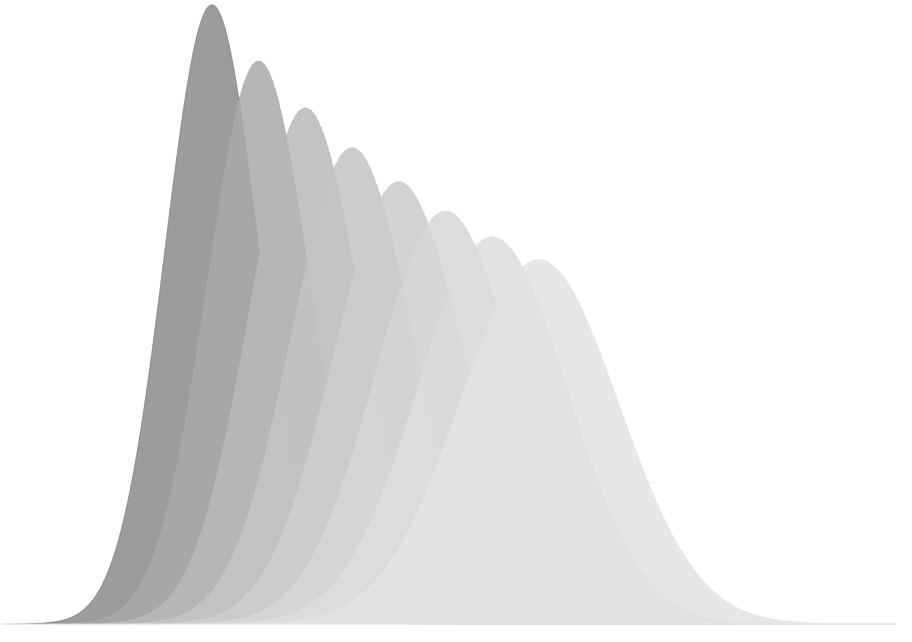
Hidden depths

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]. Novelly, 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.