

Hidden depths

Citation for published version (APA):

van der Nest, G. (2022). *Hidden depths: robustness of modelling approaches for uncovering latent classes in longitudinal data*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20221221gn>

Document status and date:

Published: 01/01/2022

DOI:

[10.26481/dis.20221221gn](https://doi.org/10.26481/dis.20221221gn)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

HIDDEN DEPTHS

ROBUSTNESS OF MODELLING

APPROACHES FOR UNCOVERING LATENT
CLASSES IN LONGITUDINAL DATA

An abstract geometric design consisting of several rectangular blocks of different colors (red, yellow, blue, and white) arranged in a non-uniform, overlapping pattern. The blocks are separated by thin black lines. The colors are primarily red, yellow, blue, and white, with some blocks being solid and others having a white background with colored borders.

GAVIN VAN DER NEST

Hidden Depths

**Robustness of modelling approaches for
uncovering latent classes in longitudinal data**

Gavin van der Nest

© Gavin van der Nest, Maastricht 2022
Printed by: Gildeprint – The Netherlands
ISBN: 978-94-6469-094-1

Hidden Depths

Robustness of modelling approaches for uncovering latent classes in longitudinal data

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof.dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday 21 December 2022, at 16.00 hours

by

Gavin van der Nest

Supervisors:

Prof. dr. G.J.P. van Breukelen

Dr. M.J.J.M. Candel

Co-supervisor:

Dr. V. Lima Passos

Assessment Committee:

Prof. dr. S.M.G. Zwakhalen (chair)

Prof. dr. D.S. Nagin (Carnegie Mellon University)

Dr. A. Oenema

Dr. S. Vanbelle

Prof. dr. R. van de Schoot (Utrecht University)

The research presented in this thesis was conducted at CAPHRI Care and Public Health Research Institute, Department of Methodology and Statistics, of Maastricht University. CAPHRI participates in the Netherlands School of Public Health and Care Research (CaRe).

CONTENTS

1. General Introduction	1
1.1. Longitudinal finite mixture models (FMMs)	3
1.2. Motivation for the use of longitudinal FMMs	4
1.3. A historical context of FMMs	5
1.4. Some challenges in the application of longitudinal FMMs	6
1.5. Aims and objectives of this thesis	7
1.6. Outline of the thesis	8
2. An overview of mixture modelling for latent evolutions in longitudinal data	11
2.1. Introduction	13
2.2. Types of longitudinal growth models and their interrelatedness	13
2.2.1. Growth Curve Models	14
2.2.2. Longitudinal FMMs	17
2.3. Criteria for model selection	22
2.3.1. Statistical fit indices for determining K	22
2.3.2. Determining the order of the polynomials and other model considerations	30
2.3.3. Past simulation studies: results and recommendation	31
2.4. Software availability	31
2.4.1. SAS	32
2.4.2. Stata	32
2.4.3. Mplus	34
2.4.4. R and associated packages	34
2.4.5. Latent GOLD	36
2.4.6. Further remarks	36
2.5. An empirical example illustrating a strategy for fitting longitudinal mixture models (GBTM, LCGA and GMM)	36
2.5.1. General strategy	36
2.5.2. An illustration	39
2.6. Concluding remarks	45
Appendix A.	49
3. Model fit criteria curve behaviour in class enumeration	51
3.1. Introduction	53
3.2. Specification of models	54
3.2.1. Class enumeration	55
3.2.2. Class separation	56
3.2.3. Covariance misspecification	58
3.3. Methods	59
3.3.1. Design of the simulation study	59

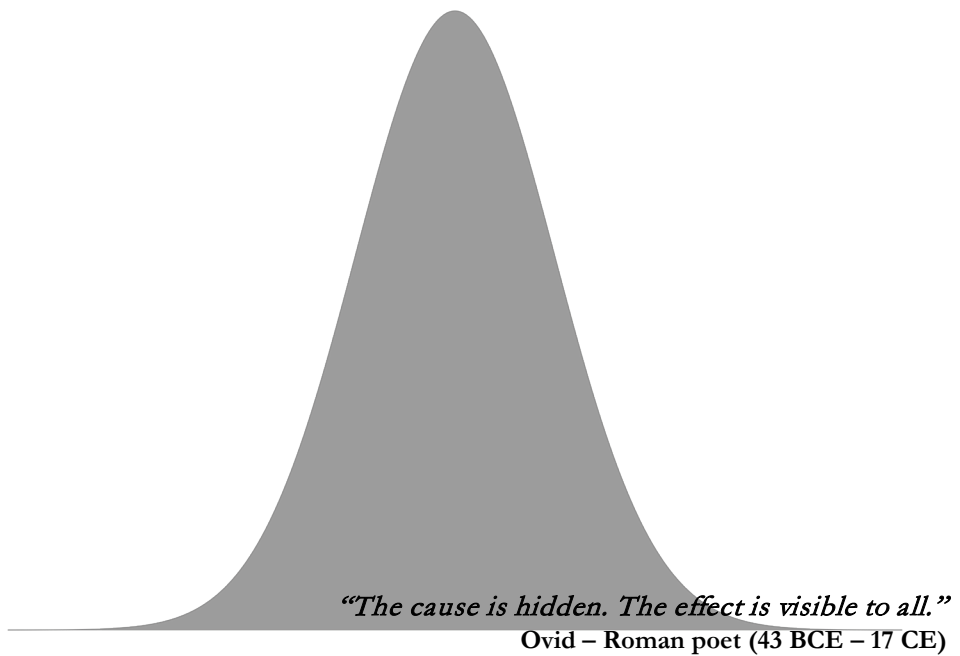
3.3.2. Simulation procedure	62
3.4. Results.....	63
3.4.1. Accuracy of class extraction in relation to design conditions and D misspecification.....	63
3.4.2. Identifiable patterns of fit statistic curves across all conditions	67
3.4.3. Unequal class sizes and small sample size	74
3.5. Application.....	74
3.6. Discussion.....	76
3.6.1. Research questions recalled.....	76
3.6.2. Fresh insights and recommendations.....	77
3.6.3. Class enumeration: Reification and validation.....	79
3.6.4. Limitations	80
3.7. Conclusion	80
Appendix B.	82
Appendix C.	88
4. Univariate versus multivariate latent trajectory modelling: a comparison of class recovery performance.....	89
4.1. Introduction.....	91
4.2. Group-based trajectory models (GBTM).....	93
4.2.1. Univariate (GBTM) trajectory model.....	93
4.2.2. Multivariate (GBMTM) multi-trajectory model	94
4.3. Methods.....	95
4.3.1. Data generation	95
4.3.2. Simulation conditions.....	96
4.3.3. Metrics for model performance evaluation.....	98
4.4. Simulation results.....	99
4.4.1. Class enumeration accuracy	100
4.4.2. The model as a classifier.....	102
4.4.3. Class recovery	106
4.4.4. Simulation results summary	111
4.5. Application.....	112
4.5.1. The dataset	112
4.5.2. Model selection	113
4.6. Discussion.....	116
4.6.1. Univariately generated data	116
4.6.2. Multivariately generated data	117
4.6.3. Guidelines for GBMTM fitting.....	117
4.6.4. The dawn of multivariate models for longitudinal data	119
4.6.5. Limitations and future research direction.....	120
Appendix D.....	122

5. Robustness of multivariate longitudinal finite mixture models to covariance misspecification	123
5.1. Introduction.....	125
5.2. Multivariate longitudinal finite mixture models	126
5.2.1. Multivariate covariance pattern growth mixture models (MCPGMM).....	128
5.2.2. Multivariate latent class growth analysis (MLCGA) and multivariate group-based trajectory models (GBMTM).....	128
5.3. Method	129
5.3.1. Data generation	129
5.3.2. Simulation conditions.....	131
5.3.3. Metrics for model performance	133
5.3.4. Technical details	134
5.4. Simulation results.....	135
5.5. Application.....	142
5.5.1. The data set.....	142
5.5.2. Model comparison	143
5.6. Discussion.....	144
5.6.1. Design condition effects on model performance.....	144
5.6.2. The effects of covariance misspecification on model performance	146
5.6.3. Practical recommendations when fitting multivariate FMMs.....	146
5.6.4. Relevance of findings	147
5.6.5. Limitations and future research direction.....	147
Appendix E.	149
Appendix F.....	153
6. General Discussion	155
6.1. Class enumeration and covariance misspecification.....	156
6.2. Nuances between univariate and multivariate models	158
6.3. The effect of data conditions.....	159
6.4. The dilemma of class separation.....	160
6.5. Ideas for future research.....	160
6.6. Concluding thoughts.....	161
6.6.1. The necessity of random effects	161
6.6.2. Reification of classes and model validation.....	162
6.6.3. Conclusion	163
7. Summary	165
7.1. Summary in English	166
7.2. Samenvatting in het Nederlands	168
8. Scientific and social impact of this thesis	171
9. References.....	177
Acknowledgements	203
About the author	206

Table of Common Abbreviations

AB	Absolute bias
ABC	Area between the curves
AIC	Akaike Information Criterion
APPA	Average posterior probability of assignment
BIC	Bayesian Information Criterion
BLRT	Bootstrap likelihood ratio test
CAIC	Consistent Akaike Information Criterion
CLC	Classification likelihood criterion
CML	Classification maximum likelihood
CPGMM	Covariance pattern growth mixture model
CVE	Cross-validation error
EM	Expectation-maximization
FMM	Finite mixture model
GBMTM	Multivariate group-based trajectory model
GBTM	Group-based trajectory model
GCM	Growth curve model
GMM	Growth mixture model
IC	Information criterion
LCGA	Latent class growth analysis
LGM	Latent growth model
LRT	Likelihood ratio test
MAD	Mean absolute deviation
MCPGMM	Multivariate covariance pattern growth mixture model
MD	Mahalanobis distance
MLCGA	Multivariate latent class growth analysis
OVL	Overlap coefficient
sE	Scaled Entropy
SM	Supplementary Material
ssBIC	Sample-size adjusted Bayesian Information Criterion
VLMR	Vuong-Lo-Mendell-Rubin test

1. General Introduction



This chapter's epigraph forms part of Ovid's musings about cause and effect. Often, it is not immediately apparent whether groups of units of observation (e.g. persons) exhibiting similar paths of temporal development exist, which could explain heterogeneity in the observed data. In the absence of a variable to distinguish between groups, only the combined effects are directly observable. This thesis deals with statistical models which can assist practitioners in uncovering such latent (hidden) profiles, and consequently gain a better understanding of dynamics within their data.

This thesis investigates, compares, and assesses the theoretical, empirical, and statistical properties of finite mixture models (FMMs) for longitudinal data i.e., for multiple sequential measures over time per subject. Longitudinal FMMs are model-based clustering methods which assume that the population studied comprises distinct, but unobserved subpopulations. These subpopulations are probabilistically defined such that clusters of subjects following similar temporal developmental profiles are identified in the absence of a known grouping variable. Such clusters are called latent (hidden) classes.

FMMs have been increasingly used in the analysis of longitudinal data. In health and medical studies, these models are useful in identifying differences in treatment response, and/or disorder or disease aetiology/development over time. Recent applications include: identifying trajectories of psychological distress during the COVID-19 pandemic and linking these to socio-demographic and health factors [1], distinguishing distinct trajectories of psychological functioning after First-Episode Psychosis and exploring their relation to several cognitive composite constructs [2], finding shared trajectories across several health outcomes in older adults [3], and uncovering trajectories of distinct eating behaviours (e.g. healthy consumption habits) throughout adolescence [4].

Notwithstanding the popularity of longitudinal FMMs in practice, their users can often be unaware of the models' underlying assumptions and the potential implications of their violations. This study seeks to elucidate some of the issues faced by practitioners, increase the accessibility of these models, and elicit a greater appreciation of these models' applicability and limitations. Various longitudinal FMMs will be compared in terms of their underlying assumptions, robustness to violations of their assumptions, applications to real-world data, and their availability in software. Further, visual and investigational statistics will be developed to address complex group-based research questions.

1.1. Longitudinal finite mixture models (FMMs)

Longitudinal FMMs are probabilistic models which combine at least two density functions to fit data and to account for heterogeneity in longitudinal processes in observed outcomes.

Longitudinal FMMs develop from the assumption that within a population, K latent classes exist with subjects within classes following similar temporal paths. Let $\mathbf{y}_i^{tr} = (y_{i0}, \dots, y_{i(T-1)})$ be a vector of repeated measures for subject i , $i = 1, \dots, N$ over time t , $t = 0, \dots, T - 1$, with superscript tr denoting vector transpose. Then, the marginal probability distribution $P(\mathbf{y}_i)$ of a randomly chosen trajectory is modelled as,

$$P(\mathbf{y}_i) = \sum_{k=1}^K \pi_k P^k(\mathbf{y}_i) \quad (1.1)$$

$P^k(\mathbf{y}_i)$ is the conditional distribution of the longitudinal sequence \mathbf{y}_i given that subject i is in class k , $k = 1, \dots, K$. $P^k(\mathbf{y}_i)$ is uniquely defined by the trajectory specification per class. π_k is the class membership probability (mixing weights) where $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$, with $K > 1$. These models assume K to be known, which is difficult to deduce directly from the data.

Various statistical and practical measures can assist in the enumeration of K , which are discussed at length in **Chapter 2**.

$P^k(\mathbf{y}_i)$ could in principle be any statistical distribution, which makes these models flexible in handling continuous, binary, ordinal, nominal, and count outcomes. For continuous data, the sole focus of this thesis, the multivariate normal density function is typically employed. Poisson may be used for count data and the binary logit for binary data [5,6]. Moreover, $P^k(\mathbf{y}_i)$ need not conform to the same distributional parameters nor even to the same density function across classes.

FMMs are categorised as model-based soft-clustering approaches. They are model-based since classes are defined by a formal statistical model (**Eq. 1.1**) with parameters estimated by conventional methods like maximum likelihood [7]. As a soft-clustering technique, each subject's sequence of measures (\mathbf{y}_i) belongs to a class k with (posterior) probability $P(k|\mathbf{y}_i)$. The individual sequence is probabilistically assigned to class k when $P(k|\mathbf{y}_i) > P(l|\mathbf{y}_i)$, $l = 1, \dots, K$, $k \neq l$ [8]. This is a stochastic assignment in the sense that uncertainty exists in the class assignment of a subject due to inter- and intra-individual variation around the class trajectory. This contrasts with hard-clustering, where \mathbf{y}_i belongs

exclusively to a single class i.e., deterministic assignment. *K-means* is a typical hard-clustering algorithm [9].

1.2. Motivation for the use of longitudinal FMMs

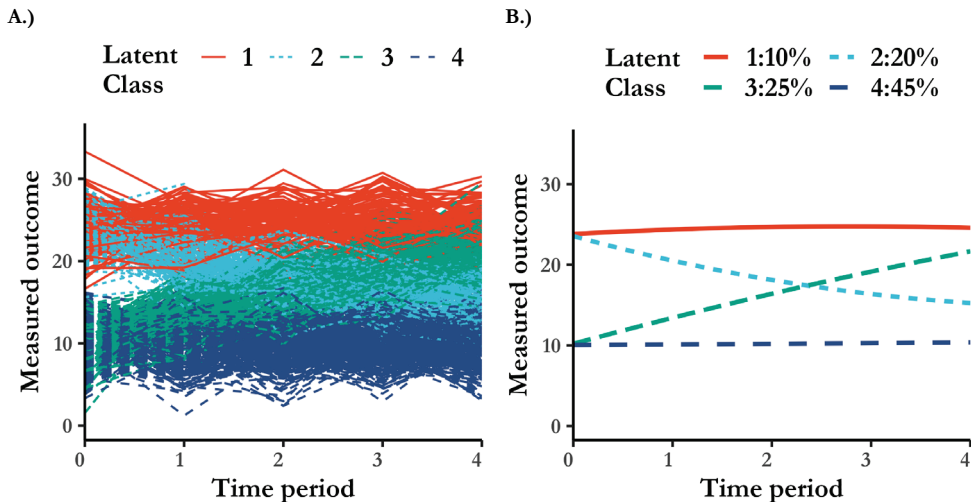
FMMs are often deployed in one of two contexts, but the divide between them is not clear-cut, and, as Titterton *et al.* [10] argue “is related more to expository convenience than to philosophical niceties”:

- *Direct applications*: the assumption is that there are K underlying classes, where \mathbf{y}_i belongs to one of these classes. K is either not directly observed or is too expensive to measure. Here, $P^k(\mathbf{y}_i)$ gives the probability distribution of \mathbf{y}_i given that it actually arises from class k , and π_k is the probability of \mathbf{y}_i emanating from class k . This application is the focus of this thesis.
- *Indirect applications*: The FMM is simply utilized as a mathematical device which allows for a tractable form of analysis [10], such as approximating any arbitrary continuous distribution using a mixture of normal densities [11]. This makes FMMs a powerful instrument for the modelling of multimodal, skewed, or asymmetrical data.

In practice, when working with longitudinal data, one is often faced with many individual trajectories (**Figure 1.1: Panel A**). As an example, such data could comprise insulin levels over time in response to treatment for patients. If one is interested to ascertain whether such data exhibit different patterns of temporal development for subsets of individuals, then the application of longitudinal FMMs may yield latent classes of subjects following distinct temporal development (**Figure 1.1: Panel B**). This could be advantageous when considering therapy tailored for specific patient profiles in the age of precision medicine or seeking to understand the potential drivers underlying the heterogeneity of treatment response.

FMMs’ flexibility allows them to be employed in a variety of fields to classify subjects or other observational units, to account for clustering, and to model unobserved heterogeneity. Moreover, FMMs’ ability to identify hidden (process) heterogeneity is an essential component in the analysis of clinical and epidemiological data [12], where patients with different risks and/or responses to medical therapies can be identified.

Figure 1.1: **Panel A:** Spaghetti plot of individual trajectories for the measured outcome (e.g. insulin levels over time), coloured by latent (unknown) class. **Panel B:** Size of latent classes and mean trajectory per latent class uncovered by applying longitudinal FMM.



Longitudinal FMMs are also valuable in examining the various aspects of disease progression by considering multiple outcomes simultaneously [13]. This may be particularly useful in establishing the dynamics of multifaceted diseases such as multiple sclerosis or Crohn’s disease. Further, as an example, the inclusion of multiple outcomes in clinical trials allows one to establish the existence of temporally heterogeneous effects (i.e. distinct patterns of responsiveness between classes) as well as whether treatment effect differs per outcome.

1.3. A historical context of FMMs

FMMs have a long history in modern statistical modelling. Among the earliest known implementations include Newcomb’s [14] use of normal mixtures to model outliers, and the work of Pearson [15] in which the method of moments was applied to fit normal mixtures to accommodate asymmetry in data. However, the method of moments approach presented a laborious challenge for Pearson (and statisticians until the advent of the computer age) who ultimately succeeded in solving a ninth-degree polynomial for a two-component mixture. Rao [16] was the first to suggest maximum likelihood estimation for normal mixtures. Wolfe [17] was the first to apply maximum likelihood estimation to mixtures of multivariate normal densities. The seminal work of Dempster *et al.* [18] formulated the expectation-maximization (EM) algorithm in general terms for the modelling of heterogeneous data by FMMs. Since then, the EM algorithm has become ubiquitous in mixture modelling applications and

literature.

Amongst social scientists, Nagin and his collaborators were the first to suggest longitudinal FMMs for the clustering of trajectories [19–21]. Their objective was to evaluate, amongst others, Moffitt's [22] theory of the development of antisocial and criminal behaviour. Here, two distinct classes of delinquent individuals were proposed: a small group engaging in antisocial behaviour at every life stage, and a large group exhibiting antisocial behaviour only during adolescence.

Following this, two separate but related methodologies of longitudinal FMMs for general use were presented [7], group-based trajectory models (GBTM) [23] and growth mixture models (GMM) [24]. Each methodology was supported by software which allowed for general dissemination, *PROC TRAJ* in *SAS* for GBTM [25] and *Mplus* for GMM [26]. The distinction between these methods is in the treatment of classes. GBTM assumes that within a latent class, all subjects follow similar within-class paths of temporal development with any deviation from the class average trend treated as intra-individual random variation. GMM assumes latent classes to contain a within-class heterogeneous set of individuals described by a probability distribution. In other words, with GBTM, differences between individual trajectories are captured by latent classes as between-class differences through the mean trajectory, and for GMM, individual differences in temporal development are split into between-class differences (the mean trajectory) and a within-class component (inter-individual variability captured through random effects components) [7]. Finally, covariance pattern growth mixture models (CPGMM) have been proposed as an alternative to GBTM and GMM. CPGMM relaxes GBTMs' conditional independence assumption and unlike GMM, models the temporal structure and association between observations within classes without variance partitioning or the inclusion of subject-specific random effects [27].

1.4. Some challenges in the application of longitudinal FMMs

Stemming from the comparatively easy accessibility of these models there are some drawbacks to their use. The relative novelty of these models combined with the fact that model fitting and validation requires an array of (often heuristic and/or *ad hoc*) decision calls between a series of lengthy iterative steps raise doubts and uncertainty on the part of the applied user. There is no one-size-fits-all approach to fitting FMMs.

Class enumeration (determining the number of latent classes (K)) remains a difficult task. Statistical fit indices assist in the choice of K , but they all have inherent weaknesses since their accuracy in determining the true K is highly dependent on the underlying data features. Thus, the question of which fit statistic is most valid remains unresolved [11].

The estimation of parameters for the covariance structure of longitudinal FMMs is challenging [28]. If a too general structure is chosen, then the model may suffer from multiple complications such as non-convergence and improper solutions. The latter includes models terminating at a local optimum on the likelihood surface, which may then fail to adequately capture the underlying class structure of the data. To aid model convergence, various constraints are often imposed on the model which may not be realistic or ideal. On the other hand, choosing a too restrictive covariance structure could also have unfavourable consequences, such as influencing: the shape and level of the growth trajectories in each class [29,30], the number of classes extracted [31–33] and the quality of class assignment [34].

To address some of these challenges, this thesis provides an exposition of popular longitudinal FMMs and offers some practical guidance in model selection strategies i.e., under what circumstances the available models (GBTM, latent class growth analysis (LCGA), GMM, or others) should be recommended. It will address questions about class enumeration and the behaviour of various fit-criteria used for this purpose. Further, the thesis will investigate what factors influence class extraction and how this is affected by violations of the model's underlying assumptions.

Finally, interpretational issues around the meaning of the identified latent classes do exist, particularly with how to adequately represent and interpret the data-driven taxonomic categories, i.e. as true entities (reification fallacy caution in exploratory studies [35]) or as properties of the data [36]. Moreover, statistical validation of the extracted classes, subsequent or possibly concurrent to the model selection, remains an issue, since this may greatly affect final findings and substantive conclusions. Although of paramount importance, this thesis will ultimately not focus on interpretational or validation issues.

1.5. Aims and objectives of this thesis

Given the value and application of model-based clustering of longitudinal data, this thesis aims to:

- (1) Identify and produce an inventory of the main conceptual, theoretical, and statistical issues of the methods of model-based clustering for longitudinal data. This includes criteria for the comparison between different methods, the pros and cons of the methods, what researchers should be cognizant of when using each method, and the issue of the fallibility of statistical criteria in class enumeration,
- (2) Explore and investigate the statistical and empirical properties of longitudinal FMM models using real and simulated datasets, first with univariate data but then focusing on multivariate settings which are largely unexplored in the literature. This includes their application and estimation, underlying assumptions, robustness to violation of assumptions, limitations in their use, as well as possible further development and refinement,
- (3) Developing accessible codes and data visualisations to facilitate understanding and increase the accessibility of these modern and complex techniques to applied users. These are provided in R and *Mplus* formats,
- (4) Ultimately, we seek to provide an accessible guide for practitioners to these methods.

1.6. Outline of the thesis

Chapters 2-5 of this thesis can be read as self-contained articles. Each chapter covers a separate topic.

Chapter 2 provides an inventory of the main conceptual, theoretical, and statistical issues pertaining to the model-based clustering of longitudinal data. It develops from an exposition of growth curve modelling to cover longitudinal FMMs with a specific focus on GBTM, LCGA, and GMM. Further, criteria specifically for the enumeration of classes are extensively discussed, including their strengths and weaknesses, when and when not they should be used, and the fallibility of statistical criteria in class enumeration. Statistical software for longitudinal FMMs is discussed and a model selection strategy is provided. This chapter serves as a self-contained introduction to longitudinal FMMs and strives to address the confusion practitioners often face when confronted with model terminology, class enumeration, model selection, and software options.

Chapter 3 considers how statistical fit-criteria behaviour, specifically that of the Bayesian Information Criterion (BIC), sample-size adjusted BIC (ssBIC), Akaike Information Criterion (AIC) and scaled Entropy (sE), in terms of their fit-criteria plot could ultimately

guide the class enumeration process in identifying possible model misspecification. Models considered here include GBTM, LCGA, and GMM. The misspecification considered is that of within-class inter-individual covariances i.e., the presence or absence of random effects. This is an important question as practitioners often apply models without considering such variation (by using software defaults or constraining covariances to improve convergence) and we show the consequences thereof. The specific fit-criteria curve behaviour identified allows for recommendations to be offered for determining which models to apply and which fit statistic to use. Finally, some practical tools are presented to practitioners to assist in their model building routines.

Chapter 4 is a comparative study of model performance between univariate GBTM and multivariate GBTM (GBMTM) when applied to univariately or multivariately generated data. We show how certain data conditions often encountered in practice (e.g. class separation) and violations of the assumptions of the underlying model drive differences in terms of class enumeration accuracy, subject classification (i.e. is subject assigned to the correct class), and class recovery (i.e. are the actual trajectory profiles and class sizes recovered). We consider whether the two-stage process of first fitting GBTM and then GBMTM is advised whilst highlighting potential differences in these models' extraction. Predicated on these findings, guidelines for GBMTM fitting, including practical recommendations are provided to assist practitioners in their research. As no previous research has comparatively evaluated the performance of GBTM and GBMTM, this study provides useful research-based information to assist the field in better understanding these two different modelling approaches.

Chapter 5 considers the implication of covariance misspecification in the presence of between-outcome correlation for multivariate longitudinal FMMs. Misspecification in the within-class within-outcome correlation and the time-dependency of the variance are considered. Models covered include GBMTM, multivariate LCGA (MLCGA), and multivariate covariance pattern growth mixture models (MCPGMM). This is a comparative study, examining the effect of misspecification under various data conditions on the class enumeration, subject classification, and class recovery performance of these models.

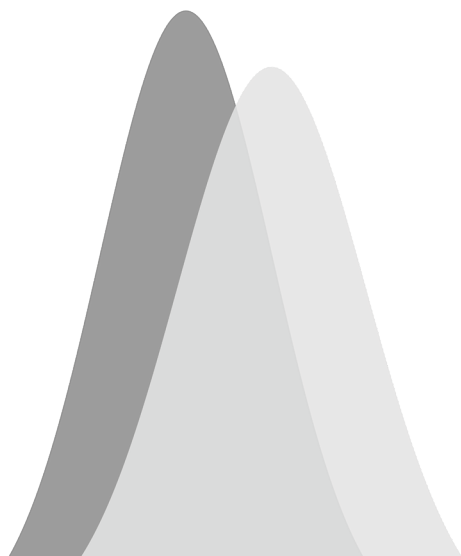
Chapter 6 reflects on research developed from the preceding four chapters. The implications of the research are discussed along with their correspondence to the practical application of longitudinal FMMs. Some thoughts on possible future research avenues are also presented. **Chapter 7** is a summary of each of the chapters in the thesis whilst **Chapter 8** provides an overview of the scientific and societal impacts of this research.

For the sake of parsimony, the online supplementary materials for **Chapters 2-5** are not included in this thesis but these are available upon request. Where relevant, a table of contents giving further details of the supplementary material is provided at the end of each chapter.

2. An overview of mixture modelling for latent evolutions in longitudinal data

Modelling approaches, fit statistics and software

Gavin van der Nest
Valéria Lima Passos
Math J.J.M. Candel
Gerard J.P. van Breukelen



Published as:

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, 43, 100323. — <https://doi.org/10.1016/j.alcr.2019.100323>

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.

Keywords: growth mixture model; latent class growth analysis; trajectory; hidden heterogeneity; repeated measures; classification

2.1. 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 [37]. 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) [38] 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 [39] over their life course. Some recent applications include uncovering distinct trajectories of treatment response for adults with obsessive-compulsive disorder [40], disparate patterns of change over time in terms of criminogenic risks of juvenile offenders [41], divergent general psychopathology trajectories and their link to social outcomes [42], examining group differences in the link between alcohol consumption evolution and cardiovascular events [43], and relating distinctive cannabis use patterns among adolescents to life satisfaction, academic achievement and other psychoactive substance usage [44].

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 strategy for fitting these models.

2.2. Types of longitudinal growth models and their interrelatedness

There are several model-based techniques for analysing outcome development over time [45], 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 [46,47], which are typically represented as time trends, time paths, growth curves or latent trajectories [46].

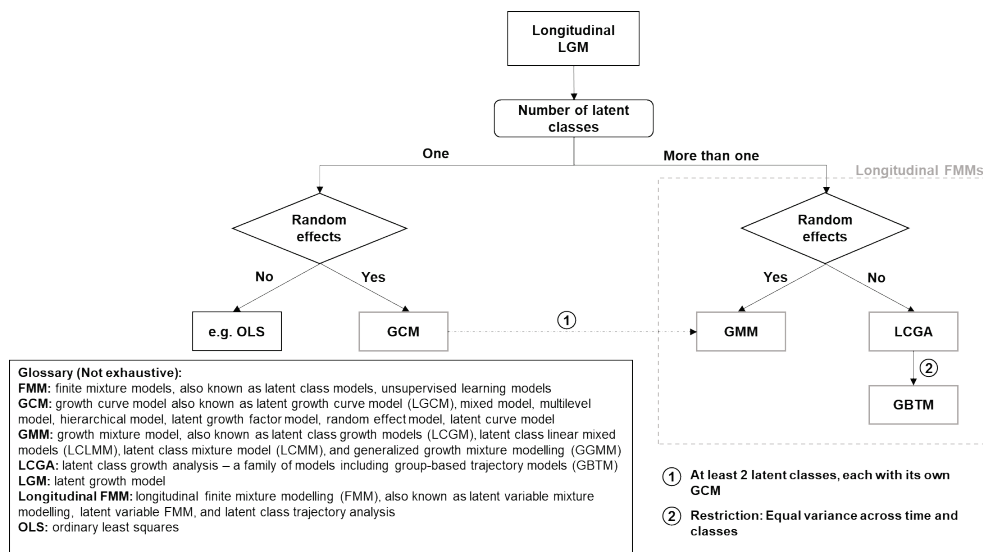
Within the family of longitudinal LGM models, the same model may be termed differently. This often creates confusion in literature and in practice, specifically concerning model commonalities and applications. For clarification, a partial glossary of these terms and their aliases are contained in **Figure 2.1**. As **Figure 2.1** 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 [24] (GMM), latent class growth analysis [48] (LCGA), and group-based trajectory models [6] (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 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.

2.2.1. Growth Curve Models

The single-class GCM models are not concerned with categorising subjects, but rather with modelling the relationship between explanatory variables and the development of a repeatedly measured outcome [49]. 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) [50,51].

Figure 2.1: Interrelatedness of longitudinal LGM models.



The equation for the single trajectory of a GCM in scalar form is presented in Eq. 2.1 in Table 2.1. A matrix formulation is provided in the **Supplementary Material (SM)**. In the scalar formulation, y_{it} is the measured outcome for subject i at time $t = 1, \dots, T$, 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 Figure 2.2 (a)) [43]. 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 Figure 2.2 (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 average trend in Figure 2.2 (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.

Table 2.1: Model trajectory specification [30,53].

General:		
$k = 1, \dots, K$ is the class		
$t = 1, \dots, T$ is the time point		
$i = 1, \dots, n$ is the subject		
X_{it} = predictor value of subject i at time point t		
Model	Trajectory Specification	Assumptions
GCM	$y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})X_{it} + (\beta_2 + b_{2i})X_{it}^2 + \epsilon_{it}$ (2.1)	$b_{ji} \sim N(0, \sigma_{b_j}^2), j = 0,1,2$ $\epsilon_{it} \sim N(0, \sigma_{\epsilon_t}^2)$ $cov(b_{ji}, b_{hi}) \neq 0, j \neq h, h = 0,1,2$
LCGA	$y_{it}^k = \beta_0^k + \beta_1^k X_{it} + \beta_2^k X_{it}^2 + \epsilon_{it}^k$ (2.2)	$\epsilon_{it}^k \sim N(0, \sigma_{\epsilon_{kt}}^2)$
GMM	$y_{it}^k = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2 + \epsilon_{it}^k$ (2.3)	$b_{ji}^k \sim N(0, \sigma_{b_j^k}^2), j = 0,1,2$ $\epsilon_{it}^k \sim N(0, \sigma_{\epsilon_{kt}}^2)$ $cov(b_{ji}^k, b_{hi}^k) \neq 0, j \neq h, h = 0,1,2$

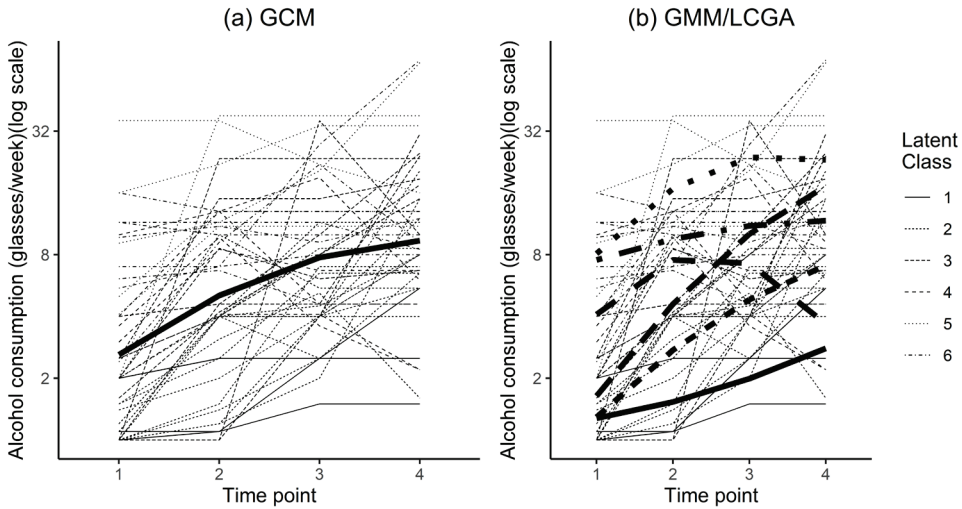
A GCM may be extended to examine differences in outcome development between known subgroups, for instance between males and females. As an example, a GCM may differentiate linear trends between sexes by adding sex and a sex by time interaction term to the model,

$$y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 sex_i + \beta_3 sex_i X_{it} + b_{0i} + b_{1i} X_{it} + \epsilon_{it} \quad (2.4)$$

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 $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{it}$ 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 [50]. Since GCM is

not designed to uncover latent classes, it will not be considered further in the remainder of this study.

Figure 2.2: 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).



2.2.2. 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) [11]. These models assume that the population under study is composed of distinct, latent subgroups or classes [54]. 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, a 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 2.3**).

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 [55]. 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 & Nestler (2019), FMMs provide for the *post hoc* identification and description of class differences in change [50]. 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 [46,47].

Typical longitudinal FMM models include; growth mixture modelling [24] (GMM), latent class growth analysis [6] (LCGA), mixture latent transition analysis (LTA) [57] (also known as mixture hidden Markov models), and survival mixture analysis (SMA) [58] amongst others [47,48]. 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 [59]. 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 [51,60,61]. 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 *Latent GOLD*. 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.

2.2.2.1. Latent class growth analysis (LCGA)

Equation 2.2 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 [62]. This is represented by the bold lines in **Figure 2.2 (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 (**Figure 2.2 (b)**). 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 [6,19,p.337].

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 [63].

2.2.2.2. Growth mixture models (GMM)

The class-specific trajectory for a GMM is represented in **Eq. 2.3**, which is an amalgamation of **Eqs. 2.1** and **2.2**. 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 **Figure 2.2 (b)**. Random effects are used to capture individual differences in trajectories within a class [51], 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 **Figure 2.2 (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.

2.2.2.3. More general formulation of LCGA and GMM

As longitudinal FMMs, 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 [6],

$$P(\mathbf{y}_i) = \sum_{k=1}^K \pi_k P^k(\mathbf{y}_i), \quad (2.5)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})^{tr}$ is the column vector of measured outcomes (e.g. alcohol consumption) for subject $i, i = 1, \dots, n$, at time $t = 1, \dots, T$, and $P^k(\mathbf{y}_i)$ is the conditional distribution of the longitudinal sequence, \mathbf{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 $K > 1$. **Equation 2.5** shows that $P(\mathbf{y}_i)$ 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 $P^k(\mathbf{y}_i)$ on the right side of **Eq. 2.5**) with the class membership probabilities (the π_k on the right side of **Eq. 2.5**) allows the mixture model to approximate any arbitrary marginal distribution (the $P(\mathbf{y}_i)$ on the left side of **Eq. 2.5**). It is this property which makes FMMs a powerful and flexible tool for the modelling of complex data [11], such as highly asymmetrical and multimodal data.

In the longitudinal context, with a repeatedly measured continuous outcome, $P^k(\mathbf{y}_i)$ could be the multivariate normal (MVN) density function (in line with the model assumptions in **Eqs. 2.1** and **2.3** in **Table 2.1**). So, for subject i in class k ,

$$\mathbf{y}_i^k \sim MVN(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k) \quad (2.6)$$

with \mathbf{y}_i^k the vector of T successive 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 [64]. For this model, the GCM of **Eq. 2.1** is assumed to hold per class.

In GMM models, the $\boldsymbol{\Sigma}^k$ 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 [30]. In GBTM and LCGA, the \mathbf{D}^k matrix is zero (so that $\boldsymbol{\Sigma}^k = \mathbf{R}^k$) 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 [6,19,p.337]. Possible specifications for the $\mathbf{\Sigma}^k$ and \mathbf{R}^k matrices are presented in the SM.

2.2.2.4. Extension beyond continuous outcomes and a polynomial trend

The MVN assumption on \mathbf{y}_i in Eq 2.6 may be relaxed to accommodate outcomes that are not continuous. $P^k(\mathbf{y}_i)$ 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 [5] and LCGA [6] studies.

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

$$p^k(y_{it}) = \frac{\lambda_{kt}^{y_{it}} e^{-\lambda_{kt}}}{y_{it}!}. \quad (2.7)$$

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:

$$p^k(y_{it} = 1) = \frac{\exp((\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2)}{1 + \exp((\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)X_{it}^2)}. \quad (2.8)$$

An alternative for binary data is the probit model, which is almost empirically indistinguishable from the logit model [65]. However, the logit is often chosen due to having a closed-form equation [6].

An alternative approach to modelling the outcome trend as a polynomial function of time is piecewise regression (see [66]) 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:

$$y_{it}^k = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)X_{it} + (\beta_2^k + b_{2i}^k)(X_{it} - c_1)D_1 + (\beta_3^k + b_{3i}^k)(X_{it} - c_2)D_2 + \epsilon_{it}^k. \quad (2.9)$$

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. 2.9 where $X_{it} = c_1$ and c_2) in which the relationship between the predictor (time) and the measured outcome (alcohol consumption) abruptly changes.

2.3. 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 [6]. 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 to test models against each other cannot be conducted. Nonetheless, a plethora of fit indices, including modified likelihood ratio tests, 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 [11,p.175,67]. Finally, these model fit statistics may be used in conjunction with Wald tests and likelihood ratio tests to determine the final polynomial order.

2.3.1. 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 $P^k(\mathbf{y}_i)$ in Eq. 2.5, and all their associated class probabilities (mixing weights), π_k , are non-zero [11,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 2.2** [68–70] presents an overview of the most frequently cited of these fit indices which will be discussed in detail.

Table 2.2: 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	$\frac{1}{n} \sum_{i=1}^n \log\left(\frac{\hat{p}(y_i m(K_1))}{\hat{p}(y_i m(K_1-1))}\right)^2$	$H_0: K = K_1 - 1$ $H_1: K = K_1$ 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
Distributional statistics	ICL-BIC	$ICL-BIC = -2 \log L(K) + \log(n)m(K) + 2E(K)$	Smallest
	MVS MVK		$H_0: K$ class model $H_1: \text{Not } K$ class model

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

2.3.1.1. Log-likelihood criteria

The log-likelihood information criteria (LLIC) statistics have the general form [71],

$$-2 \log[L(K)] + a(n)m(K) \quad (2.10)$$

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. 2.10** 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 [69]. 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 [68].

The Bayes Factor [72] (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 [11,p.210,73],

$$BF_{10} = \frac{P(\mathbf{y}|K_1)}{P(\mathbf{y}|K_0)} \quad (2.11)$$

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 $BF_{10} = \exp(\Delta BIC_{01} / 2)$ [73,p.796,804,74–76], 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 [77].

The Vuong-Lo-Mendell-Rubin test (VLMR) is a modified likelihood ratio test (LRT) [78]. 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 [79]. 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 [78]. Moreover, the VLMR and aLMR have not escaped scrutiny as their original proof has been shown to contain mathematical errors [80]. Nevertheless, they appear to work well in detecting homoscedastic (equal variance across classes) normal mixtures [78].

The bootstrap likelihood ratio test (BLRT) is a parametric bootstrap alternative approach to estimate the distribution of the LRT statistic [81]. The BLRT addresses distributional issues with the LRT [11,p.186] which has no closed-form distribution under mixture models [80] 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 [67]. However, studies show that with complex growth trajectory shapes and large sample size conditions, the BLRT tends to outperform other likelihood-based enumeration indexes [67,82], 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 [67].

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

2.3.1.2. Classification-based criteria

Classification statistics based on the classification maximum likelihood are complementary to the log-likelihood statistics [83]. 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,

$$E(K) = - \sum_{k=1}^K \sum_{i=1}^n pp_{ik} \log[pp_{ik}] \geq 0 \quad (2.12)$$

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,

$$pp_{ik} = P(k|y_i) = \frac{\widehat{\pi}_k \widehat{P}^k(y_i)}{\sum_{h=1}^K \widehat{\pi}_h \widehat{P}^h(y_i)} \quad (2.13)$$

where $\widehat{P}^k(y_i)$ is the estimated probability of observing the data if i is a member of class k . $\widehat{\pi}_k$ is the estimated proportion of the population in class k [6] (where class membership follows a multinomial distribution) with the constraint that $\sum_{k=1}^K \widehat{\pi}_k = 1$.

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 [83].

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 [83]. 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 [84]. 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 2.2**) 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 [55] 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 [85], the VLMR showed good performance under conditions of low Entropy.

The average posterior probability of assignment (APPA) [86] and the odds of correct classification (OCC) [6] 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 $\widehat{\pi}_k$ represents the probability of a randomly selected individual belonging to class k) [87]. 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 [6,69]. OCC close to 1 is indicative of the maximum probability assignment rule having predictive power not beyond random chance [6].

2.3.1.3. Likelihood-Classification Hybrids

The classification likelihood criterion (CLC) incorporates $E(K)$ as a classification uncertainty penalty term in LLIC [88] (see **Section 2.3.1.1**). The objective is to choose a K which minimises the CLC [11]. 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

[11,p.214].

The integrated classification likelihood (BIC approximation) (ICL-BIC) was developed to address shortcomings in the BIC and CLC [89]. 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).

2.3.1.4. 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 [90]. 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 non-normal population from a mixture of multiple normal populations [68,90]. However, more research is required to determine the viability of this approach.

2.3.1.5. 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) [91]. If the model predicts well, then it is seen as a good and appropriate model [92].

Cross-validation error (CVE) [92] is a measure of the predictive accuracy of a fitted model. The CVE for individual i is measured as,

$$CVE(i) = \frac{1}{T} \sum_{t=1}^T (y_{it} - \hat{y}_{it}^{[-i]})^2 \quad (2.14)$$

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,

$$CVE = \frac{1}{n} \sum_{i=1}^n CVE(i) \quad (2.15)$$

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 [92].

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 $P_1 \cup \dots \cup P_M = \{1, \dots, n\}$. Then for each $m = 1, \dots, M$, 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,

$$CVE(m) = \frac{1}{n_m} \frac{1}{T} \sum_{i \in P_m} \sum_{t=1}^T (y_{it} - \hat{y}_{it}^{[-m]})^2 \quad (2.16)$$

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 ,

$$CVE = \frac{1}{M} \sum_{m=1}^M CVE(m) \quad (2.17)$$

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 minimises the CVE. A recent study [93] 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 [94].

2.3.2. 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 [6,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 [6,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 non-significant 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 [72,73,95].

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 [55]. 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 [30].

Several studies [30,32,79] 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 [30].

2.3.3. 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 [67]. In general, practitioners often employ the least computationally intensive statistics (or the most familiar) in determining an appropriate solution [68].

The relative performance of a selection of these fit indices has been compared in a variety of simulation studies [55,67,78,79,85,90,96–98]. 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 [85]. 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 [6,94,99]. 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 summarise distinctive features of the data in as parsimonious and as sensible a manner as possible [6,p.77].

2.4. 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 (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 summarised in **Table 2.3**. 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 the **SM**.

2.4.1. SAS

Proc traj [25] 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 estimation following the general quasi-Newton procedure. *Proc traj* can handle multivariate outcome models [13].

Proc NLMIXED is another *SAS* procedure which allows for multiple classes, the inclusion of random effects and a variety of link functions [100]. 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 [100,101].

2.4.2. Stata

Traj [102] 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 [103].

Gllamm [104] is capable of handling more complex longitudinal FMMs, including random effects [105]. 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 [106]. The

Table 2.3: Features of popular software for longitudinal FMM (as at May 2019) (*Only those mentioned in Section 2.3.1 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)	X	Traj: X 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, Zero inflated Poisson, Negative binomial
Trajectory specification					
Random effects	Censored normal only	Traj: X GLLAMM: ✓ Traj: No random effects	✓	✓	✓
Covariance structure of random effects (D matrix)	Censored normal: Equal between classes	GLLAMM: Covariance structure may be specified by user Traj: Fixed to be the same across classes and time	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	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	X	Traj: X 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

only default model fit criteria output of *gllamm* is the log-likelihood but the likelihood ratio test, AIC and BIC are easily computed by other procedures using *gllamm*'s exported log-likelihood.

2.4.3. Mplus

Mplus [26], built upon the structural equation modelling framework, is often cited in latent trajectory studies [63,92,99,107]. 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 [26]. 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 [26,107,108] and is the only program of the five considered here which provides the MVS and MVK tests (**Table 2.2**). 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.

2.4.4. 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 [109–113].

The *LCMM* package [110] can accommodate most outcomes (but excludes count responses) using non-linear 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 non-linear 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 [114,115] 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 [116]. *OpenMX* provides support for modelling autocorrelation. The AIC, BIC, sample-size corrected AIC, and ssBIC [115,p.315] are part of the default output and the LRT may be requested.

Flexmix [117] is capable of estimating longitudinal FMMs. It has support for normal, binomial and Poisson link functions [111]. 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 [109] may also be used in longitudinal FMM estimation [30], 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 [118] has the capacity to estimate longitudinal FMM for both parametric and semiparametric 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.

2.4.5. Latent GOLD

Latent GOLD [119] has the capacity to model trajectory specifications which differ between classes and the use of B-splines instead of polynomials [120]. 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, CLC, and ICL-BIC.

2.4.6. 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 2.2**, 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 [87] (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.

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

2.5.1. 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 [99] 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 [99]. Furthermore, a framework by Lennon *et al.* [121] 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.* [121] and van de Schoot *et al.* [99], which cover this extensively.

Figure 2.3 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 [50,121]. From the figure, it is apparent that several decisions need to be made when selecting the best fitting model.

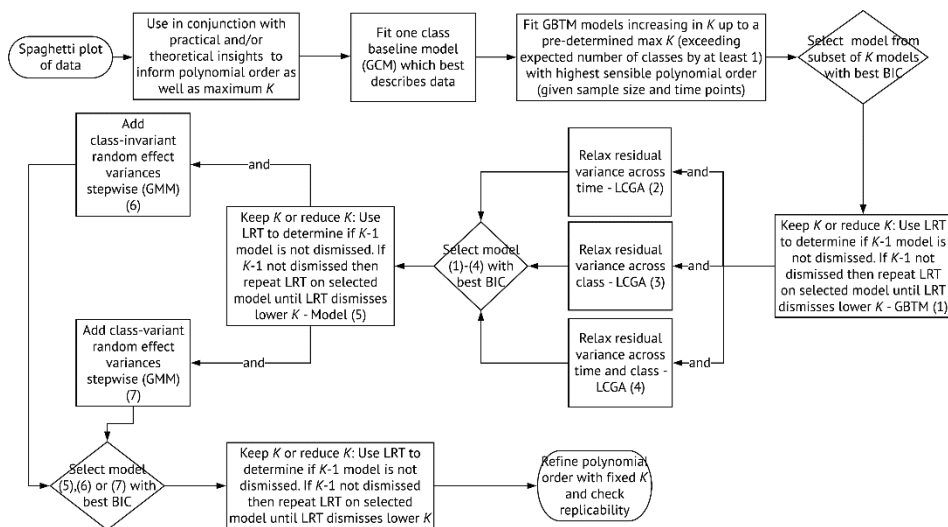
It is good practice to first plot a random selection of subjects to provide a visual representation of whether enough heterogeneity of development is evident in the data to justify the use of mixture modelling (spaghetti plot). One then needs to select a maximum K and polynomial order given the number 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 [121]. Ideally, this should be confirmed by looking at other available fit statistics which are discussed in **Section 2.3**.

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 [50].

Figure 2.3: Model selection flowchart.



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 non-significant iteratively per

class using a Wald test. Lower order non-significant 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).

2.5.2. An illustration

2.5.2.1. The dataset

We expanded upon the methodology of a GBTM study [43] 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 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 [43]. 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 the **SM** 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:

$$AC_{it}^{*k} = (\beta_0^k + b_{0i}^k) + (\beta_1^k + b_{1i}^k)time_{it} + (\beta_2^k + b_{2i}^k)time_{it}^2 + \epsilon_{it}^k \quad (2.18)$$

where $i = 1, \dots, n$, $t = 1, 2, 3, 4$, $k = 1, \dots, K$, $time$ 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. 2.3**. However, as said in **Section 2.5.1**, we start the modelling with the GCM and the GBTM (which are both special cases of the GMM) for reasons explained there.

2.5.2.2. Model selection

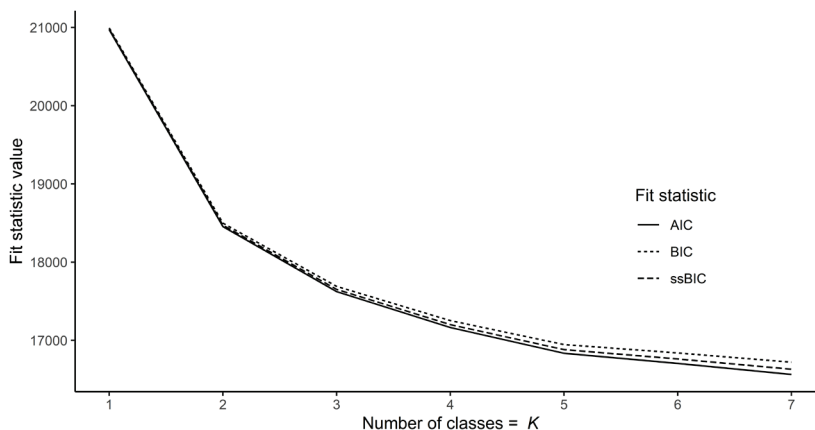
We used *Mplus* v7.3 for our analysis with some selected code provided in the **SM**. We first fitted the best one-class model (GCM) to the data with which we will compare subsequent models to justify multiple class solutions. Using the BIC as an aid and considering a variety of **D** and **R** specifications, 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 $AC_{it}^* = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})time_{it} + (\beta_2 + b_{2i})time_{it}^2 + \epsilon_{it}$ with constant residual variance over time (i.e. $\epsilon_{it} \sim N(0, \sigma^2)$). 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 [43]) plus 1. For the GBTM, the random effects in **Eq. 2.18** were set to zero (i.e. $\mathbf{D}^k = \mathbf{0}$) with the restriction of equal residual variances across time and classes (i.e. $\Sigma^k = \sigma^2 \mathbf{I}$ 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 **Figure 2.4**). The improvement in these fit statistics with an increase in K is a known issue [122] and may motivate model extension i.e. freer estimation of \mathbf{D} and \mathbf{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 (**Figure 2.4**)). These results are reported in **Table 2.4** for further analysis.

Figure 2.4: Fit-criteria performance GBTM.



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 [67]. 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 2.4**, 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 **Figure 2.5**.

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.

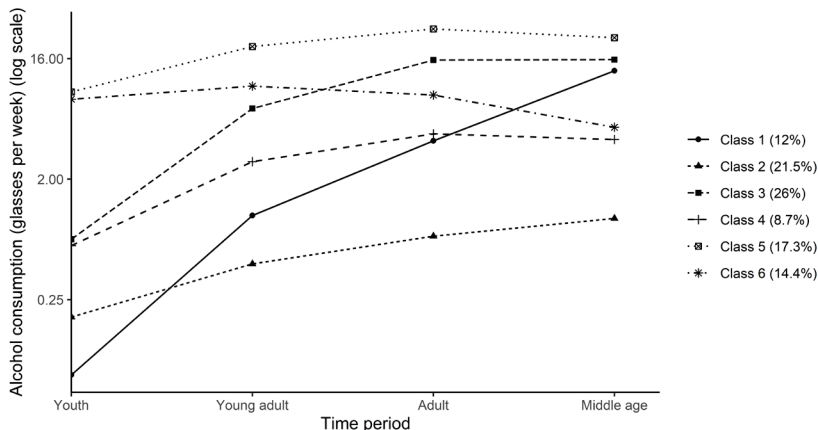
Table 2.4: Fit statistics for considered GBTM.

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

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-variant random intercept variance (GMM2) (see **Table 2.5**). 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 2.5**).

Figure 2.5: Estimated trajectories of 6-class GBTM model.

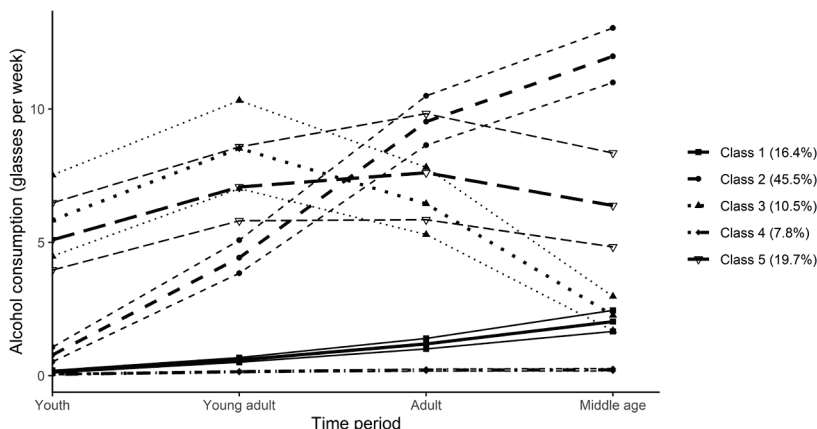


Next, the significance of the various polynomial fixed effects terms of the selected model were checked. Discarding non-significant higher-order polynomial terms led to a marginally better model fit (GMM3) and we settled on this as our final solution. Its estimated trajectories with confidence intervals (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 **Figure 2.6**. The estimated mean equations are given in the **SM**.

Table 2.5: 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

Figure 2.6: Final 5-class GMM estimated trajectories and 95% confidence intervals.



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 pre-defined 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 \mathbf{D} and \mathbf{R} specifications. Practitioners should be guided by statistical criteria as outlined in **Section 2.3** 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.

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 [86,123,124]. This may assist in understanding the processes generating the heterogeneity of developmental paths and their potential implications.

2.6. 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 2.3** 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 [13,45,125,126] and from addressing questions related to important steps subsequent (or possibly concurrent) to model selection, referred to as model validation [123,124]. Readers should consult a recent overview article [123] 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) [127], a small number of time points (<4) [127], 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 [54]. 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 [128]. Sample size studies for GMM are rather limited, but a simulation study [129] 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 [130] 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 [127] 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 [131]. 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 [131].

Violations of the assumptions of the underlying conditional distribution of the longitudinal sequence [132] (see **Eq. 2.5**) have been shown to lead to class over-extraction when using penalized likelihood criteria [36,133,134] as discussed in **Section 2.3**. 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 [133,134]. Moreover, researchers are particularly cautioned to be careful in the specification of the \mathbf{D}^k and \mathbf{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 [32].

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 [55,64,135] 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) [55]. 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 [85], 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 [90] 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 [58,61]) but our hope is that this paper will serve as an introductory guide to the discussed methods for applied studies.

Appendix A.

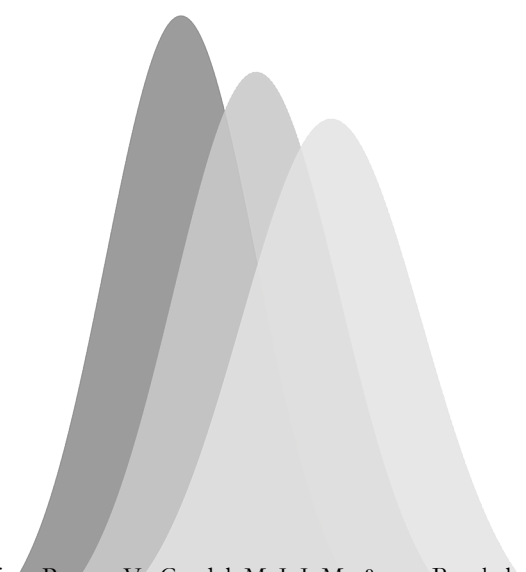
Table of Contents of Supplementary Material

- S.1 Matrix specification of model trajectories
- S.2 Residual error constraint specifications
- S.3 Expanded features of software packages
- S.4 Additional application output
- S.5 *Mplus* exemplar code
- S.6 References

3. Model fit criteria curve behaviour in class enumeration

*A diagnostic tool for model
(mis)specification in longitudinal mixture
modelling*

Gavin van der Nest
Valéria Lima Passos
Math J.J.M. Candel
Gerard J.P. van Breukelen



Published as:

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2021). Model fit criteria curve behaviour in class enumeration – a diagnostic tool for model (mis)specification in longitudinal mixture modelling. *Journal of Statistical Computation and Simulation*, 1–33. <https://doi.org/10.1080/00949655.2021.2004141>

Abstract

The use of longitudinal finite mixture models (FMMs) to identify latent classes of individuals following similar paths of temporal development is gaining traction in applied research. However, FMM's users may be unaware of how data features as well as the inappropriate specification of the model's covariance structure impacts class enumeration. To elucidate this, we investigated model fit-criteria curve behaviour across an array of data conditions and covariance structures. Fit statistic patterns were variable among the fit-criteria and across a range of data conditions. This variability was greatly attributable to the level of class separation and the presence/absence of random effects. Our findings support some widely held notions (e.g. BIC outperforms other criteria) whilst debunking others (adding random effects is not always the solution). Based on the obtained results, we present guidelines on how the behaviour of fit-criteria curves can be used as a diagnostic aid during class enumeration.

Keywords: growth mixture model; latent class growth analysis; trajectory; repeated measures; covariance misspecification; class extraction

3.1. Introduction

Longitudinal finite mixture models (FMMs) are model-based clustering approaches designed to uncover latent heterogeneity in longitudinal profiles of the repeated measures type. This heterogeneity is usually represented as developmental trajectories, which comprise both inter-individual (between-subjects) and intra-individual (within-subjects) variability over time. These methods assist in identifying distinct latent classes of subjects within the population that show similar (within-class) temporal development. Assignment of subjects to such classes is typically done according to where their posterior probability (of the parameters) given the data is highest. Popular longitudinal FMMs include growth mixture models (GMM) [24], latent class growth analysis (LCGA) [48], and group-based trajectory models (GBTM) [6].

Longitudinal FMMs are increasingly used in applied sciences, particularly health sciences to understand differences in the development and aetiology of a variety of disorders and diseases, as well as subject responses to treatment. Recent studies include whether group differences in alcohol consumption are related to cardiovascular disease [43], understanding different treatment responses for adults with obsessive-compulsive disorder [40], and establishing the link between cannabis use in adolescents and a variety of health factors [44].

Nonetheless, it is often overlooked in practice that analysis results obtained with FMMs are sensitive to violations of their underlying assumptions, in particular the variance-covariance structure of the outcome variables in each class. This paper will investigate the impact of between-subject covariance misspecification on fit statistic behaviour during class extraction and, ultimately, on the choice of the number of classes. We examine, for instance, whether an inconclusive behaviour (e.g. continual improvement) of the considered model fit statistics (AIC, BIC, ssBIC and scaled Entropy) as a function of increasing the number of fitted classes, a recurring phenomenon in practice [122], is evidence of such covariance misspecification. We ascertain if identified fit statistic behaviour under such misspecifications may be used as a diagnostic tool in finding an adequate covariance structure.

We conduct a simulation study in which several data features (design conditions) are manipulated (e.g. number of repeated measures, degree of class separation, trajectory shape, true covariance structure), conforming to a specific GMM, LCGA or GBTM model. We then fit models misspecified in terms of the covariance to the data to investigate (1) whether a plateauing behaviour (or other peculiar behaviour) of the fit statistics under the fitted model is a relic of covariance misspecification, (2) how sensitive in terms of class enumeration are these

fit statistics to covariance misspecification under various data features (e.g. class separation, number of time points), and (3) whether identified fit statistic patterns may assist in finding the correct model. Moreover, an empirical example using alcohol consumption data is used to illustrate the fit-criteria curves as a diagnostic aid during class enumeration and model specification. Such a diagnostic tool may be useful since covariance misspecification has important consequences for both class extraction and classification performance [30,32,79,136–139].

3.2. Specification of models

Longitudinal FMMs develop from the premise that within the population, K latent classes (subgroups) exist with subjects within classes following similar paths of development over time (trajectories). The marginal probability distribution $P(\mathbf{y}_i)$ of a randomly chosen trajectory is then modelled as,

$$P(\mathbf{y}_i) = \sum_{k=1}^K \pi_k P^k(\mathbf{y}_i) \quad (3.1)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})^{tr}$ is a column vector of repeated measures for subject i , $i = 1, \dots, n$ at time t , $t = 0, \dots, T - 1$, and $P^k(\mathbf{y}_i)$ is the conditional distribution of the longitudinal sequence, \mathbf{y}_i , given that the subject i is in class k , $k = 1, \dots, K$. Further, π_k is the class membership probability and conforms to $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$, with $K > 1$. These models assume K to be known, but this is difficult to deduce directly from the data.

For continuous outcomes data, $P^k(\mathbf{y}_i)$ is assumed multivariate normal (MVN) within classes, that is,

$$\mathbf{y}_i^k \sim MVN(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k) \quad (3.2)$$

with \mathbf{y}_i^k a $T \times 1$ vector of continuous outcomes for subject i , and $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ are the model-implied mean vector and covariance matrix for class k respectively. $P^k(\mathbf{y}_i)$ is uniquely defined by the trajectory specification per class.

A GMM is the most general of our considered longitudinal FMMs. It includes both fixed effects to quantify class-specific average growth curves and random effects to allow for individual differences (inter-individual differences) from the average growth curve within classes. Its class-specific trajectories may be expressed as,

$$\mathbf{y}_i^k = \mathbf{X}\boldsymbol{\beta}^k + \mathbf{Z}\mathbf{b}_i^k + \mathbf{e}_i^k \quad (3.3)$$

where the superscript k specifies the class, \mathbf{X} is a $T \times b$ design matrix for the fixed effects, $\boldsymbol{\beta}^k$ is a $b \times 1$ vector of fixed effects, \mathbf{Z} is a $T \times q$ design matrix for the random effects, \mathbf{b}_i^k is a $q \times 1$ vector of random effects, and \mathbf{e}_i^k is a $T \times 1$ residual vector. It is assumed that $\mathbf{b}_i^k \sim \mathcal{N}(\mathbf{0}, \mathbf{D}^k)$ and $\mathbf{e}_i^k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^k)$. \mathbf{D}^k is the model-implied $q \times q$ random effects covariance matrix (inter-individual variation) and \mathbf{R}^k is the $T \times T$ residual covariance matrix (intra-individual variation) of the k -th class. Ultimately, a GMM is specified where $\boldsymbol{\mu}^k = \mathbf{X}\boldsymbol{\beta}^k$ and $\boldsymbol{\Sigma}^k = \mathbf{R}^k + \mathbf{Z}\mathbf{D}^k\mathbf{Z}'$ in Eq. (3.2).

LCGA and GBTM are special cases of Eq. 3.3 in which there are no random effects i.e. $\mathbf{Z}\mathbf{b}_i^k = \mathbf{0}$, such that $\boldsymbol{\Sigma}^k = \mathbf{R}^k$, with $\boldsymbol{\Sigma}^k$ diagonal [30]. Diagonal $\boldsymbol{\Sigma}^k$ implies independence between the repeated measures within a given individual. LCGA models allow for the residual variance to differ between classes and time points. The GBTM, a popular special case of the LCGA, makes the explicit assumption of the residual variance being equal for all classes and all time points i.e. $\mathbf{R}^k = \mathbf{R} = \sigma^2\mathbf{I}$, where \mathbf{I} is the identity matrix [6,19,140].

3.2.1. Class enumeration

A key outcome of FMM analysis is to identify the optimal number of classes K which adequately describe the data. Several statistical fit indices can assist in selecting K [140], a process known as class enumeration (synonymous with extraction). However, no fit statistic has yet emerged as the clear best performer [67,78,85,90,97,98]. Therefore, practitioners are often advised to use a variety of fit statistics as well as a substantive interpretation of their models during class extraction [6,94,99].

In this study, we restrict ourselves to the likelihood-based, information criterion (IC) model fit indices most often encountered in practice (and widely available by default in most software), that is, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and sample-size adjusted BIC (ssBIC). Scaled Entropy (sE), a statistic derived from Entropy $E(K)$, is included as a complement to the IC indices which is customarily reported as a measure of classification certainty. **Table 3.1** presents their equations. The first term of the AIC, BIC and ssBIC reward models for having better log-likelihoods. The second term

penalizes models for lack of model parsimony. sE is scaled to be bound between zero and one [83] and is higher when models show clear classification into classes.

Table 3.1: Summary of fit statistic calculations.

Measure	Equation	Model selection*
AIC	$-2 \log[L(K)] + 2[m(K)]$	Smallest value
BIC	$-2 \log[L(K)] + \log(n) [m(K)]$	Smallest value
ssBIC	$-2 \log[L(K)] + \log\left(\frac{n+2}{24}\right) [m(K)]$	Smallest value
sE	$1 - E(K)/n \log(K)$, where $E(K) = -\sum_{k=1}^K \sum_{i=1}^n pp_{ik} \log[pp_{ik}] \geq 0$	Largest value

*Fit statistic calculation may differ per software

$L(K)$: Maximum likelihood of K -class model

$m(K)$: Number of parameters of K -class model

pp_{ik} : posterior probability of subject i belonging to class k given the data

$E(K)$: Entropy of K -class model

$\log(x)$: The natural logarithm of x

n : Sample size

In an ideal situation, one would expect a clear minimum of the IC and an sE close to 1 at the true number of classes. However, in practice, these ICs often do not exhibit clear-cut behaviour (e.g. a minimum value) as a function of increasing K . For instance, a ‘plateauing’ curve is frequently observed [122] in that the IC continues to improve marginally as the fitted number of classes increases. It is to be established whether sE elicits similar behaviour. We hypothesise that such behaviour is evidence of random effect (between-subject) covariance misspecification, which can have serious consequences for class enumeration accuracy, classification performance, and model interpretability [30,32]. We ascertain whether aforesaid identified behaviour may assist in finding the correct covariance specification.

3.2.2. Class separation

Class separation in longitudinal FMMS typically refers to the degree of overlap between growth trajectories for latent classes [141]. This may be quantified in terms of the amount of overlap between the latent classes’ growth trajectory intercept and slope or the degree of overlap between the observed repeated measures [142].

Low class separation has been shown to play a substantial role in decreasing estimation accuracy in GMMs [85,143]. However, to date, there is no consensus on the best definition of class separation, and indeed which measure of class separation to utilize (See e.g. Nowakowska *et al.* [144]). As such, it is largely dependent on the researcher to decide upon given the investigation at hand [141]. This study employs the Cohen’s D (CD), which is often

used to quantify effect sizes [145], as a class separation measure. We report a time-averaged version, calculated as,

$$CD_{ave} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{|\mu_{at} - \mu_{bt}|}{\sigma_t} \quad (3.4)$$

where $t = 0, \dots, T - 1$ is the time point, μ_{at} and μ_{bt} is the class mean (of the observed outcome variable) at time point t for class a and b respectively, and σ_t is the square root of the diagonal element of the total covariance matrix (Σ) corresponding to time point t .

Additionally, we report the multivariate Mahalanobis distance (MD) [146], a popular class separation measure in longitudinal FMM studies [82,127,147,148]. The pairwise MD in terms of the observed repeated measures is calculated as,

$$MD = \sqrt{(\boldsymbol{\mu}^a - \boldsymbol{\mu}^b)^T \Sigma^{-1} (\boldsymbol{\mu}^a - \boldsymbol{\mu}^b)} \quad (3.5)$$

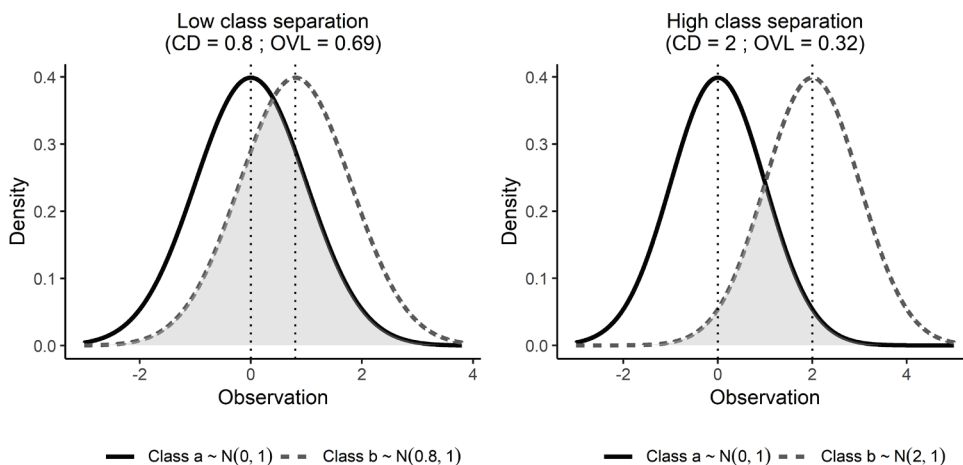
where $\boldsymbol{\mu}^a$ and $\boldsymbol{\mu}^b$ are the mean vectors of the observed repeated measures for class a and b respectively, and Σ^{-1} corresponds to the inverse of the covariance matrix of \mathbf{y} which is assumed equal in both latent classes [11,82]. An MD of one and three usually reflects small and large class separation respectively found in the literature [93,97,147,149,150].

Lastly, we provide the overlap coefficient (OVL) [144] for class separation. We calculate this as the average over all time points of the overlap of two class distributions at each time point,

$$OVL = \frac{1}{T} \sum_{t=0}^{T-1} \left[\int_{-\infty}^{\infty} \min[f_{at}(x; \mu_{at}, \Sigma_{at}), f_{bt}(x; \mu_{bt}, \Sigma_{bt})] dx \right] \quad (3.6)$$

where $f_{at}(x; \mu_{at}, \Sigma_{at})$ and $f_{bt}(x; \mu_{bt}, \Sigma_{bt})$ correspond to the class density function (univariate normal) at time point t of class a and b respectively. **Figure 3.1** shows low and high separation in terms of the OVL (**Eq. 3.6**) for two univariate normal densities for a single time point. The OVL is the common area under the lower of the two density functions. The greater the overlap between the densities, the broader the x -axis range is where the minimum of the two densities is high. For the example of low separation, a CD of 0.8 corresponds to an OVL of 0.69 (see the grey area in the leftmost plot). For high separation, a CD of 2 results in an OVL of 0.32 (see the grey area in the rightmost plot).

Figure 3.1: An illustration of class separation for two univariate normal densities.



3.2.3. Covariance misspecification

Covariance misspecification implies assuming an incorrect structure for the random effects' covariance matrix \mathbf{D}^k and/or for the residual covariance matrix \mathbf{R}^k during model estimation. Such misspecification may be broadly classified into three categories [32].

Covariance *underspecification* can occur when the true model includes class-specific covariance matrices (i.e. \mathbf{D}^k and \mathbf{R}^k), but the fitted model specification constrains within-class covariance matrices to be equal across classes (i.e. $\mathbf{D}^k = \mathbf{D}$ or $\mathbf{R}^k = \mathbf{R}$) or even equal to zero (i.e. $\mathbf{D}^k = \mathbf{0}$). It can also be that the true model includes equal within-class covariance matrices (i.e. $\mathbf{D}^k = \mathbf{D}$ and $\mathbf{R}^k = \mathbf{R}$), a GMM, but is specified such that $\mathbf{D}^k = \mathbf{0}$, an LCGA or GBTM. Covariance *overspecification* can occur when the true model contains equal random effect covariance matrices across classes (i.e. $\mathbf{D}^k = \mathbf{D}$) or equal residual covariance matrices across classes (i.e. $\mathbf{R}^k = \mathbf{R}$), whilst the model selected for analysis allows for the estimation of class-specific matrices (i.e. \mathbf{D}^k and/or \mathbf{R}^k). Additionally, overspecification also arises when the true model has no random effect variability within classes (i.e. $\mathbf{D}^k = \mathbf{0}$) but is estimated with such variability. In this context, the true model is an LCGA or GBTM, but the assumed model is a GMM. *General covariance misspecification* can occur when fitting a mixture model when one is not needed, that is, where the true model consists of a single population (i.e. growth curve model),

but the analysis proceeds assuming population heterogeneity (i.e. LCGA, GBTM or GMM) [32].

In our paper, we will examine the effects of *under-* and *overspecification* on fit statistic behaviour, more specifically the effect of incorrectly specifying the \mathbf{D}^k matrix.

3.3. Methods

3.3.1. Design of the simulation study

To imitate model specifications frequently and currently used in practice [43,151–153], we limit ourselves to the case of underspecification where the true model has $\mathbf{D}^k = \mathbf{D}$ and $\mathbf{R}^k = \mathbf{R}$ (i.e. a GMM), but is estimated such that $\mathbf{D}^k = \mathbf{0}$ (i.e. an LCGA or GBTM). In the case of overspecification, we investigate the impact where the true model has $\mathbf{D}^k = \mathbf{0}$ with $\mathbf{R}^k = \mathbf{R}$ (i.e. an LCGA or GBTM), but is estimated such that $\mathbf{D}^k = \mathbf{D}$ (i.e. a GMM). Such equal within-class covariance matrices is the default specification of most software [140] which is often inadvertently selected by practitioners.

The (true) models for data simulation are:

- Model 1 (GBTM): With $\mathbf{D}^k = \mathbf{0}$, and $\mathbf{R}^k = \mathbf{R} = \sigma^2 \mathbf{I}$
- Model 2 (LCGA): With $\mathbf{D}^k = \mathbf{0}$, and $\mathbf{R}^k = \mathbf{R} = \sigma_{\varepsilon_t}^2 \mathbf{I}$, that is, class-invariant but time-variant residual variance
- Model 3 (GMM-I): With class-invariant random intercept and random linear slope allowed to covary $\mathbf{D}^k = \mathbf{D}$, and $\mathbf{R}^k = \mathbf{R} = \sigma^2 \mathbf{I}$
- Model 4 (GMM-II): With class-invariant random intercept and random linear slope allowed to covary $\mathbf{D}^k = \mathbf{D}$, and $\mathbf{R}^k = \mathbf{R} = \sigma_{\varepsilon_t}^2 \mathbf{I}$

We then study the effect of the chosen misspecifications by considering various fitted on true model combinations. These are shown in **Table 3.2**. The misspecification of the \mathbf{R} matrix in terms of time-dependency (either time-variant or time-invariant) is beyond the scope of this paper, but we do note that GMM with random slopes generates heterogeneity of variance across time points and thus may resemble a time-variant in \mathbf{R} LCGA.

Table 3.2: True with fitted models considered (Misspecification: a: D underspecified, b: D over-specified, c: D correctly specified).

True Model	Fitted Model	
GBTM	GBTM ^c	GMM-I ^b
LCGA	LCGA ^c	GMM-II ^b
GMM-I	GMM-I ^c	GBTM ^a
GMM-II	GMM-II ^c	LCGA ^a

The design conditions underlying the data generating process are informed by previous simulation studies and applied research [79,154,155], and are summarised in **Table 3.3** and described below.

The choice of a sample size of 1000 reflects the median condition in applied studies [155]. Furthermore, a minimum sample size of 900 is suggested under conditions of multiple classes and low class separation [129]. We also briefly investigate a sample size of 260 for a subset of models, as small sample sizes have a demonstrably negative impact on class enumeration [127], with $N = 200$ being the recommended minimum for complete case data and high class separation [129].

Five repeated measures are chosen to be the lower bound at which to detect non-linear growth trajectories and to ensure model identifiability [79], especially when including full rank covariance matrices and larger K . This is expanded to eight to mirror the higher number of repeated measures seen in applied GMM research [127]. Moreover, it has been shown that increasing the number of time points has a positive effect on classification performance [30]. Equally spaced time values over a fixed time interval from zero to seven are chosen, and so for $T = 5, t = 0,1,75,3.5,5.25,7$ and for $T = 8, t = 0,1,2,3,4,5,6,7$. The time interval is the same for both T -values to prevent confounding of the effect of number of time points with the effect of a change of total follow-up time.

Table 3.3: Primary design conditions investigated.

Design condition	Choice	Additional features
Sample size	$N = 1000$	
Number of repeated measures	$T = 5,8$	Equally spaced
Number of classes	$K = 4$	
Class sizes	Equal	
Class separation	Low, high	$CD = 0.5,2$
Fixed effects	Same intercept and different slope (NS) Different intercept and different slopes (CC)	Quadratic trend
Random effects	None, or intercept and linear slope that covary	Class-invariant
Errors	Time-variant or time-invariant	Uncorrelated across time, class-invariant

Most simulation studies in the literature consider two or three true classes. We expand upon this by including four classes. We focus primarily on equal class sizes. However, we also explore unequal class sizes ($k = 1$ (35%)/ $k = 2$ (15%)/ $k = 3$ (15%)/ $k = 4$ (35%)) for a subset of models since a substantial decrease in the class enumeration accuracy of the BIC compared to the AIC and ssBIC has been noted when one class is considerably smaller [85].

We will impose a CD_{ave} of approximately 0.5 and 2 to reflect low and high class separation respectively. The data will be constructed in such a way that each class will be at least CD_{ave} units away from each other. MD and OVL are also reported.

Fixed effects' parameters are altered according to the degree of class separation (low or high) corresponding with our chosen Cohen's D separation metric. We have chosen a second-order polynomial in the fixed effects as it is a flexible function which can capture many patterns across time, including monotonic trends, and u- and n-shaped trends as well as parts thereof. Two conditions of trajectory growth are studied. One in which trajectories comprise the same intercept but different slopes between classes (natural starting (NS) point) and the second includes both different intercepts and different slopes between classes. The second condition's functional form mimics the "cat's cradle" (CC) phenomenon often identified in applied health research with a small number of time points [27,156,157]. Sher *et al.* [156] present this pattern empirically in terms of alcohol use over time. Subjects' alcohol consumption in one class starts high and remains high (chronically bad), in a second class it starts low and remains low (unaffected/non-drinkers), in a third class it starts high but reduces over time (recovery), and in the fourth class it starts low but increases over time (delayed onset).

Lastly, we impose an \mathbf{R} that is diagonal, equal across classes, and either time-variant or time-invariant. For the \mathbf{R} matrix of the GBTM model, each diagonal element is set to equal the average of the sum of the diagonals of the full $\mathbf{\Sigma}$ matrix of the GMM. For the LCGA specification, the diagonal elements of the \mathbf{R} matrix are set equal to the corresponding diagonal elements of the $\mathbf{\Sigma}$ of the GMM. This strategy is effected so that the total average diagonal variation is similar across the design conditions. For conditions with a non-zero \mathbf{D} matrix (i.e. where the true model is a GMM), the proportion of total average diagonal variation explained by the random effects was set to a fixed proportion of approximately 0.5. We enforce a weak

positive correlation of 0.1 between random intercept and random linear slope, in line with previous studies [36,79,136,141,142].

The data generated are of the following general form,

$$y_{it}^k = (b_{0i}^k + \beta_0^k) + (b_{1i}^k + \beta_1^k)t + \beta_2^k t^2 + \varepsilon_{it}^k \quad (3.7)$$

with $k = 1, \dots, K$, $i = 1, \dots, n$, β_0^k, β_1^k and β_2^k are fixed effects quantifying the population average growth curve for class k , and b_{0i}^k and b_{1i}^k are random effects that allow for individual differences from the average growth curve of class k . In the case of LCGA and GBTM, random effects are not included. **Figures 3.2** and **3.3** show the different trajectory shapes of selected true GMM-I models for different parameter sets. All considered models' parameters are found in the **Supplementary Material (SM)**.

3.3.2. Simulation procedure

Longitudinal repeated measures data conforming to our true models were generated in R v3.6.3. The fitted models were estimated using the R package *Mplus Automation* [107], which interfaces directly with *Mplus* [26]. We used *Mplus* v7.3 for our analysis and *ggplot2* in R [158] for the plotting of figures.

Subjects were first assigned to classes according to the chosen class size, e.g. for $N = 1000$, $K = 4$ and equal classes, there were exactly 250 subjects in each class. Then, a vector of random effects for each subject in a class was generated according to $\mathbf{b}_i^k \sim \text{MVN}(\mathbf{0}, \mathbf{D}^k)$. A vector of continuous repeated measures (\mathbf{y}_i) for that subject within a class was then generated according to **Eq. 3.7**. This process was repeated 200 times for each of the 32 design conditions in **Table 3.3** to generate independent datasets, giving a total of 6400 simulated datasets. Each generated dataset was used as input in the subsequent *Mplus Automation* step where both the true and misspecified \mathbf{D}^k models, as given in **Table 3.2**, were fitted over $K = 1, 2, \dots, 10$ producing $6400 \times 2 \times 10 = 128\,000$ estimated models. Anticipating that 200 replications may be too few, we ran 1000 replications for select conditions but did not observe marked differences in the results. We, therefore, adhered to 200 replications, which is in line with other published FMM research [82,159,160].

For model estimation, we bore in mind that longitudinal FMMs are notoriously sensitive to starting values for model parameters [160]. Selecting too few starting values may negatively impact the chance of finding the global solution, whilst too many may return

improbable combinations, likely leading to nonconverged solutions and zero class sizes. Therefore, in line with research [79,160] and practical [26] recommendations, for a thorough investigation of the likelihood surface, we instructed *Mplus* to use 100 random sets of starting values for all model parameters of a given model on a given run. The program was then ordered to run through 20 iterations on each of these sets. Next, the program was directed to use the 10 sets yielding the highest log-likelihood from the first stage as starting values in the final stage optimisation until convergence criteria were met. The model with the highest log-likelihood from this stage was used as the basis for further analysis. Any nonconverged solutions were discarded, with the proportion of non-convergence out of all 200 runs computed per design combination per true model per fitted number of classes never exceeding 5%. For 96% of the cases, non-convergence was below 1% (See **SM Section S.2**). If non-convergence exceeded 2%, this was always for true or fitted model GMM, T=8, low separation, and number of fitted classes exceeding 5 (See **SM Table S.3**).

3.4. Results

3.4.1. Accuracy of class extraction in relation to design conditions and **D** misspecification

The impact of design conditions and **D** misspecification on the probability of class extraction was investigated with two logistic regression analyses; one using as outcome correct versus incorrect extraction and including all cases (logistic model 1 – **LM-I**), and one using as outcome over- versus underextraction and including only cases of incorrect extraction (logistic model 2 - **LM-II**). The K chosen from the fitted models corresponded to that K at which the IC value was lowest or the sE highest. The abundance of interactions found in these analyses between true model and every other design factor in **Table 3.3** justified separate logistic regression analyses on subsets of the data to facilitate interpretation. Subset (a) included true GMM-I and GMM-II, whilst subset (b) considered true GBTM and LCGA. These subsets corresponded to examining under- and overspecification of **D** respectively (See **SM Section S.3.1** for full logistic regression results).

In the subset analyses, two-way interactions were included to test whether the effect of **D** misspecification on K extraction depended on the level of the design condition (e.g. low versus high class separation) and also on the fit statistic (e.g. AIC versus BIC). Further, interactions between the fit statistic and design conditions were also considered. Likelihood

Figure 3.2: True GMM-I with NS scenario for fixed effects, $T = 8$ and time-invariant R .

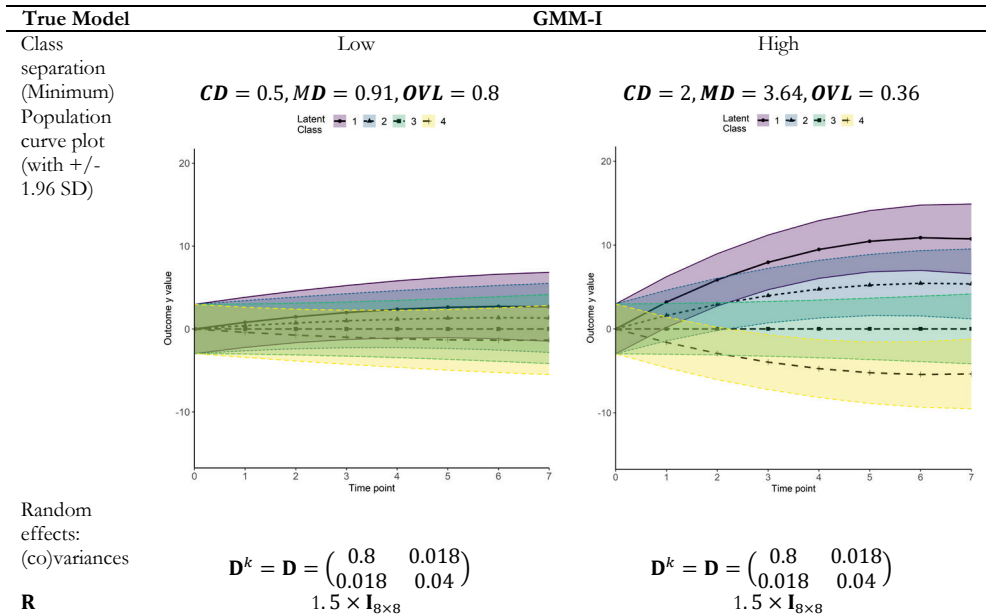
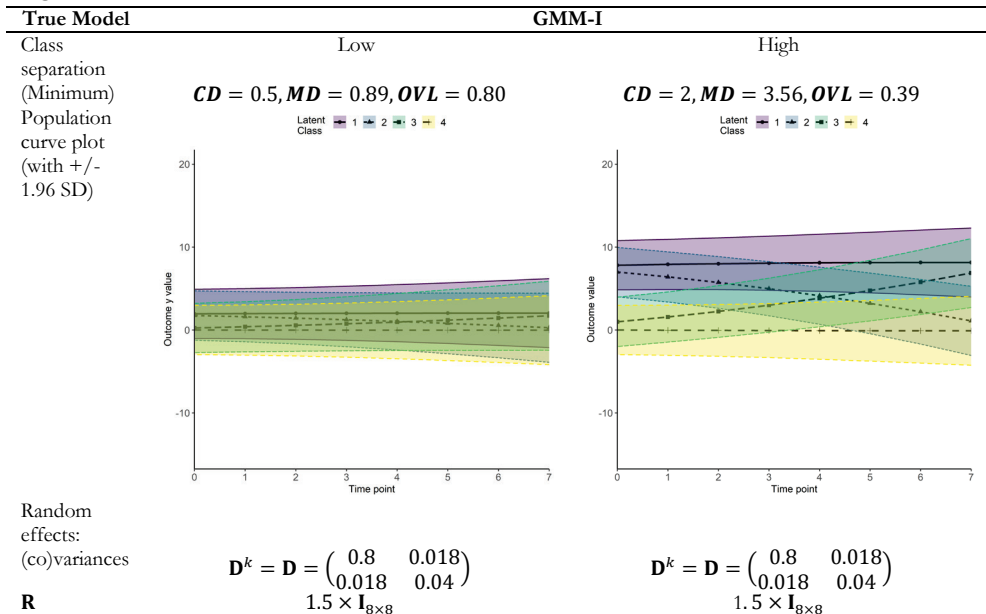


Figure 3.3: True GMM-I with CC scenario for fixed effects, $T = 8$ and time-invariant R .



ratio tests were conducted to ascertain whether the included interactions yielded a significantly better fit. Many of these interactions were significant. Of particular note were; the interactions between (1) the fit statistic and class separation across all logistic regression analyses, (2) the fitted \mathbf{D} and fit statistic in all logistic models (except LM-II(b)), and (3) the fitted \mathbf{D} and class separation level across all logistic models (except for LM-II (a)). These interactions are displayed in **Figure 3.4** (to be discussed in greater detail) and **Appendix Figure B.1-Figure B.3**, which show the patterns and sizes of the effects of the design conditions on class enumeration performance for each criterion.

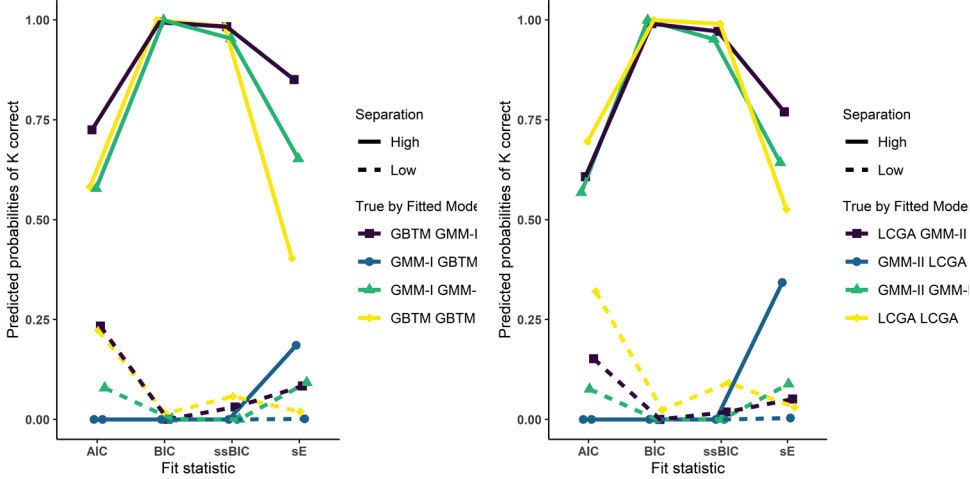
Figure 3.4 presents the estimated probabilities of K extraction for the fit statistics under natural starting point (NS) and eight time points ($T=8$) for all logistic models considered. The remaining combinations of NS/CC and $T=5/T=8$ are presented in the **Appendix**. These figures were chosen since they respect the prominent interactions of fitted \mathbf{D} with class separation and fit statistic.

The findings of LM-I (outcome: correct class extraction) displayed in **Figure 3.4** (a) are multifaceted. First, it shows that, irrespective of class separation, all IC fit statistics had a low probability of selecting the true K when \mathbf{D} was underspecified (i.e. true model is GMM, fitted model is GBTM/LCGA). Further, under high class separation, BIC and ssBIC performed almost perfectly for fitted models with \mathbf{D} overspecification (i.e. true model is GBTM/LCGA, fitted model is GMM) or correct specification, whereas the AIC and sE performed substantially worse. By contrast, all fit statistics performed poorly under low class separation irrespective of the \mathbf{D} specification, although the AIC generally performed slightly better than the other fit statistics if \mathbf{D} was correctly specified or over-specified. The effect of the number of time points, time-variant versus time-invariant \mathbf{R} for all fitted models and fit statistics on correct class extraction was inconspicuous compared to the effect of class separation (See the **Appendix**).

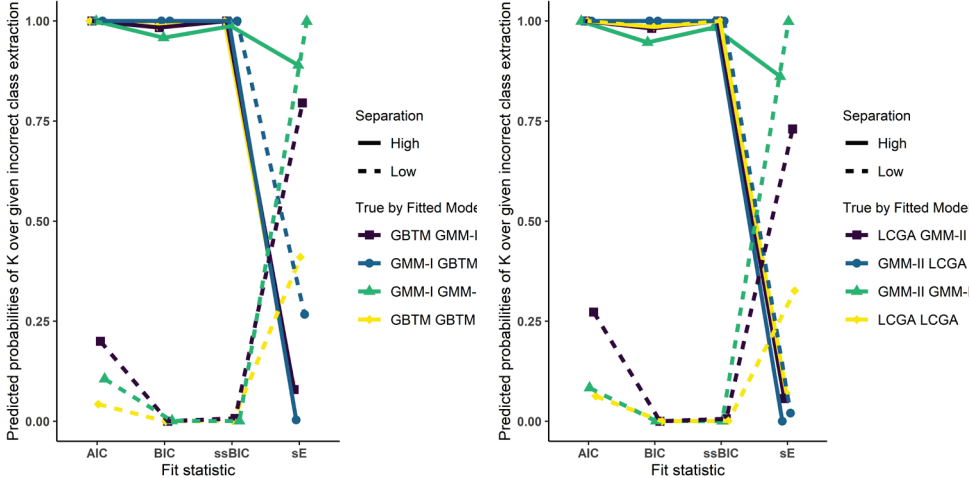
Figure 3.4 (b) shows the results of LM-II (over- versus underextraction). Here, regardless of class separation, for underspecified \mathbf{D} , the IC fit statistics had a 100% probability of over-extraction (given incorrect extraction). For \mathbf{D} over- or correct specification, all IC fit statistics showed a high probability to overextract under high separation and under-extract under low separation. sE, however, showed converse behaviour, under- and over-extracting under high and low separation respectively.

Figure 3.4: Estimated probabilities of K correct (upper half) or of over-extraction given incorrect K (lower half) for true by fitted model under low/high class separation given conditions: Natural starting point, $T=8$ repeated measures. Left half concerns models GMM-I and GBTM, right half concerns models GMM-II and LCGA.

a.) K correct ($Y=1$ if K is correct, $Y=0$ if $K \neq 4$): Results from LM-I(a)-(b).



b.) K over ($Y=1$ if $K > 4$, $Y=0$ if $K < 4$): Results from LM-II(a)-(b).



To conclude, all fit statistics performed poorly in terms of correct K extraction under low separation, with BIC and ssBIC performing the worst. Under high separation, BIC performs best, followed by the ssBIC. Furthermore, under high separation, underspecification of \mathbf{D} is associated with a high risk of incorrect class extraction for the IC fit statistics, whereas overspecification of \mathbf{D} for all fit statistics shows little risk of incorrect K extraction, particularly for the BIC and ssBIC. Moreover, among the cases that were incorrectly extracted, underspecification of \mathbf{D} is associated with a high risk of over-extraction by the IC statistics

regardless of class separation. For over- and correct specification of \mathbf{D} , IC fit statistics tended to overextract under high separation and under-extract under low separation, among the subset of cases with incorrect extraction.

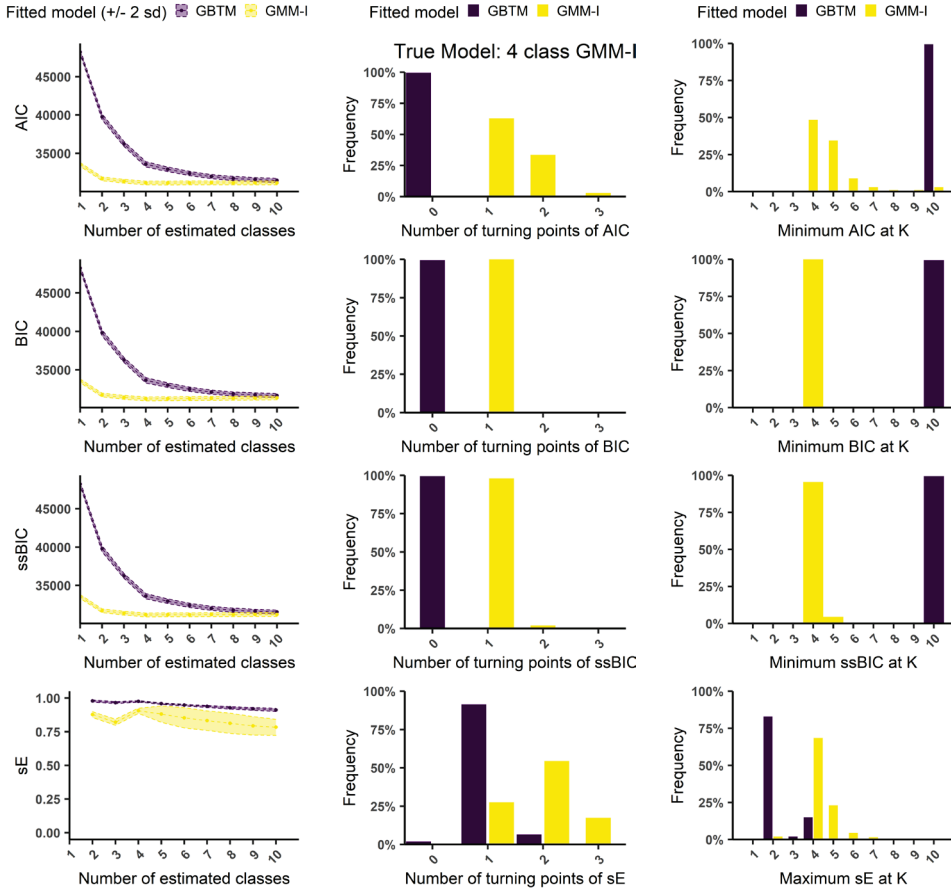
3.4.2. Identifiable patterns of fit statistic curves across all conditions

3.4.2.1. Screeplot behaviour

Each of the 32 simulation design conditions considered in **Table 3.3** for the true model yielded fit statistic curves and summary bar charts. Given space constraints, only two of these conditions are presented in **Figures 3.5** and **3.6**, but the corresponding figures for all design conditions are provided in **SM Section S.4**. Each figure condenses the output information of 200 simulations (runs) for one design condition. For each fit criterion (rows), three distinct plots are shown (columns): the fit statistic curve given as the average over 200 runs at each number of classes (left), frequency distribution of the number of turning points of the fit statistics' curve of a single run (middle), and frequency distribution of the final selected K (right). This information is provided in each subplot separately for each fitted model. A turning point for AIC, BIC and ssBIC is defined as a point K where $IC(K) < IC(K - 1)$ and $IC(K) < IC(K + 1)$. For sE this is defined as a point K where $sE(K) > sE(K - 1)$ and $sE(K) > sE(K + 1)$.

Figure 3.5 (true model = GMM-I) shows that when a GBTM is fitted (i.e. underspecified \mathbf{D}), the IC statistics exhibit clear plateauing behaviour given their continual improvement as K increases (left column). Furthermore, the general absence of turning points (middle column) highlights their proclivity to overextract as the maximum considered $K = 10$ is always selected (right column). This conforms with the findings in **Figure 3.4**, showing the ICs' high probability of incorrect extraction for underspecified in \mathbf{D} models (**Figure 3.4** (a)), see True: GMM-I Fitted: GBTM, specifically over-extraction (**Figure 3.4** (b)). This pattern is repeated throughout conditions where underspecified models are fitted (see **SM Section S.4**). For a correctly specified GMM, both the BIC and ssBIC are highly accurate and stable showing a large majority of one turning point at $K = 4$. sE exhibits erratic and inaccurate performance compared to the BIC and ssBIC for the true model. AIC shows a tendency to overextract, even under the correct model. Again, these observations conform to the logistic regression results (**Figure 3.4**) which highlights the poor accuracy of the AIC and sE relative to the BIC and ssBIC.

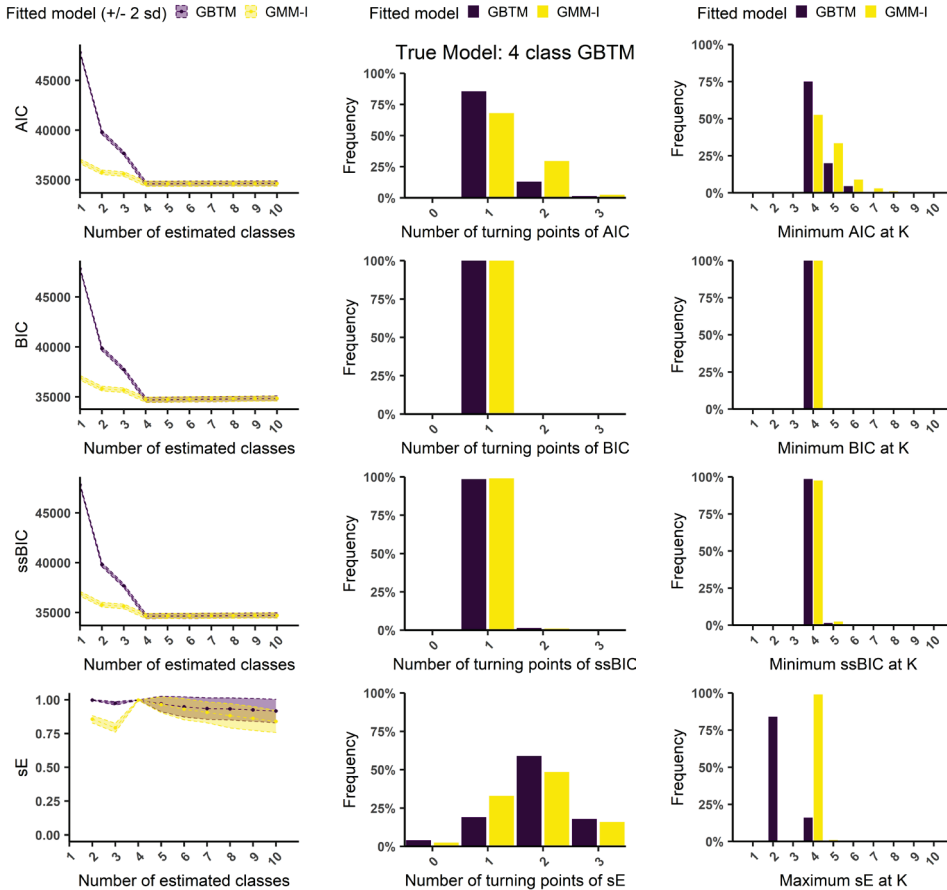
Figure 3.5: Fit statistic behaviour for true 4 class GMM-I with natural starting point, high class separation and $T = 8$. Left column: The average fit statistic value over all runs (ordinate axis) against the number of estimated classes (abscissa); middle column: Frequency of the number of turning points in the individual fit statistic curves ($n=200$ runs); right column: frequency of specific K being selected.



Note: Turning point for AIC, BIC and ssBIC is defined as a point K where both $IC(K) < IC(K - 1)$ and $IC(K) < IC(K + 1)$. For sE, a turning point is defined as a point K where both $sE(K) > sE(K - 1)$ and $sE(K) > sE(K + 1)$.

In **Figure 3.6** (true model = GBTM), the BIC and ssBIC of both correct and over-specified models extract the correct K . In contrast, the sE for correct \mathbf{D} and the AIC for both correct \mathbf{D} and over-specified \mathbf{D} shows lower accuracy. The BIC and ssBIC do not show plateauing behaviour as there is a single turning point. This pattern recurs in similar cases (see **SM Section S.4**) indicating that the risk of an incorrect K under overspecification appears

Figure 3.6: Fit statistic behaviour for true 4 class GBTM with natural starting point, high class separation and $T = 8$. Left column: The average fit statistic value over all runs (ordinate axis) against the number of estimated classes (abscissa); middle column: Frequency of the number of turning points in the individual fit statistic curves ($n=200$ runs); right column: frequency of specific K being selected.



Note: Turning point for AIC, BIC and ssBIC is defined as a point K where both $IC(K) < IC(K - 1)$ and $IC(K) < IC(K + 1)$. For sE, a turning point is defined as a point K where both $sE(K) > sE(K - 1)$ and $sE(K) > sE(K + 1)$.

small with high separation.

It is noticeable that the average scree plot of the various IC fit statistics (which approximately matched the individual scree plots, one per simulated dataset) (See **SM Section S.4**) is smooth (i.e. gradual improvement in IC) for underspecified models, that is, when the true model is a GMM and a GBTM or LCGA is fitted. By contrast, the curves are jagged (i.e. quick uneven improvement in IC to an elbow, with no or hardly any improvement in the IC beyond the true K) for correct or over-specified models, that is, when a GBTM or LCGA is

the true model underlying the data and a GBTM, LCGA or GMM is fitted. This noticeable pattern may assist practitioners in refining their model's covariance structure as the smoothness indicates the necessity for random effects or a respecification of the covariance structure.

3.4.2.2. Fit statistic behaviour across all simulation conditions

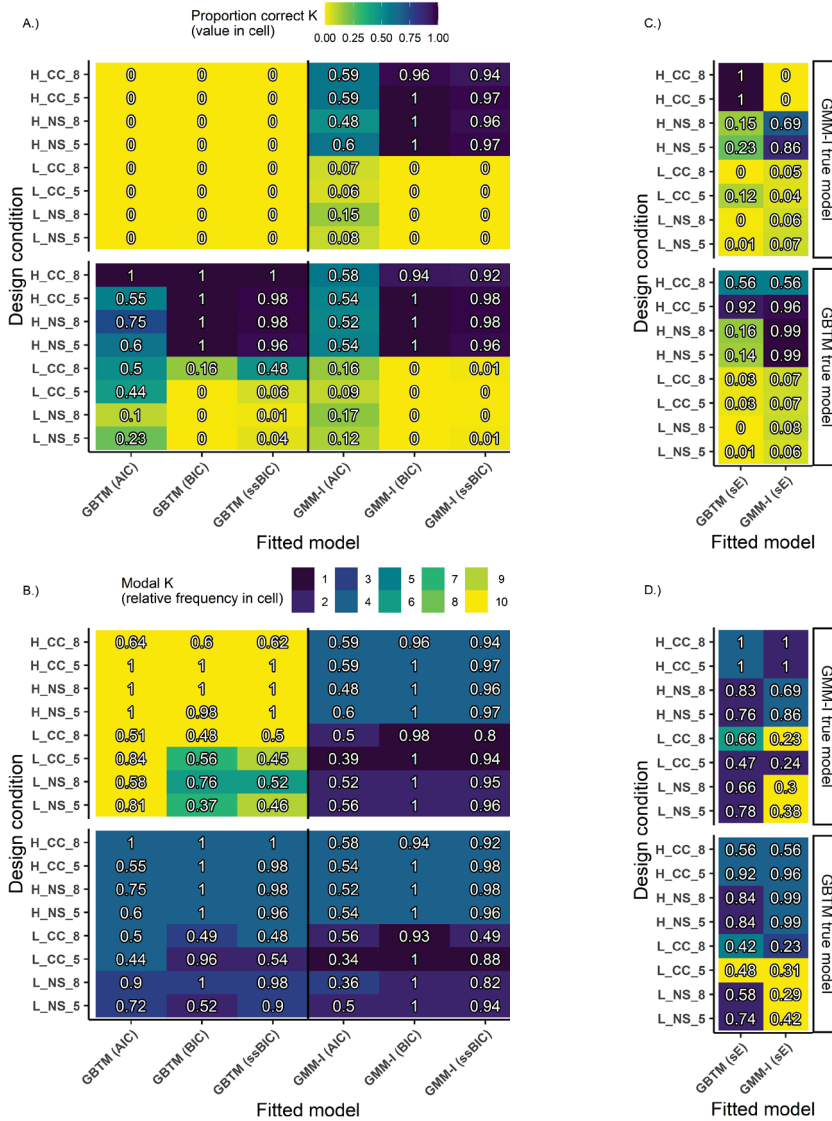
Figure 3.4 summarizes the results of only four of all 32 simulation conditions whilst **Figures 3.5** and **3.6** do so for one condition each. Therefore, to visualise patterns within the data for all 32 simulation conditions, heatmaps [161] are presented. These heatmaps summarise the outputs of all the crossed true model by fitted model simulations for time-invariant (**Figure 3.7**) and time-variant (**Appendix Figure B.4**) **R** conditions respectively. The results of the 200 simulations per design condition (rows), per fitted model and per fit criterion (columns) are summarised by two cells in separate but complementary heatmaps within **Figure 3.7** (respectively **Appendix Figure B.4**):

- each cell in **panel A** (for IC fit statistics) or **panel C** (for scaled Entropy) displays the proportion of correct K extracted by the fit statistic out of all runs for each design condition,
- each cell in **panel B** (for IC fit statistics) or **panel D** (for scaled Entropy) shows the modal K (i.e. the most frequent K extracted by a fit statistic) for each design condition.

Combined, the two cells for a given condition inform whether the fit statistics performed well in terms of extracted K accuracy (**panels A** and **C**) under each design by fitted model condition, whilst simultaneously hinting at their underlying fit statistic curve behaviour (**panels B** and **D**).

Each panel is divided into 4 quadrants. The top left quadrant corresponds to underspecification of **D**, the bottom right quadrant represents overspecification of **D**, and the remaining two quadrants correspond to correct specification. Within panels, the x -axis corresponds to the fitted model and its associated fit statistic, e.g. GBTM (ssBIC) shows that a GBTM was fitted with its ssBIC output given. The y -axis shows the true model and its underlying design conditions, with the naming convention of **TrueModel_Trajectory**

Figure 3.7: Heatmaps of the proportion correct $K=4$ (panels A and C) and modal K (panels B and D) extracted by different fit statistics for different fitted models under time-invariant R conditions (GBTM, GMM-I). Ordinate axis coded as H/L_NS/CC_8/5 indicating: Class separation: H(igh) or L(ow), Trajectory shape: N(atural) S(start) or C(at's) C(radle), and Time points: 8 or 5. All panels: Quadrants clockwise from upper left: 1.) Underspecified GBTM, 2.) Correctly specified GMM, 3.) Over-specified GMM, 4.) Correctly specified GBTM.



shape_Degree of class separation_Number of repeated measures. For example, the performance of the AIC for a GBTM fitted to a true GMM-I with a natural starting point and high class separation for 8 measurements (row 1 in **Figure 3.5**) corresponds to the coordinate (Panel: A and C, Row: GMM-I_NS_H(igh)_T=8, Column: GBTM(AIC)) in **Figure 3.7**. The results in the heatmaps are arranged such that the upper half of each panel corresponds to true GMM models, and the lower half to either true GBTM (**Figure 3.7**) or LCGA (**Appendix Figure B.4**) models. Within each half, the results are further divided by class separation, then trajectory shape, and finally the number of repeated measures.

To facilitate interpretation, consider light cells in **panel A** of **Figure 3.7**. These cells have low proportions of correct K extraction. However, these cells on their own do not indicate whether the fit statistics extracted more or fewer classes than the correct K , just that $K = 4$ was selected hardly ever in all the 200 runs. The additional nuance of under- or over-extraction is found in the corresponding cell in **panel B**. Here, given A, if the cell in B is also light, e.g. for (GMM-I_NS_H_T=8, GBTM(AIC)) the modal $K = 10$, this signifies class over-extraction. This occurs mainly if the true model is GMM and the fitted model is GBTM, that is, if the covariance matrix \mathbf{D} is underspecified. The fact that this over-extraction is accompanied by a plateauing behaviour of the fit statistic curve is confirmed by considering both the left and middle columns of **Figure 3.5** (or associated Supplemental figures) as discussed previously. By contrast, if the cell is light in A, but dark in B (bottom right quadrant), this is an indication of underextraction by the fit statistic i.e. the fit statistic curve reached a minimum point before the correct K . This occurs if \mathbf{D} is correctly specified or over-specified, combined with low class separation.

Consider now the darkest cells in **panel A** which display the highest accuracy of correct K extracted (upper half of top right quadrant and of both bottom quadrants). Their counterparts in **panel B** confirm that the fit statistics excelled in selecting a modal $K = 4$. This optimal extraction performance again transpires in exemplar **Figure 3.6** (considering GBTM_NS_H_T=8, GBTM(BIC,ssBIC)): these IC fit statistics curve had a clear (elbow) minimum turning point at the correct $K = 4$, with no improvement in the curve beyond the true K (left) and highest frequency of one turning point (middle).

The heatmaps thus encapsulate unfolding fit statistic patterns over increasing K , whilst directing attention to specific combinations of design conditions and fitted models (x - and y -axes), in which standout curve behaviours are observed.

Information criteria (IC) for time-invariant R (Figure 3.7 A and B): What is immediately apparent is the large number of zero proportion correct K in **panel A**. These are seen when fitted models are underspecified in **D** (the upper left quadrant of **panel A**) or for low class separation (rows 5-8 and 13-16 from the top) independent of whether **D** is over-, under- or correctly specified. For high separation (rows 1-4 and 9-12), we notice with correct specification of **D** (upper half of upper right quadrant and of lower left quadrant of **panel A**) and with overspecification of **D** (upper half of the bottom right quadrant of **panel A**) high accuracy of the ICs, which in most cases exceeds 98% correct. However, the AIC is considerably less accurate than the BIC and ssBIC.

Linking the above-identified behaviour to **panel B**, we see that under high class separation and underspecified **D** this is associated with a modal $K = 10$ in a vast majority of cases – in line with plateauing behaviour (see **SM** scree and turning point bar plots) and the associated risk of over-extraction. Under low class separation, underspecified **D** is still associated with over-extraction exhibiting a high modal K where most cases exceed $K = 8$. Therefore, when the IC of fitted GBTM selects considerably more classes than a fitted GMM, this may point to underspecification in terms of **D**. For high separation, the over-specified or correctly specified in **D** models (associated with the darker regions) show a low risk of over-extraction where they almost always have a modal K at the true K . Moreover, under conditions of low separation for correctly specified and over-specified models we notice the tendency of ICs to under-extract classes.

sE for time-invariant R: sE also performs better under high class separation, but its performance is inconsistent. In particular, the sE does not appear to perform better in class extraction if the correct model is fitted than if an under- or over-specified model is fitted.

The findings of time-variant **R** (shown in **Appendix Figure B.4**) are similar to time-invariant **R**. This is in line with our logistic regression results, which confirms that the effect of the level of **R** is small relative to those of class separation and of **D** specification.

To conclude, it appears that when the true model is fitted under high class separation, the best IC fit statistic is most often observed at or close to the true K showing a clear elbow. If an underspecified model, GBTM respectively LCGA, is fitted to a true GMM-I and GMM-II respectively, it is frequently observed that the AIC, BIC and ssBIC fit statistics continue to improve as K increases (plateauing behaviour). No useful fit curve behaviour for sE can be found. Lastly, fit statistic class enumeration behaves poorly under low class separation, but the

patterns identified under high separation repeats in the low separation conditions, however with the modal K being lower (see **SM Section S.4**).

3.4.3. Unequal class sizes and small sample size

Whether the patterns identified above also hold for select models with unequal class sizes (two classes of 35%, two of 15%) or small sample size ($N=260$) was briefly investigated. Only conditions of high separation (as low separation has already been shown to be detrimental to class extraction), NS and CC, $T=8$ and time-invariant \mathbf{R} , were considered.

The ICs perform similarly under unequal classes (**Appendix Figure B.5**). However, compared to equal classes, the over-extraction behaviour of underspecified models in \mathbf{D} is more pronounced under unequal classes given cat's cradle as they show a higher relative frequency of modal $K = 10$. The BIC remains accurate with correctly specified and over-specified models. sE again performs poorly and erratically under correctly fitted models. Finally, the plateauing and elbow behaviour identified previously also holds under unequal class sizes (see **SM Section S.4.3**).

Under small sample conditions (**Appendix Figure B.6**), the previously identified behaviour of the ICs is retained. The BIC performs better than the other ICs, but does suffer a decrease in accuracy under correctly specified GMM-I and over-specified GMM-I under CC compared to the large sample condition. We also take note of the decrease in accuracy of the ssBIC under small samples, which is noticeable given that it is meant to perform better under small samples [68]. The sE appears to perform better in small samples than in large samples, with 8/12 of the crossed models having a modal $K = 4$ frequency of above 50%, but in general, remains an unreliable class enumeration measure. Moreover, plateauing and elbow behaviour of the IC fit statistics associated with the level of \mathbf{D} specification is still clearly evident under small sample sizes (**SM Section S.4.4**).

3.5. Application

In this Section, we will show how an appropriate covariance structure for a longitudinal finite mixture model can be selected using the fit statistic behaviour as an aid. We consider the log-transformed self-reported alcohol consumption ($AC_{it}^* = \log(AC_{it} + 1)$) of $n=908$ individuals from a former longitudinal FMM study [43]. AC_{it} is the total volume of weekly consumption (in glasses) of subject i measured at four time intervals ($t = 1$ Youth: 12–18 years, $t = 2$ Young

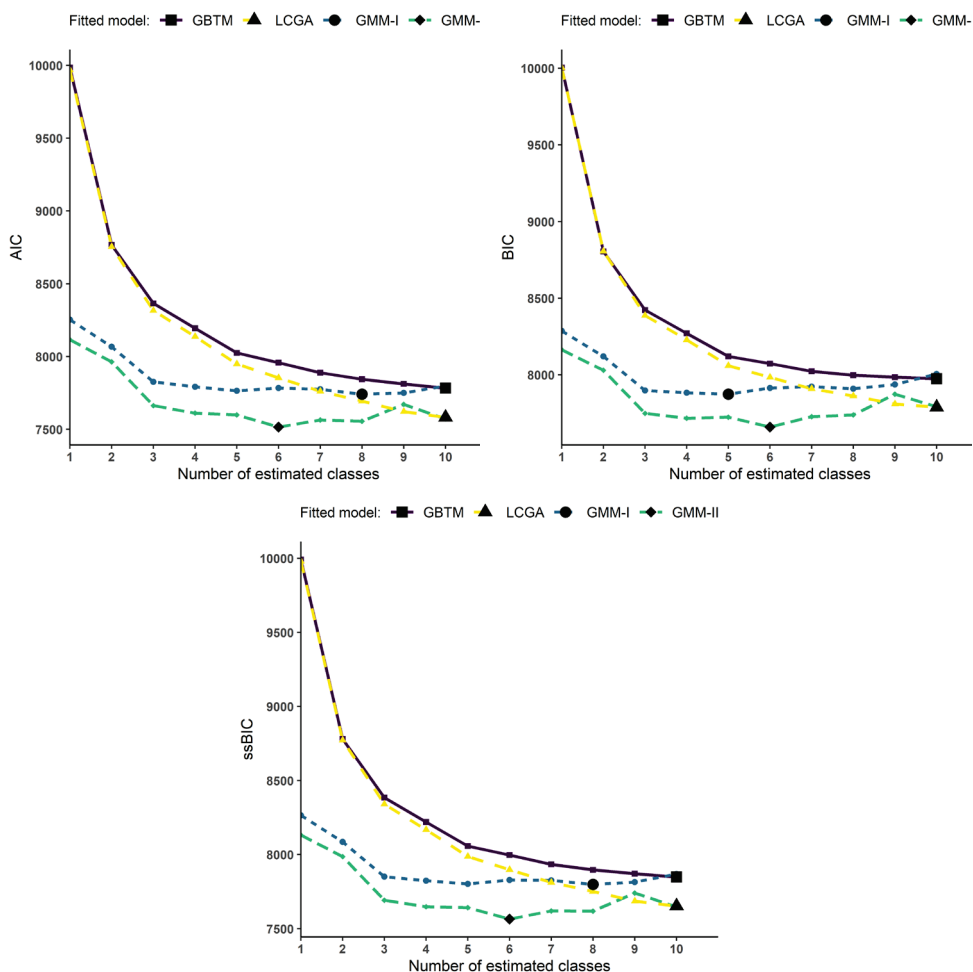
adult: 19–27 years, $t = 3$ Adult: 28–44 years, $t = 4$ Middle age: 45–60 years). The specifications of the models fitted to the data conform to GBTM, LCGA, GMM-I and GMM-II considered in this paper. Model pairings for comparison in line with our study would be GBTM with GMM-I and LCGA with GMM-II. In line with a previous study [140], each class trajectory is modelled as a quadratic function of time, such that,

$$AC_{it}^{*k} = (\beta_0^k + b_{0i}^k) + (b_{1i}^k + \beta_1^k)time_{it} + \beta_2^k time_{it}^2 + \varepsilon_{it}^k \quad (3.8)$$

with the specific polynomial and equation parameters conforming to **Eq. 3.7**. The objective of such an exercise is to identify classes of individuals following distinct trajectories of alcohol consumption over time.

The IC fit statistic curves of the estimated models are displayed in **Figure 3.8**. We exclude sE as we have established that it exhibits no discernible pattern in identifying covariance misspecification. A plateauing behaviour of the IC statistics for the fitted GBTM and LCGA is evident, as all curves have a minimum value at $K = 10$ (which is the maximum K examined). For the ICs, both GMM models show clear turning points at a K consistently lower than the associated GBTM or LCGA. The ICs for GMM-I suggest a K between 5 and 8 classes, whilst for GMM-II they all point to 6 classes. This preliminary evidence, taken together, hints at the LCGA and GBTM being underspecifications of the covariance structure of the data. They are, therefore, inconclusive. When presented with such fit statistic behaviour, the researcher is advised to explore models that allow for a more complex covariance structure. Accordingly, good candidate models to explore further include the GMM-I and GMM-II with their more general covariance structure. These models can then be further refined in sequential steps as has been suggested by other authors [50,99,121,140]. Such refinements would include (amongst others) inspection of models for non-convergence, non-identifiability, checking the significance of fixed effects parameters, class separation, and ascertaining distinctiveness of trajectories [140]. As an illustration, using the available OVL R code (in the **SM**), we computed the magnitude of the class separation among the $K = 6$ trajectories (shown in the **SM**) for the GMM-II model. This yielded OVLs ranging from 0.197 (between trajectories 2 and 4) to 0.73 (between 3 and 4), indicative of high to moderately low class separation levels (See **Appendix Table B.1**).

Figure 3.8: Fit statistic curves of estimated models (optimum fit statistic value at bold shape).



3.6. Discussion

3.6.1. Research questions recalled

Given the above results, we can answer our research questions:

- (1) Is a plateauing behaviour (or other peculiar behaviour) of the fit statistics under the fitted model a relic of covariance misspecification?

We find that underspecification of the **D** matrix (random effects structure) across all considered design conditions leads to a continual improvement, and associated plateauing

behaviour in the IC fit statistic (AIC, BIC, ssBIC) as fitted K increases. These underspecified models do not adequately capture the underlying variability (contained in \mathbf{D}), which increases the likelihood of over-extraction. This covariance misspecification is, thus, encapsulated as spurious latent classes [33]. No useful consistent pattern for sE fit curves across fitted models is easily identifiable.

- (2) How sensitive in terms of class enumeration are these fit statistics to covariance (\mathbf{D}) misspecification under various data features (e.g. class separation, number of time points)?

The ICs, especially the BIC and ssBIC, of over-specified and correctly specified models enumerate accurately under high separation. The sE under similar conditions performs worse than the BIC and ssBIC. However, under low separation, all fit statistics perform poorly with the ICs tending to under-extract whilst the sE overextracts. For all levels of \mathbf{D} misspecification, the effect of the number of repeated measures, time-variant versus time-invariant \mathbf{R} , and NS versus CC on correct K extraction by IC fit statistics is considerably lower than the effect of class separation.

- (3) Do identified fit statistic patterns assist in finding the correct model?

We posit that if the ICs of a fitted GBTM or LCGA continually improve as K increases, then this is indicative of covariance underspecification. This position is even more compelling if a GMM fitted to the same data yields a better fit in terms of IC fit statistics at a considerably lower K . In this case, the guidance provided by the ICs (namely the number of classes) for the GBTM or LCGA may be misleading and prone to over-extraction. A thorough investigation of the proposed covariance structure and model is then warranted.

If an over-specified GMM is fitted where an LCGA or GBTM would suffice, the value of the IC fit statistics for all three fitted models tends to be lowest and similar at the true K motivating the selection of the more parsimonious (i.e. GBTM or LCGA) model. This can be confirmed using likelihood ratio tests (LRTs) [140] such as the adjusted Lo-Mendell-Rubin LRT (aLMR) [78]. Finally, no identifiable diagnostic pattern for the sE was found.

3.6.2. Fresh insights and recommendations

Some further insights can be gleaned from our results, which debunk, confirm and/or

complement several widely held opinions about FMM class extraction:

- Firstly, although the ICs of an over-specified GMM are less likely to extract spurious classes than the ICs of an underspecified GBTM or LCGA under high separation, they, all perform poorly under low class separation. Here, the ICs of the over-specified GMM underextracts, whilst the ICs of the underspecified GBTM and LCGA continue to overextract, which complements established research [82,85,162]. Crucially, with low separation, the addition of random effects is not a panacea and could potentially collapse clinically meaningful classes with distinct patterns of change into a single class.
- Secondly, under high separation, the AIC shows a greater tendency to incorrectly extract classes (compared to BIC and ssBIC), even under correctly specified models. The BIC was the most accurate class enumeration fit statistic, followed by the ssBIC, but as with AIC, they tend to overextract when \mathbf{D} is underspecified. Under low separation, all fit statistics perform poorly.
- Thirdly, the use of sE for model selection during class enumeration has been cautioned against [163]. Our findings warrant this cautionary tone.
- Fourthly, the notion that the risk of over-extraction is high in particular for larger samples ($N > 1000$) [27,32,67,159] is not fully correct. We have shown that underspecification of \mathbf{D} can lead to over-extraction even for smaller sample sizes ($N = 260$).

Additionally, it must be emphasised that the fit statistic criteria only serve as a guide in determining the number of classes. The final decision of how many classes to extract is not an automatic process and demands considerable involvement from the researcher at every step of model fitting. This includes the judicious use of statistical analysis and substantive interpretation [99,121]. Considering our research findings, we recommend that:

- If a plateauing behaviour of ICs for GBTM and LCGA is evident, a visual inspection of the estimated mean trajectories within each class is warranted (particularly if practitioners do not have access to GMM capable software). Higher K solutions showing classes not substantively different from each other (e.g. trajectories that are either parallel, have especially low class separation, or exhibit very small or null class sizes) should be discarded and a more parsimonious model selected. Example code

for OVLs, i.e. the class separation index, is provided in the **SM**. Researchers are advised to compute the OVLs among the trajectories as an adjunct to assess the quality of class extraction.

- If there are multiple candidates for K based on the BIC, then likelihood ratio tests such as the adjusted Lo-Mendell-Rubin LRT (aLMR), substantive interpretation and visual inspection of trajectories could assist in further refining K [50,99,121,140].
- In scenarios suggesting underextraction, in particular under low class separation, researchers are advised to carefully evaluate the distinctiveness of the longitudinal profiles of candidate models, whilst considering their theoretical relevance. One could check for multimodality or a wide mode of the residual distribution per time point (for GBTM), or of the random effect distribution of the intercept and slope (for GMM), with deviations from normality being indicative of possible underextraction. Failure to address this may lead to wrong inferences such as an incorrect standard error of the class trajectory slope and the slope itself may be biased. We have, however, not explored this possibility in this paper.

3.6.3. Class enumeration: Reification and validation

In empirical sciences, FMMs are widely used for clustering purposes. Lesser known is that FMMs can be used to approximate oddly shaped distributions using a mixture of normal distributions, with specific applications in handling non-normal data including missing values [164,165] and outlier detection [166]. In a clustering context, however, this ability to approximate a non-normal distribution becomes a liability. In 2003, Bauer and Curran [36] drew attention to this by demonstrating that FMMs uncover spurious latent classes in one-class, non-normal data. Since then, other studies have replicated GMMs' over-extracting tendency [167] within a clustering context, with a solution for that developed and implemented using robust non-normal skewed distributions [103,168,169]. We did not address models' and fit-criteria' performances under violations of distributional assumptions, and further simulations should explore whether our findings can be replicated under such circumstances.

Moreover, the vicissitudes of class enumeration make the 'reification fallacy' [35] admonition as relevant as ever. The caution that one should refrain from interpreting latent classes as true entities, particularly in exploratory studies, is seldom misplaced. However, this issue pertains more to the external (as against internal) validation of the classes. For instance,

two recent developments in FMMS applications substantiate a more theoretically founded interpretation of identified trajectories, specifically through criterion validity (genotyping) [170,171] or replicability of findings (meta-analyses) [172]. In these cases, a more lenient posture towards classes' reification may be justified.

3.6.4. Limitations

We acknowledge the limitations of this study, which includes the focus on continuous repeated measures as encapsulated by a multivariate normal link function (as in **Eq. 3.2**). It would be instructive to investigate such emergent fit statistic patterns for different data types (e.g. count and binary data). Parameter recovery of the extracted trajectories when the correct K is selected was not considered as our focus was on establishing fit statistic curve behaviour under different \mathbf{D} misspecification. However, preliminary visual inspection of the trajectories when the known simulated K is selected (under high separation, correct and over-specified \mathbf{D}) suggests that the recovery of trajectories' temporal paths and class sizes is sound.

3.7. Conclusion

This paper has shown via extensive simulation that fit statistic curve behaviour can be a valuable diagnostic tool assisting model selection. Hence, practitioners of longitudinal FMM are advised to plot and inspect the fit statistics' patterns of change as a function of increasing K during class enumeration. These plots engender a better understanding of data features which underlie problematic behaviour of model fit statistical indices, helping to identify possible covariance misspecification. Notably, a continual improvement of the IC for fitted GBTM and LCGA as the researcher increases the number of classes is a clear indication of the models not adequately capturing the underlying covariance structure (underspecification), which then manifests into spurious latent classes. As a tool, these plots represent an additional step in following a transparent and methodological approach when fitting longitudinal FMM [6,99,121,140].

Finally, the OVL may serve an ancillary role in model fitting by first establishing the level of class separation between extracted trajectories, and thus the quality of class extraction before further refinements in the model fitting process. For cases of low separated classes, researchers will need to go beyond fit-criteria to transparently substantiate their choices, such as using complementary criteria (e.g. theoretical justification, residual plots).

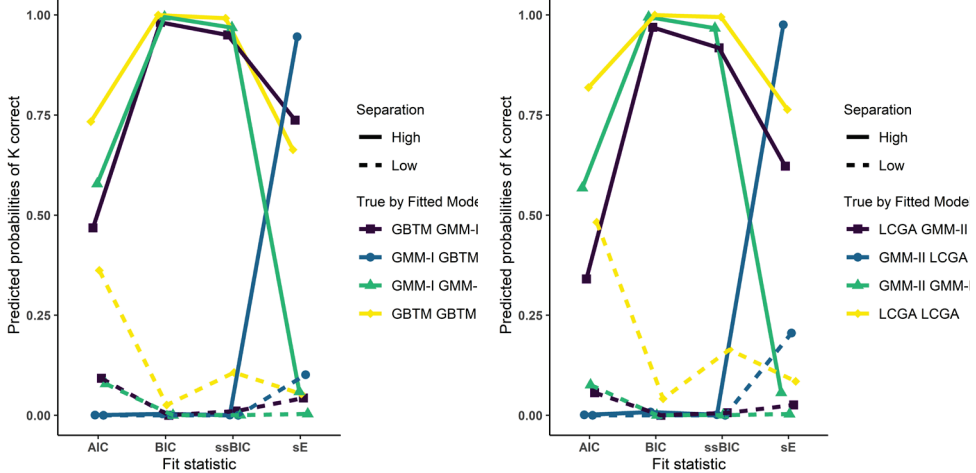
Software acknowledgements

Additional packages not previously cited but used to conduct our research includes; *brglm2* [173], *dplyr* [174], *ggeffects* [175], *lme4* [176], *stargazer* [177], *viridis* [178]

Appendix B.

Figure B.1: Estimated probabilities of K correct (upper half) or of over-extraction given incorrect K (lower half) for true by fitted model under low/high class separation given conditions: Cat's cradle, $T=5$ repeated measures. Left half concerns models GMM-I and GBTM, right half concerns models GMM-II and LCGA.

a.) K correct ($Y=1$ if K is correct, $Y=0$ if $K \neq 4$): Results from LM-I(a)-(b).



b.) K over ($Y=1$ if $K > 4$, $Y=0$ if $K < 4$): Results from LM-II(a)-(b).

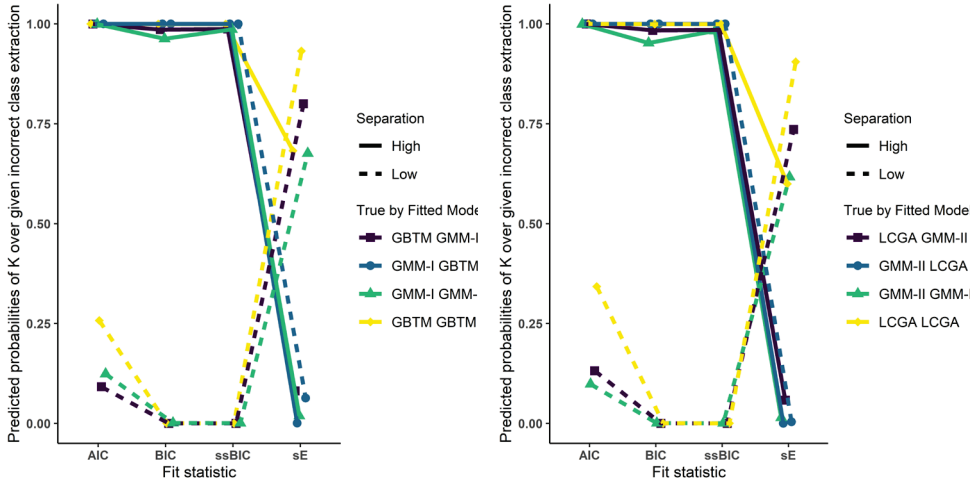
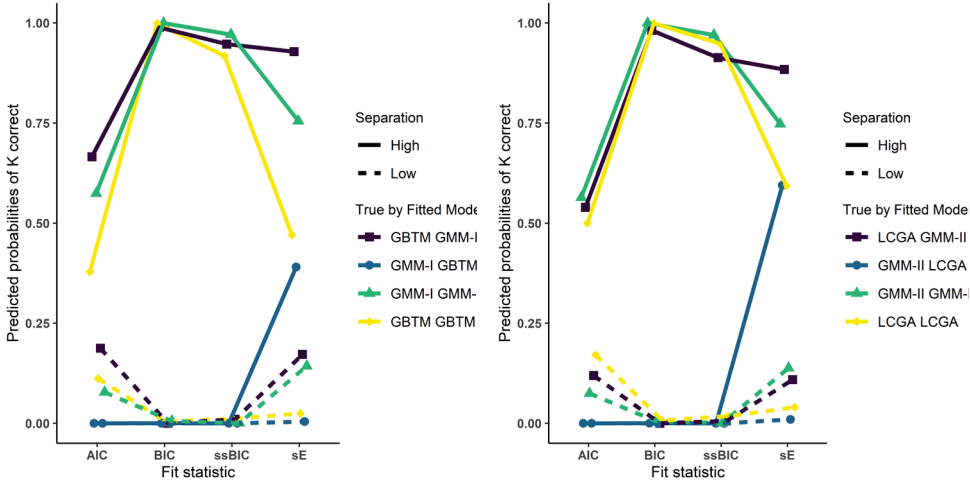


Figure B.2: Estimated probabilities of K correct (upper half) or of over-extraction given incorrect K (lower half) for true by fitted model under low/high class separation given conditions: Natural starting point, $T=5$ repeated measures. Left half concerns models GMM-I and GBTM, right half concerns models GMM-II and LCGA.

a.) K correct ($Y=1$ if K is correct, $Y=0$ if $K \neq 4$): Results from LM-I(a)-(b).



b.) K over ($Y=1$ if $K > 4$, $Y=0$ if $K < 4$): Results from LM-II(a)-(b).

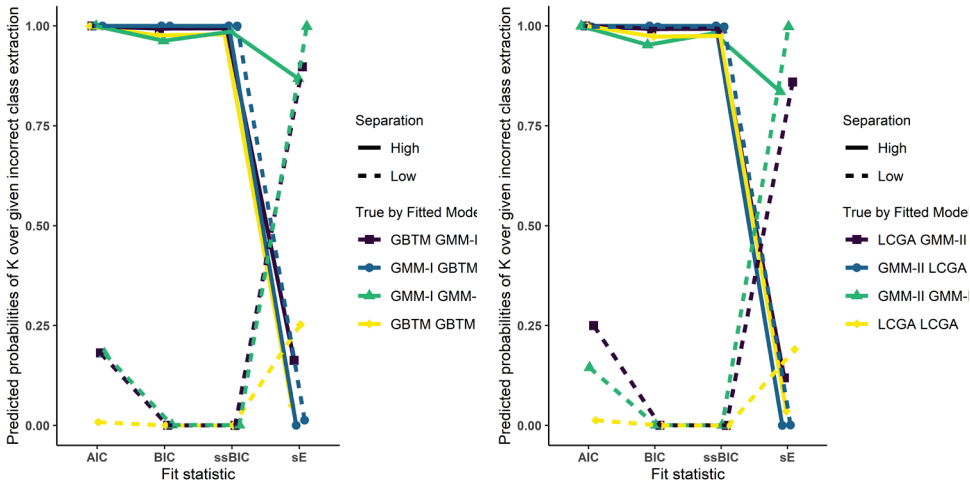
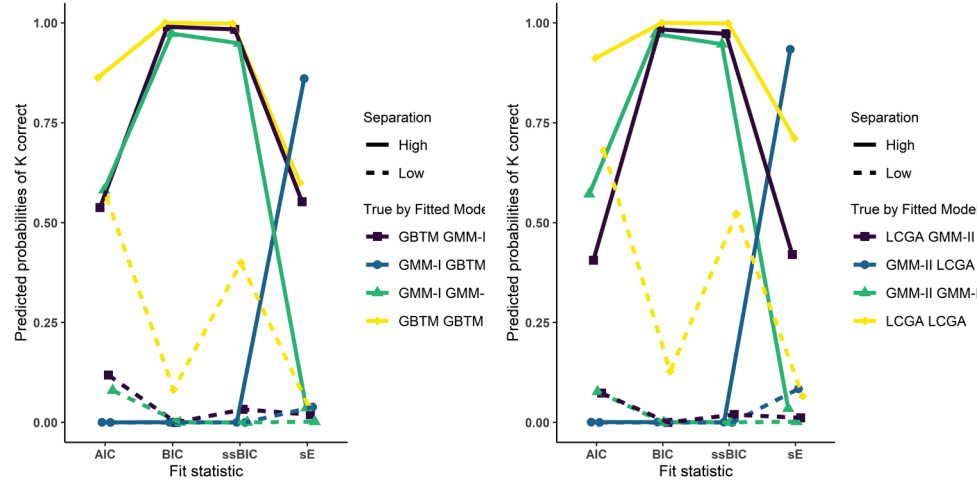


Figure B.3: Estimated probabilities of K correct (upper half) or of over-extraction given incorrect K (lower half) for true by fitted model under low/high class separation given conditions: Cat's cradle, $T=8$ repeated measures. Left half concerns models GMM-I and GBTM, right half concerns models GMM-II and LCGA.

a.) K correct ($Y=1$ if K is correct, $Y=0$ if $K \neq 4$): Results from LM-I(a)-(b).



b.) K over ($Y=1$ if $K > 4$, $Y=0$ if $K < 4$): Results from LM-II(a)-(b).

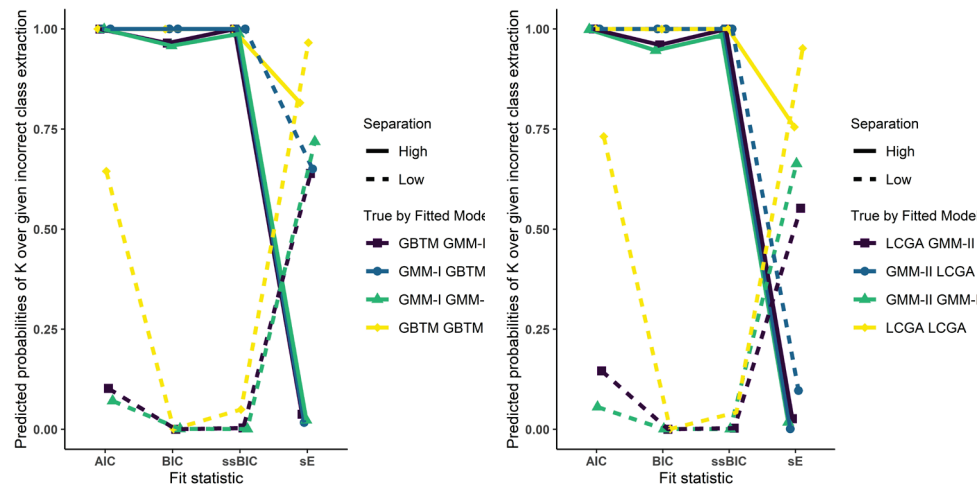


Figure B.4: Heatmaps of the proportion correct $K=4$ (panels A and C) and modal K (panels B and D) extracted by different fit statistics for different fitted models under time-variant R conditions (LCGA, GMM-II). Ordinate axis coded as H/L_NS/CC_8/5 indicating: Class separation: H(igh) or L(ow), Trajectory shape: N(atural) S(tart) or C(at's) C(radle), and Time points: 8 or 5. All panels: Quadrants clockwise from upper left: 1.) Underspecified LCGA 2.) Correctly specified GMM 3.) Over-specified GMM 4.) Correctly specified LCGA.

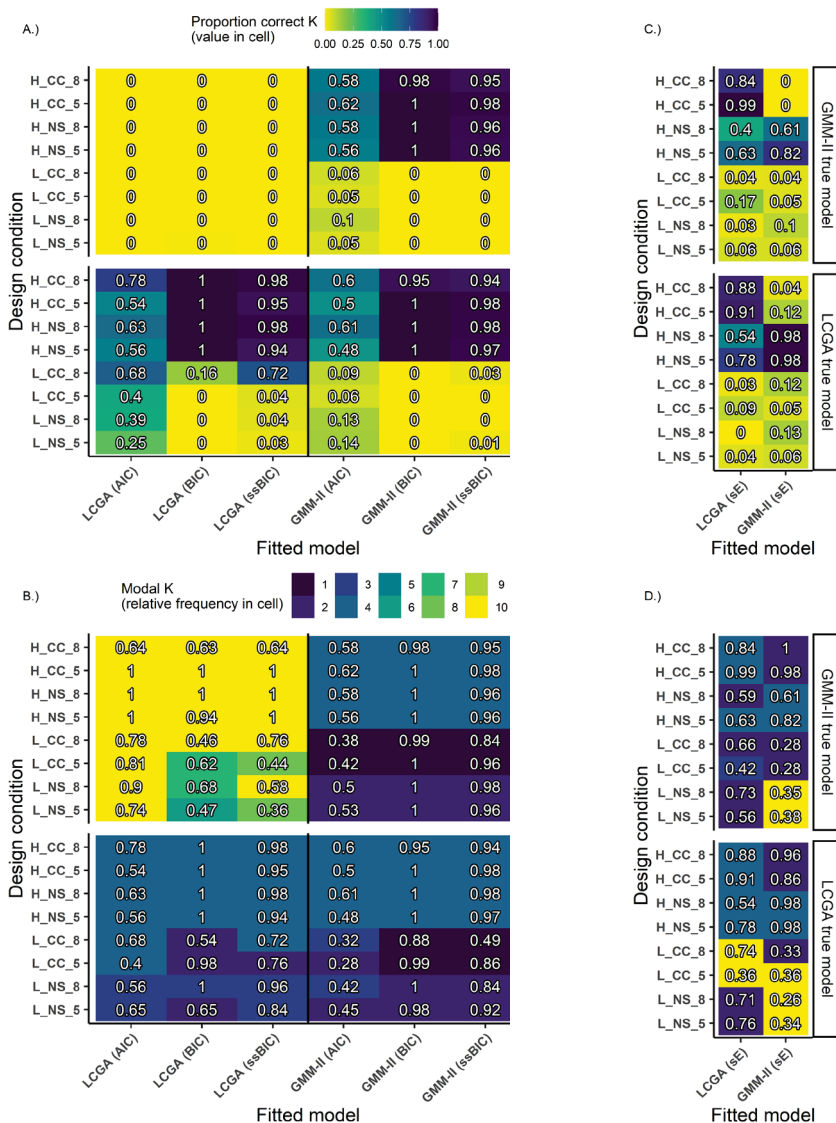


Figure B.5: Heatmaps of the proportion correct $K=4$ (panels A and C) and modal K (panels B and D) extracted by different fit statistics for time-invariant R with unequal class sizes (35%/15%/15%/35%) (GBTM, GMM-I). Ordinate axis coded as GMM-I/GBTM_NS/CC_H/L_T=8/5 indicating: True Model: GMM-I or GBTM, Trajectory shape: N(atural) S(tart) or C(at's) C(radle), Class separation: H(igh) or L(ow), and Time points: 8 or 5. All panels: Quadrants clockwise from upper left: 1.) Underspecified GBTM, 2.) Correctly specified GMM, 3.) Over-specified GMM, 4.) Correctly specified GBTM.

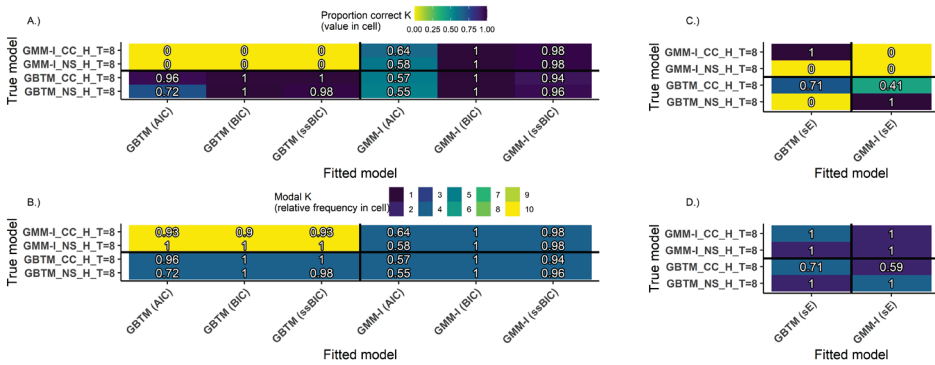


Figure B.6: Heatmaps of the proportion correct $K=4$ (panels A and C) and modal K (panels B and D) extracted by different fit statistics for time-invariant R for $N=260$ (GBTM, GMM-I). Ordinate axis coded as GMM-I/GBTM_NS/CC_H/L_T=8/5 indicating: True Model: GMM-I or GBTM, Trajectory shape: N(atural) S(tart) or C(at's) C(radle), Class separation: H(igh) or L(ow), and Time points: 8 or 5. All panels: Quadrants clockwise from upper left: 1.) Underspecified GBTM, 2.) Correctly specified GMM, 3.) Over-specified GMM, 4.) Correctly specified GBTM.

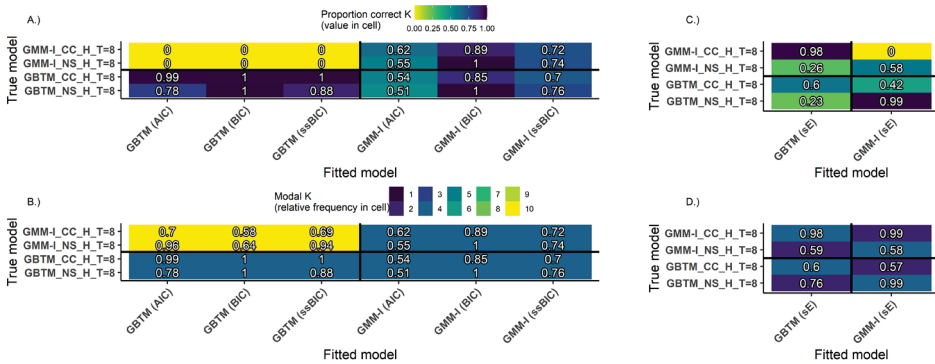


Table B.1: OVL for 6-class GMM-II. Each cell corresponds to the OVL between corresponding classes given in row and column headings.

Class	k=2	k=3	k=4	k=5	k=6
k=1	0.68	0.291	0.29	0.664	0.676
k=2		0.316	0.197	0.495	0.394
k=3			0.73	0.582	0.585
k=4				0.519	0.532
k=5					0.514

Appendix C.

Table of Contents of Supplementary Material

S.1	Model Parameters
S.2	Model convergence
S.3	Logistic regression results
S.3.1	Analysis results per true by fitted model combinations
S.3.2	Instances of data separation in simulation results
S.4	Results of simulations
S.4.1	Natural starting points
S.4.2	Cat's cradle
S.4.3	Unequal classes
S.4.4	Small sample cases
S.5	Overlap coefficient (OVL) exemplar code
S.6	Extracted trajectories of $K=6$ GMM-II in application
S.7	References

4. Univariate versus multivariate latent trajectory modelling: a comparison of class recovery performance

Gavin van der Nest
Math J.J.M. Candel
Gerard J.P. van Breukelen
Valéria Lima Passos


EMBARGOED



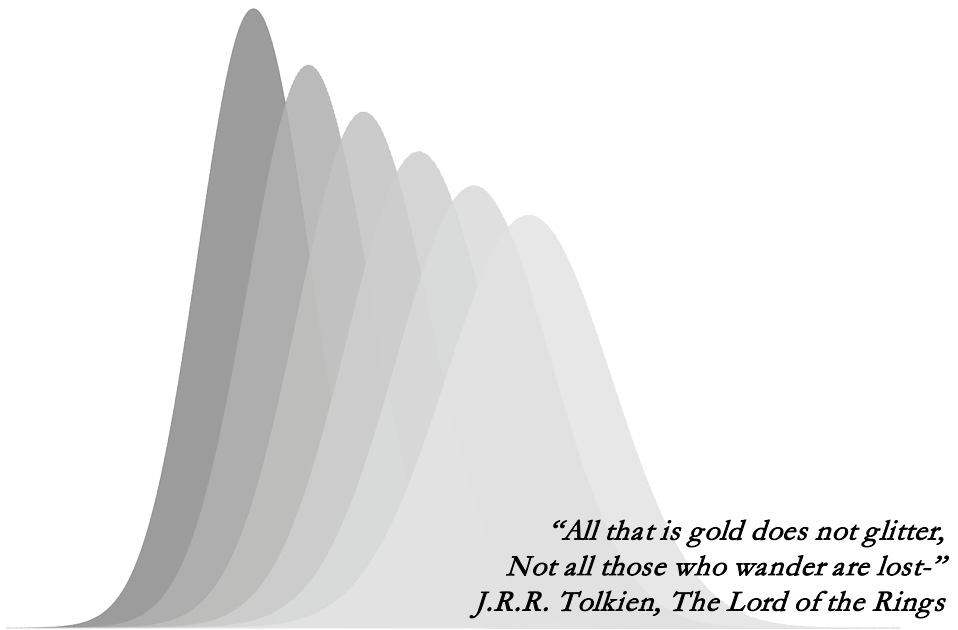
5. Robustness of multivariate longitudinal finite mixture models to covariance misspecification

Gavin van der Nest
Math J.J.M. Candel
Gerard J.P. van Breukelen
Valéria Lima Passos

EMBARGOED



6. General Discussion



Although finite mixture models (FMMs) have been in existence for well over a century, longitudinal FMMs are being increasingly used in applied sciences, particularly in medical and social sciences. Their popularity stems from their ability to uncover classes of subjects exhibiting different temporal development. This could, for example, be useful in establishing patient response to treatment [212], elucidating differences in mental health [213], or with the growing availability of multivariate data, exploring associations in temporal development between different outcomes (e.g. smoking frequency and drug use [184]). However, longitudinal FMM model fitting and selection is notoriously challenging as they can be affected by a multitude of factors. This thesis explores some of these factors and their effects, with the most striking being highlighted below.

6.1. Class enumeration and covariance misspecification

Class enumeration and the identification of proper structures to model the covariance pattern (or covariances) between repeated measures remain one of the greatest challenges in fitting longitudinal FMMs. Although multiple fit statistics were used throughout this thesis for these purposes, no one fit statistic performed best under all studied data conditions. Further, all fit statistics performed poorly when classes were lowly separated or when models had an underspecified covariance structure. The former condition was generally associated with the underextraction of classes, whilst the latter was associated with over-extraction. Although the Bayesian Information Criterion (BIC) was most robust given the models and data studied (**Chapters 2-5**), this thesis was not exhaustive and the same cannot be claimed for all possible data conditions. However, patterns recognised in this thesis offered some guidance in class enumeration and covariance structure selection.

It was demonstrated that the fit curve statistic behaviour of the Akaike Information Criterion (AIC), BIC, and sample-size adjusted BIC (ssBIC) could serve as a diagnostic tool for covariance misspecification during the process of class enumeration (**Chapter 3**). Specifically, a continual improvement and related plateauing (asymptotic) behaviour in the fit statistic curve as fitted K increased was associated with covariance underspecification and the possibility of class over-extraction. The behaviour of the fit statistic curve is consistent with unaccounted variability within the data which translates into spurious classes to capture heterogeneity by the misspecified models. This was evident in both univariate (**Chapter 3**) and multivariate models (**Chapter 5**). In univariate models, such underspecification was related to fitting group-based

trajectory models (GBTM) and latent class growth analysis (LCGA) to data generated by growth mixture models (GMM). In these instances, the underspecification was in assuming no random effects when in fact there were. Yet, it was established that the addition of random effects is not necessarily a panacea, since their inclusion, particularly under low separation, had the potential for underextraction, leading to the collapsing of what may be clinically meaningful classes with distinct patterns of change into a single class. For multivariate models, the same asymptotic behaviour was found when underspecified multivariate GBTM (GBMTM) and multivariate LCGA (MLCGA) were fitted to multivariate covariance pattern growth mixture models (MCPGMM) (**Chapter 5**). Such misspecification stems from the GBMTM's and MLCGA's assumption of conditional independence which ignores within-outcome associations across time. For these underspecified multivariate models, increases in the within-outcome correlation not only translated into poor class enumeration (i.e. asymptotic fit-criteria behaviour and potential over-extraction) but also a decline in the class (trajectory and size) recovery performance even when the true K was given (**Chapter 5**).

Thus, by carefully studying the fit statistic curve, a proper accounting of the covariance structure could be enforced. This includes remedial adjustments (i.e. relaxing covariance constraints or trimming classes) which could lead to better class enumeration accuracy and model accuracy. However, if the ability to generalise the covariance structure is absent in the software, extra caution on the side of the researcher when analysing results must be exercised. Unaccounted variability potentially results in spurious classes. Therefore, if classes for a selected higher K solution are not measuredly different (both in terms of development profiles over time and substantively) or cannot be justified theoretically or validated, then they should be discarded, and a more parsimonious model chosen. Here, the overlap coefficient (OVL) introduced in **Chapter 3** could serve an ancillary role as a determinant of class extraction quality. Nevertheless, as **Chapter 5** demonstrated empirically, even when the same K is selected for multivariate models with different covariance structures, different trajectories and markedly different class sizes can emerge from these models. It is therefore important that the validity of the chosen model is confirmed to lend credence to the extracted trajectories, such as through cross-validation, linking to a distal outcome, or meta-analyses.

6.2. Nuances between univariate and multivariate models

GBTM and GBMTM applied to the same data can yield considerably different results. Therefore, practitioners following the usual two-step process of first fitting GBTM and then GBMTM should handle the information obtained from the univariate models with care. The same univariate delineation may not be recovered by the multivariate model either in terms of class enumeration, or in trajectory profiles of development, or subject classification or class sizes, which would have interpretational consequences in practice.

We have illustrated that class extraction for multivariate models given univariate data can be considerably impacted by patterns of correlation, in terms of conditional class assignment, among outcomes (**Chapter 4**). As a rule, it is recommended for univariate models to be fitted before multivariate models to get a semblance of:

- (1) The underlying heterogeneity in each outcome as well as class separation,
- (2) The cross-combinations of univariate classes, which is indicative of the strength of association between outcomes, and
- (3) How (1) and (2) combined could inform the number of classes to expect in the multivariate model and whether the multivariate trajectories would reflect the univariate trajectories.

All meaningful cross-class combinations from the GBTM should be considered when determining the K of the GBMTM. The strength of such cross-class assignments could be quantified using Cramér's V (CV) [145,187]. Weak cross-class associations (low CV) would signal to fit a GBMTM up to K equal to the number of potential cross-classes from the GBTM analysis, whilst a higher CV would be indicative of a GBMTM requiring fewer classes than the number of potential univariate cross-classes to adequately capture the distinct patterns of co-development (joint evolution) between outcomes. GBMTM clustering may also be driven by the highest separated outcome, which could distort trajectory profiles in the low separated outcomes (**Chapter 4**). When this is suspected, it would be useful to calculate the OVL for a GBTM fitted to each outcome individually, with a high OVL (low class separation) signalling caution to the practitioner for further data exploration, including assessing the quality of class extraction.

Further, as **Chapter 4** shows, if outcomes are studied in isolation, then a univariate model, even under high separation, may fail to capture all the nuances of the data specifically

that of co-occurring temporal processes. The GBMTM uses the mutual information and potential correlations across multiple outcomes to improve model performance, even in cases of low class separation. Nevertheless, **Chapter 4** showed that outcomes need not be (linearly) correlated to yield patterns of joint development. Thus, there are interpretational consequences for extracted multivariate trajectories. The decision remains whether they should be interpreted as patterns of co-occurrences (no meaningful association, quirks of the data) or as co-development (joint evolution established as statistical association).

Moreover, the GBMTM in terms of class enumeration accuracy, subject classification performance, and class recovery (regarding trajectory and class size bias) was shown to be generally more robust than GBTM (**Chapter 4**). Specifically, GBMTM fitted to univariate data performed better than GBTM did on multivariate data. Even with low class separation in at most one of the outcomes in multivariate data, the GBMTM performed well and greatly outperformed the GBTM. This contrasts with univariate data where both models performed poorly under low class separation in at least one outcome. Such observations may motivate the application of multivariate models where feasible.

Although the use of specific models should be practically and theoretically motivated, the presence of (statistically significant) estimated within-class between-outcome correlation would lend some credence to the application of multivariate models. Moreover, the existence of statistically significant within-class within-outcome correlation supports an expanded covariance structure (**Chapter 5**).

6.3. The effect of data conditions

Throughout this thesis, class separation was shown to have had the largest impact on model performance across most scenarios. Models performed poorer as the level of class separation decreased. Not only did class enumeration accuracy tend to decrease (**Chapters 3-5**), but so too did classification accuracy and class recovery (**Chapter 4-5**). Although small sample sizes were generally not associated with decreases in class enumeration accuracy for the BIC (**Chapters 3-5**), they were associated with generally worse and more variable model performance specifically in classification accuracy and class recovery (trajectory and class size bias) (**Chapter 4-5**). Compared to class separation, the effect on model performance of between-outcome correlation (**Chapter 4-5**), number of repeated measures (**Chapter 3**), heterogeneity of the residual variance (**Chapters 3 and 5**), and trajectory fixed effect growth specification (i.e. different intercept and slopes) (**Chapter 3**) were inconspicuous.

6.4. The dilemma of class separation

All of the considered fit statistics (AIC, BIC, ssBIC and scaled Entropy (sE)) performed poorly in class enumeration in instances of low class separation studied in this thesis. Moreover, low class separation was also associated with worse classification accuracy, trajectory bias, and class size recovery (**Chapters 4-5**).

However, two distinct issues related to class separation must be highlighted. That is, the poor performance of the information criteria fit indices (and models in trajectory and class size recovery) because of low separation between true classes is an issue separate from low separation of extracted classes. For the first issue, we can only know the true separation if we have perfect knowledge of the data generating process including the true number of classes and trajectory specification. For the second issue, at least two possibilities exist; either 1.) the extracted classes are so close that to distinguish between them would not make substantive sense and in this case, one could do with fewer classes, or 2.) if the extracted classes already show low separation, even more classes may exist, but happen to be omitted due to still lower separation. For the issue regarding low separation of the extracted classes, the OVL could serve an ancillary role as a first check in establishing class separation and quality of the extracted trajectories, before further exploration of the data. Especially in instances where underextraction is suspected because of low separated extracted classes, researchers will have to go beyond fit-criteria to transparently substantiate a higher K selection. This would include using complementary criteria such as evaluating the distinctiveness of the trajectories, the theoretical justification thereof, and residual plot inspection (e.g. multimodal within-class distributions).

6.5. Ideas for future research

This thesis focused exclusively on continuous outcomes with the multivariate normal density function used to model the conditional distribution of the longitudinal data. It would be instructive to ascertain whether the results obtained in this thesis could be generalised to binary and count outcomes.

Another possibility is to study the effect of more than two outcomes on multivariate model performance. It would be interesting to see whether model performance deteriorates as the number of outcomes increases since this would necessitate the estimation of more

parameters. Moreover, in such cases, it would be useful to establish the maximum number of outcomes which can be reliably fitted by these models.

A further pursuit would be relaxing the class-invariant covariance structure constraints considered in this thesis and studying the effects thereof on multivariate models. Is there a point where such relaxations compromise multivariate model performance and/or convergence to a solution? One could also study the effect of different within- and between-outcome correlations among more than two outcomes e.g. weak association between two outcomes and a strong association between the others. How would such associations impact multivariate model enumeration accuracy and specifically could differences in such associations distort trajectory recovery?

Only linear associations between repeated measures and outcomes were studied in this thesis. It would be instructive to explore the consequences of non-linear associations on model performance. In doing so, it would be useful to develop guidelines for practitioners to identify such situations and how to properly account for these instances.

Missing data, and the degree and type of missingness, are avenues worthy of exploration. Specifically, one could study whether multivariate models outperform univariate models in terms of robustness to types of missingness under a variety of data conditions.

Longitudinal FMMs are notoriously sensitive to starting values for estimated parameters due to the complexity of the likelihood surface. Specifying a too narrow or too wide a range may either lead to a local instead of a global optimum solution or inadmissible solutions (e.g. negative variance estimates and/or null classes). Therefore, it may be useful to repeat the simulation studies conducted in this thesis to establish guidelines for determining an appropriate number of random starts, or what initial parameter values should be given, to improve the chances of reaching a global optimum.

6.6. Concluding thoughts

6.6.1. The necessity of random effects

GBTM and LCGA were developed to provide for the discretization of continuous random effects by approximating heterogeneity through the means of latent classes [214]. As advocates of these methods have argued, the inclusion of random effects, particularly when class average trajectories are of sole interest, unnecessarily complicates the model, which often leads to convergence and identifiability issues [27]. However, as we have shown, ignoring random



effects or a complex (residual) covariance structure can lead to class over-extraction.

Now, if the focus is exclusively on class average trajectories and not individuals, an alternative to random effects is to relax the independent and identically distributed assumption of the residual covariance for the GBTM/GBMTM and LCGA/MLCGA. Then, a covariance structure that directly relates the residuals of the repeated measures over time can be selected, which is the covariance pattern mixture models (CPMM) [27] (its multivariate counterpart, the MCPGMM was explored in this thesis). We have shown that MCPGMM, as with the univariate covariance pattern growth mixture model, can properly handle extra within-class variability and the issue of non-independence of repeated measures over time, without requiring random effects. This greatly reduces the added complexity and associated convergence issues of within-class random effect estimation in GMMs [32]. In GMMs, the covariance parameters are often set to be equal across classes to reach convergence (which is the default setting in *Mplus* [116]), but such constraints may have an impact on the biasedness of class trajectories and class enumeration [27].

6.6.2. Reification of classes and model validation

Throughout this thesis, we have followed the direct approach for the application of FMMs i.e., the assumption of K underlying classes to which a subject's longitudinal sequence belongs. However, there is an ongoing discussion of the interpretation of classes in FMMs, as either theoretical entities (direct approach) or as statistical summaries of the data (indirect approach) [215,216]. Furthermore, FMMs by design are capable of approximating oddly shaped distributions using a mixture of normal distributions and have been shown to extract spurious classes in one-class, non-normal data [36,167]. In these cases, classes should not be interpreted as theoretical entities, but as properties of the data, but distinguishing between these interpretations remains an outstanding challenge.

Thus, the eccentricities of class enumeration call special attention to the 'reification fallacy' [35] admonition, i.e., positing latent classes as real entities. As shown throughout this thesis, the process of class enumeration can face diminutions in accuracy resulting in either underextraction of classes or spurious class extraction. Therefore, during the exploratory stages of class enumeration and model selection, it is advised to not reify the classes. Once classes and the model are settled, it is of paramount importance that these classes are externally validated. Some recent examples of this validation with theoretically founded interpretations of

identified trajectories include criterion validity (genotyping, i.e. genetic associations with observed phenotypic trajectories) [170,171] or the replicability of findings through meta-analyses [172]. Here, where classes can be theoretically confirmed, a more lenient attitude to class reification may be warranted.

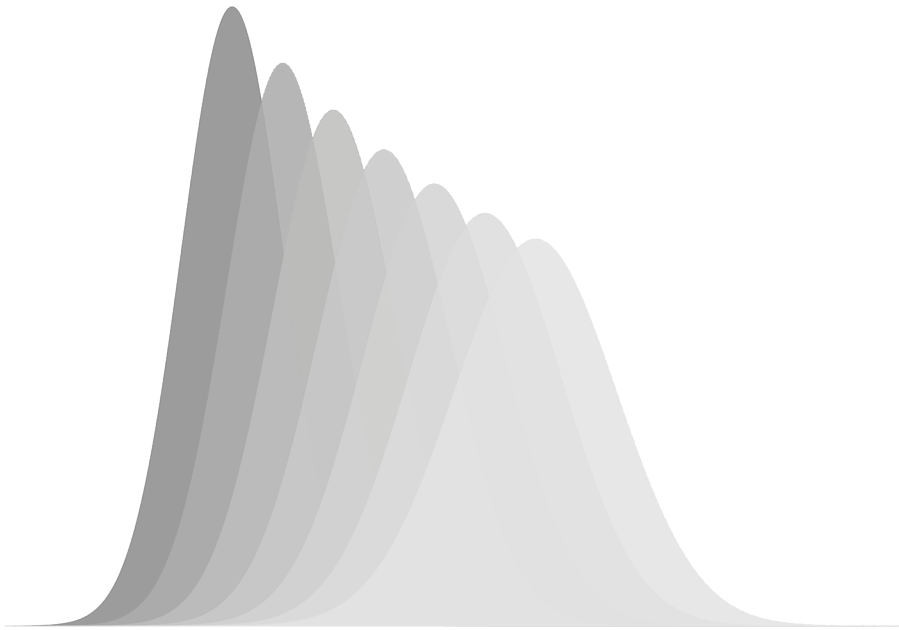
6.6.3. Conclusion

The golden thread of this thesis has interlaced two central aims. Firstly, this thesis sought to increase the accessibility to, appreciation for, and understanding of longitudinal FMMs by applied researchers and practitioners. Secondly, it studied the robustness of longitudinal FMMs performance to several forms of covariance misspecification under varying data conditions commonly found in practice. In pursuing these positions, several conceptual, theoretical, and statistical issues of longitudinal FMMs were explored in detail which shaped the proposed model fitting strategies. Moreover, this thesis has gone beyond typical univariate studies [27,28,30,32] which generally focus on class enumeration performance, by not only extending to studying multivariate FMMs but also covering the full ambit of model performance. This ranged from the accuracy of class enumeration and fit statistic criteria to the correct classification of subjects and class recovery (including trajectory and class size bias).

The fitting of longitudinal FMMs is a balance between two approaches; the science of statistical objectivity (guided by statistical criteria) and the art of model building (guided by domain knowledge, practical insights, and theoretical rigour). Both approaches certainly have value, but they are intertwined. There is no automated model selection strategy where some fit statistic should be optimised. Although statistical criteria lend some degree of objectivity to the process of model building, the blind application thereof can have the unattended consequence of not capturing the nuances in the data (through under- or over-extraction of classes, decreases in model performance, distorted patterns of co-development). As always, practitioners should be guided by the sensible application of these methods. This art-science divide is encapsulated in this thesis' cover. The *De Stijl* movement, founded by Dutch artists Theo van Doesburg (1883-1931) and Pieter Mondrian (1872-1944), prescribed the reduction of nature into the essentials of shape and colour. The analogue for longitudinal FMMs? Looking deeper into the data and simplifying it into the basic underlying elements of latent classes and profiles of temporal development to capture the underlying heterogeneity. To conclude, the epigraph of this chapter reminds us that not everything that is valuable or useful (hidden

profiles of development) is immediately apparent and that a healthy inquisitiveness in model selection does not imply a misguided path.

7. Summary



7.1. Summary in English

This thesis pertains to longitudinal finite mixture models (FMMs), which can identify classes of individuals following similar profiles of development over time (trajectories). These models are particularly useful in identifying distinct patterns of development when a grouping variable is either unknown (such as disease diagnosis given clinical measurements) or is expensive to measure (a rare genetic marker given phenotype). Thus, FMMs have great applicability in the age of precision medicine as identifying distinct latent classes of temporal development could assist practitioners in early diagnosis and/or tailored treatments. Although these models are gaining popularity in applied research, practitioners are often unaware of their underlying assumptions and/or fit them given software defaults. This thesis explores what the implications for model fitting are when models are improperly specified, particularly in the covariance structure, as well as provides guidance for practitioners to properly employ these models in their research.

Chapter 1 briefly discusses longitudinal FMMs and provides motivations for their use. A short historical context of FMMs is provided along with a discussion of the challenges in the application of longitudinal FMMs. The aims and objectives of the thesis are given, along with the general outline of the thesis.

Chapter 2 introduces commonly used longitudinal FMMs which comprise latent class growth analysis (LCGA), group-based trajectory models (GBTM), and growth mixture modelling (GMM). This chapter aims to address the confusion experienced by practitioners new to these methods by discussing the various available techniques in-depth and providing an overview of their interrelatedness and applicability. Criteria for model selection, specifically for class enumeration, and often encountered challenges and unresolved issues in model fitting are highlighted. Finally, model availability in software is showcased, and a model selection strategy using an applied example is illustrated.

Chapter 3 explores how data features as well as the inappropriate specification of an FMM's covariance structure impact class enumeration. To elucidate this, model fit criteria curve behaviour across an array of data conditions and covariance structures was investigated. Variable fit statistic patterns among the fit-criteria and across a range of data conditions were observed. This variability was greatly attributable to the level of class separation and the presence/absence of random effects. These findings support some widely held notions (e.g. the Bayesian Information Criterion outperforms other criteria) whilst debunking others

(adding random effects is not always the solution). Based on the obtained results, guidelines on how the behaviour of fit-criteria curves can be used as a diagnostic aid during class enumeration are presented.

Chapter 4 examines multivariate group-based trajectory models (GBMTM), which are gaining traction in empirical sciences. These models identify subjects following similar paths of temporal development across multiple outcomes. Customary analysis of multivariate data proceeds first with fitting univariate GBTM to each outcome and then fitting a multivariate GBMTM to capitalize on patterns of co-dependencies between outcomes. This procedure may yield differing univariate and multivariate trajectories, in one or several outcomes, in terms of the number and size of latent classes and the level and shape of trajectories. This chapter primarily investigates the impact of longitudinal data features on class enumeration and parameter recovery of GBMTM and GBTM when the data generating model is either GBMTM or GBTM. Consequently, the aim was to understand and elucidate the dynamics driving the discrepancies and similarities of these models' results. Based on the simulation findings, guidelines for the fitting of GBMTM are provided. Finally, this model fitting approach is illustrated, along with salient differences between the models, using an empirical data set.

Chapter 5 ascertains, through simulation, the effects of within-outcome covariance misspecification for GBMTM, multivariate LCGA (MLCGA) and multivariate covariance pattern growth mixture models (MCPGMM) under data conditions typically faced in practice. This is motivated by the fact that practitioners often restrict covariance structures to aid model convergence or run models according to software defaults (which usually constrain the covariance). The relative performances of these models are compared in terms of enumeration, classification, and class recovery. This chapter shows that restricted covariances, exacerbated by low class separation, can potentially lead to poor class enumeration. Moreover, despite the correct number of classes being chosen, variations in model performance across conditions emerged. Salient differences between the models, in terms of enumeration and class recovery, on an empirical data set are also illustrated.

Chapter 6 discusses the salient findings of this thesis. Further, the implications of the results are discussed, including the choice of fit statistic in class enumeration, the class separation dilemma, model selection and covariance specification. Ideas for future research are also presented. Finally, the chapter highlights the topical issues of the necessity for random effects, the reification of classes along with the challenge of model validation.

Chapter 8 considers the scientific and societal impact of this thesis.

7.2. Samenvatting in het Nederlands

Verborgen diepten: robuustheid van modelleringsbenaderingen voor het blootleggen van latente klassen in longitudinale data

Dit proefschrift behandelt longitudinale *finite mixture* modellen (FMMs), die klassen van individuen kunnen identificeren die vergelijkbare ontwikkelingsprofielen over de tijd hebben (trajecten). Deze modellen zijn met name nuttig bij het identificeren van verschillende ontwikkelingspatronen wanneer een classificatievariabele ofwel onbekend is (zoals de ziektediagnose bij klinische metingen) of duur is om te meten (een zeldzame genetische marker bij fenotypes). FMMs hebben dus een grote toepasbaarheid in het tijdperk van precisiegeneeskunde, aangezien het identificeren van verschillende latente klassen van ontwikkeling over de tijd kan helpen bij vroege diagnose en/of op maat gemaakte behandelingen. Hoewel deze modellen steeds populairder worden in toegepast onderzoek, zijn onderzoekers zich vaak niet bewust van hun onderliggende aannames en/of passen ze deze toe met de standaardinstellingen van software. Dit proefschrift onderzoekt wat de implicaties zijn voor modelschatting wanneer modellen onjuist zijn gespecificeerd, met name wat betreft de covariantiestructuur, en biedt toegankelijke richtlijnen voor onderzoekers om deze modellen op de juiste manier in hun onderzoek te gebruiken.

Hoofdstuk 1 bespreekt kort longitudinale FMMs en geeft een motivatie voor het gebruik ervan. Er wordt een korte historische context van FMMs gegeven, als ook een bespreking van de uitdagingen bij de toepassing van longitudinale FMMs. De doelstellingen van het proefschrift worden gegeven, alsmede een algemene schets van het proefschrift.

Hoofdstuk 2 introduceert veelgebruikte longitudinale FMM's: *latent class growth analysis* (LCGA), *group-based trajectory models* (GBTM) en *growth mixture models* (GMM). Dit hoofdstuk is bedoeld om helderheid te geven aan onderzoekers, voor wie deze methoden nieuw zijn, door de verschillende beschikbare technieken diepgaand te bespreken en een overzicht te geven van hun onderlinge samenhang en toepasbaarheid. Criteria voor modelselectie, met name voor het bepalen van het aantal klassen, en vaak voorkomende issues en onopgeloste problemen bij het schatten van modellen worden besproken. Tenslotte wordt de beschikbaarheid van modellen in software beschreven en wordt een strategie voor modelselectie geïllustreerd met behulp van een empirisch voorbeeld.

Hoofdstuk 3 onderzoekt hoe kenmerken van de data en een onjuiste specificatie van de covariantiestructuur van een FMM van invloed zijn op het vaststellen van het aantal klassen. Om dit op te helderen, werd het gedrag van de modelfit-criteria curve voor verschillende data condities en covariantiestructuren onderzocht. Verschillende patronen in fit-grootheden werden waargenomen voor verschillende fit-criteria en voor verschillende data condities. Deze variatie was in hoge mate toe te schrijven aan de mate van klassenscheiding en de aanwezigheid van random effecten. Deze bevindingen ondersteunen enkele wijdverbreide opvattingen (bijv. het Bayesian Information Criterion presteert beter dan andere criteria), terwijl andere worden ontkracht (het toevoegen van random effecten is niet altijd de oplossing). Op basis van de verkregen resultaten worden richtlijnen gepresenteerd over hoe het gedrag van fit-criteria curves kan worden gebruikt als diagnostisch hulpmiddel om het aantal klassen te bepalen.

Hoofdstuk 4 onderzoekt multivariate *group-based trajectory models* (GBMTM), welke steeds meer terrein winnen in de empirische wetenschappen. Deze modellen identificeren klassen die vergelijkbare paden van temporele ontwikkeling volgen over meerdere uitkomsten. De gebruikelijke analyse van multivariate gegevens schat als eerste een univariate GBTM voor elke uitkomstvariabele en vervolgens een multivariate GBMTM gebruikmakend van patronen van afhankelijkheden tussen de uitkomsten zoals gevonden in de univariate analyse. Deze procedure kan verschillende univariate en multivariate trajecten opleveren, in een of meerdere uitkomsten, in termen van het aantal en de grootte van latente klassen en het niveau en de vorm van trajecten. Dit hoofdstuk onderzoekt voornamelijk de impact van de kenmerken van longitudinale data op de bepaling van het aantal klassen en het terugvinden van de ware parameters van GBMTM en GBTM terwijl het data-model GBMTM of GBTM is. Het doel is om de dynamiek die de discrepanties en overeenkomsten tussen de resultaten van deze modellen veroorzaakt, te begrijpen en op te helderen. Op basis van de simulatieresultaten worden richtlijnen gegeven voor de toepassing van GBMTM. Tenslotte wordt deze richtlijnen geïllustreerd en worden opvallende verschillen tussen de analyseresultaten van beide modellen besproken aan de hand van een empirische dataset.

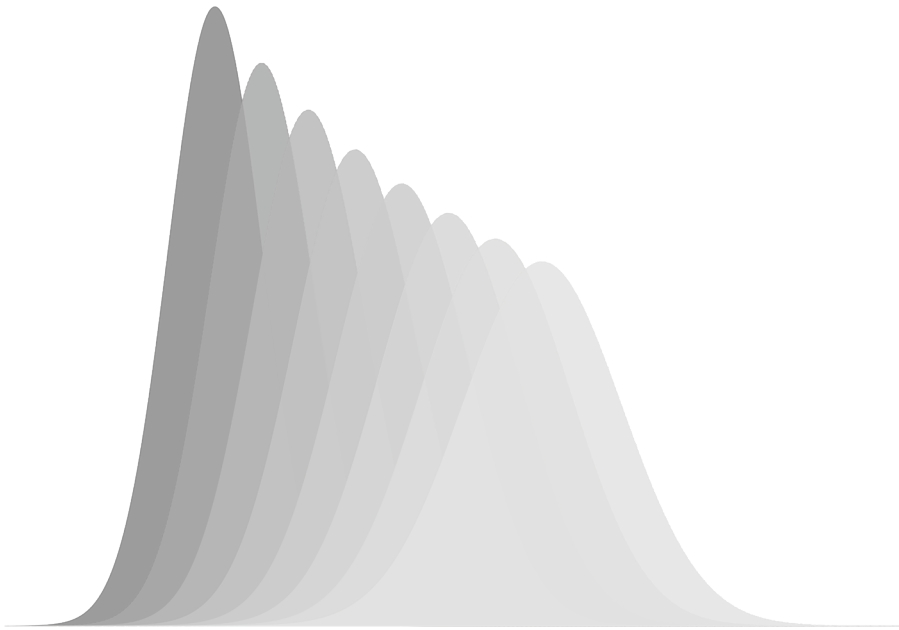
Hoofdstuk 5 stelt, door middel van simulatie, de effecten vast van misspecificatie van de covariantiestructuur voor de uitkomstvariabele voor GBMTM, multivariate LCGA (MLCGA) en multivariate *covariance pattern growth mixture models* (MCPGMM) onder data condities die typisch zijn voor de praktijk. Dit wordt gemotiveerd door het feit dat onderzoekers vaak eenvoudige covariantiestructuren specificeren om modelconvergentie te

bevorderen of modellen schatten volgens standaardinstellingen van de software (die gewoonlijk uitgaan van eenvoudige covariantiestructuren). De relatieve prestaties van deze modellen worden vergeleken wat betreft de bepaling van het aantal klassen, de classificatie van personen en het terugvinden van de klassen. Dit hoofdstuk laat zien dat eenvoudige covariantiestructuren, verergerd door geringe scheiding tussen klassen, kunnen leiden tot vaststelling van een onjuist aantal klassen. Bovendien, zelfs als het juiste aantal klassen wordt gekozen, komen er verschillen in modelprestaties naar voren voor de onderzochte condities. Opvallende verschillen tussen de modellen, in termen van de vaststelling van het aantal klassen en het terugvinden van de klassen, worden ook op een empirische dataset geïllustreerd.

Hoofdstuk 6 bespreekt de belangrijkste bevindingen van dit proefschrift. Verder worden de implicaties van de resultaten besproken, waaronder de keuze van de fit-grootheid bij de bepaling van het aantal klassen, het dilemma wanneer klassen dicht bij elkaar liggen, modelselectie en het specificeren van de covariantiestructuur. Ook worden ideeën voor toekomstig onderzoek gepresenteerd. Tenslotte belicht het hoofdstuk de noodzaak van random effecten, de reïficatie van klassen en het belang van modelvalidatie.

Hoofdstuk 8 gaat in op de wetenschappelijke en maatschappelijke impact van dit proefschrift.

8. Scientific and social impact of this thesis



For longitudinal data (i.e. multiple measurements recorded per subject over time), the sole focus of this thesis, finite mixture models (FMMs) can assist practitioners in identifying different classes/groups of subjects following distinct paths of temporal development (trajectories) in the absence of a known grouping variable. This is advantageous in situations where a grouping variable is either unknown (e.g. disease diagnosis given clinical measurements) or expensive to measure (e.g. a rare epigenetic marker given observed phenotypes). Some recent examples of fields in which FMMs have been applied include psychology [217–219], public health [220,221], and medicine [222–224]. FMMs are of practical importance since by identifying distinct groups, a better understanding may arise of differences in: development between groups (i.e. alcohol consumption over time), possible outcome events (e.g. by linking classes to a distal outcome such as the occurrence of myocardial infarction) and/or risk factors (e.g. by linking to covariates such as sodium intake). Moreover, by considering multivariate trajectories, the association between and the development of several outcomes can be simultaneously explored. This is helpful when studying the natural progression of complex, multi-dimensional diseases, where multivariate trajectories could account for various biomarkers over time, the occurrence of clinical endpoints (e.g. a distal outcome such as death), and heterogeneity over time between patients [225]. Finally, insights gleaned from a longitudinal FMM analysis could have important policy or treatment implications. An example would be developing targeted interventions as a result of uncovering trajectories of childhood diet and their link to cigarette usage in adulthood.

Hence, because of these potential implications and to develop an accurate understanding of dynamics driving differences in outcomes between individuals, it is important that when fitting these models, the underlying classes are well-defined and correctly extracted. This is more likely achieved by ensuring that the chosen statistical model is correctly specified such that it is an accurate representation of the underlying process that generated the observed data. Large differences between true and extracted classes could have direct consequences. Minimising these differences is important for several reasons. Firstly, enumeration (that is the number of classes extracted) accuracy ensures that the correct number of classes are identified such that targeted interventions/treatments are provided for the correct number of groups. Secondly, accurate classification may yield improvements in personalised treatment and intervention quality [226]. Thirdly, correct trajectory recovery is important since it provides for an accurate depiction of the development over time which again could have practical implications including the nature of the treatment provided and/or gaining a proper

understanding of the underlying temporal development. Lastly, accurate class size recovery ensures that the proportion of individuals within each class gives an accurate composition of the population under study. This could have utility when establishing the occurrence of rare behavioural conditions or gaining insights into the developmental profile frequency of specific behaviours over time.

Considering the above, the main objective of this thesis was to study the effects of various longitudinal FMM (mis)specifications under differing data conditions (commonly found in practice) on model performance. In so doing, model selection strategies were developed and presented to ensure good model performance and to assist practitioners and applied researchers in their understanding and application of longitudinal FMMs in their research. Model performance was gauged according to class enumeration accuracy (i.e. correctly identifying the underlying number of classes), and by extension classification accuracy (i.e. whether subjects are assigned to the correct class), trajectory (i.e. shape and level of the development profile over time) and class size recovery (i.e. the proportion of subjects comprising each latent class). Factors explored in this thesis which potentially impact model performance included class separation levels, sample size, different trajectory specifications (in the shape and level), number of repeated measures, and model specifications including random effects, time-variant variances, and within- and between-outcome correlation.

This thesis studied both univariate (i.e. considering one measure such as alcohol consumption) and multivariate (i.e. considering multiple measures simultaneously such as alcohol consumption and marijuana use) outcomes, the latter of which is gaining prominence as increasing numbers of studies consider the developmental dynamics of several outcomes simultaneously. Across all studied outcomes and conditions, low class separation (i.e. high overlap between classes) and covariance underspecification (i.e. fitting a model which does not account for all of the heterogeneity and dependencies in the data) were identified as major factors affecting model performance. The former factor was generally associated with the underextraction of classes (i.e. too few), whilst the latter tended to be associated with class over-extraction. The fit statistic curve, which shows how the value of a certain statistical criterion used for class enumeration changes as the fitted number of classes increases, was identified as a useful diagnostic tool for potential underspecification of the covariance. If the curve continued to improve as the fitted number of classes increased whilst showing a so-called plateauing (asymptotic behaviour), then this could be taken as evidence of possible covariance underspecification. In such cases, it is suggested to either relax constraints on the



covariance if the software allows, or if not, a more parsimonious model (i.e. fewer classes) should be chosen if the extracted trajectories are not substantively different, or cannot be theoretically justified or validated (such as through cross-validation).

Further, we studied the comparative performance of multivariate and univariate longitudinal FMMs. Multivariate models were found to be generally more robust than univariate models in that they performed better on univariate data than univariate models did on multivariate data in class enumeration accuracy, classification accuracy, and trajectory and class recovery. Additionally, we showed that for multivariate trajectories, the clustering may be driven by the outcome with higher separated classes which might distort classes in the lowly separated outcome, and thus could have interpretational and practical implications.

Moreover, we studied the relative performance of several multivariate FMMs which differed in the restrictions placed on their covariance structure. We showed that multivariate models with restricted covariance structures, exacerbated by low class separation can potentially lead to poor class enumeration. Moreover, even when classes were correctly enumerated, variations in model performance across conditions emerged. These results provided for a better understanding of factors driving good multivariate model performance which allowed us to establish some model selection guidelines for practitioners to follow in their research.

The scientific impact of this thesis includes presenting the behaviour of the fit-criteria curve as a diagnostic tool for covariance misspecification and remediation. Further, not only was class enumeration accuracy studied, but the full ambit of model performance covering classification accuracy, and trajectory and class size recovery were studied, which goes beyond typical univariate studies [30,32,85,97,98] and as far as we are aware of the first of its kind for multivariate studies [27,28,30,32]. Novelty, the area between the curves (ABC) was presented as a measure of trajectory recovery. The ABC is especially useful in Monte Carlo simulation studies, where statisticians may be interested in the bias of trajectory recovery, with low ABC signalling good trajectory profile recovery. Additionally, the ABC was also employed for the first time to address the class label switching problem (where class labels switch between successive simulation runs), an issue especially prolific in longitudinal FMM simulation studies [190]. We intend to further develop the ABC to mitigate class label switching and present it as a statistical package for interested parties to use in their research. The overlap coefficient (OVL) was also suggested as a new measure for the quality of class extraction with a high OVL

(indicative of high class overlap) signalling caution on the side of the user and calling for a deeper exploration of the data and classes extracted.

The societal impact is derived from the guidelines and results established in this thesis. In proposing model selection and remediation strategies, we hope that these will lead to better model specification by practitioners which could have a direct impact on the veracity of inferences derived from the longitudinal FMMs fitted, particularly by minimising extraneous class enumeration and poor model performance. Ensuring such veracity has great utility, specifically in the field of health and life sciences, where accurate diagnostic and prognostic conclusions derived from statistical models are essential: for advancing science, for when patient care decisions need to be made by clinical practitioners, and for improving the quality and reducing the costs of healthcare through informed decision making by administrators and policy makers[227].

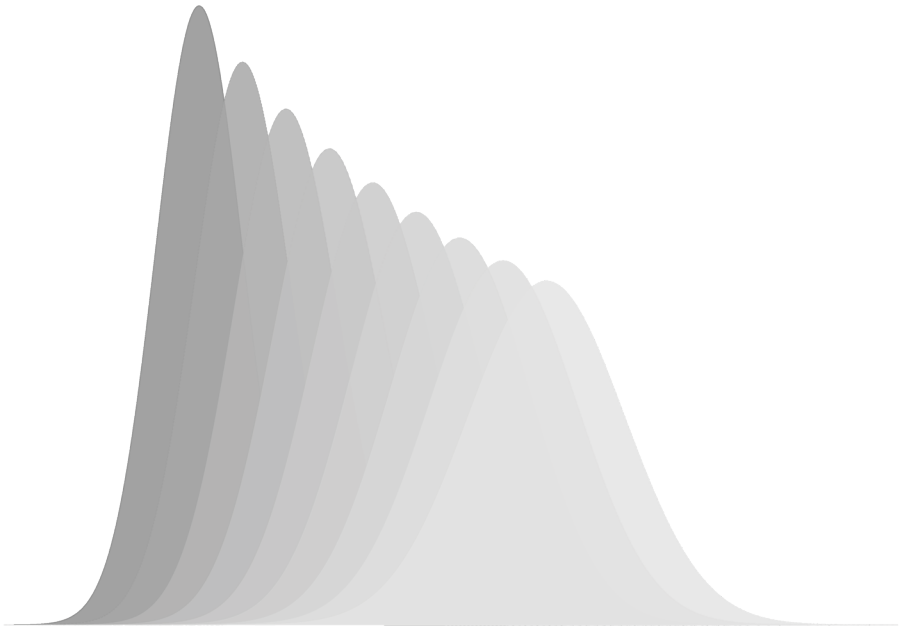
These research results are interesting for practitioners of longitudinal FMMs who apply such models to better understand the heterogeneity in their datasets and where an accurate representation of such heterogeneity is a necessity for correct treatment and/or interventions. The application sections of this thesis, along with the provided *R* and *Mplus* code for the models considered and investigational statistics employed, may serve as a starting point for practitioners and applied statisticians in their research. Also, the in-depth discussion and exposition of the various longitudinal FMMs may serve as teaching material for advanced courses in applied classification statistics and/or investigational statistics. Moreover, the OVL and ABC presented in this thesis may be useful tools for statisticians to employ during their own Monte Carlo simulation studies.

To increase the accessibility of this research, Chapter 2 has been presented at a statistical colloquium, presented as a poster at an (online) international conference, and was also presented during an online seminar at an international university. Chapter 3 has also been presented at a statistical colloquium. Further, Chapter 2 and 3 have both been published in international scientific journals, with Chapter 2 garnering citations in diverse fields including epidemiology [228], gerontology [3], nutrition [4], psychiatry [2], finance [229] and public health [230]. Chapter 4 is in the process of submission to a scientific journal. Chapter 5 will thereafter be submitted for publication in an international scientific journal. Moreover, the various statistical visualisations and investigational statistics developed in this thesis could be contained in a user-friendly package for *R* so that researchers may freely and easily use it in



their research. This package could be accompanied by a tutorial or non-technical paper published in an international journal to make these methods accessible to applied researchers.

9. References



- [1] Ellwardt L, Präg P. Heterogeneous mental health development during the COVID-19 pandemic in the United Kingdom. *Sci Rep* [Internet]. 2021;11:15958. Available from: <https://doi.org/10.1038/s41598-021-95490-w>.
- [2] Salagre E, Grande I, Solé B, et al. Exploring Risk and Resilient Profiles for Functional Impairment and Baseline Predictors in a 2-Year Follow-Up First-Episode Psychosis Cohort Using Latent Class Growth Analysis. *J Clin Med* [Internet]. 2020;10:73. Available from: <https://www.mdpi.com/2077-0383/10/1/73>.
- [3] Duim E, Lima Passos V. Highways to Ageing - Linking life course SEP to multivariate trajectories of health outcomes in older adults. *Arch Gerontol Geriatr* [Internet]. 2020;91:104193. Available from: <https://doi.org/10.1016/j.archger.2020.104193>.
- [4] Doggui R, Ward S, Johnson C, et al. Trajectories of Eating Behaviour Changes during Adolescence. *Nutrients* [Internet]. 2021;13:1313. Available from: <https://www.mdpi.com/2072-6643/13/4/1313>.
- [5] Reinecke J, Seddig D. Growth mixture models in longitudinal research. *AStA Adv Stat Anal* [Internet]. 2011;95:415–434. Available from: <http://link.springer.com/10.1007/s10182-011-0171-4>.
- [6] Nagin DS. *Group-Based Modeling of Development* [Internet]. Cambridge, MA and London, England: Harvard University Press; 2005. Available from: <http://www.degruyter.com/view/books/9780674041318/9780674041318/9780674041318.xml>.
- [7] Bauer DJ. Observations on the Use of Growth Mixture Models in Psychological Research. *Multivariate Behav Res* [Internet]. 2007;42:757–786. Available from: <https://www.tandfonline.com/doi/full/10.1080/00273170701710338>.
- [8] Theodoridis S, Koutroumbas K. *Clustering Algorithms III: Schemes Based on Function Optimization*. *Pattern Recognit* [Internet]. Elsevier; 2009. p. 701–763. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9781597492720500165>.
- [9] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning* [Internet]. New York, NY: Springer New York; 2009. Available from: <http://link.springer.com/10.1007/b94608>.
- [10] Titterton DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley Ser. Probab. Math. Stat. Chichester: Wiley; 1985.
- [11] McLachlan G, Peel D. *Finite Mixture Models* [Internet]. Wiley Ser. Probab. Stat. Appl.

- Probab. Stat. Sect. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2000. Available from: <http://doi.wiley.com/10.1002/0471721182>.
- [12] Ng JWY, Barrett LM, Wong A, et al. The role of longitudinal cohort studies in epigenetic epidemiology: Challenges and opportunities. *Genome Biol.* 2012;13:1–13.
- [13] Nagin DS, Jones BL, Passos VL, et al. Group-based multi-trajectory modeling. *Stat Methods Med Res* [Internet]. 2018;27:2015–2023. Available from: <http://journals.sagepub.com/doi/10.1177/0962280216673085>.
- [14] Newcomb S. A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *Am J Math* [Internet]. 1886;8:343. Available from: <https://www.jstor.org/stable/2369392?origin=crossref>.
- [15] Pearson K. Contributions to the Mathematical Theory of Evolution. *Philos Trans R Stat Soc London A* [Internet]. 1894;185:71–110. Available from: <https://www.jstor.org/stable/90667>.
- [16] Rao CR. The Utilization of Multiple Measurements in Problems of Biological Classification. *J R Stat Soc Ser B* [Internet]. 1948;10:159–193. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1948.tb00008.x>.
- [17] Wolfe JH. PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS. *Multivariate Behav Res* [Internet]. 1970;5:329–350. Available from: http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr0503_4.
- [18] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J R Stat Soc Ser B* [Internet]. 1977;39:1–22. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x>.
- [19] Nagin DS, Land KC. Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model. *Criminology.* 1993;31:327.
- [20] Land KC, Nagin DS. Micro-models of criminal careers: A synthesis of the criminal careers and life course approaches via semiparametric mixed poisson regression models, with empirical applications. *J Quant Criminol* [Internet]. 1996;12:163–191. Available from: <https://www.taylorfrancis.com/books/9781351552554/chapters/10.4324/9781315089256-17>.
- [21] Nagin D, Tremblay RE. Trajectories of Boys' Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency.

- Child Dev [Internet]. 1999;70:1181–1196. Available from:
<https://onlinelibrary.wiley.com/doi/10.1111/1467-8624.00086>.
- [22] Moffitt TE. Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. Psychol Rev [Internet]. 1993;100:674–701. Available from:
<https://www.taylorfrancis.com/books/9781351573610/chapters/10.4324/9781315096278-3>.
- [23] Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. Psychol Methods [Internet]. 1999;4:139–157. Available from:
<http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.4.2.139>.
- [24] Muthén BO, Shedden K. Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. Biometrics [Internet]. 1999;55:463–469. Available from:
<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.1999.00463.x>.
- [25] Jones BL, Nagin DS, Roeder K. A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. Sociol Methods Res. 2001;29:374–393.
- [26] Muthén LK, Muthén BO. MPlus User’s Guide. Eighth Edi. Los Angeles, CA: Muthén & Muthén; 2017.
- [27] McNeish D, Harring J. Covariance pattern mixture models: Eliminating random effects to improve convergence and performance. Behav Res Methods [Internet]. 2019; Available from: <http://link.springer.com/10.3758/s13428-019-01292-4>.
- [28] McNeish D, Harring JR. Improving convergence in growth mixture models without covariance structure constraints. Stat Methods Med Res. 2021;1–19.
- [29] Heggeseth B. Longitudinal Cluster Analysis with Applications to Growth Trajectories [Internet]. [Berkeley, CA]; 2013. Available from:
http://digitalassets.lib.berkeley.edu/etd/ucb/text/Heggeseth_berkeley_0028E_13164.pdf.
- [30] Davies CE, Glonek GFV, Giles LC. The impact of covariance misspecification in group-based trajectory models for longitudinal data with non-stationary covariance structure. Stat Methods Med Res [Internet]. 2017;26:1982–1991. Available from:
<http://journals.sagepub.com/doi/10.1177/0962280215598806>.
- [31] van der Nest G, Lima Passos V, Candel MJJM, et al. Model fit criteria curve behaviour in class enumeration – a diagnostic tool for model (mis)specification in longitudinal mixture modelling. J Stat Comput Simul [Internet]. 2021;1–33. Available from:
<https://www.tandfonline.com/doi/full/10.1080/00949655.2021.2004141>.

- [32] Diallo TMO, Morin AJS, Lu HZ. Impact of Misspecifications of the Latent Variance–Covariance and Residual Matrices on the Class Enumeration Accuracy of Growth Mixture Models. *Struct Equ Model* [Internet]. 2016;23:507–531. Available from: <http://www.tandfonline.com/doi/full/10.1080/10705511.2016.1169188>.
- [33] Kreuter F, Muthén BO. Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *J. Quant. Criminol.* 2008.
- [34] Infurna FJ, Luthar SS. Resilience to Major Life Stressors Is Not as Common as Thought. *Perspect Psychol Sci* [Internet]. 2016;11:175–194. Available from: <http://journals.sagepub.com/doi/10.1177/1745691615621271>.
- [35] Nagin DS, Tremblay RE. Developmental trajectory groups: Fact or fiction? *Criminology.* 2005;43:873–904.
- [36] Bauer DJ, Curran PJ. Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes. *Psychol Methods* [Internet]. 2003;8:338–363. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.8.3.338>.
- [37] Burton-Jeangros C, Blane D, Howe LD, et al. A Life Course Perspective on Health Trajectories and Transitions [Internet]. Burton-Jeangros C, Cullati S, Sacker A, et al., editors. Springer. Cham: Springer International Publishing; 2015. Available from: <http://link.springer.com/10.1007/978-3-319-20484-0>.
- [38] Nagin DS, Odgers CL. Group-Based Trajectory Modeling (Nearly) Two Decades Later. *J Quant Criminol* [Internet]. 2010;26:445–453. Available from: <http://link.springer.com/10.1007/s10940-010-9113-7>.
- [39] Nagin DS, Odgers CL. Group-Based Trajectory Modeling in Clinical Research. *Annu Rev Clin Psychol* [Internet]. 2010;6:109–138. Available from: <http://www.annualreviews.org/doi/10.1146/annurev.clinpsy.121208.131413>.
- [40] Falkenstein MJ, Nota JA, Kropfing J, et al. Empirically-derived response trajectories of intensive residential treatment in obsessive-compulsive disorder: A growth mixture modeling approach. *J Affect Disord* [Internet]. 2019;245:827–833. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0165032718314083>.
- [41] Hilterman ELB, Bongers IL, Nicholls TL, et al. Supervision trajectories of male juvenile offenders: growth mixture modeling on SAVRY risk assessments. *Child Adolesc Psychiatry Ment Health* [Internet]. 2018;12:15. Available from: <https://doi.org/10.1186/s13034-018-0222-7>.

- [42] Lee TK, Wickrama KAS, O'Neal CW, et al. Social stratification of general psychopathology trajectories and young adult social outcomes: A second-order growth mixture analysis over the early life course. *J Affect Disord* [Internet]. 2017;208:375–383. Available from: <http://dx.doi.org/10.1016/j.jad.2016.08.037>.
- [43] Lima Passos V, Klijn S, van Zandvoort K, et al. At the heart of the problem - A person-centred, developmental perspective on the link between alcohol consumption and cardio-vascular events. *Int J Cardiol* [Internet]. 2017;232:304–314. Available from: <http://dx.doi.org/10.1016/j.ijcard.2016.12.094>.
- [44] Grevenstein D, Kröninger-Jungaberle H. Two Patterns of Cannabis Use Among Adolescents: Results of a 10-Year Prospective Study Using a Growth Mixture Model. *Subst Abuse* [Internet]. 2015;36:85–89. Available from: <http://www.tandfonline.com/doi/abs/10.1080/08897077.2013.879978>.
- [45] Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Stat Methods Med Res* [Internet]. 2014;23:42–59. Available from: <http://journals.sagepub.com/doi/10.1177/0962280212445834>.
- [46] Curran PJ, Obeidat K, Losardo D. Twelve Frequently Asked Questions About Growth Curve Modeling. *J Cogn Dev* [Internet]. 2010;11:121–136. Available from: <http://www.tandfonline.com/doi/abs/10.1080/15248371003699969>.
- [47] Muthén BO. Latent variable hybrids: Overview of old and new models. In: Hancock GR, Samuelsen KM, editors. *Adv latent Var Mix Model*. Charlotte, NC: Information Age Publishing, Inc.; 2008. p. 1–24.
- [48] Berlin KS, Parra GR, Williams NA. An Introduction to Latent Variable Mixture Modeling (Part 2): Longitudinal Latent Class Growth Analysis and Growth Mixture Models. *J Pediatr Psychol* [Internet]. 2014;39:188–203. Available from: <https://academic.oup.com/jpepsy/article-lookup/doi/10.1093/jpepsy/jst084>.
- [49] Laursen BP, Hoff E. Person-Centered and Variable-Centered Approaches to Longitudinal Data. *Merrill Palmer Q* [Internet]. 2006;52:377–389. Available from: http://muse.jhu.edu/content/crossref/journals/merrill-palmer_quarterly/v052/52.3laursen01.html.
- [50] Ram N, Grimm KJ. Methods and Measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *Int J Behav Dev* [Internet]. 2009;33:565–576. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718544/pdf/nihms482397.pdf>.

- [51] Muthen BO, Muthen LK. Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes. *Alcohol Clin Exp Res* [Internet]. 2000;24:882–891. Available from: <http://doi.wiley.com/10.1111/j.1530-0277.2000.tb02070.x>.
- [52] Verbeek M. *A guide to modern econometrics*. 4th editio. Chichester: Wiley; 2012.
- [53] Demidenko E. *Mixed Models: Theory and Applications with R*. 2nd ed. Wiley Ser. Probab. Stat. John Wiley & Sons SE - 758 s; 2013.
- [54] Berlin KS, Williams NA, Parra GR. An Introduction to Latent Variable Mixture Modeling (Part 1): Overview and Cross-Sectional Latent Class and Latent Profile Analyses. *J Pediatr Psychol* [Internet]. 2014;39:174–187. Available from: <https://academic.oup.com/jpepsy/article-lookup/doi/10.1093/jpepsy/jst084>.
- [55] Diallo TMO, Morin AJS, Lu H. The impact of total and partial inclusion or exclusion of active and inactive time invariant covariates in growth mixture models. *Psychol Methods* [Internet]. 2017;22:166–190. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27643403>.
- [56] Mund M, Nestler S. Beyond the Cross-Lagged Panel Model: Next-generation statistical tools for analyzing interdependencies across the life course. *Adv Life Course Res* [Internet]. 2019;41:100249. Available from: <https://doi.org/10.1016/j.alcr.2018.10.002>.
- [57] Collins LM, Lanza ST. *Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley; 2010.
- [58] Muthén BO, Masyn K. *Discrete-Time Survival Mixture Analysis*. *J Educ Behav Stat* [Internet]. 2005;30:27–58. Available from: <http://journals.sagepub.com/doi/10.3102/10769986030001027>.
- [59] Killian MO, Cimino AN, Weller BE, et al. A Systematic Review of Latent Variable Mixture Modeling Research in Social Work Journals. *J Evid Based Soc Work* [Internet]. 2019;16:192–210. Available from: <https://doi.org/10.1080/23761407.2019.1577783>.
- [60] Piccarreta R, Studer M. Holistic analysis of the life course: Methodological challenges and new perspectives. *Adv Life Course Res* [Internet]. 2019;41:100251. Available from: <https://doi.org/10.1016/j.alcr.2018.10.004>.
- [61] Magidson J, Vermunt JK, Tran B. Using a mixture latent Markov model to analyze longitudinal U.S. employment data involving measurement error. In: Shigemasa K,

- Okada A, Imaizumi T, et al., editors. *New trends Psychom*. Tokyo: Universal Academy Press; 2009. p. 235–242.
- [62] Pennoni F, Romeo I. Latent Markov and growth mixture models for ordinal individual responses with covariates: A comparison. *Stat Anal Data Min ASA Data Sci J* [Internet]. 2017;10:29–39. Available from: <http://doi.wiley.com/10.1002/sam.11335>.
- [63] Jung T, Wickrama KAS. An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling. *Soc Personal Psychol Compass* [Internet]. 2008;2:302–317. Available from: <http://doi.wiley.com/10.1111/j.1751-9004.2007.00054.x>.
- [64] Davies CE, Giles LC, Glonek GF. Performance of methods for estimating the effect of covariates on group membership probabilities in group-based trajectory models. *Stat Methods Med Res* [Internet]. 2018;27:2918–2932. Available from: <http://journals.sagepub.com/doi/10.1177/0962280216689580>.
- [65] Chen G, Tsurumi H. Probit and Logit Model Selection. *Commun Stat - Theory Methods* [Internet]. 2010;40:159–175. Available from: <https://www.tandfonline.com/doi/full/10.1080/03610920903377799>.
- [66] Chamroukhi F. Piecewise Regression Mixture for Simultaneous Functional Data Clustering and Optimal Segmentation. *J Classif* [Internet]. 2016;33:374–411. Available from: <http://arxiv.org/abs/1312.6974>.
- [67] Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Struct Equ Model A Multidiscip J* [Internet]. 2007;14:535–569. Available from: <https://www.tandfonline.com/doi/full/10.1080/10705510701575396>.
- [68] Henson JM, Reise SP, Kim KH. Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Struct Equ Model*. 2007;14:202–226.
- [69] Blaze TJ. Enumerating the correct number of classes in a semiparametric group-based trajectory model. University of Pittsburgh; 2013.
- [70] Mardia K V. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika*. 1970;57:519–530.
- [71] Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* [Internet]. 1987;52:333–343. Available from: <https://doi.org/10.1007/BF02294360>.
- [72] Kass RE, Raftery A. Bayes Factors. *J Am Stat Assoc*. 1995;90:773–795.

- [73] Wagenmakers EJ. A practical solution to the pervasive problems of p values. *Psychon Bull Rev.* 2007;14:779–804.
- [74] Faulkenberry TJ. Computing Bayes factors to measure evidence from experiments: An extension of the BIC approximation. 2018;55:31–43. Available from: <http://arxiv.org/abs/1803.00360><http://dx.doi.org/10.2478/bile-2018-0003>.
- [75] Kass RE, Wasserman L. A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *J Am Stat Assoc* [Internet]. 1995;90:928–934. Available from: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476592>.
- [76] Raftery AE. Bayesian Model Selection in Social Research. *Sociol Methodol* [Internet]. 1995;25:111. Available from: <https://www.jstor.org/stable/271063?origin=crossref>.
- [77] Wasserman L. Bayesian Model Selection and Model Averaging. *J Math Psychol* [Internet]. 2000;44:92–107. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0022249699912786>.
- [78] Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika.* 2001;88:767–778.
- [79] McNeish D, Harring JR. The Effect of Model Misspecification on Growth Mixture Model Class Enumeration. *J Classif* [Internet]. 2017;34:223–248. Available from: <http://link.springer.com/10.1007/s00357-017-9233-y>.
- [80] Jeffries NO. A Note on “Testing the Number of Components in a Normal Mixture.” *Biometrika* [Internet]. 2003;90:991–994. Available from: <http://www.jstor.org/stable/30042105>.
- [81] Tekle FB, Gudicha DW, Vermunt JK. Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Adv Data Anal Classif.* 2016;10:209–224.
- [82] Peugh J, Fan X. How Well Does Growth Mixture Modeling Identify Heterogeneous Growth Trajectories? A Simulation Study Examining GMM’s Performance Characteristics. *Struct Equ Model A Multidiscip J* [Internet]. 2012;19:204–226. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10705511.2012.659618>.
- [83] Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *J Classif* [Internet]. 1996;13:195–212. Available from: <http://link.springer.com/10.1007/BF01246098>.

- [84] Ramaswamy V, Desarbo WS, Reibstein DJ, et al. An Empirical Pooling Approach for Estimating Marketing Mix Elasticities with PIMS Data. *Mark Sci.* 1993;12:103–124.
- [85] Tofighi D, Enders CK. Identifying the correct number of classes in growth mixture models. In: Hancock GR, Samuelsen KM, editors. *Adv Latent Var Mix Model*. Greenwich, CT: Information Age; 2008. p. 317–341.
- [86] Asparouhov T, Muthén BO. Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using M plus. *Struct Equ Model A Multidiscip J* [Internet]. 2014;21:329–341. Available from: <http://dx.doi.org/10.1080/10705511.2014.915181>.
- [87] Klijn SL, Weijenberg MP, Lemmens P, et al. Introducing the fit-criteria assessment plot – A visualisation tool to assist class enumeration in group-based trajectory modelling. *Stat Methods Med Res* [Internet]. 2017;26:2424–2436. Available from: <http://journals.sagepub.com/doi/10.1177/0962280215598665>.
- [88] Hathaway RJ. Another interpretation of the EM algorithm for mixture distributions. *Stat Probab Lett* [Internet]. 1986;4:53–56. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0167715286900167>.
- [89] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell* [Internet]. 2000;22:719–725. Available from: <http://ieeexplore.ieee.org/document/865189/>.
- [90] Muthén BO. Statistical and Substantive Checking in Growth Mixture Modeling: Comment on Bauer and Curran (2003). *Psychol Methods* [Internet]. 2003;8:369–377. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.8.3.369>.
- [91] Hélié S. An Introduction to Model Selection: Tools and Algorithms. *Tutor Quant Methods Psychol* [Internet]. 2006;2:1–10. Available from: <https://doaj.org/article/183474b2fed44d5a7d853a5888b7f0a>.
- [92] Nielsen JD, Rosenthal JS, Sun Y, et al. Group-based criminal trajectory analysis using cross-validation criteria. *Commun Stat - Theory Methods*. 2014;43:4337–4356.
- [93] He J, Fan X. Evaluating the Performance of the K-fold Cross-Validation Approach for Model Selection in Growth Mixture Modeling. *Struct Equ Model* [Internet]. 2018;00:1–14. Available from: <https://doi.org/10.1080/10705511.2018.1500140>.
- [94] Grimm KJ, Mazza GL, Davoudzadeh P. Model Selection in Finite Mixture Models: A k -Fold Cross-Validation Approach. *Struct Equ Model A Multidiscip J* [Internet]. 2017;24:246–256. Available from: <http://dx.doi.org/10.1080/10705511.2016.1250638>.
- [95] Jeffreys H. *The theory of probability* [Internet]. Third edit. Oxford Class. texts Phys.

- Sci. Oxford: Clarendon Press ; 2004. Available from:
<http://www.loc.gov/catdir/enhancements/fy0606/99175168-t.html> LK -
<https://maastrichtuniversity.on.worldcat.org/oclc/959889309>.
- [96] Xu P, Peng H, Huang T. Unsupervised learning of mixture regression models for longitudinal data. *Comput Stat Data Anal*. 2018;125:44–56.
- [97] Kim ES, Wang Y. Class Enumeration and Parameter Recovery of Growth Mixture Modeling and Second-Order Growth Mixture Modeling in the Presence of Measurement Noninvariance between Latent Classes. *Front Psychol* [Internet]. 2017;8. Available from: <http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01499/full>.
- [98] Brame R, Nagin DS, Wasserman L. Exploring Some Analytical Characteristics of Finite Mixture Models. *J Quant Criminol* [Internet]. 2006;22:31–59. Available from: <http://link.springer.com/10.1007/s10940-005-9001-8>.
- [99] van de Schoot R, Sijbrandij M, Winter SD, et al. The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. *Struct Equ Model* [Internet]. 2017;24:451–467. Available from: <http://dx.doi.org/10.1080/10705511.2016.1247646>.
- [100] Grimm KJ, Ram N. Nonlinear Growth Models in M plus and SAS. *Struct Equ Model A Multidiscip J* [Internet]. 2009;16:676–701. Available from: <http://dx.doi.org/10.1016/j.neulet.2011.03.010>.
- [101] Lin H, Han L, Peduzzi PN, et al. A dynamic trajectory class model for intensive longitudinal categorical outcome. *Stat Med* [Internet]. 2014;33:2645–2664. Available from: <http://doi.wiley.com/10.1002/sim.6109>.
- [102] Jones BL, Nagin DS. A Note on a Stata Plugin for Estimating Group-based Trajectory Models. *Sociol Methods Res* [Internet]. 2013;42:608–613. Available from: <http://journals.sagepub.com/doi/10.1177/0049124113503141>.
- [103] Elmer J, Jones BL, Nagin DS. Using the Beta distribution in group-based trajectory models. *BMC Med Res Methodol* [Internet]. 2018;18:152. Available from: <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0620-9>.
- [104] Rabe-Hesketh S, Skrondal A, Pickles A. GLLAMM Manual [Internet]. U.C. Berkeley Div. Biostat. Work. Pap. Ser. Berkeley, CA, CA; 2004. Available from: <https://biostats.bepress.com/ucbbiostat/paper160/>.
- [105] Palardy GJ, Vermunt JK. Multilevel Growth Mixture Models for Classifying Groups. *J Educ Behav Stat* [Internet]. 2010;35:532–565. Available from:

- <http://journals.sagepub.com/doi/10.3102/1076998610376895>.
- [106] Rabe-Hesketh S, Skrondal A. *GLLAMM Companion. Multilevel Longitud Model Using Stata*. 3rd ed. College Station, TX, TX: Stata Press; 2012.
- [107] Hallquist MN, Wiley JF. *MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus*. *Struct Equ Model A Multidiscip J* [Internet]. 2018;25:621–638. Available from: <https://www.tandfonline.com/doi/full/10.1080/10705511.2017.1402334>.
- [108] Asparouhov T, Muthén BO. Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes*. 2012.
- [109] Scrucca L, Fop M, Murphy T, et al. Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R J*. 2016;8:289–317.
- [110] Proust-Lima C, Philipps V, Lique B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *J Stat Softw* [Internet]. 2017;78. Available from: <http://arxiv.org/abs/1503.00890><http://dx.doi.org/10.18637/jss.v078.i02>.
- [111] Gruen B, Leisch F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J Stat Softw* [Internet]. 2008;28. Available from: <http://www.jstatsoft.org/v28/i04/>.
- [112] Grimm KJ, Ram N, Estabrook R. Nonlinear Structured Growth Mixture Models in M plus and OpenMx. *Multivariate Behav Res* [Internet]. 2010;45:887–909. Available from: <http://www.tandfonline.com/doi/abs/10.1080/00273171.2010.531230>.
- [113] Baker E, Iqbal E, Johnston C, et al. Trajectories of dementia-related cognitive decline in a large mental health records derived patient cohort. *PLoS One*. 2017;12.
- [114] Neale MC, Hunter MD, Pritikin JN, et al. OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika* [Internet]. 2016;81:535–549. Available from: <http://link.springer.com/10.1007/s11336-014-9435-8>.
- [115] Boker SM, Maes HH, Spiegel M, et al. *OpenMx User Guide* [Internet]. Release 2. 2018. Available from: <https://vipbg.vcu.edu/vipbg/OpenMx2/docs//OpenMx/latest/OpenMxUserGuide.pdf>.
- [116] Infurna FJ, Grimm KJ. 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. *Journals Gerontol Ser B* [Internet]. 2018;73:148–

159. Available from:
<https://academic.oup.com/psychogerontology/article/73/1/148/3063820>.
- [117] Leisch F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J Stat Softw* [Internet]. 2004;11:1–18. Available from: <https://cran.r-project.org/web/packages/flexmix/vignettes/flexmix-intro.pdf>.
- [118] Benaglia T, Chauveau D, Hunter DR, et al. mixtools : An R Package for Analyzing Finite Mixture Models. *J Stat Softw* [Internet]. 2009;32. Available from: <http://www.jstatsoft.org/v32/i06/>.
- [119] Vermunt JK, Magidson J. *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.; 2016.
- [120] Francis B, Elliott A, Weldon M. Smoothing Group-Based Trajectory Models Through B-Splines. *J Dev Life-Course Criminol* [Internet]. 2016;2:113–133. Available from: <http://link.springer.com/10.1007/s40865-016-0025-6>.
- [121] Lennon H, Kelly S, Sperrin M, et al. Framework to construct and interpret latent class trajectory modelling. *BMJ Open* [Internet]. 2018;8. Available from: <http://bmjopen.bmj.com/>.
- [122] Erosheva EA, Matsueda RL, Telesca D. Breaking Bad: Two Decades of Life-Course Data Analysis in Criminology, Developmental Psychology, and Beyond. *Annu Rev Stat Its Appl* [Internet]. 2014;1:301–332. Available from: <http://www.annualreviews.org/doi/10.1146/annurev-statistics-022513-115701>.
- [123] Nylund-Gibson K, Grimm RP, Masyn KE. Prediction from Latent Classes: A Demonstration of Different Approaches to Include Distal Outcomes in Mixture Models. *Struct Equ Model A Multidiscip J* [Internet]. 2019;26:967–985. Available from: <https://doi.org/10.1080/10705511.2019.1590146>.
- [124] Lanza ST, Tan X, Bray BC. Latent Class Analysis With Distal Outcomes: A Flexible Model-Based Approach. *Struct Equ Model A Multidiscip J* [Internet]. 2013;20:1–26. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10705511.2013.742377>.
- [125] Jones BL, Nagin DS. Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. *Sociol Methods Res* [Internet]. 2007;35:542–571. Available from: <http://journals.sagepub.com/doi/10.1177/0049124106292364>.
- [126] Lai D, Xu H, Koller D, et al. A multivariate finite mixture latent trajectory model with application to dementia studies. *J Appl Stat* [Internet]. 2016;43:2503–2523. Available

- from: <https://www.tandfonline.com/doi/full/10.1080/02664763.2016.1141181>.
- [127] Diallo TMO, Morin AJS, Lu HZ. Performance of growth mixture models in the presence of time-varying covariates. *Behav Res Methods* [Internet]. 2017;49:1951–1965. Available from: <http://dx.doi.org/10.3758/s13428-016-0823-0>.
- [128] Muthén LK, Muthén BO. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Struct Equ Model A Multidiscip J* [Internet]. 2002;9:599–620. Available from: http://www.tandfonline.com/doi/abs/10.1207/S15328007SEM0904_8.
- [129] Kim SY. Sample Size Requirements in Single- and Multiphase Growth Mixture Models: A Monte Carlo Simulation Study. *Struct Equ Model*. 2012;19:457–476.
- [130] Tan X, Dierker L, Rose J, et al. How spacing of data collection may impact estimates of substance use trajectories. *Subst Use Misuse*. 2011;46:758–768.
- [131] Heggseth BC, Jewell NP. How Gaussian mixture models might miss detecting factors that impact growth patterns. *Ann Appl Stat* [Internet]. 2018;12:222–245. Available from: <https://projecteuclid.org/euclid.aoas/1520564471>.
- [132] Fruhwirth-Schnatter S, Celeux G, Robert CP. *Handbook of Mixture Analysis*. Fruhwirth-Schnatter S, Celeux G, Robert CP, editors. Boca Raton, FL: Chapman and Hall/CRC; 2019.
- [133] Fruhwirth-Schnatter S, Pyne S. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* [Internet]. 2010;11:317–336. Available from: <https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxp062>.
- [134] Lee SX, McLachlan GJ. EMMIXuskew : An R Package for Fitting Mixtures of Multivariate Skew t Distributions via the EM Algorithm. *J Stat Softw* [Internet]. 2013;55. Available from: <http://www.jstatsoft.org/v55/i12/>.
- [135] Kim M, Vermunt J, Bakk Z, et al. Modeling Predictors of Latent Classes in Regression Mixture Models. *Struct Equ Model A Multidiscip J* [Internet]. 2016;23:601–614. Available from: <http://www.tandfonline.com/doi/full/10.1080/10705511.2016.1158655>.
- [136] Enders CK, Tofighi D. The Impact of Misspecifying Class-Specific Residual Variances in Growth Mixture Models. *Struct Equ Model A Multidiscip J* [Internet]. 2008;15:75–95. Available from:

- <https://www.tandfonline.com/doi/full/10.1080/10705510701758281>.
- [137] Morin AJS, Maïano C, Nagengast B, et al. General Growth Mixture Analysis of Adolescents' Developmental Trajectories of Anxiety: The Impact of Untested Invariance Assumptions on Substantive Interpretations. *Struct Equ Model A Multidiscip J* [Internet]. 2011;18:613–648. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10705511.2011.607714>.
- [138] Sijbrandij JJ, Hoekstra T, Almansa J, et al. Identification of developmental trajectory classes: Comparing three latent class methods using simulated and real data. *Adv Life Course Res* [Internet]. 2019;42:100288. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1040260818301898>.
- [139] Sijbrandij JJ, Hoekstra T, Almansa J, et al. Variance constraints strongly influenced model performance in growth mixture modeling: a simulation and empirical study. *BMC Med Res Methodol* [Internet]. 2020;20:276. Available from: <https://bmcmredmethodol.biomedcentral.com/articles/10.1186/s12874-020-01154-0>.
- [140] van der Nest G, Lima Passos V, Candel MJJM, et al. An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Adv Life Course Res* [Internet]. 2020;43:100323. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1040260819301881>.
- [141] Depaoli S. Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychol Methods* [Internet]. 2013;18:186–219. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0031609>.
- [142] Tolvanen A. *hideLatent growth mixture modeling: A simulation study* [Internet]. Unpublishe. University of Jyvaskyla, Finland; 2007. Available from: <http://urn.fi/URN:ISBN:951-39-2971-8>.
- [143] Li M, Harring JR. Investigating Approaches to Estimating Covariate Effects in Growth Mixture Modeling: A Simulation Study. *Educ Psychol Meas* [Internet]. 2017;77:766–791. Available from: <http://journals.sagepub.com/doi/10.1177/0013164416653789>.
- [144] Nowakowska E, Koronacki J, Lipovetsky S. Tractable Measure of Component Overlap for Gaussian Mixture Models. 2014;1–24. Available from: <http://arxiv.org/abs/1407.7172>.
- [145] Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ,

- NJ: Lawrence Erlbaum Associates; 1988.
- [146] McLachlan GJ. Mahalanobis distance. *Resonance* [Internet]. 1999;4:20–26. Available from: <https://link.springer.com/content/pdf/10.1007/BF02834632.pdf>.
- [147] Lubke G, Neale MC. Distinguishing Between Latent Classes and Continuous Factors: Resolution by Maximum Likelihood? *Multivariate Behav Res* [Internet]. 2006;41:499–532. Available from: http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr4104_4.
- [148] Martin DP, von Oertzen T. Growth Mixture Models Outperform Simpler Clustering Algorithms When Detecting Longitudinal Heterogeneity, Even With Small Sample Sizes. *Struct Equ Model A Multidiscip J* [Internet]. 2015;22:264–275. Available from: <http://dx.doi.org/10.1080/10705511.2014.936340>.
- [149] Tueller S, Lubke G. Evaluation of Structural Equation Mixture Models: Parameter Estimates and Correct Class Assignment. *Struct Equ Model A Multidiscip J* [Internet]. 2010;17:165–192. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10705511003659318>.
- [150] Liu Y, Luo F, Liu H. Factors of Piecewise Growth Mixture Model: Distance and Pattern. *Acta Psychol Sin* [Internet]. 2014;46:1400. Available from: <http://pub.chinasciencejournal.com/article/getArticleRedirect.action?doiCode=10.3724/SP.J.1041.2014.01400>.
- [151] Mattsson M, Maher GM, Boland F, et al. Group-based trajectory modelling for BMI trajectories in childhood: A systematic review. *Obes Rev* [Internet]. 2019;20:998–1015. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/obr.12842>.
- [152] Watson L, Belcher J, Nicholls E, et al. Latent Class Growth Analysis of Gout Flare Trajectories: A Three-Year Prospective Cohort Study in Primary Care. *Arthritis Rheumatol* [Internet]. 2020;72:1928–1935. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/art.41476>.
- [153] Coyne SM, Padilla-Walker LM, Holmgren HG, et al. Instagrowth: A Longitudinal Growth Mixture Model of Social Media Time Use Across Adolescence. *J Res Adolesc*. 2019;29:897–907.
- [154] Hu J, Leite WL, Gao M. An evaluation of the use of covariates to assist in class enumeration in linear growth mixture modeling. *Behav Res Methods* [Internet]. 2017;49:1179–1190. Available from: <http://link.springer.com/10.3758/s13428-016->

- 0778-1.
- [155] Morgan GB, Hodge KJ, Baggett AR. Latent profile analysis with nonnormal mixtures: A Monte Carlo examination of model selection using fit indices. *Comput Stat Data Anal* [Internet]. 2016;93:146–161. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0167947315000602>.
- [156] Sher KJ, Jackson KM, Steinley D. Alcohol Use Trajectories and the Ubiquitous Cat’s Cradle: Cause for Concern? *J Abnorm Psychol*. 2011;120:322–335.
- [157] Bonanno GA. Loss, Trauma, and Human Resilience: Have We Underestimated the Human Capacity to Thrive After Extremely Aversive Events? *Am Psychol* [Internet]. 2004;59:20–28. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.59.1.20>.
- [158] Wickham H. *ggplot2: Elegant Graphics for Data Analysis* [Internet]. New York: Springer-Verlag; 2016. Available from: <https://ggplot2.tidyverse.org>.
- [159] Li L, Hser Y-I. On Inclusion of Covariates for Class Enumeration of Growth Mixture Models. *Multivariate Behav Res* [Internet]. 2011;46:266–302. Available from: <http://www.tandfonline.com/doi/abs/10.1080/00273171.2011.556549>.
- [160] Hipp JR, Bauer DJ. Local solutions in the estimation of growth mixture models. *Psychol Methods* [Internet]. 2006;11:36–53. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.11.1.36>.
- [161] Engle S, Whalen S, Joshi A, et al. Unboxing cluster heatmaps. *BMC Bioinformatics* [Internet]. 2017;18:63. Available from: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1442-6>.
- [162] Lubke G, Muthén BO. Performance of Factor Mixture Models as a Function of Model Size, Covariate Effects, and Class-Specific Parameters. *Struct Equ Model A Multidiscip J* [Internet]. 2007;14:26–47. Available from: http://www.leaonline.com/doi/abs/10.1207/s15328007sem1401_2.
- [163] Masyn KE. Latent Class Analysis and Finite Mixture Modeling. Little TD, editor. *Oxford Handb Quant Methods*. 2013;2:784.
- [164] Dodge HH, Shen C, Ganguli M. Application of the Pattern-Mixture Latent Trajectory Model in an Epidemiological Study with Non-Ignorable Missingness. *J Data Sci* [Internet]. 2008;6:247–259. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20401339>.
- [165] Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. *J Am Stat Assoc*

- [Internet]. 1993;88:125. Available from:
<https://www.jstor.org/stable/2290705?origin=crossref>.
- [166] Bouguessa M. A Mixture Model-Based Combination Approach for Outlier Detection. *Int J Artif Intell Tools* [Internet]. 2014;23:1460021. Available from:
<https://www.worldscientific.com/doi/abs/10.1142/S0218213014600215>.
- [167] Guerra-Peña K, Steinley D. Extracting Spurious Latent Classes in Growth Mixture Modeling With Nonnormal Errors. *Educ Psychol Meas* [Internet]. 2016;76:933–953. Available from: <http://journals.sagepub.com/doi/10.1177/0013164416633735>.
- [168] Depaoli S, Winter SD, Lai K, et al. Implementing continuous non-normal skewed distributions in latent growth mixture modeling: An assessment of specification errors and class enumeration. *Multivariate Behav Res* [Internet]. 2019;54:795–821. Available from: <https://www.tandfonline.com/doi/full/10.1080/00273171.2019.1593813>.
- [169] Nam Y, Hong S. Growth Mixture Modeling With Nonnormal Distributions: Implications for Data Transformation. *Educ Psychol Meas* [Internet]. 2021;81:698–727. Available from: <http://journals.sagepub.com/doi/10.1177/0013164420976773>.
- [170] Lubke GH, Miller PJ, Verhulst B, et al. A powerful phenotype for gene-finding studies derived from trajectory analyses of symptoms of anxiety and depression between age seven and 18. *Am J Med Genet Part B Neuropsychiatr Genet* [Internet]. 2016;171:948–957. Available from:
<https://onlinelibrary.wiley.com/doi/10.1002/ajmg.b.32375>.
- [171] Hall TO, Stanaway IB, Carrell DS, et al. Unfolding of hidden white blood cell count phenotypes for gene discovery using latent class mixed modeling. *Genes Immun* [Internet]. 2019;20:555–565. Available from: <http://www.nature.com/articles/s41435-018-0051-y>.
- [172] De Rubeis V, Andreacchi AT, Sharpe I, et al. Group-based trajectory modeling of body mass index and body size over the life course: A scoping review. *Obes Sci Pract* [Internet]. 2021;7:100–128. Available from:
<https://onlinelibrary.wiley.com/doi/10.1002/osp4.456>.
- [173] Kosmidis I. *brglm2: Bias Reduction in Generalized Linear Models* [Internet]. 2020. Available from: <https://cran.r-project.org/package=brglm2>.
- [174] Wickham H, François R, Henry L, et al. *dplyr: A Grammar of Data Manipulation* [Internet]. 2020. Available from: <https://cran.r-project.org/package=dplyr>.

- [175] Lüdtke D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *J Open Source Softw* [Internet]. 2018;3:772. Available from: <http://joss.theoj.org/papers/10.21105/joss.00772>.
- [176] Bates D, Mächler M, Bolker B, et al. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* [Internet]. 2015;67. Available from: <http://www.jstatsoft.org/v67/i01/>.
- [177] Hlavac M. stargazer: Well-Formatted Regression and Summary Statistics Tables [Internet]. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI); 2018. Available from: <https://cran.r-project.org/package=stargazer>.
- [178] Garnier S. viridis: Default Color Maps from “matplotlib” [Internet]. 2018. Available from: <https://cran.r-project.org/package=viridis>.
- [179] Latham-Mintus K, Clarke PJ. Linking Mastery Across the Life Course to Mobility Device Use in Later Life. Carr D, editor. *Journals Gerontol Ser B* [Internet]. 2019;74:1222–1232. Available from: <https://academic.oup.com/psychogerontology/article/74/7/1222/3796261>.
- [180] Fieuws S, Verbeke G. Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Stat Med* [Internet]. 2004;23:3093–3104. Available from: <http://doi.wiley.com/10.1002/sim.1885>.
- [181] Halfon N, Forrest CB. The Emerging Theoretical Framework of Life Course Health Development. In: Halfon N, Forrest CB, Lerner RM, et al., editors. *Handb Life Course Heal Dev* [Internet]. Cham: Springer International Publishing; 2018. p. 19–43. Available from: https://doi.org/10.1007/978-3-319-47143-3_2.
- [182] Koochi F, Khalili D, Mansournia MA, et al. Multi-trajectories of lipid indices with incident cardiovascular disease, heart failure, and all-cause mortality: 23 years follow-up of two US cohort studies. *J Transl Med* [Internet]. 2021;19:286. Available from: <https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-021-02966-4>.
- [183] Mukherjee N, Arathimos R, Chen S, et al. DNA methylation at birth is associated with lung function development until age 26 years. *Eur Respir J* [Internet]. 2021;57:2003505. Available from: <http://erj.ersjournals.com/lookup/doi/10.1183/13993003.03505-2020>.
- [184] Hix-Small H, Duncan TE, Duncan SC, et al. A Multivariate Associative Finite Growth Mixture Modeling Approach Examining Adolescent Alcohol and Marijuana Use. *J Psychopathol Behav Assess* [Internet]. 2004;26:255–270. Available from:

- <http://link.springer.com/10.1023/B:JOB A.0000045341.56296.f a>.
- [185] LACOURSE E, NAGIN D, TREMBLAY RE, et al. Developmental trajectories of boys' delinquent group membership and facilitation of violent behaviors during adolescence. *Dev Psychopathol* [Internet]. 2003;15:183–197. Available from: https://www.cambridge.org/core/product/identifier/S0954579403000105/type/journal_article.
- [186] Cramér H. *Mathematical methods of statistics*. Princet. Math. Ser. Princeton: Princeton University Press; 1946.
- [187] Kotrlik J, Williams H, Jabor K. Reporting and Interpreting Effect Size in Quantitative Agricultural Education Research. *J Agric Educ* [Internet]. 2011;52:132–142. Available from: <http://www.jae-online.org/vol-52-no-1-2011/1536-reporting-and-interpreting-effect-size-in-quantitative-agricultural-education-research.html>.
- [188] R Core Team. *R: A Language and Environment for Statistical Computing* [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: <https://www.r-project.org/>.
- [189] Spinhoven P, Batelaan N, Rhebergen D, et al. Prediction of 6-yr symptom course trajectories of anxiety disorders by diagnostic, clinical and psychological variables. *J Anxiety Disord* [Internet]. 2016;44:92–101. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0887618516301232>.
- [190] Tueller SJ, Drotar S, Lubke GH. Addressing the Problem of Switched Class Labels in Latent Variable Mixture Model Simulation Studies. *Struct Equ Model A Multidiscip J* [Internet]. 2011;18:110–131. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10705511.2011.534695>.
- [191] Livingston G. *Managing Agitation and Raising Quality of Life Study, 2014-2019* [Internet]. Data Collect. Colchester, Essex: UK Data Service; 2021. Available from: 10.5255/UKDA-SN-854856.
- [192] Laybourne A, Livingston G, Cousins S, et al. Carer coping and resident agitation as predictors of quality of life in care home residents living with dementia: Managing Agitation and Raising Quality of Life (MARQUE) English national care home prospective cohort study. *Int J Geriatr Psychiatry* [Internet]. 2019;34:106–113. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/gps.4994>.
- [193] Smith S, Lamping D, Banerjee S, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an

- evaluation of current methodology. *Health Technol Assess (Rockv)* [Internet]. 2005;9:1–93, iii–iv. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15774233>.
- [194] Szabó Á, Hyde M, Towers A. One slope does not fit all: longitudinal trajectories of quality of life in older adulthood. *Qual Life Res* [Internet]. 2021;30:2161–2170. Available from: <https://doi.org/10.1007/s11136-021-02827-z>.
- [195] Saunders R, Buckman JEJ, Cape J, et al. Trajectories of depression and anxiety symptom change during psychological therapy. *J Affect Disord* [Internet]. 2019;249:327–335. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0165032718324443>.
- [196] Venables WN, Ripley BD. *Modern Applied Statistics with S* [Internet]. Fourth. New York: Springer; 2002. Available from: <https://www.stats.ox.ac.uk/pub/MASS4/>.
- [197] Wickham H. *tidyr: Tidy Messy Data* [Internet]. 2021. Available from: <https://cran.r-project.org/package=tidyr>.
- [198] Ratz T, Pischke CR, Voelcker-Rehage C, et al. Distinct physical activity and sedentary behavior trajectories in older adults during participation in a physical activity intervention: a latent class growth analysis. *Eur Rev Aging Phys Act* [Internet]. 2022;19:1. Available from: <https://eurapa.biomedcentral.com/articles/10.1186/s11556-021-00281-x>.
- [199] Wickrama KAS, Lee TK 1979-, O’Neal CW 1985-, et al. *Higher-order growth curves and mixture modeling with Mplus : a practical guide* [Internet]. 2 nd editi. New York, NY: Routledge; 2022. Available from: <https://www.taylorfrancis.com/books/9781003158769>.
- [200] Shiyko M, Ram N, Grimm K. An overview of growth mixture modeling: a simple nonlinear application in OpenMx. *Handb Struct Equ Model* [Internet]. New York, NY: The Guilford Press; 2012. Available from: <http://catalogue.bnf.fr/ark:/12148/cb44396522j>.
- [201] Bauer DJ, Curran PJ. The Integration of Continuous and Discrete Latent Variable Models: Potential Problems and Promising Opportunities. *Psychol Methods* [Internet]. 2004;9:3–29. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.9.1.3>.
- [202] Cole VT, Bauer DJ. A Note on the Use of Mixture Models for Individual Prediction. *Struct Equ Model* [Internet]. 2016;23:615–631. Available from: <http://dx.doi.org/10.1080/10705511.2016.1168266>.

- [203] Sterba SK, Bauer DJ. Predictions of Individual Change Recovered With Latent Class or Random Coefficient Growth Models. *Struct Equ Model* [Internet]. 2014;21:342–360. Available from: <http://dx.doi.org/10.1080/10705511.2014.915189>.
- [204] Sterba SK, Bauer DJ. Matching method with theory in person-oriented developmental psychopathology research. *Dev Psychopathol*. 2010;22:239–254.
- [205] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data [Internet]. Springer Ser. Stat. New York SE - xxii, 568 pages : illustrations ; 25 cm.: Springer; 2000. Available from: <http://swbplus.bsz-bw.de/bsz086402706cov.htm>.
- [206] Fitzmaurice GM, Ware JH, Laird NM. Applied longitudinal analysis [Internet]. Wiley Ser. Probab. Stat. TA - TT -. Hoboken, N.J. SE - xix, 506 p. : illustrations ; 25 cm.: Wiley-Interscience; 2004. Available from: http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=013141757&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.
- [207] Gao F, Philip Miller J, Xiong C, et al. Estimating correlation between multivariate longitudinal data in the presence of heterogeneity. *BMC Med Res Methodol* [Internet]. 2017;17:124. Available from: <http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0398-1>.
- [208] Harring JR, Blozis SA. Fitting correlated residual error structures in nonlinear mixed-effects models using SAS PROC NLMIXED. *Behav Res Methods* [Internet]. 2014;46:372–384. Available from: <http://link.springer.com/10.3758/s13428-013-0397-z>.
- [209] Wolfinger RD. Heterogeneous Variance-Covariance Structures for Repeated Measures. *J Agric Biol Environ Stat* [Internet]. 1996;1:205–230. Available from: <https://www.jstor.org/stable/1400366>.
- [210] Wolfinger R. Covariance structure selection in general mixed models. *Commun Stat - Simul Comput* [Internet]. 1993;22:1079–1106. Available from: <http://www.tandfonline.com/doi/abs/10.1080/03610919308813143>.
- [211] Xiao N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for “ggplot2” [Internet]. 2018. Available from: <https://cran.r-project.org/package=ggsci>.
- [212] Stull DE, Wiklund I, Gale R, et al. Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for COPD. *Contemp Clin Trials* [Internet]. 2011;32:818–828. Available from:

- <https://linkinghub.elsevier.com/retrieve/pii/S1551714411001492>.
- [213] Hawrilenko M, Masyn KE, Cerutti J, et al. Individual Differences in the Stability and Change of Childhood Depression: A Growth Mixture Model With Structured Residuals. *Child Dev* [Internet]. 2021;92. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/cdev.13502>.
- [214] Nagin DS, Tremblay RE. FROM SEDUCTION TO PASSION: A RESPONSE TO SAMPSON AND LAUB*. *Criminology* [Internet]. 2005;43:915–918. Available from: <http://doi.wiley.com/10.1111/j.1745-9125.2005.00028.x>.
- [215] NAGIN DS, TREMBLAY RE. DEVELOPMENTAL TRAJECTORY GROUPS: FACT OR A USEFUL STATISTICAL FICTION?*. *Criminology* [Internet]. 2005;43:873–904. Available from: <http://doi.wiley.com/10.1111/j.1745-9125.2005.00026.x>.
- [216] Sterba SK, Baldasaro RE, Bauer DJ. Factors Affecting the Adequacy and Preferability of Semiparametric Groups-Based Approximations of Continuous Growth Trajectories. *Multivariate Behav Res* [Internet]. 2012;47:590–634. Available from: <http://www.tandfonline.com/doi/abs/10.1080/00273171.2012.692639>.
- [217] Lewis AJ, Sae-Koew JH, Toumbourou JW, et al. Gender differences in trajectories of depressive symptoms across childhood and adolescence: A multi-group growth mixture model. *J Affect Disord* [Internet]. 2020;260:463–472. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0165032719307906>.
- [218] Padilla-Walker LM, Son D, Nelson LJ. A longitudinal growth mixture model of child disclosure to parents across adolescence. *J Fam Psychol* [Internet]. 2018;32:475–483. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/fam0000369>.
- [219] Witt A, Münzer A, Ganser HG, et al. The impact of maltreatment characteristics and revictimization on functioning trajectories in children and adolescents: A growth mixture model analysis. *Child Abuse Negl* [Internet]. 2019;90:32–42. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S014521341930033X>.
- [220] Bahabin Boroujeni M, Mehrabani K, Racisi Shahraki H. Clustering Trend Changes of Lung Cancer Incidence in Europe via the Growth Mixture Model during 1990–2016. Radfar A, editor. *J Environ Public Health* [Internet]. 2021;2021:1–9. Available from: <https://www.hindawi.com/journals/jep/2021/8854446/>.
- [221] Yu W, Chen R, Zhang M, et al. Cognitive decline trajectories and influencing factors in China: A non-normal growth mixture model analysis. *Arch Gerontol Geriatr*

- [Internet]. 2021;95:104381. Available from:
<https://linkinghub.elsevier.com/retrieve/pii/S0167494321000443>.
- [222] Inglin L, Lavikainen P, Jalkanen K, et al. LDL-cholesterol trajectories and statin treatment in Finnish type 2 diabetes patients: a growth mixture model. *Sci Rep* [Internet]. 2021;11:22603. Available from: <https://www.nature.com/articles/s41598-021-02077-6>.
- [223] Lee JY, Walton DM, Tremblay P, et al. Defining pain and interference recovery trajectories after acute non-catastrophic musculoskeletal trauma through growth mixture modeling. *BMC Musculoskelet Disord* [Internet]. 2020;21:615. Available from: <https://bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/s12891-020-03621-7>.
- [224] Deakin CT, Papadopoulou C, McCann LJ, et al. Identification and prediction of novel classes of long-term disease trajectories for patients with juvenile dermatomyositis using growth mixture models. *Rheumatology* [Internet]. 2021;60:1891–1901. Available from: <https://academic.oup.com/rheumatology/article/60/4/1891/5955720>.
- [225] Proust-Lima C, Saulnier T, Philipps V, et al. Describing complex disease progression using joint latent class models for multivariate longitudinal markers and clinical endpoints. 2022;1–32. Available from: <http://arxiv.org/abs/2202.05124>.
- [226] Jo B, Findling RL, Wang C-P, et al. Targeted use of growth mixture modeling: a learning perspective. *Stat Med* [Internet]. 2017;36:671–686. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/sim.7152>.
- [227] Henley SS, Golden RM, Kashner TM. Statistical modeling methods: challenges and strategies. *Biostat Epidemiol* [Internet]. 2020;4:105–139. Available from: <https://www.tandfonline.com/doi/full/10.1080/24709360.2019.1618653>.
- [228] Hetherington E, Plamondon A, Williamson T. Trajectory Modeling with Latent Groups: Potentials and Pitfalls. *Curr Epidemiol Reports* [Internet]. 2020;7:171–178. Available from: <https://link.springer.com/10.1007/s40471-020-00242-5>.
- [229] Heo W, Rabbani A, Grable JE. An Evaluation of the Effect of the COVID-19 Pandemic on the Risk Tolerance of Financial Decision Makers. *Financ Res Lett* [Internet]. 2021;41:101842. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1544612320316561>.
- [230] Ellyson AM, Gause EL, Oesterle S, et al. Trajectories of Handgun Carrying in Rural Communities From Early Adolescence to Young Adulthood. *JAMA Netw Open*

[Internet]. 2022;5:e225127. Available from:

<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2790618>.

Acknowledgements

This thesis is the product of the support and encouragement of many inspiring people. As such, it is impossible to acknowledge them all, but this does not discount the role they played or my appreciation of their support.

I owe an immense debt of gratitude to my supervisors Prof. dr. Gerard van Breukelen, Dr. Math Candel, and Dr. Valéria Lima Passos. This thesis is a reflection of their years of support, reflection, encouragement and challenges provided. Gerard, thank you for your detailed feedback, commitment to this project, and encouraging me to look deeper than face value. You have instilled in me a fond appreciation of the unambiguity of equations in light of superfluous words. Math, your sharp eye for finding errors and omissions where ordinary eyes simply gloss over, your deep commitment, and extensive feedback have been invaluable. Valéria, it has been a pleasure and a privilege working together. Thank you for helping me navigate the currents of a PhD, encouraging my development as an independent thinker, for all the great ideas which challenged my coding and writing skills, and for being my protector. Finally, to all my supervisors, thank you for allowing me to grow as a researcher.

This thesis is dedicated to my family. They have been there every step of the way in the development of my academic and personal pursuits. To my husband Zander, thank you for allowing me to pursue my dreams, for every word and act of love and encouragement, and for uprooting and establishing yourself far away from the shores of our homeland. Adjusting to life abroad has not always been easy (especially during the isolation of COVID-19), but I truly cherish your selfless sacrifice. To my parents, Alice and Roelf, thank you for your support and love and for giving me every opportunity to an education. To my grandmother, Yvonne, I stand here today in no small way due to your love, encouragement, and protection. To my late grandfather, Hennie, thank you for always having the patience to answer my many (many) questions. To my sister, Lee-Anne, thank you for always being there for me and showing me that there is a life beyond the books. To my grandmother, Dorothy, thank you for awakening my interest in the natural sciences through gardening. To my late grandfather, Jock, thank you for the interesting conversations which challenged my young mind. To my parents-in-law, Suzette and Hannes, thank you for your acceptance, love, and support. To Lawrie, Maureen, and Nicholas, your hospitality, guidance, and being our family pillar in Europe have been invaluable. To my extended family, I am grateful for all your constant encouragement and understanding.

To all my colleagues at the M&S department, I thank you for making it such a welcoming and engaging place to work. Edith, you are the glue that holds us all together. I am

so grateful for the friendship that we have developed. To the PhD candidates and graduates of the department as well as special visitors (in short, my friends), I want to say thank you. Etienne, Francesco, Mutamba and Zaheer, it has been a great pleasure working together with such inspiring people and sharing the challenges of knowledge creation. COVID-19 robbed us of many opportunities to socialise, but I look forward to our future collaborations.

To my dear friends, you have all played an important role in establishing the person that I am today. To Marina, thank you for always having a listening ear and being a constant companion in the virtual worlds of online gaming. To Morgan, thank you for helping me find the courage to be myself and for our two decades of friendship. To Charl and Tomás, thank you for the great times exploring Europe together, for your friendship, and for maintaining my sanity. Nadia, our friendship has come a long way from that picnic on the “French Riviera”. Thank you for being a source of inspiration and encouragement during difficult times, and for reminding me to take a deep breath. To Thomas, my paranymp, thank you for the many opportunities to smile, be happy, and put everything into perspective. To Tatenda, my paranymp, our meeting was indeed fortuitous and a reflection of the power of networking. I thank you for the many stimulating conversations and your invaluable friendship.

I am fortunate to have you all in my life.

Finally, thank you to the assessment committee of this thesis for their dedication, input, and sacrifice of their precious time during the summer.

About the author

Gavin van der Nest was born on May 12, 1986, in Benoni, South Africa. In 2011, he graduated with a BCom degree in Econometrics (with distinction) and in 2012 with a BCom (Honours) in Statistics (with distinction) both at the University of Pretoria, South Africa. In 2012, he was awarded a Commonwealth Scholarship to read for his MSc Economics degree at the University of Edinburgh, Scotland which was awarded with merit in 2013. During 2013-2018, he worked in several industries in South Africa, including banking, nongovernmental organizations, and consultancy. In April 2018, he began his PhD project at the department of Methodology and Statistics, Maastricht University. In September 2022, he will join the same department as an assistant professor.

“If a machine is expected to be infallible, it cannot also be intelligent.” — Alan Turing