriving the appropriate distribution of the difference between the likelihoods of these non-nested models. The asymptotic distribution of the VLMR test statistic is that of a weighted sum of $m(K_1 - 1) + m(K_1)$ independent chi-square random variables. However, in simulation studies, the VLMR showed inflated Type I error rates, particularly in small samples, and the adjusted VLMR (aLMR, known as the Lo-Mendell-Rubin adjusted LRT test) was proposed to address this by correcting for sample size and the number of estimated parameters (Lo et al., 2001). Moreover, the VLMR and aLMR have not escaped scrutiny as their original proof has been shown to contain mathematical errors (Jeffries, 2003). Nevertheless, they appear to work well in detecting homoscedastic (equal variance across classes) normal mixtures (Lo et al., 2001).

The bootstrap likelihood ratio test (BLRT) is a parametric bootstrap alternative approach to estimate the distribution of the LRT statistic (Tekle, Gudicha, & Vermunt, 2016). The BLRT addresses distributional issues with the LRT (McLachlan & Peel, 2000, p. 186) which has no closed-form distribution under mixture models (Jeffries, 2003) and seeks to address the shortcomings of the VLMR and aLMR tests. A bootstrap p-value is obtained and is used to test the null hypothesis of a $K_1 - 1$ class model against the alternative hypothesis of a K_1 class model. Violation of the multivariate normality assumption under the BLRT was shown to lead to class over-extraction (Nylund et al., 2007). However, studies show that with complex growth trajectory shapes and large sample size conditions, the BLRT tends to outperform other likelihood-based enumeration indexes (Nylund et al., 2007; Peugh & Fan, 2012), but this needs to be balanced against its computational intensity. It is recommended to first select a plausible subset of models using the

BIC and VLMR before refining the selection using the BLRT (Nylund et al., 2007).

Of these fit statistics considered, the BIC tends to be the most frequently used in practice and is widely available in commercial software packages.

3.1.2. Classification-based criteria

Classification statistics based on the classification maximum likelihood (CML) are complementary to the log-likelihood statistics (Celeux & Soromenho, 1996). They use entropy, $E(K)$, which is a measure of classification uncertainty in class assignment, as a penalty term in ascertaining model fit. In formula,

$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} pp_{ik}\log[pp_{ik}] \geq 0 \tag{12}$$

where higher values for $E(K)$ signify greater classification uncertainty, and pp_{ik} is the posterior probability of subject i belonging to class k given the data. This, in turn, is obtained by applying Bayes' law,

$$pp_{ik} = P(k|\boldsymbol{y}_i) = \frac{\hat{\pi}_k \hat{P}^k(\boldsymbol{y}_i)}{\sum_{h=1}^{K} \hat{\pi}_h \hat{P}^h(\boldsymbol{y}_i)} \tag{13}$$

where $\hat{P}^k(\boldsymbol{y}_i)$ is the estimated probability of observing the data if i is a member of class k. $\hat{\pi}_k$ is the estimated proportion of the population in class k (Nagin, 2005) (where class membership follows a multinomial distribution) with the constraint that $\sum_{k=1}^{K} \hat{\pi}_k = 1$.

In mixture modelling, individuals are customarily assigned to classes with the highest posterior probability (pp_{ik}). Posterior probabilities are also used to assess model fit. If the posterior probability for every individual approaches 1 for one class and 0 for all other classes (signifying high classification confidence) then $E(K)$ approaches 0. In the case of estimated classes being distinct and well-defined, then each individual will have a single large posterior probability. Solutions showing unambiguous classification from posterior probabilities (e.g. $pp_{ik} > 0.80$) are posited to represent better models (Celeux & Soromenho, 1996).

It must be noted that $E(K)$ cannot be used directly to evaluate the number of classes in an FMM, since $E(K) \geq E(1) = 0$ for any $K > 1$ and $E(K)$ is an increasing function of K (Celeux & Soromenho, 1996). This renders $E(K)$ uncomparable across different K, since by definition $E(K+1) > E(K)$. To address the shortcomings of $E(K)$ for a $K > 1$ component model, the scaled entropy (sE) and normalized entropy criterion (NEC) were introduced. The sE rescales $E(K)$ to be bounded by 0 and 1, where higher values of sE designate better classification (Ramaswamy, Desarbo, Reibstein, & Robinson, 1993). This allows for direct comparisons between models. The NEC is the ratio of the classification uncertainty in class assignment (expressed by the numerator of the NEC in Table 2) relative to the change in the log-likelihood between models (expressed by the denominator of the NEC). It has the advantage of being comparable across non-nested models but has the drawback that when $K = 1$, $NEC(1)$ is not defined. Therefore, it cannot be used to compare one class with more than one class solutions, in which case other fit statistics should be considered. Smaller $NEC(K)$ values are indicative of a more precise classification of individuals (since $E(K)$ approaches 0).

A study (Diallo et al., 2017a) recommends using threshold values for entropy as a first step in informing the choice of fit statistic for model selection. Under conditions of high entropy (low scaled entropy (< 0.5), high NEC) the ssBIC and BLRT were found to outperform the BIC and CAIC. Under low entropy (high scaled entropy (> 0.8), low NEC) conditions, the CAIC, and BIC performed better than the ssBIC and BLRT. In another study (Tofighi & Enders, 2008), the VLMR showed good performance under conditions of low entropy.

The average posterior probability of assignment (APPA) (Asparouhov & Muthén, 2014) and the odds of correct classification (OCC) (Nagin, 2005) are additional classification statistics. The APPA is calculated as the average posterior probability of belonging to class k over all the individuals assigned to class k. It may be thought of as the average latent class probabilities for the most likely latent class membership. The OCC is the ratio of the odds of classifying subjects into class k based on the maximum probability classification rule (as used in the APPA) to the odds based on random assignment (where $\hat{\pi}_k$ represents the probability of a randomly selected individual belonging to class k) (Klijn, Weijenberg, Lemmens, van den Brandt, & Lima Passos, 2017). These statistics are class-specific and ideally, all classes should exceed a minimum threshold value. APPA close to 1 (ideally > 0.7) and higher values of the OCC are indicative of a good fit (Blaze, 2013; Nagin, 2005). OCC close to 1 is indicative of the maximum probability assignment rule having predictive power not beyond random chance (Nagin, 2005).

3.1.3. Likelihood-classification hybrids

The classification likelihood criterion (CLC) incorporates $E(K)$ as a classification uncertainty penalty term in LLIC (Hathaway, 1986) (see Section 3.1.1). The objective is to choose a K which minimizes the CLC (McLachlan & Peel, 2000). The CLC works well when class probabilities are restricted to being equal but has a tendency to overestimate the number of classes when no such restrictions exist (McLachlan & Peel, 2000, p. 214).

The integrated classification likelihood (BIC approximation) (ICL-BIC) was developed to address shortcomings in the BIC and CLC (Biernacki, Celeux, & Govaert, 2000). It is more robust when the underlying mixture model assumptions are violated (leading to boundary of parameter space issues) and addresses issues where the BIC tends to over extract classes when the fit of the data to the mixture model is poor. The ICL-BIC is equivalent to the BIC when $E(K) = 0$ (the case of perfect classification).

3.1.4. Distributional statistics criteria

These tests seek to identify the most appropriate K-class model by comparing the multivariate skew (MVS) and kurtosis (MVK) values derived from the proposed mixture model to the actual sample quantities. The skew and kurtosis (SK) tests compute the multivariate skew and kurtosis values across a large number of simulated (bootstrapped) samples from the mixture model being tested (Muthén, 2003). These simulations provide an empirical sampling distribution against which the actual sample values are compared. The SK test yields two p-values (for the multivariate skew and kurtosis) with a significant p-value indicating that the actual skew and kurtosis are not likely to be sampled from the K class model being tested. It is claimed that this test has sufficient power in small samples ($n \geq 200$) and works well in distinguishing a single class nonnormal population from a mixture of multiple normal populations (Henson et al., 2007; Muthén, 2003). However, more research is required to determine the viability of this approach.

3.1.5. Cross-validation

Cross-validation has also recently been considered to assist in class enumeration, but literature on its use in longitudinal FMM is limited and equivocal. Cross-validation involves splitting data into an independent training set (for model estimation) and test set (to test the model's predictive power) (Hélie, 2006). If the model predicts well, then it is seen as a good and appropriate model (Nielsen et al., 2014).

Cross-validation error (CVE) (Nielsen et al., 2014) is a measure of the predictive accuracy of a fitted model. The CVE for individual i is measured as,

$$CVE(i) = \frac{1}{T}\sum_{t=1}^{T} (y_{it} - \hat{y}_{it}^{[-i]})^2 \tag{14}$$

where T is the number of time points. $CVE(i)$ is the average squared difference between the observed values (y_{it}) and the predicted values

($\hat{y}_{it}^{[-i]}$), with the latter obtained by fitting the model on all data except those of individual i. This is known as leave-one-out cross-validation. Averaging over all n individuals, the overall CVE is given as,

$$CVE = \frac{1}{n}\sum_{i=1}^{n} CVE(i) \tag{15}$$

The best number of classes K is selected as that number which minimises this CVE. When applied to observational data, the CVE reached a minimum, whereas the BIC and AIC improved monotonically, seemingly without a practical limit, with an increase in K (Nielsen et al., 2014).

M-fold cross-validation is an alternative method which involves randomly dividing the sample into M partitions of equal size (n/M), using one of the M partitions as a test set and the remaining M-1 partitions as the training set, and repeating that M times, using another test set each time. The division of the data into partitions may be represented as $P_1 \cup ...\cup P_M = \{1, ...,n\}$. Then for each $m = 1, ...,M$, a prediction function is fit on the training set, which is then used to predict outcomes in the m-th test set ($\hat{y}_{it}^{[-m]}$). Then the error on the points in the m-th partition is evaluated as,

$$CVE(m) = \frac{1}{n_m}\frac{1}{T}\sum_{i\in P_m}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it}^{[-m]})^2 \tag{16}$$

where n_m and T are the number of subjects and time points in the m-th partition respectively. Finally, the obtained $CVE(m)$ values are averaged over all m,

$$CVE = \frac{1}{M}\sum_{m=1}^{M} CVE(m) \tag{17}$$

Note that the leave-one-out cross-validation is a special case of this where $M = n$. Again, the best number of classes is that value which minimizes the CVE. A recent study (He & Fan, 2018) suggests that M-fold cross-validation for class enumeration in GMMs only works well under high class separation. Again, when applied to observational data, the M-fold cross-validation enumerated a limited number of classes whereas the AIC and BIC continued to improve monotonically with an increase in K (Grimm, Mazza, & Davoudzadeh, 2017).

3.2. Determining the order of the polynomials and other model considerations

Once the number of classes has been established in the first step (where all classes would have some pre-set polynomial order informed by expert opinion, number of time points, previous studies or visual inspection), then the best order of the polynomial describing each class may be determined (Nagin, 2005, p. 66). The choice of the order of the trajectory for each class is considered as less important than the choice of the number of classes (Nagin, 2005, p. 67). Note that in the first step it is safe to choose a polynomial order that is too large, but in that case, model convergence may become a problem.

Visually inspecting the shape and size of the various trajectories could assist in pruning polynomial terms. Additionally, Wald tests for individual parameter significance (e.g. H_0: $\beta_1^k = 0$ vs. H_1: $\beta_1^k \neq 0$) within classes may be used and are usually reported in software. The highest polynomial order nonsignificant terms should be dropped in one class per iteration. The BIC is then also typically inspected to see if this leads to an improved model fit. If the BIC improves by more than 4.6 (leading to a Bayes Factor greater than 10), then there is strong evidence in favour of the simpler model (Jeffreys, 2004; Kass & Raftery, 1995; Wagenmakers, 2007).

Similarly, the choice of the covariance structure for the model is informed by practical experience, statistical inspection of data and model output, and running a series of models with various specifications. If the model fails to converge to a solution and/or produces severely out of bounds parameter estimates or a degenerate solution with empty classes, then users should simplify the model. This is done by fixing various model parameters such as assuming residual variance to be the same across time points and/or classes (Diallo et al., 2017a). In situations with few time points and where the covariance structure is to be determined, the BIC is suggested to ascertain whether various model constraints or relaxing of constraints leads to a better model fit (Davies et al., 2017).

Several studies (Davies et al., 2017; Diallo, Morin, & Lu, 2016; McNeish & Harring, 2017) highlight the detrimental impact of model misspecification, particularly covariance misspecification, on class enumeration and model fit. They caution that models should be flexible to account for different covariance structures across time points and between classes as this could have a significant impact on estimation, classification, and class enumeration. Furthermore, it was found that although misclassification resulting from inappropriate same variance across classes assumptions was much greater than from inappropriate same variance across time assumptions, neither should be ignored (Davies et al., 2017).

3.3. Past simulation studies: results and recommendation

Determining the best model fit index for the correct number of classes under a variety of different scenarios remains an outstanding issue in mixture modelling. To date, there is no one commonly accepted statistical indicator for class enumeration in mixture models (Nylund et al., 2007). In general, practitioners often employ the least computationally intensive statistics (or the most familiar) in determining an appropriate solution (Henson et al., 2007).

The relative performance of a selection of these fit indices has been compared in a variety of simulation studies (Brame, Nagin, & Wasserman, 2006; Diallo et al., 2017a; Kim & Wang, 2017; Lo et al., 2001; McNeish & Harring, 2017; Muthén, 2003; Nylund et al., 2007; Tofighi & Enders, 2008; Xu, Peng, & Huang, 2018). These studies investigated the performance of fit indices to find the true number of classes as assessed across a variety of different scenarios. These scenarios included variations of class probabilities, within-class distribution of outcomes, class separation (variously defined in terms of distinctness between parameters defining classes, entropy, Mahalanobis distance), sample size, and covariance structure. These studies show that fit statistic class enumeration performance is highly dependent on data-specific characteristics, with low class separation, small sample sizes, and covariance misspecification having particularly detrimental effects. Furthermore, no one fit statistic consistently emerges as superior in class enumeration across all studied data conditions.

Given the inconclusive findings of the simulation studies and the fact that there is no one commonly accepted fit statistic for class enumeration in mixture models, such decisions need to be made based upon a variety of evidence (Tofighi & Enders, 2008). It is recommended to use multiple fit indices to add some statistical objectivity to the class enumeration and model selection process, as well as a substantive interpretation of the estimated model (Grimm et al., 2017; Nagin, 2005; van de Schoot, Sijbrandij, Winter, Depaoli, & Vermunt, 2017). Such interpretation should consider whether the emergent trajectories are distinct, and whether they are theoretically relevant. Furthermore, users are reminded that the objective of model selection should not be the maximization of some specific fit statistic but rather to summarize distinctive features of the data in as parsimonious and as sensible a manner as possible (Nagin, 2005, p. 77).

4. Software availability

Several software packages exist for the estimation of longitudinal FMMs. Popular packages range from licenced software such as SAS, Stata, Mplus, and Latent GOLD to the open-source R platform and its associated packages. They vary in their capacity to run the various models, ability to extend beyond standard and default specifications

Table 3
Features of popular software for longitudinal FMM (as at May 2019) (*Only those mentioned in Section 3 in this paper are reported).

Software	SAS	Stata	Mplus	R	Latent GOLD
Relevant package/ procedure	Proc Traj	Traj, GLLAMM	TYPE = MIXTURE	LCMM, OpenMX, flexMix, mclust, mixtools	FM Regression
Model types	GBTM	Traj: GBTM, GLLAMM: GMM, LCGA	GMM, LCGA, GBTM	GMM, LCGA, GBTM	GMM, LCGA, GBTM
Outcome types and link function					
Continuous	Censored normal	Censored normal/ beta	Normal/ censored normal	Normal/ censored normal	Multivariate/ censored/ truncated normal
Categorical (ordinal and nominal)	✗	Traj: ✗ GLLAMM: Multinomial logit	Multinomial logit	Multinomial logit	Multinomial logit
Binary	Logit	Probit/ logit	Probit/ logit	Probit/ logit	Probit/ logit
Count	Poisson, Zero inflated Poisson	Zero inflated Poisson	Poisson, Zero inflated Poisson, Negative binomial	Poisson	Truncated/ overdispersed Poisson, truncated/ overdispersed binomial, Zero inflated Poisson, Negative binomial
Trajectory specification					
Random effects	Censored normal only	Traj: ✗ GLLAMM: ✓	✓	✓	✓
Covariance structure of random effects (D matrix)	Censored normal: Equal between classes	Traj: No random effects GLLAMM: Covariance structure may be specified by user	Covariance structure may be specified by user	Covariance structure may be specified by user	Covariance structure may be specified by user
R matrix	Fixed to be the same across classes and time	Traj: Fixed to be the same across classes and time GLLAMM: Structure may be specified by user	Structure may be specified by user	Structure may be specified by user	Structure may be specified by user
Allows for first-order autoregressive term in R	✗	Traj: ✗ GLLAMM: ✓	✓	Package dependent LCMM, OpenMX: ✓	✓
Fit criteria and test statistics					
Fit and test statistics*	AIC, APPA, BIC, log-likelihood, Wald test	AIC, BIC, log-likelihood	AIC, APPA, aLMR, BIC, BLRT, MVK, MVS, ssBIC, VLMR, Wald test	AIC, APPA, BIC, BLRT, CAIC, CVE, ssBIC	AIC, BIC, BLRT, CAIC, CLC, ICL-BIC, ssBIC

(such as varying covariance structures and the inclusion of random effects), and standard model fit criteria output.

An outline of the various features of popular software packages used in applied studies is summarized in Table 3. This list is not intended to be exhaustive, but provides a starting point for researchers. This section delineates the various capabilities of the software packages, including types of outcomes supported, trajectory specification, inclusion of random effects, constraints on the covariance structure, and default fit criteria provided. An expanded features list for the packages, such as model extensions accounting for non-random attrition, time-variant and -invariant predictors, multivariate outcomes is presented in Table 8 in the appendix.

4.1. SAS

Proc traj (Jones, Nagin, & Roeder, 2001) is a procedure in SAS to primarily estimate GBTM, but random effects are possible with the censored normal specification. It supports binary, continuous and count outcomes. Regarding trajectory specification, the procedure can accommodate up to quintic polynomial orders. The covariance structure (**Σ**) is restricted to a common diagonal covariance structure across classes and time. Proc traj assumes conditional independence and thus can use maximum likelihood (ML) estimation following the general quasi-Newton procedure. Proc traj can handle multivariate outcome models (Nagin, Jones, Lima Passos, & Tremblay, 2018).

Proc NLMIXED is another SAS procedure which allows for multiple classes, the inclusion of random effects and a variety of link functions (Grimm & Ram, 2009). From our investigation, it has not often been used in applied research concerning longitudinal FMMs. However, for SAS practitioners it may be worthwhile to consider a selection of studies as a reference (Grimm & Ram, 2009; Lin et al., 2014).

4.2. Stata

Traj (Jones & Nagin, 2013) is a package developed for Stata by the creators of Proc Traj for SAS. As such, it has most of the salient features of Proc Traj. However, it is not able to accommodate random effects of any type and is only able to estimate GBTM models. Traj is able to include the beta distribution for continuous data poorly fit by the normal distribution (Elmer, Jones, & Nagin, 2018).

Gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2004) is capable of handling more complex longitudinal FMMs, including random effects (Palardy & Vermunt, 2010). In contrast to Traj, it can handle ordered and unordered categorical outcomes as well as splines in the trajectory specification. The covariance structure (of random effects, **D** matrix) may also be specified by the user to vary across time and class (Rabe-Hesketh & Skrondal, 2012). The only default model fit criteria output of gllamm is the log-likelihood but the LRT, AIC and BIC are easily computed by other procedures using gllamm's exported log-likelihood.

4.3. Mplus

Mplus (Muthén & Muthén, 2017), built upon the structural equation modelling framework, is often cited in latent trajectory studies (Hallquist & Wiley, 2018; Jung & Wickrama, 2008; Nielsen et al., 2014; van de Schoot et al., 2017). It can handle multiple outcome types and is technically unconstrained in trajectory specification (bearing in mind model convergence and performance).

Mplus has flexibility in modelling outcomes such as allowing for differences in residual variances over time, correlated residuals over time, and allowing for different covariance matrices of the random effects per class. The default specification for the residuals of outcome variables (**R** matrix) is to allow their variance to differ between time points and not to allow autocorrelation. The default for variances and covariances of random effects (**D** matrix) is equality across classes. These restrictions can be relaxed, but this adds to the computational complexity and may prevent convergence of the model.

The software has the capacity to model combinations of outcome types for multivariate growth processes (Muthén & Muthén, 2017). Moreover, Mplus may accommodate time points in measurement that differ between individuals, linear and non-linear parameter constraints, as well as providing bootstrap standard errors and confidence intervals.

Mplus provides an extensive selection of model fit criteria (Asparouhov & Muthén, 2012; Hallquist & Wiley, 2018; Muthén & Muthén, 2017) and is the only program of the five considered here which provides the MVS and MVK tests (Table 2). Classification quality measures provided include entropy, average latent class probabilities for most likely latent class membership, and individual classification probabilities for most likely latent class membership. In addition, Wald chi-square test of parameter equalities, and tests of whether fixed effects differ across latent classes using posterior probability-based multiple imputations, amongst other features, are provided.

4.4. R and associated packages

R is an open-source software consisting of many packages, ranging in their capabilities and default specifications for handling longitudinal FMM. The most often cited packages include: LCMM, OpenMX, flexMix, mclust and mixtools, and have been applied in a variety of developmental trajectory studies (Baker et al., 2017; Grimm, Ram, & Estabrook, 2010; Gruen & Leisch, 2008; Proust-Lima, Philipps, & Liquet, 2017; Scrucca, Fop, Murphy, & Raftery, 2016).

The **LCMM package** (Proust-Lima et al., 2017) can accommodate most outcomes (but excludes count responses) using nonlinear link functions. In addition to higher order polynomials in modelling the trajectory, LCMM can accommodate splines or the Beta cumulative distribution function in modelling the trajectory. Random effects are handled in LCMM with their default variance-covariance matrix being non-structured, but a diagonal matrix can be set. It can be allowed to vary over latent classes. Correlation between errors may also be modelled.

The parameters of the nonlinear link functions and of the latent process are estimated simultaneously using the ML method and may be extended to non-linear fixed effects using splines and the Beta link function. Model fit criteria provided include the log-likelihood, posterior probability of assignment, AIC and BIC. Additional features include the capacity to test for conditional independence.

OpenMX (Boker et al., 2018; Neale et al., 2016) is a versatile and comprehensive package capable of estimating longitudinal FMMs. It has the same capability as Mplus in handling outcomes of various types, various trajectory specifications as well as support for splines. The package allows for the free estimation of variances, intercepts, and non-diagonal covariances. However, the user must define the means and variance parameters as there is no default setting (Infurna & Grimm, 2017). OpenMX provides support for modelling autocorrelation. The AIC, BIC, ssAIC and ssBIC (Boker et al., 2018, p. 315) are part of the default output and the LRT may be requested.

Flexmix (Leisch, 2004) is capable of estimating longitudinal FMMs. It has support for normal, binomial and Poisson link functions (Gruen & Leisch, 2008). Users may set diagonal or unconstrained covariance matrix models. Model estimation is with ML-EM (Expectation Maximization). Fit statistics provided include the AIC, BIC, ICL, and bootstrapped *p*-value.

Mclust (Scrucca et al., 2016) may also be used in longitudinal FMM estimation (Davies et al., 2017), particularly of the Gaussian mixture modelling type. Users can specify different covariance structures. It has support for CVE, and outputs the BIC, BLRT, ICL and log-likelihood for model selection.

Mixtools (Benaglia, Chauveau, Hunter, & Young, 2009) has the capacity to estimate longitudinal FMM for both parametric and semi-parametric settings. It operates within a mixtures-of-regressions setting and has the capacity to handle linear regression, logistic regression,

Poisson regression, linear regression with change points, predictor-dependent class probabilities as well as including random effects regressions. Model fit statistics provided include AIC, BIC, BLRT, CAIC and ICL.

4.5. Latent GOLD

Latent GOLD (Vermunt & Magidson, 2016) has the capacity to model trajectory specifications which differ between classes and the use of B-splines instead of polynomials (Francis, Elliott, & Weldon, 2016). It can handle count, continuous, binary and categorical outcomes. It is as flexible as Mplus and R in its available features. Model fit statistics which can be output include the log-likelihood, AIC, BIC, CAIC, ssBIC, estimated proportion of classification errors, Entropy, Classification Likelihood Criterion (CLC), and ICL-BIC.

4.6. Further remarks

The software packages discussed vary considerably in their capabilities, output and default model specifications. It is for the user to decide which is best suited for their purposes, bearing in mind their own model's underlying assumptions, flexibility, and limitations.

Despite the ever-increasing list of fit criteria and their importance for class enumeration, their integration into software and software capability is limited. The AIC and BIC are often the only default statistics provided, meaning that, if a user is interested in using other fit criteria as outlined in Table 2, they will have to be calculated separately by the software user.

One of the attempts to remedy this is given by the fit-criteria assessment plot (Klijn et al., 2017) (F-CAP). F-CAP is a tool available for GBTM in SAS and Stata which exports the log-likelihood and other fit statistics directly from the software package. It includes several goodness-of-fit (AIC, BIC, log-likelihood) and model-adequacy criteria (APPA, OCC) and displays these visually. The user can then gain informative insight into how these criteria change through increasing the number of latent trajectories, which assists in class enumeration.

5. An empirical example illustrating a strategy for fitting longitudinal mixture models (GBTM, LCGA and GMM)

5.1. General strategy

The absence of an automated model selection process makes the user's involvement fundamental. Nonetheless, some best practice guidelines for latent class trajectory modelling exist. The GRoLTS-checklist (van de Schoot et al., 2017) provides a list of which key components in latent trajectory studies should be reported, such as whether alternative specifications of within-class heterogeneity have been considered, alternative specifications for between-class variance-covariance matrices, alternative shapes and functional forms of the trajectories, as well as model fit statistics used in model selection. The complete checklist consists of 16 items which are intended to increase the uniformity of reporting in latent trajectory studies such that presented results are transparent. It is designed to assist researchers during the modelling and write-up process as well as in the interpretation, critical assessment, replicability and comparison of models (van de Schoot et al., 2017). Furthermore, a framework by Lennon et al. (2018) details the modelling steps, considerations, and interpretation of latent growth models. These range from establishing an initial exploratory model, the inclusion of random effects, covariance structures, use of model fit statistics for model selection, graphical analysis, to sensitivity analysis for the generalisability of results. Since our aim is to illustrate model selection choices and not the writing up and reporting of results, we refer the reader to Lennon et al. (2018) and van de Schoot et al. (2017), which cover this extensively.

Fig. 3 illustrates the model selection path undertaken in our application. It should be noted that this is neither definitive nor binding, but accords, to some extent with frameworks suggested in previous studies (Lennon et al., 2018; Ram & Grimm, 2009). From the figure, it is apparent that several decisions need to be made when selecting the best fitting model.

It is good practice to first plot a random selection of subjects to provide a visual representation of whether enough heterogeneity of development is evident in the data to justify the use of mixture modelling (spaghetti plot). One then needs to select a maximum K and polynomial order given the number of time points, sample size, previous theoretical and/or practical insights, and the spaghetti plot, for the initial scoping of potential models. If deciding upon a maximum K is a challenge, users are reminded that more than 10 classes rarely emerge in applied studies. Additionally, we also recommend fitting a GCM to the data to present the best single-class depiction of change and with which we will compare our models (using relevant fit statistics).

To illustrate model selection in the application of longitudinal FMM, we begin with the most constrained of the considered models, the GBTM. The GBTM should converge the quickest to a solution given its lower number of free parameters when compared to LCGA and GMM. We suggest finding a range of plausible K's for the GBTM and selecting the K-class GBTM model within that subset with the best BIC (Lennon et al., 2018). Ideally, this should be confirmed by looking at other available fit statistics which are discussed in Section 3.

We then use modified likelihood ratio tests (LRTs) to assess whether that K-class GBTM model is dismissed in favour of a $K-1$ model. If it is dismissed, the lower K model is chosen. The LRT is then repeated on that selected model to ascertain whether a lower K model may be selected. The LRT is repeated until it dismisses a lower K model (Ram & Grimm, 2009).

We then extend the model for the selected K by dropping one constraint at a time (by allowing for the dependence of residual variance on time and/or class), which is an LCGA. We then select the LCGA or GBTM model with the lowest BIC. If it is an LCGA, we then use the LRT to determine whether that selected model's K can be reduced further. This same strategy is used when refining the model during the subsequent steps of relaxing the model constraints (by allowing for class-variant or class-invariant random effect variances), that is, select the model with the best BIC and then check how much K can be reduced using the LRT.

The strategy of fitting consecutively more lenient models is motivated by the fact that the cause of non-convergence if it occurs, will be easier to identify. It is recommended to inspect the trajectory plots at each step to ensure that the emergent patterns are sensible by considering their empirical implications and whether the trajectories are distinct. Once a K is selected, we do not consider $K+1$ models in subsequent steps in order to narrow down our possible choices and to preserve the principle of parsimony.

When a model extension does not lead to a lower BIC and K cannot be further reduced by an LRT, the polynomial order may be pruned subject to significance. This is achieved by deleting the highest order polynomial term that is nonsignificant iteratively per class using a Wald test. Lower order nonsignificant polynomial terms are not removed if the highest order polynomial term is significant. Finally, it is advised that the replicability of the chosen model be tested through cross-validation on a new sample (but this is beyond the scope of our illustration).

5.2. An illustration

5.2.1. The dataset

We expanded upon the methodology of a GBTM study (Lima Passos et al., 2017) comprising a data set ($n = 1907$)of log-transformed self-reported retrospective alcohol consumption ($AC^{*}_{it} = \log(AC_{it} + 1)$) by including a GMM analysis. AC_{it} is the total volume expressing the weekly consumption (in glasses) of subject i and was measured at 4

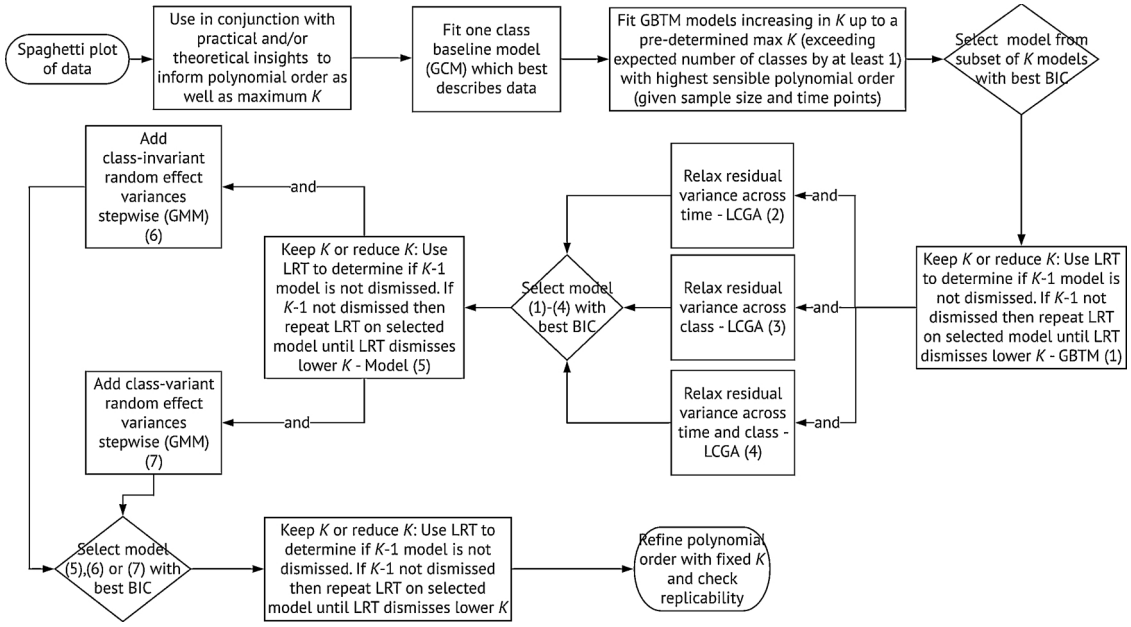

Fig. 3. Model selection flowchart.

time intervals (t = 1 Youth: 12–18 years, t = 2 Young adult: 19–27 years, t = 3 Adult: 28–44 years, t = 4 Middle age: 45–60 years). Skewness and kurtosis measures for outcomes showed highly non-normal data and motivated the log transform in the referenced study (Lima Passos et al., 2017). Even with the log transform, some skewness remains, but for the purpose of illustration we chose to follow the same methodology of the referenced study.

A spaghetti plot of a random selection of subjects is presented in Fig. 7 in the appendix and may motivate the choice of a quadratic function to model the trajectories. Therefore, we will assume that each of the trajectories for alcohol consumption in distinct classes, AC_{it}^{*k}, may be modelled by a quadratic function of the GMM general form:

$$AC_{it}^{*k} = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)time_{it} + (\beta_2^k + b_{2i}^k)time_{it}^2 + \varepsilon_{it}^k \quad (18)$$

where $i = 1, \ldots, n$, $t = 1,2,3,4$, $k = 1, \ldots, K$, *time* is the time period considered, and β_0^k, b_{0i}^k, β_1^k, β_2^k, b_{1i}^k, b_{2i}^k and ε_{it}^k are as defined in Eq. (3). However, as said in Section 5.1, we start the modelling with the GCM and the GBTM (which are both special cases of the GMM) for reasons explained there.

5.2.2. Model selection

We used Mplus v7.3 for our analysis with some selected code provided in the appendix. We first fitted the best one-class model (GCM) to the data with which we will compare subsequent models to justify multiple class solutions. Using the BIC as an aid and considering a variety of **D** and **R** specifications, in addition to ensuring that estimated parameters make mathematical sense (i.e. non-negative variances, correlations between -1 and 1), we settled on a GCM model of the form $AC_{it}^{*} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})time_{it} + (\beta_2 + b_{2i})time_{it}^2 + \varepsilon_{it}$ with constant residual variance over time (i.e. $\varepsilon_{it} \sim N(0, \sigma^2)$). This model exhibited a BIC of 17 167.811.

In the next step of our analysis, we fitted GBTMs from $K = 2$ until a maximum $K = 8$. The maximum K is usually set as the expected number of classes (which was informed by the referenced study (Lima Passos et al., 2017)) plus 1. For the GBTM, the random effects in Eq. (18) were set to zero ((i.e. $\mathbf{D}^k = 0$) with the restriction of equal residual variances across time and classes (i.e. $\mathbf{\Sigma}^k = \sigma^2\mathbf{I}$where σ^2 is the residual variance).

In the GBTM step, the BIC continued to improve as K increased. The AIC and ssBIC showed similar behaviour (see Fig. 4). The improvement in these fit statistics with an increase in K is a known issue (Erosheva, Matsueda, & Telesca, 2014) and may motivate model extension i.e. freer estimation of **D** and **R** matrices. Nonetheless, the AIC, BIC, and ssBIC of the K = 5, 6, 7, and 8 GBTM's were less than the GCM's (which was close to the BIC of the 4-class GBTM (Fig. 4)). These results are reported in Table 4 for further analysis.

We then used the VLMR and aLMR LRTs to establish whether K could be reduced further since the BLRT bootstrap draws did not converge to a reliable solution. Moreover, the BLRT is particularly sensitive to model misspecification and is advised against using during initial model exploration (Nylund et al., 2007). Our goal was to ensure that the information criteria decreased (improved) with model extension and that the LRT supported the lowest possible K class model.

From Table 4, the 8-class GBTM had the best BIC (GBTM4). However, the VLMR and aLMR p-values led us to not dismiss the $K = 7$ class GBTM (GBTM3). In turn, for GBTM3, the VLMR and aLMR led us to not dismiss the $K = 6$ class model (GBTM2). We, therefore, settled on a $K = 6$ class quadratic GBTM model (GBTM2), since the VLMR and aLMR both led to the dismissal of a $K = 5$ solution at the 5% significance level. The plot for the estimated trajectories for the $K = 6$ GBTM model is shown in Fig. 5.

Given the $K = 6$ quadratic GBTM model, we extended the model to allow different residual variance error structures i.e. same over class but different across time (LCGA1), same over time but different over class (LCGA2), and different across time and over class. This last extension was not possible as it led to singularity of the information matrix. We then compared these models' (LCGA1 and LCGA2) BIC value to that of the $K = 6$ GBTM2 model and selected the model with the best BIC. This happened to be the LCGA2 model which had a BIC of 15,885.010. Furthermore, its VLMR and aLMR p-values led us to not dismiss a $K = 5$ solution. We, therefore, estimated a lower $K = 5$ model (LCGA3), which was not dismissed by the VLMR and aLMR tests. Thus, we retained the LCGA3 model.

We then expanded the LCGA3 into a GMM by adding class-invariant random effect variances stepwise, and class-variant random effect variances stepwise, respectively (i.e. first for the intercept, then for the intercept and linear slope, and finally for the intercept, linear slope and quadratic slope, allowing for covariance between the random effects). Of the 5-class GMM specifications investigated, only two converged to a solution and did not obtain negative variances (which is indicative of an inappropriate model). These were a 5-class GMM with class-invariant random intercept variance (GMM1) and a 5-class GMM with class-

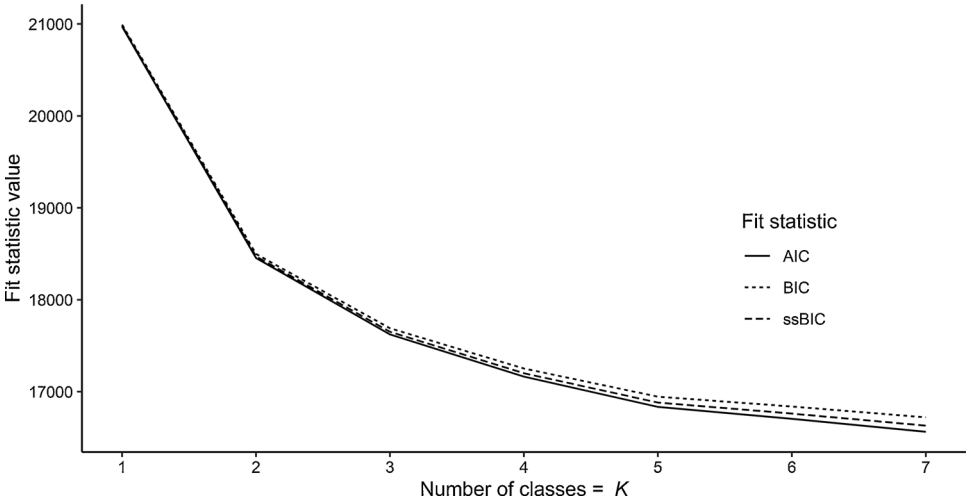

Fig. 4. Fit criteria performance GBTM.

Table 4
Fit statistics for considered GBTM.

Model	GBTM1	GBTM2	GBTM3	GBTM4
Classes	5	6	7	8
Average APPA	0.8518	0.8262	0.8247	0.8178
Lowest APPA	0.821	0.747	0.765	0.765
AIC	16834.988	16706.057	16564.953	16424.918
BIC	16946.054	16839.336	16720.445	16602.623
ssBIC	16882.514	16763.088	16631.489	16500.959
Scaled Entropy	0.783	0.77	0.784	0.78
VLMR p-value	<0.0001	0.0270	0.1478	0.2372
aLMR p-value	<0.0001	0.0303	0.1567	0.2436

variant random intercept variance (GMM2) (see Table 5). Both models exhibited a BIC better than the LCGA3 with the GMM2 having the best BIC, and we selected this for further refinement. Finally, the VLMR and aLMR showed that the 5-class GMM2 could not be reduced to a 4-class GMM (Table 5).

Next, the significance of the various polynomial fixed effects terms of the selected model were checked. Discarding nonsignificant higher order polynomial terms led to a marginally better model fit (GMM3) and we settled on this as our final solution. Its estimated trajectories with confidence intervals (to display class separation for the final model in terms of the fixed effects i.e. $\hat{\beta}_0^k + \hat{\beta}_1^k time_{it} + \hat{\beta}_2^k time_{it}^2 \pm 1.96\sqrt{Var(\hat{\beta}_0^k + \hat{\beta}_1^k time_{it} + \hat{\beta}_2^k time_{it}^2)}$) are shown in Fig. 6. The estimated mean equations are given in Table 9 in the appendix.

Finally, the model should be replicated on more data (known as model validation), but due to space limitations and the absence of a second independent data set is beyond the scope of this illustration.

It is important to note that slight deviations in the modelling strategy could result in different best-fit models. Unfortunately, one is forced to choose a certain strategy, as it is almost impossible to investigate all mixture models within the chosen range for K, where the set of possible models is a multiplicative function of the number of possible covariance structures, the number of classes, and the polynomial order per class.

We have used this empirical example to illustrate a possible predefined model selection procedure, but this is in no means definitive since not all possible model specifications were considered, such as higher K and higher order polynomials with alternate **D** and **R** specifications. Practitioners should be guided by statistical criteria as outlined in Section 3 as well as practical experience with data and results from previous studies when estimating such models. The rote application of model selection in mixture modelling without careful consideration of the practical and/or theoretical implications of the emergent trajectories and classification of individuals must be strongly discouraged.

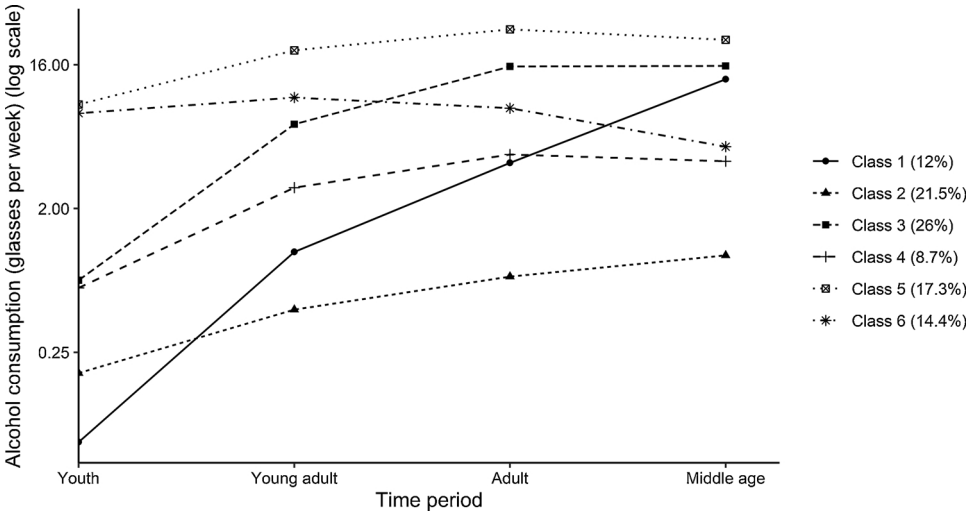

Fig. 5. Estimated trajectories of 6-class GBTM model.

Table 5
Final LCGA and GMM solution.

Model	GBTM2	LCGA1	LCGA2	LCGA3	GMM1	GMM2	GMM3
Classes	6	6	6	5	5	5	5
Specification	Same residual variance over class and over time	Same residual variance over class, different over time	Same residual variance over time, different over class	Same residual variance over time, different over class	Class-invariant random intercept variance, same residual variance over time and different over class	Class-variant random intercept variance, same residual variance over time and different over class	Class-variant random intercept variance, same residual variance over time and different over class, all fixed effects significant
Average APPA	0.8262	0.851	0.882	0.867	0.865	0.839	0.840
Lowest APPA	0.747	0.817	0.826	0.803	0.811	0.759	0.753
AIC	16706.06	16499.75	15723.965	16035.771	15651.773	15413.507	15413.256
BIC	16839.34	16649.69	15885.01	16169.05	15790.606	15574.552	15568.748
ssBIC	16763.09	16563.91	15792.877	16092.802	15711.18	15482.419	15479.792
Scaled entropy	0.77	0.811	0.832	0.804	0.778	0.764	0.762
VLMR p-value	0.027	0.5819	0.0566	0.0005	0.0012	0.0251	0.0104
aLMR p-value	0.0303	0.5888	0.0583	0.0006	0.0014	0.0268	0.0115

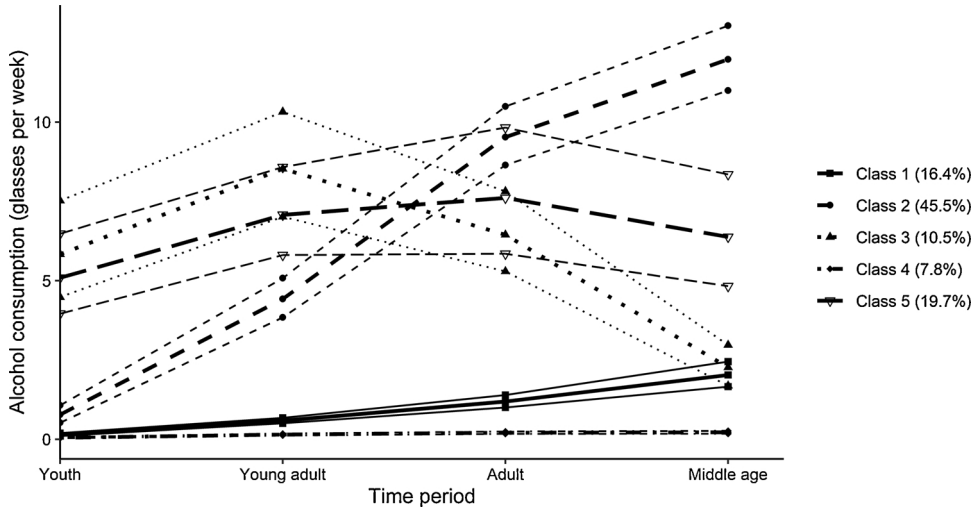

Fig. 6. Final 5-class GMM estimated trajectories and 95% confidence intervals.

Furthermore, model selection is not the final step. Additional steps are routinely undertaken to determine the interpretational and/or conceptual meaningfulness of emergent trajectories. This includes ascertaining which members' characteristics are associated with class membership and/or linking the emergent developmental patterns to distal outcomes (Asparouhov & Muthén, 2014; Lanza, Tan, & Bray, 2013; Nylund-Gibson, Grimm, & Masyn, 2019). This may assist in understanding the processes generating the heterogeneity of developmental paths and their potential implications.

6. Concluding remarks

This paper has given an instructive overview of longitudinal FMM models, specifically GMM, LCGA, GBTM, and their interrelatedness. Of the models considered, the GMM is the most versatile. It allows inter-individual variability between subjects within latent classes through the inclusion of random effects, and a complex covariance structure. By contrast, LCGA and GBTM do not have random effects. They make the restrictive assumption of independent errors, with LCGA allowing time and class-variant error variances, and GBTM imposing the same residual variance over classes and over time. Furthermore, we provided an overview of various software available for the estimation of longitudinal FMM which all vary in their capabilities, particularly of fit statistics reported and allowable covariance structures.

We described and illustrated the important first step of model selection, which is determining the number of classes *K*. The use of statistical fit indices for class enumeration introduces some statistical vigour to the process, but remains to some extent also heuristic. Our review, together with the empirical example reiterates the consensus that there is of yet no one best fit statistic for class enumeration, as their performance is largely dependent on the underlying data properties. Therefore, it is recommended to use as many of these fit statistics as practical to determine the best model whilst bearing in mind their limitations as detailed in Section 3 in addition to a vigorous inspection of the emergent trajectories.

In our illustrative example, we offered a possible but, by no means, binding model selection strategy for class enumeration and polynomial order determination. We followed the path of going from simpler (GBTM) to more complex models (GMM), whereas the opposite direction was chosen for the polynomial order (from higher to lower). These choices were made to enable identification of the cause of model non-convergence, if it occurs, as well as to restrict the set of models investigated. However, it is apparent that there are many possible choices and pathways for researchers to follow. Researchers should be guided by parsimony, model fit, and their research question as well as being cognisant of possible software limitations.

We refrained from expanding trajectories to multiple outcomes simultaneously (Jones & Nagin, 2007; Lai, Xu, Koller, Foroud, & Gao, 2016; Nagin et al., 2018; Verbeke et al., 2014) and from addressing questions related to important steps subsequent (or possibly concurrent) to model selection, referred to as model validation (Lanza et al., 2013; Nylund-Gibson et al., 2019). Readers should consult a recent overview article (Nylund-Gibson et al., 2019) for more details on best practice guidelines for model validation. Space restrictions also

precluded us from addressing in detail further modelling issues for longitudinal FMM, many of which to date are still unresolved. Below we briefly discuss some issues relating to data features, which have a marked impact on class enumeration accuracy, trajectory shape detectability, and classification performance.

Data features known to negatively impact the quality of class enumeration in longitudinal FMMs are small sample sizes (< 250) (Diallo, Morin, & Lu, 2017), a small number of time points (< 4) (Diallo et al., 2017b), the lack of a natural starting point in the longitudinal measurements (e.g. birth), and a misspecified covariance structure.

An insufficient sample size is known to underlie model convergence issues, improper solutions and the inability to identify small but meaningful subgroups (Berlin, Williams et al., 2014). However, adequate sample size calculations are often difficult as these depend on a variety of factors, including the complexity of the model, distribution of the variables, the amount of missing data, number of repeated measures and the strength of the relationship between variables in the model (Muthén & Muthén, 2002). Sample size studies for GMM are rather limited, but a simulation study (Kim, 2012) found a minimum sample size of 200 is required in the case of complete data, high class separation and 2 classes, and a required sample size of 900 for the case of 20% missing data, low class separation and 6 classes.

The impact of the number of, and the spacing between, time points on class enumeration, classification and parameter estimates is understudied. An empirical GBTM study (Tan, Dierker, Rose, & Li, 2011) showed that, although adding time points within a given time interval did not have a marked impact on the estimation of trajectory curves, it did have a marked impact on the correct classification of individuals. In a simulation study, Davies et al. (2017) showed that increasing the number of time points (from 4 to 8) by expanding the time interval had a modest positive effect on classification performance, particularly for GMMs with residual and random effect variances free to vary between classes. Furthermore, a simulation study (Diallo et al., 2017b) for a GMM with, next to time as a predictor, also a time-varying predictor investigated the impact of increasing the number of time points by expanding the interval from 4, to 6, to 8 measurements. They found that of the design factors considered (number of time points, sample size, class probabilities, constraints on the error variances, and proportion of explained variance in repeated measures due to time-varying predictor), a small sample size, a small number of time points, and especially their combination had a considerable impact on the presence of bias in the estimation of random effect variances and covariances.

Another underinvestigated issue is the case where data exhibits no natural starting point and as a result show high onset variability reflected by markedly different intercepts. In this case, extracted trajectories may be dominated by level effects (Heggeseth & Jewell, 2018). More precisely, intercept variance dominance may lead to important small classes, which differ significantly in shape and growth over time, not being detected. A sometimes-used solution is pre-processing the data by subtracting each subject's average from their repeated measures which removes the level effect. However, this has important implications for the covariance and dependency structure, especially if the number of time points is small or if individuals are not all observed at fixed time intervals (Heggeseth & Jewell, 2018).

Violations of the assumptions of the underlying conditional distribution of the longitudinal sequence (Frühwirth-Schnatter, Celeux, & Robert, 2019) (see Eq. (5)) have been shown to lead to class over-extraction when using penalized likelihood criteria (Bauer & Curran, 2003; Fruhwirth-Schnatter & Pyne, 2010; Lee & McLachlan, 2013) as discussed in Section 3. This may be addressed by choosing more flexible probability density functions for the classes, which in many cases provide a better estimate of the true number of classes than the normal (Gaussian) approach (Fruhwirth-Schnatter & Pyne, 2010; Lee & McLachlan, 2013). Moreover, researchers are particularly cautioned to be careful in the specification of the $\boldsymbol{D}^k$ and $\boldsymbol{R}^k$ matrix, since it has been shown that using too restrictive models far outweighs other design conditions such as sample size, prior class probabilities and class separation in terms of class enumeration accuracy (Diallo et al., 2016).

In recent years, the issue of whether to estimate models with or without predictors (besides time) during class enumeration has emerged as a major consideration and remains controversial. Of relevance is the question to what extent the predictors (and the conditions under which they are added) change the trajectories' shape and class assignment. Currently, there is no simple solution on how and when to include predictors of latent classes, but some consensus has emerged. Simulation studies (Davies et al., 2018; Diallo et al., 2017a; Kim, Vermunt, Bakk, Jaki, & Van Horn, 2016) have investigated the influence of various predictor specifications (such as absence of predictor effects, predictor effects on class membership, and predictor by time interactions in the trajectory) on class enumeration. They generally recommend including predictors after class enumeration, because, even when the true model for data generation included predictors, they found that including correctly specified predictors in the enumeration phase only led to small improvements in class enumeration accuracy. Improvements in enumeration accuracy had limited practical significance, particularly since the models were found to be highly sensitive (in terms of class enumeration and parameter estimates) to predictor misspecification (specifying a relationship when in fact there is none) (Diallo et al., 2017a). This is important since in practice it is often impossible to know beforehand the precise predictor effects. Once a stable solution in terms of class enumeration is found, class predictors may then be introduced into the model with a fixed number of latent classes to examine their effects on parameter estimates and class enumeration. In another study (Tofighi & Enders, 2008), where predictors were specified to have an impact on both the trajectory and class membership, it was found that the inclusion of predictors during class extraction led to substantial class enumeration inaccuracies, especially when the sample size was less than 1000.

To conclude, we have shown throughout this paper that there are many considerations to be taken and issues to be aware of when conducting analyses based on longitudinal FMM models. Typically, a combination of fit statistics, the research question, model parsimony, domain knowledge, and model interpretability should all play a role, not only in the motivation and use of longitudinal FMM (Muthén, 2003) but also in the model selection procedure. It is imperative that researchers keep this in mind and that they clearly document their studies to ensure transparency, replicability, and defensibility. We have attempted to provide a broad introduction to these techniques to increase their accessibility to practitioners. Our paper is not exhaustive, as other mixture FMMs including mixture LTA and SMA exist which practitioners are encouraged to investigate (See Magidson et al., 2009; Muthén and Masyn, 2005), but our hope is that this paper will serve as an introductory guide to the discussed methods for applied studies.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.alcr.2019.100323.

References

Asparouhov, T., & Muthén, B. O. (2012). *Using Mplus TECH11 and TECH14 to test the number of latent classes. Mplus web notes, Vol. 14*.

Asparouhov, T., & Muthén, B. O. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181.

Baker, E., Iqbal, E., Johnston, C., Broadbent, M., Shetty, H., Stewart, R., ... Dobson, R. J. B. (2017). Trajectories of dementia-related cognitive decline in a large mental health records derived patient cohort. *PLoS One, 12*(6), https://doi.org/10.1371/journal.

pone.0178562.
Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338.
Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software, 32*(6), https://doi.org/10.18637/jss.v032.i06.
Berlin, K. S., Parra, G. R., & Williams, N. A. (2014). An introduction to latent variable mixture modeling (Part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology, 39*(2), 188–203. https://doi.org/10.1093/jpepsy/jst085.
Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (Part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology, 39*(2), 174–187. https://doi.org/10.1093/jpepsy/jst084.
Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725. https://doi.org/10.1109/34.865189.
Blaze, T. J. (2013). *Enumerating the correct number of classes in a semiparametric group-based trajectory model.* University of Pittsburgh.
Boker, S. M., Maes, H. H., Spiegel, M., Brick, T. R., Bates, T. C., Mehta, P., ... Kirkpatrick, R. M. (2018). *OpenMx user guide (release 2).* Retrieved fromhttps://vipbg.vcu.edu/vipbg/OpenMx2/docs//OpenMx/latest/OpenMxUserGuide.pdf.
Brame, R., Nagin, D. S., & Wasserman, L. (2006). Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology, 22*(1), 31–59. https://doi.org/10.1007/s10940-005-9001-8.
Burton-Jeangros, C., Blane, D., Howe, L. D., Firestone, R., Tilling, K., & Lawlor, D. A. (2015). In C. Burton-Jeangros, S. Cullati, A. Sacker, & D. Blane (Eds.). *A Life Course perspective on health trajectories and transitions*Springerhttps://doi.org/10.1007/978-3-319-20484-0.
Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195–212. https://doi.org/10.1007/BF01246098.
Chamroukhi, F. (2016). Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification, 33*(3), 374–411. https://doi.org/10.1007/s00357-016-9212-8.
Chen, G., & Tsurumi, H. (2010). Probit and logit model selection. *Communications in Statistics - Theory and Methods, 40*(1), 159–175. https://doi.org/10.1080/03610920903377799.
Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences.* Hoboken, NJ: Wiley.
Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development, 11*(2), 121–136. https://doi.org/10.1080/15248371003699969.
Davies, C. E., Glonek, G. F. V., & Giles, L. C. (2017). The impact of covariance misspecification in group-based trajectory models for longitudinal data with non-stationary covariance structure. *Statistical Methods in Medical Research, 26*(4), 1982–1991. https://doi.org/10.1177/0962280215598806.
Davies, C. E., Giles, L. C., & Glonek, G. F. (2018). Performance of methods for estimating the effect of covariates on group membership probabilities in group-based trajectory models. *Statistical Methods in Medical Research, 27*(10), 2918–2932. https://doi.org/10.1177/0962280216689580.
Demidenko, E. (2013). *Mixed models: Theory and applications with RWiley series in probability and statistics* (2nd ed.). John Wiley & Sons SE - 758 s.
Diallo, T. M. O., Morin, A. J. S., & Lu, H. Z. (2016). Impact of misspecifications of the latent variance–covariance and residual matrices on the class enumeration accuracy of growth mixture models. *Structural Equation Modeling, 23*(4), 507–531. https://doi.org/10.1080/10705511.2016.1169188.
Diallo, T. M. O., Morin, A. J. S., & Lu, H. (2017a). The impact of total and partial inclusion or exclusion of active and inactive time invariant covariates in growth mixture models. *Psychological Methods, 22*(1), 166–190. https://doi.org/10.1037/met0000084.
Diallo, T. M. O., Morin, A. J. S., & Lu, H. Z. (2017b). Performance of growth mixture models in the presence of time-varying covariates. *Behavior Research Methods, 49*(5), 1951–1965. https://doi.org/10.3758/s13428-016-0823-0.
Elmer, J., Jones, B. L., & Nagin, D. S. (2018). Using the Beta distribution in group-based trajectory models. *BMC Medical Research Methodology, 18*(1), 152. https://doi.org/10.1186/s12874-018-0620-9.
Erosheva, E. A., Matsueda, R. L., & Telesca, D. (2014). Breaking bad: Two decades of life-course data analysis in criminology, developmental psychology, and beyond. *Annual Review of Statistics and Its Application, 1*(1), 301–332. https://doi.org/10.1146/annurev-statistics-022513-115701.
Falkenstein, M. J., Nota, J. A., Krompinger, J. W., Schreck, M., Garner, L. E., Potluri, S., ... Elias, J. A. (2019). Empirically-derived response trajectories of intensive residential treatment in obsessive-compulsive disorder: A growth mixture modeling approach. *Journal of Affective Disorders, 245*(July), 827–833. https://doi.org/10.1016/j.jad.2018.11.075.
Faulkenberry, T. J. (2018). Computing Bayes factors to measure evidence from experiments: An extension of the BIC approximation. *Biometrical Letters, 55*(1), 31–43. https://doi.org/10.2478/bile-2018-0003.
Francis, B., Elliott, A., & Weldon, M. (2016). Smoothing group-based trajectory models through B-splines. *Journal of Developmental and Life-Course Criminology, 2*(1), 113–133. https://doi.org/10.1007/s40865-016-0025-6.
Fruhwirth-Schnatter, S., & Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics, 11*(2), 317–336. https://doi.org/10.1093/biostatistics/kxp062.
Frühwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). In S. Frühwirth-Schnatter, G. Celeux, & C. P. Robert (Eds.). *Handbook of mixture analysis*. Boca Raton, FL: Chapman and Hall/CRC.
Grevenstein, D., & Kröninger-Jungaberle, H. (2015). Two patterns of cannabis use among adolescents: Results of a 10-year prospective study using a growth mixture model. *Substance Abuse, 36*(1), 85–89. https://doi.org/10.1080/08897077.2013.879978.
Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in m plus and SAS. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 676–701. https://doi.org/10.1080/10705510903206055.
Grimm, K. J., Ram, N., & Estabrook, R. (2010). Nonlinear structured growth mixture models in M plus and OpenMx. *Multivariate Behavioral Research, 45*(6), 887–909. https://doi.org/10.1080/00273171.2010.531230.
Grimm, K. J., Mazza, G. L., & Davoudzadeh, P. (2017). Model selection in finite mixture models: A k-fold cross-validation approach. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 246–256. https://doi.org/10.1080/10705511.2016.1250638.
Gruen, B., & Leisch, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software, 28*(4), https://doi.org/10.18637/jss.v028.i04.
Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25*(4), https://doi.org/10.1080/10705511.2017.1402334.
Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters, 4*(2), 53–56. https://doi.org/10.1016/0167-7152(86)90016-7.
He, J., & Fan, X. (2018). Evaluating the performance of the K-fold cross-validation approach for model selection in growth mixture modeling. *Structural Equation Modeling, 0*(0), 1–14. https://doi.org/10.1080/10705511.2018.1500140.
Heggeseth, B. C., & Jewell, N. P. (2018). How Gaussian mixture models might miss detecting factors that impact growth patterns. *The Annals of Applied Statistics, 12*(1), 222–245. https://doi.org/10.1214/17-AOAS1066.
Hélie, S. (2006). An introduction to model selection: Tools and algorithms. *Tutorials in Quantitative Methods for Psychology, 2*(1), 1–10. Retrieved from https://doaj.org/article/183474b2fede44d5a7d853a5888b7f0a.
Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling, 14*(2), 202–226. https://doi.org/10.1080/10705510709336744.
Hilterman, E. L. B., Bongers, I. L., Nicholls, T. L., & van Nieuwenhuizen, C. (2018). Supervision trajectories of male juvenile offenders: Growth mixture modeling on SAVRY risk assessments. *Child and Adolescent Psychiatry and Mental Health, 12*(1), 15. https://doi.org/10.1186/s13034-018-0222-7.
Infurna, F. J., & Grimm, K. J. (2017). The use of growth mixture modeling for studying resilience to major life stressors in adulthood and old age: Lessons for class size and identification and model selection. *The Journals of Gerontology: Series B, 73*(1), 148–159. https://doi.org/10.1093/geronb/gbx019.
Jeffreys, H. (2004). *The theory of probabilityOxford classic texts in the physical sciences* (third ed.). Retrieved from http://www.loc.gov/catdir/enhancements/fy0606/99175168-t.html LK -.
Jeffries, N. O. (2003). A note on "Testing the number of components in a normal mixture". *Biometrika, 90*(4), 991–994. Retrieved from http://www.jstor.org/stable/30042105.
Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods and Research, 35*(4), 542–571. https://doi.org/10.1177/0049124106292364.
Jones, B. L., & Nagin, D. S. (2013). A note on a stata plugin for estimating group-based trajectory models. *Sociological Methods & Research, 42*(4), 608–613. https://doi.org/10.1177/0049124113503141.
Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29*(3), 374–393.
Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x.
Kass, R. E., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.
Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928–934. https://doi.org/10.1080/01621459.1995.10476592.
Killian, M. O., Cimino, A. N., Weller, B. E., & Hyun Seo, C. (2019). A systematic review of latent variable mixture modeling research in social work journals. *Journal of Evidence-Based Social Work, 16*(2), 192–210. https://doi.org/10.1080/23761407.2019.1577783.
Kim, S. Y. (2012). Sample size requirements in single- and multiphase growth mixture models: A Monte Carlo simulation study. *Structural Equation Modeling, 19*(3), 457–476. https://doi.org/10.1080/10705511.2012.687672.
Kim, E. S., & Wang, Y. (2017). Class enumeration and parameter recovery of growth mixture modeling and second-order growth mixture modeling in the presence of measurement noninvariance between latent classes. *Frontiers in Psychology, 8*(September), https://doi.org/10.3389/fpsyg.2017.01499.
Kim, M., Vermunt, J., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 601–614. https://doi.org/10.1080/10705511.2016.1158655.
Klijn, S. L., Weijenberg, M. P., Lemmens, P., van den Brandt, P. A., & Lima Passos, V. (2017). Introducing the fit-criteria assessment plot—A visualisation tool to assist

class enumeration in group-based trajectory modelling. *Statistical Methods in Medical Research, 26*(5), 2424–2436. https://doi.org/10.1177/0962280215598665.
Lai, D., Xu, H., Koller, D., Foroud, T., & Gao, S. (2016). A multivariate finite mixture latent trajectory model with application to dementia studies. *Journal of Applied Statistics, 43*(14), 2503–2523. https://doi.org/10.1080/02664763.2016.1141181.
Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(1), 1–26. https://doi.org/10.1080/10705511.2013.742377.
Laursen, B. P., & Hoff, E. (2006). Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly, 52*(3), 377–389. https://doi.org/10.1353/mpq.2006.0029.
Lee, S. X., & McLachlan, G. J. (2013). EMMIXuskew: An R package for fitting mixtures of multivariate skew t distributions via the EM algorithm. *Journal of Statistical Software, 55*(12), https://doi.org/10.18637/jss.v055.i12.
Lee, T. K., Wickrama, K. A. S., O'Neal, C. W., & Lorenz, F. O. (2017). Social stratification of general psychopathology trajectories and young adult social outcomes: A second-order growth mixture analysis over the early life course. *Journal of Affective Disorders, 208*(August), 375–383. https://doi.org/10.1016/j.jad.2016.08.037.
Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software, 11*(8), 1–18. https://doi.org/10.18637/jss.v011.i08.
Lennon, H., Kelly, S., Sperrin, M., Buchan, I., Cross, A. J., Leitzmann, M., … Renehan, A. G. (2018). Framework to construct and interpret latent class trajectory modelling. *BMJ Open, 8*. https://doi.org/10.1136/bmjopen-2017-020683.
Lima Passos, V., Klijn, S., van Zandvoort, K., Abidi, L., & Lemmens, P. (2017). At the heart of the problem—A person-centred, developmental perspective on the link between alcohol consumption and cardio-vascular events. *International Journal of Cardiology, 232*, 304–314. https://doi.org/10.1016/j.ijcard.2016.12.094.
Lin, H., Han, L., Peduzzi, P. N., Murphy, T. E., Gill, T. M., & Allore, H. G. (2014). A dynamic trajectory class model for intensive longitudinal categorical outcome. *Statistics in Medicine, 33*(15), 2645–2664. https://doi.org/10.1002/sim.6109.
Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767–778. https://doi.org/10.1093/biomet/90.4.991.
Magidson, J., Vermunt, J. K., & Tran, B. (2009). Using a mixture latent Markov model to analyze longitudinal U.S. employment data involving measurement error. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.). *New trends in psychometrics* (pp. 235–242). Tokyo: Universal Academy Press.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519–530.
McLachlan, G., & Peel, D. (2000). *Finite mixture models. Wiley series in probability and statistics. Applied probability and statistics section*https://doi.org/10.1002/0471721182.
McNeish, D., & Harring, J. R. (2017). The effect of model misspecification on growth mixture model class enumeration. *Journal of Classification, 34*(2), 223–248. https://doi.org/10.1007/s00357-017-9233-y.
Mund, M., & Nestler, S. (2019). Beyond the Cross-Lagged Panel Model: Next-generation statistical tools for analyzing interdependencies across the life course. *Advances in Life Course Research, 41*(October), 100249. https://doi.org/10.1016/j.alcr.2018.10.002.
Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods, 8*(3), 369–377. https://doi.org/10.1037/1082-989X.8.3.369.
Muthén, B. O. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock, & K. M. Samuelsen (Eds.). *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing, Inc.
Muthén, B. O., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics, 30*(1), 27–58. https://doi.org/10.3102/10769986030001027.
Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882–891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x.
Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8.
Muthén, L. K., & Muthén, B. O. (2017). *MPlus user's guide* (eighth edi). Los Angeles, CA: Muthén & Muthén.
Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463–469. https://doi.org/10.1111/j.0006-341X.1999.00463.x.
Nagin, D. S. (2005). *Group-based modeling of development.* https://doi.org/10.4159/9780674041318.
Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*(3), 327.
Nagin, D. S., & Odgers, C. L. (2010a). Group-based trajectory modeling (Nearly) two decades later. *Journal of Quantitative Criminology, 26*(4), 445–453. https://doi.org/10.1007/s10940-010-9113-7.
Nagin, D. S., & Odgers, C. L. (2010b). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology, 6*(1), 109–138. https://doi.org/10.1146/annurev.clinpsy.121208.131413.
Nagin, D. S., Jones, B. L., Lima Passos, V., & Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Statistical Methods in Medical Research, 27*(7), 2015–2023. https://doi.org/10.1177/0962280216673085.
Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., … Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8.
Nielsen, J. D., Rosenthal, J. S., Sun, Y., Day, D. M., Bevc, I., & Duchesne, T. (2014). Group-based criminal trajectory analysis using cross-validation criteria. *Communications in Statistics - Theory and Methods, 43*(20), 4337–4356. https://doi.org/10.1080/03610926.2012.719986.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (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), 535–569. https://doi.org/10.1080/10705510701575396.
Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 0*(0), 1–19. https://doi.org/10.1080/10705511.2019.1590146.
Palardy, G. J., & Vermunt, J. K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics, 35*(5), 532–565. https://doi.org/10.3102/1076998610376895.
Pennoni, F., & Romeo, I. (2017). Latent Markov and growth mixture models for ordinal individual responses with covariates: A comparison. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 10*(1), 29–39. https://doi.org/10.1002/sam.11335.
Peugh, J., & Fan, X. (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), 204–226. https://doi.org/10.1080/10705511.2012.659618.
Piccarreta, R., & Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research, 41*(November), 100251. https://doi.org/10.1016/j.alcr.2018.10.004.
Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software, 78*(2), https://doi.org/10.18637/jss.v078.i02.
Rabe-Hesketh, S., & Skrondal, A. (2012). *GLLAMM companionMultilevel and longitudinal modeling using stata: Vol. I* (3rd ed.). College Station, TX: Stata Press.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual. U.C. Berkeley division of biostatistics working paper series*. Retrieved from https://biostats.bepress.com/ucbbiostat/paper160/.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111. https://doi.org/10.2307/271063.
Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576. https://doi.org/10.1177/0165025409343765.
Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*(1), 103–124. https://doi.org/10.1287/mksc.12.1.103.
Reinecke, J., & Seddig, D. (2011). Growth mixture models in longitudinal research. *AStA Advances in Statistical Analysis, 95*(4), 415–434. https://doi.org/10.1007/s10182-011-0171-4.
Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360.
Scrucca, L., Fop, M., Murphy, T., & Raftery, A. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289–317.
Tan, X., Dierker, L., Rose, J., & Li, R. (2011). How spacing of data collection may impact estimates of substance use trajectories. *Substance Use and Misuse, 46*(6), 758–768. https://doi.org/10.3109/10826084.2010.537731.
Tekle, F. B., Gudicha, D. W., & Vermunt, J. K. (2016). Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Advances in Data Analysis and Classification, 10*(2), 209–224. https://doi.org/10.1007/s11634-016-0251-0.
Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock, & K. M. Samuelsen (Eds.). *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.
van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GRoLTS-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling, 24*(3), 451–467. https://doi.org/10.1080/10705511.2016.1247646.
Verbeek, M. (2012). *A guide to modern econometrics* (4th ed.). Chichester: Wiley.
Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research, 23*(1), 42–59. https://doi.org/10.1177/0962280212445834.
Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax.* Belmont, MA: Statistical Innovations Inc.
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105.
Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278.
Xu, P., Peng, H., & Huang, T. (2018). Unsupervised learning of mixture regression models for longitudinal data. *Computational Statistics and Data Analysis, 125*, 44–56. https://doi.org/10.1016/j.csda.2018.03.012.