SECTION 2

METHODOLOGY AND DATASETS

2

Methodology and modelling procedures

Also partly based on:
Jos Twisk & Trynke Hoekstra

Classifying developmental trajectories over time should be done with great caution:
A comparison between methods
Journal of Clinical Epidemiology 65: 1078-1087
2012

This chapter provides an overview of the most common latent class models, including important methodological considerations (such as model assumptions and estimation procedures) that should be taken into account before, during and after the latent class modelling process. To introduce latent class models, a short description of a latent growth model (LGM) [35, 36] is first provided because latent class models are often extensions of LGM. The chapter will end with a description of the datasets used in the next chapters of this thesis.

Latent growth models

A latent growth model (LGM) is a special case of a structural equation model [30–32]; a regression-based model which allows the analysis of observed or unobserved (latent) variables. This longitudinal statistical technique, depicted in general form in **figure 1**, incorporates (continuous) latent- and observed variables. The observed variables are the repeated measures over time and the

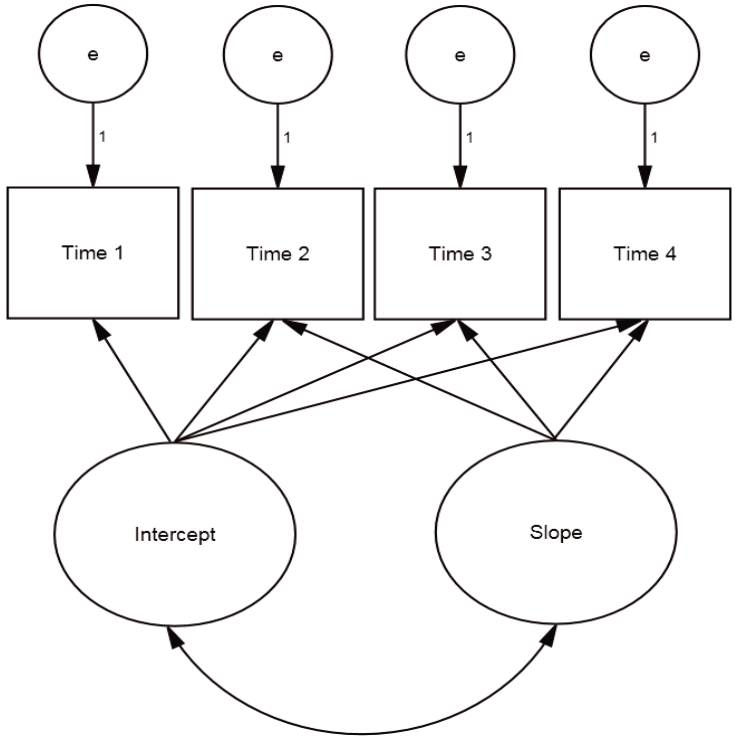

Figure 1 An example of a (linear) latent growth model with a continuous outcome measured at 4 timepoints

unobserved (or latent) variables represent aspects of the repeated measures. The most important latent variables in a LGM are the intercept parameter and the (linear) slope parameter. Both parameters should sound familiar to researchers who frequently conduct regression analyses; the intercept generally represents the initial status of the variable that is analysed over time and the slope represents the rate at which this variable changes over time. For example, if we are interested in the development of the amount of tobacco smoking over time, the intercept would denote the average amount of tobacco smoked at baseline (i.e. the start of the follow-up period) and the slope would give us the mean change (how and in which direction) in the amount of tobacco smoking over time.

In addition to an intercept and a slope parameter, a LGM also estimates the variation, or variance, around both these parameters. These measures tell us about the inter-individual differences; the larger the variances around the intercept and/or slope, the more individuals in the sample differ according to their initial status and/or development over time. The intercept and slope therefore can also be viewed as random intercepts and random slopes, indicating the relation to the earlier mentioned mixed models.

A linear LGM is a model representing upward or downward development over time. As in mixed models, one (average) intercept and one (average) slope with corresponding variance estimates are sufficient to summarise the development of study population. The modelling framework, however, is very flexible and also permits the more complex analysis of other shapes (e.g. quadratic or cubic).

Latent class models

Latent class models [33, 37] in a longitudinal context are an extension of the conventional latent growth model (LGM) described before and are depicted in general form in **figure 2**. Where in LGM one single trajectory is assumed to be sufficient to describe the trajectories for all participants in the study, in latent class models one single trajectory is considered insufficient: participants in the study might come from multiple, underlying (latent) subpopulations, with corresponding heterogeneity in trajectory (shapes). Because latent class models are clustering techniques, the main aim is to maximise homogeneity *within* classes and heterogeneity *between* classes [38, 39]. The application of latent class models allows the (statistical) identification of the number, and characteristics (intercepts, slopes) of underlying subpopulations, indicating that a wider range of research questions can be answered by applying latent

class models compared to LGM. The flexibility of the techniques in this respect allows us to specify class-specific trajectory shapes, class-specific variation (i.e. allowing more within-class variation in one, but not another class) and classifies study participants in their most probable class based on posterior probabilities (this will be explained in more detail at a later stage). Therefore, besides the continuous latent variables (i.e. the intercepts and slopes, similar to a LGM), the models now also incorporate a categorical latent variable denoting the number of unobserved subpopulations, or classes.

The two most common latent class models in a longitudinal setting are 1) a latent class growth model developed by Nagin [4, 29, 40, 41] and colleagues and 2) a latent class growth mixture model developed by Muthén [6, 33, 42, 43] and colleagues. Both techniques will be explained in detail next.

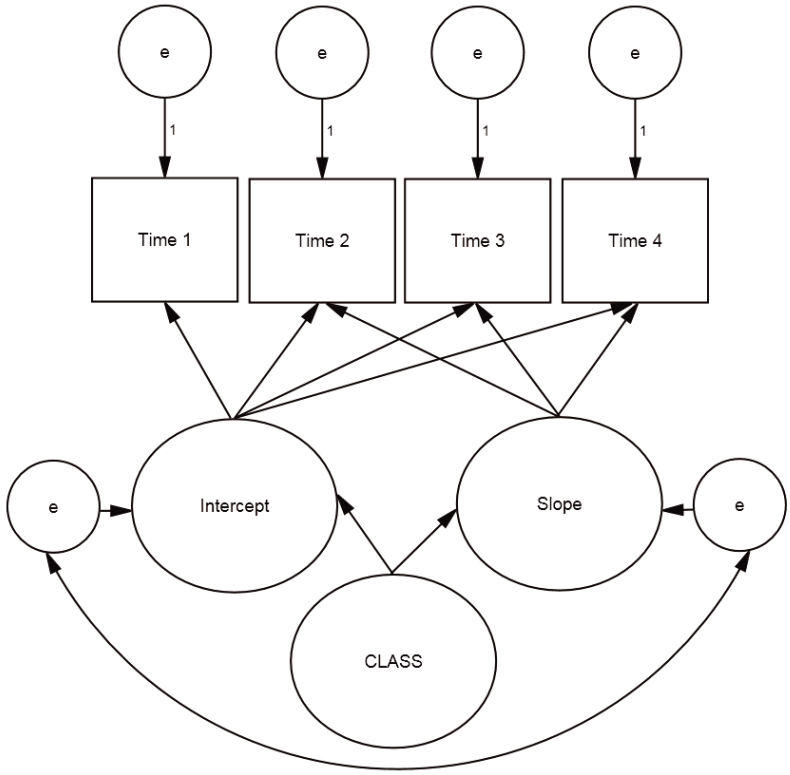

Figure 2 An example of a (linear) latent class growth (mixture) model with a continuous outcome measured at 4 timepoints

Latent class growth model (also called group-based modelling)
Latent class growth models (LCGM) are also referred to as group-based modelling [4, 41, 44]. LCGM were first developed as the counterpart for the mixed model; both techniques aim to model individual-level heterogeneity in developmental trajectories, but do so in different ways. In mixed models, we aim to describe the development over time, incorporating heterogeneity within the sample. The main objective of LCGM is to statistically classify the study participants in distinct subgroups, each with their own growth parameters. There are important restrictions being placed on the within-class variation however [4]. Both the within-class intercept- and -slope variation are set to zero, implying that participants classified into a class are very similar with regard to their individual trajectories.

Latent class growth mixture modelling
Latent class growth mixture models (LCGMM) [5, 6, 33, 42], are in a way less restricted variations of a LCGM. Although both LCGM and LCGMM aim to statistically derive distinct subgroups, LCGMM allow for a certain level of variation in intercept and slope in one or more classes, implying larger within-class heterogeneity.

Cross-sectional techniques
Besides the fact that latent class models can be used in longitudinal studies, i.e. to identify distinct developmental trajectories, they can also be used in cross-sectional studies. Latent class analysis [7, 45] (LCA), also called latent profile analysis [7, 45] in the case of continuous variables, allows for the identification of distinct subgroups of study participants with similar characteristics, for example in symptom profiles [46], problem behaviours [47] or in dietary patterns [48]. The most important model parameters are the class membership probabilities (analogous to latent class growth (mixture) models, this will be explained later in this thesis) and the class characteristics in terms of for example their symptom profile.

Model estimation
In latent class models, the model parameters are estimated based on maximum likelihood (ML) procedures [35, 49, 50]. ML provides a statistical framework to assess all information available in the dataset and estimate the parameter values which best fit the data. These parameter estimates maximise the likelihood

of the model, making the model as close as possible to the observed values. Variants of ML designed for situations when underlying assumptions are not met (e.g. in the case of non-normal data) include mean and variance adjusted maximum likelihood MLMV or robust maximum likelihood (MLR) [35].
Latent class models are sensitive to convergence problems [18, 51], starting values [18, 51] and local maxima [18, 51], for which statistical software packages have designed useful tools to evaluate these issues [34, 43]. Common tools within the Mplus statistical package include a (large) number of random starting values and detailed evaluation of specific solutions through the OPTSEED procedure [34, 43].

Model assumptions
Most assumptions underlying latent class models are fairly similar to assumptions underlying standard regression modelling. Assumptions should be carefully assessed, because if violated, latent class models have been shown to easily lead to an over extraction [38, 39, 52] of classes or even to the extraction of "wrong" classes [52]. Essential assumptions of latent class models include 1) within-class conditional normality; 2) overall goodness-of-fit; 3) missing data are missing at random; 4) independent observations between study participants; and 5) local independence.

1. Within-class conditional normality
Latent class models assume normally distributed measures within classes [52]. This assumption can fairly easily be relaxed when dealing with binary, categorical or count outcome variables amongst others, similar to regression modelling [42, 43, 52]. This assumption also implies non-normality of the marginal (pooled over the classes) distribution of the data, which has been shown to be somewhat confusing; non-normality of the data is generally regarded as evidence for the existence of multiple underlying subpopulations, whereas non-normality can of course also be caused by other factors [1, 38]. In such cases, it has been shown that a model with multiple classes is able to capture this non-normality well even when only one subpopulation exists in the dataset [53]. In practice, researchers should take caution in the selection of the correct latent class model by assessing the distribution of the outcome variables carefully [43], similar to regression modelling, but also taking into account other issues, described later in this thesis.

2. Overall goodness-of-fit

Overall goodness-of-fit implies that the specified model should fit the data well, i.e. the discrepancy between the observed and expected values should be small and means, variances and covariances should be accurately reproduced [31, 54]. Although this assumption is one of the most widely acknowledged assumptions in the structural equation modelling literature [31, 54], within the latent class modelling literature, there is some difficulty with it, mainly because checks are not readily available [53, 55]. However, several graphical diagnostics [56] have been developed to detect possible misspecification of the latent class model, but they are rarely presented [52, 55] because of their complexity. Moreover, although it is possible to let the assessment of the model fit of a latent growth model guide subsequent latent class modelling steps, this also posses difficulties as both an LGM with excellent- and poor fit can justify the modelling of more classes [52].

3. Missing at random missing data

Evaluation of missing data patterns is important in any (statistical) analysis [57, 58]. Assessing whether missing data are in fact missing at random is difficult to test [59] and not often reported [60]. It has, however, been shown that when this assumption is violated, the correctness of results of statistical analyses can be questioned. Therefore, researchers should be aware of the pitfalls of their missing data, in particular in longitudinal studies where loss to follow-up can be selective (e.g. the most frail study participants drop out of the study) [1]. Mplus provides the opportunity to assess missing data patterns through the PATTERNS option [34, 43] and in all latent class models, data (assumed to be) missing at random can be elegantly handled according to the Expectation-Maximation Algorithm used in the estimation procedure [61–63], in principle averting the need to impute missing data.

4. Independent observations between study participants

Complex study designs with clustered data (e.g. patients clustered within general practices or hospitals) can cause bias in model estimates when not accordingly dealt with, in particular in latent class models [52]. This assumption is easy to test and to relax if not met; Mplus amongst others allows for the estimation of latent class models for clustered data, analogous to a multilevel model [2, 34, 43].

5. Local independence

For cross-sectional models (LCA), particular attention should be paid to the local independence assumption [7, 45]. Local independence implies that within each class, the items, symptoms or questions are statistically independent of each other [7, 45]. In other words, this assumption states that the latent classes should fully explain the relationships between these items, symptoms or questions [45, 64]. Although researchers agree on the importance of this assumption, and testing this assumption is relatively straightforward in many software packages (e.g. through the evaluation of any statistically significant bivariate residuals in the TECH10 output in Mplus [34, 43]), it is also relatively simple to relax this assumption when conducting LCA, indicating that two items contain some overlapping information that should not be used in the modelling of the latent classes.

Model building

Determining the number of classes

In the modelling procedure, several latent class models are generally estimated and compared. Determining the final model is the main aim of latent class modelling and depends on a combination of factors including reviewing multiple model fit indices, model parsimony, theoretical background and (clinical) interpretation of the classes. Although researchers often have a strong hypothesis about the number and characteristics of the classes, there is still a great deal of discussion on how to decide on the final number of classes based on the information provided by (statistical) model fit indices [17]. The main point of discussion is the inconsistency between these model fit indices. Currently, assessing and comparing multiple fit indices in a stepwise procedure [17, 18] is advised [17, 18], starting with a one-class model. Next, step by step more classes are added to the model until either the model fit does not improve or theoretically, the maximum number of classes has been reached. Although the choice for an LCA is relatively straightforward (i.e. in the case of cross-sectional data), the choice between LCGM and LCGMM is more difficult to make. The best way to decide between the two is by assessing the model fit indices described next, within-class variance, as well as by "clinical" need. However, the final model is often a combination of a LCGM and a LCGMM with random intercepts or -slopes in some classes only [5, 18].

Several fit indices are described in the literature which aids the comparison of a model with k number of classes with a model with k-1 number of classes. It

should be noted that the available model fit indices are all examples of relative model fit indices, designed to compare models with each other only. Fit indices to asses overall goodness of fit are currently not available [55].

1. Likelihood ratio test

The likelihood ratio test (LRT) [30] might be the most simple way to compare models with a different number of classes with each other. LRT calculates the difference in -2 log-likelihood (-2LL) between two competing models. This difference is assumed to follow a Chi-squared distribution with the number of degrees of freedom denoted as the difference in the number of parameters estimated with the two models. However, from simulation studies [17, 18, 55], it has become clear that in latent class analyses, this difference in -2LL does not follow a Chi-squared distribution and is therefore considered inappropriate when comparing models. Other fit indices therefore are used to choose the best model.

2. Bootstrapped likelihood ratio test

The bootstrapped likelihood ratio test (BLRT) [65] is a likelihood-based test, which overcomes the problems with the traditional likelihood ratio test. The BLRT uses bootstrap samples to estimate the distribution of the log-likelihood difference test statistic. A significant *P*-value derived from this test would favour the *k* class model over the *k*-1 class model and a non-significant *P*-value would favour the *k*-1 class model over the *k* class model. The BLRT has been shown to be a very consistent indicator of the optimal number of classes [17, 18, 34] and can be obtained in Mplus by adding the TECH14 option to the syntax.

3. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) [66] is a commonly used fit index, also in other statistical techniques, considering both the likelihood of the model and the number of parameters in the model; a lower BIC value indicates a better model fit (a decrease of at least 10 points is regarded a sufficient improvement) [67]. Several simulation studies have shown that the BIC is a consistent fit index; often correctly pointing towards the optimal model [17, 18]. The BIC is automatically obtained in the Mplus output.

4. Entropy

Entropy is a measure that indicates the "fuzziness" (or uncertainty in classification) of the model and is defined by the posterior probabilities [68], indicating that this is a measure of how well – or how precise – study participants are classified into their most likely class. Entropy values can range from 0-1, where values closer to 1 indicate greater precision. Although this uncertainty criterion is a common criterion and can be used in the model selection process, the general advice is to not use entropy as the primary index of choice [17]. The entropy measure is automatically obtained in the Mplus output.

Other criteria to decide on the optimal number of classes

Besides the model fit indices, model parsimony, successful convergence, a minimum of 1% of the study sample in a class, high agreement between observed- and estimated class values, high posterior probabilities (close to 1, but at least 0.7 or 0.8 is advised [4, 20]), theoretical background and (clinical) interpretation of the classes are criteria that should also be used to guide the choice of the final model [17, 18]. Taking into account model parsimony implies that the choice for the final model should be guided by the "simplest" model [31, 67]. Theoretical background as well as the clinical interpretation of the classes should be criteria in the choice for the final model, meaning that clinically uninterpretable, or theoretically impossible solutions should be rejected (even if the statistical model fit is better).

Assigning individuals to classes

The decision of which class an individual best belongs to is based on Bayesian statistics [69]. As a result, for each individual a posterior probability of belonging to each of the modelled classes is provided. Posterior probabilities can range between 0 and 1; a probability of 0 indicates that it is impossible that the individual belongs to that class; a probability of 1 indicates that the individual definitely belongs to that class.

Conditional versus unconditional latent class models

In the steps described before, unconditional latent class models are estimated and compared. Once the number of classes has been decided on, a next step can be to assess the conditional versions of this model [5, 18]. Unconditional models are in a way similar to the crude models in standard regression modelling. In that setting, researchers may wish to assess the specific influence

of confounding factors in the relationship that they are interested in. A model including confounders, the adjusted model in this case, is fairly analogous to the conditional model in latent class models, where in fact adjusted classes are modelled. In the literature, there is some discussion about this topic [5, 20], where some authors argue only to use conditional classes [5], especially when the inclusion of covariates in the model greatly influences the clinical interpretation of the latent classes. Others, however, argue that conditional models cause difficulty in assessing the true characteristics of the latent classes [20, 70].

Predictors and consequences of latent class membership

Once the final number and characteristics of the latent classes are decided on, evaluating predictors or consequences of latent class membership are often the next step (i.e. what factors distinguish the classes from each other; examples include pre-existing characteristics, subsequent outcomes or treatment responses [29]). Straightforward regression modelling (or ANOVA) can be conducted to answer research questions dealing with predictors and consequences of latent class membership. In epidemiology, this step is often a separate step, mainly because individuals are generally classified into their most likely class (i.e. the one with the highest posterior probability) resulting in an observed (categorical) variable denoting class membership. This approach, however, ignores the possible uncertainty in class assignment [20]. Because of this uncertainty (and subsequent incorrect standard errors), in other fields of research, particularly psychology and criminology, these steps are undertaken in one process, hereby taking into account this uncertainty [20, 71]. This approach, however, using for example weighted probabilities or pseudo-class draws is fairly complex and if the posterior probabilities approach 1 (a general rule is that they should exceed 0.7 or even 0.8 [4, 20]), the uncertainty is generally low. Moreover, it has been shown that this one-step approach can influence the class formation process also, which is not always desirable [70]. Therefore, in this thesis the two-step approach will only be used.