

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.

1.1. 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.

1.2. 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 aan-/afwezigheid 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.

